UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

UNIVERSITY OF PRETORIA

THE VARIABLE SELECTION PROBLEM AND THE APPLICATION OF THE ROC CURVE FOR BINARY OUTCOME VARIABLES

James M Matshego

Prepared in partial fulfilment of the requirements for the degree of

Master of Science

in

Applied Statistics

Supervisor: **Prof H T Groeneveld**

External Examiner: **Prof A J Van der Merve (U. O. F)**

2006

DECLARATION

I, **James Moeng Matshego**, hereby testify that the work presented in this study is my own original work and that all the resources used have been indicated and reflected by means of complete references. I further hereby declare that the dissertation that hereby submit for the degree in Applied Statistics at the University of Pretoria has not previously been submitted by me for degree purpose at any other university.

Signed..

Acknowledgements

I sincerely thank

- My supervisor Prof H T Groeneveld for his encouragement, guidance and support.
- Department of Statistics for having been patient with me.
- Mr E Sibanda from Research and Development at TUT for availing the data set used in this study.
- My family for having afforded me time to study.
- Prof A J van der Merwe for his valuable comments and advices.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Variable selection refers to the problem of selecting input variables that are most predictive of a given outcome. Variable selection problems are found in all machine learning tasks, supervised or unsupervised, classification, regression, time series prediction , two - class or multi-class, posing various levels of challenges.

Variables selection problems are related to the problems of input dimensionality reduction and of parameter planning. It has practical and theoretical challenges of its own. From the practical point of view, eliminating variables may reduce the cost of producing the outcome and increase its speed, while space dimensionality does not address these problems. Theoretical challenges include estimating with what confidence one can state that a variable is relevant to the concept when it is useful to the outcome and providing a theoretical understanding of the stability of selected variables subsets. As the probability cut-points increase in value, the more likely it becomes that an observation is classified as a non-event by the selected variables.

The mathematical statement of the problem is not widely agreed upon and may depend on the application. One typically distinguishes:

i) The problem of discovering all the variables relevant to the outcome variable and determine HOW relevant they are and how they are related to each other.

ii) The problem of finding a minimum subset of variables that is useful to the outcome variable.

Logistic regression is an increasingly popular statistical technique used to model the probability of discrete binary outcome. Logistic regression applies maximum likelihood estimation after transforming the outcome variable into a logit variable. In this way, logistic regression estimates the probability of a certain event. When properly applied, logistic regression analyses yield a very powerful insight in to what variables are more or less likely to predict event outcome in a population of interest. These models also show the extent to which changes in the values of the variable may increase or decrease the predicted probability of event outcome. Variable selection, in all its facets is similarly important with logistic regression.

The receiver operating characteristics (ROC) curve is a graphic display that gives a measure of the predictive accuracy of a logistic regression model. It is a measure of classification performance, the area under the ROC curve (AUC) is a scalar measure gauging one facet of performance. Another measure of predictive accuracy of a logistic regression model is a classification table. It uses the model to classifying observations as events if their estimated probability is greater or equal to a given probability cut-point, otherwise events are classified as non-events. This technique, as it appears in the literature, is also studied in this thesis.

In this thesis the issue of variable selection, both for continuous and binary outcome variables, is investigated as it appears in the statistical literature. It is clear that this topic has been widely researched and still remains a feature of modern research. The last word certainly hasn’t been spoken.

CHAPTER 1

ORIENTATION

1.1 INTRODUCTION

The problem of variable selection is one of the most pervasive problems in statistical models. As stated by Guyon and Elisseeff (2002), variable selection problems are found in all machine learning, supervised or unsupervised, classification, regression, time series prediction tasks, and are posing challenges. Owing to the current availability of high speed computors, this problem has received enormous attention in recent statistical literature. Often referred to as the problem of subset selection, it arises when one wants to model the relationship between a variable of interest and a subset of potential explanatory variables or predictors, but there is uncertainty about which subset to use. A common situation is that in which the explanatory or predictor variables, which will be denoted by **X (**nxp) measured at one time can be used to predict a variable of interest or response variable denoted by **Y(**1xn) at some future time. Unless the true form of the relationship between **X** and **Y** variables is known, it will be necessary for the data to be used to select the variables and to calibrate the relationship to be representative of the conditions in which the relationship will be used for prediction.

In prediction we are usually looking for a small subset of variables which gives adequate prediction accuracy for a reasonable cost of measurement. On the other hand, in trying to understand the effect of one variable on another, particularly when the only data available are observational or survey data rather than experimental data, it may be desirable to include many potential variables which are either known or believed to have an effect (Miller (1990)).

The problem of selecting a subset of predictor variables is usually described in an idealised setting. That is, it assumes that (a) all predictors are available for inclusion or exclusion from the model, though this is not always the situation in practice. In many cases, the original set of measured variables will be augmented with other variables from them such as a product of two variables and (b) a 'good' data set is available on which to base the conclusions. The lack of these assumptions may make a detailed subset selection analysis a futile exercise.

The rationale for minimizing the number of variables in the model is that the resultant model is more likely to be numerically stable, and is more easily generalised. The more variables included in a model, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data (Hosmer and Lemeshow (1989)).

1.2 VARIABLE SELECTION

It will be assumed that there are $n \geq p+1$ observations on a matrix of predictor variables, **X** = ($\mathbf{x}_1 .... \mathbf{x}_p$), and a scalar response, y, such that the j^{th} response, $j = 1,....n$ is determined by

$$y_j = \beta_0 + \sum_{i=1}^{p} \beta_j x_{ij} + \xi_j \qquad (1.1)$$

The residuals, ξ_j are assumed identically and independently distributed, usually normal, with mean zero and unknown variance, σ^2. (The predictors, x_{ij} are frequently taken to be specified design variables, but in many cases it is more appropriate to consider them as random variables and assume a joint distribution on y and x, say, multivariate normal). Implicit in these assumptions is the assumption that the variables $\mathbf{x}_1 .... \mathbf{x}_p$ include all relevant variables though extraneous variables may be included.

The model (1.1) is frequently expressed in matrix notation as

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta} + \varepsilon \qquad (1.2)$$

where **Y** is the n vector of observed responses, **X** is the design matrix dimension $n \times (p+1)$ as defined by (1.2), assumed to have rank $p+1$ and $\boldsymbol{\beta}$ is the ($p+1$) – vector of unknown regression coefficients.

The variable selection problem is most familiar in the linear regression context where attention is restricted to normal linear models. Let γ index the subsets of $\mathbf{x}_1 \ldots \mathbf{x}_p$ and letting q_γ be the size of the γ^{th} subset, the problem is to select and fit a model of the form

$$\mathbf{Y}=\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \varepsilon \tag{1.3}$$

where $\mathbf{X}_\gamma$ is an $n \times q_\gamma$ matrix whose columns correspond to the γ^{th} subset, $\boldsymbol{\beta}_\gamma$ is a $q_\gamma \times 1$ vector of regression coefficients and $\varepsilon \sim N(0, \sigma^2 I)$. More, generally, the variable selection problem is a special case of model selection problem, where each model under consideration corresponds to a distinct subset of $\mathbf{x}_1 \ldots \mathbf{x}_p$.

The fundamental developments in variable selection seem to have occurred either directly in the context of linear model (1.3) or in the context of general model selection frameworks. Historically, the focus began with the linear model in the 1960s when the first wave of important developments occurred and computing was expensive (George (2000)).

1.3 SCOPE OF THIS WORK

This manuscript consists of six chapters. In Chapter 2, methods and procedures for selecting variables in respect of continuous outcome variables for different regressions are described. In addition, statistics for comparison of models are discussed. Chapter 3 introduces and defines the logistic regression model, a model for a binary outcome variable. Various selection procedures for this model are also discussed. The Receiver Operating Characteristic (ROC) curve, a curve representing a diagnostic test with binary outcome, is presented in Chapter 4. Chapter 5 covers a model building exercise. All selection procedures discussed with regard to binary outcome variable are applied to an available data set. We also look into the possibility of using the area under the ROC curve as a variable selection criterion by doing a test with the same data set used for other procedures. Chapter 6 wraps up this study with Discussions and Conclusions.

CHAPTER 2

SELECTION PROCEDURES FOR CONTINUOUS OUTCOME VARIABLES

This chapter will look at the problem of finding one or more subsets of variables which give models that fit a set of data fairly well. However, there is no unique statistical procedure or technique selecting the best regression equation. If there are p potential independent variables there are 2^p possible equations to be considered.

According to Miller (1984), reasons for using only some of the variables or possible predictor variables include:

I. to estimate or predict at lower cost by reducing the number of variables on which predictions can be made.

II. to predict accurately by eliminating uninformative variables.

III. to describe a multivariate data set parsimoniously.

IV. to estimate regression coefficients with small standard errors (particularly when some of the predictors are highly correlated with others).

2.1 VARIABLE SELECTION IN LINEAR REGRESSION

In linear regression an F-test is used since errors are assumed to be normally distributed (Hosmer and Lemeshow (1989)).

2.1.1 Forward Selection

Hocking (1976) suggests a technique that starts with no variable in the equation and adds one variable at a time until either all variables are in or until a stopping criterion is satisfied. The variable considered for inclusion at any step is the one yielding the largest single degree of freedom (d.f) F-ratio among those eligible for inclusion. That is : variable i is added to the r-term equation if $F_i = \max_i (\frac{RSS_r - RSS_{r+i}}{\hat{\sigma}^2_{r+i}}))\rangle F_{in}$ where RSS_r , RSS_{r+i} are residual sum of squares for r-term and ($r + i$) – term models and r the number of terms which are retained in the final equation. The subscript ($r + i$) refers to quantities computed when the variable i is adjoined to the current r-term equation.

Beale (1970) describes a method that requires the least amount of computation. In this method, all results are obtained as a by-product of solving the problem with all variables selected: if there are p regression variables, the covariance matrix of these variables is inverted by pivoting on each of the p diagonal elements in turn, and after each pivot step the results for the regression on those variables for which the corresponding diagonal elements have already been chosen as pivots, can be read off. With regard to this method there are no dependencies among the independent variables. If an element is less than some tolerance times its original value, pivoting is not done where the tolerance is normally 10^{-3} in single precision code or 10^{-7} in a double precision code. Theoretically this approach has a weakness when independent variables are correlated. Two (or more) variables may be individually useless but many together give a very good fit.

Draper and Smith (1981) use the partial correlation coefficient as a measure of the importance of variables not yet in the equation. Assume Z_1 Z_2, …Z_k, are all functions of one or more of the X's, represent the complete set of variables from which the equation is to be chosen and that this set includes any functions, such as squares, cross products, logarithms, inverses, and powers thought to be desirable and necessary. The procedure starts by first selecting the Z most correlated with Y . Suppose this Z is Z_1, the first–order linear regression equation is found to be $\hat{Y} = f(Z_1)$. We check the significance of the variable and if it is not, we quit and the model $Y = \overline{Y}$ is adopted as best, otherwise we search for the second predictor variable to enter the regression. The partial correlation coefficients of all predictors not in regression at this stage, namely Z_j, $j \neq 1$ with Y is examined. In other words, Y and Z_j, are both adjusted for their straight line relationships with Z_1, and the correlation between these adjusted values is calculated for all $j \neq 1$. Z_j with the highest partial correlation coefficient with Y is now selected, say it is Z_2. So the second regression equation $\hat{Y} = f(Z_1, Z_2)$ is fitted. The overall regression is checked for significance with the improvement in R^2 value noted, and the partial F - values for both variables now in the equation are examined. The smaller of these two partial $F's$ is then compared with an appropriate F percentage point and the corresponding predictor variable is retained in the equation or rejected according to whether the test is significant or not. The testing of the "least useful predictor currently in the equation" is done at every stage of the procedure. Thus a predictor that may have been the best entry candidate at an earlier stage may, at a later stage become redundant as a result of the relationship between it and other variables now in the regression. Such a variable will be removed from the model upon testing non-significant and the appropriate fitted regression equation is then computed for all the remaining variables still in the model. Eventually, when no variables in the current equation can be removed and the next best candidate variable cannot hold its place in the equation, the process stops. As each variable is entered into the regression, its effect on R^2 is noted. However, the correct choice of the α- levels is necessary to avoid cycling effect.

Miller (1990) suggests a method that finds a subset $r <$ p of variables $X_{(1)}, X_{(2)}, ...X_{(p)}$ from a set of variables $X_1, X_2, ...X_p$ which minimises or gives a suitably small value for

$$S = \sum_{i=1}^{n} (y_i - b_j x_{ij})^2 .$$

Since the value of b_j is given by

$$b_j = \sum_{i=1}^{n} x_{ij} y_i \bigg/ \sum_{i=1}^{n} x_{ij}{}^2$$

it follows that

$$S = \sum_{i=1} y_i^2 - \left(\sum_{i=1}^{n} x_{ij} y_i \right)^2 \bigg/ \sum_{i=1}^{n} x_{ij}^2 . \qquad (2.1.1)$$

If we let the first variable be denoted by $X_{(1)}$, this variable is then forced into further subsets. The residuals $\mathbf{Y} - \mathbf{X}_{(1)}\, \mathrm{b}_{(1)}$ are orthogonal to $X_{(1)}$, and to reduce the sum of squares by adding further variables, the space orthogonal to $X_{(1)}$ must be searched. From each variable X_j, other than the one already selected, we could form

$\mathbf{X}_{j.(1)} = \mathbf{X}_j - \mathrm{b}_{j.(1)}\mathbf{X}_{(1)}$ where $b_{j.(1)}$ is the least squares regression coefficient of X_j upon X_1, which maximises (2.1.1) when **Y** is replaced with $\mathbf{Y}\text{-}\mathbf{X}_{(1)}\mathrm{b}_{(1)}$ and X_j is replaced with $X_{j\cdot(1)}$.

The variables $X_{(1)}, X_{(2)},....X_{(r)}$ are progressively added to the prediction equation, each variable being chosen because it minimises the residual sum of squares when added to those already selected.

2.1.2 Backward elimination

The backward elimination method is more economical than the "all regressions" method in the sense that it tries to examine only the "best" regression containing a certain number of variables (Draper and Smith (1981)).

We start with all p variables, including a constant if there is one, in the selected set. Thus, a regression equation containing all variables is computed and variables are eliminated one at the time.

At any step, the variable with the smallest F- ratio as computed from the current regression is eliminated if this F- ratio does not exceed a specified value. That is, variable i deleted from the p-term equation if

$$F_i = \min_i\left(\frac{RSS_{p-i} - RSS_p}{\hat{\sigma}_P{}^2}\right) < F_{out} \ .$$

Here RSS_{p-i} denotes the residual sum of squares obtained when variable i is deleted from the current p-term equation, and RSS_p is the residual sum of squares for a p-term equation.

Draper and Smith (1981) proposed a method with the following steps applied to the regression equation with all variables:

1. the partial F - value, which is associated with test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ for any particular regression coefficient, is calculated for every predictor variable treated as though it were the last to enter the regression equation.
2. The lowest partial F - value say F_L say, is compared with pre- selected significance level F_0 say. If $F_L < F_0$, the variable which gave rise to F_L is removed and the regression

equation is calculated with the remaining variable and step 1 is performed. If $F_L > F_0$ the regression equation is adopted as calculated.

A rather simpler approach by Miller (1990) uses the residual sum of squares. If RSS_p is the corresponding residual sum for regression will all p variables, a variable is chosen for deletion if it yields the smallest value of RSS_{p-1} after deletion. Then that variable from the remaining p-1 variables which yields the smallest RSS_{p-2} is deleted. The process continues until one variable is left or a stopping criterion is satisfied.

According to Mantel (1970) the advantageous property of the backward elimination regression procedure is that it drops regressive variables, or sets of regressor variables, only when one can afford to discard without seriously impairing the goodness of fit. Thus many variables can be discarded without abruptly worsening the regression.

On the other hand, backward elimination is usually not feasible when there are more variables than observations. It also requires far more computation than forward selection.

2.1.3 Conventional Stepwise or Efroymson's Algorithm

This is a variation on forward selection. After each variable (except the first one) is added to the set of selected variables, a test is made to ascertain if any of the previously selected variables can be deleted without appreciably increasing the residual sum of squares. This algorithm incorporates criteria for the addition and deletion of variables.

2.1.3.1 Criterion for addition

If RSS_r denotes the residual sum of squares with r variables and a constant in the model and the smallest RSS which can be obtained by adding another variable to the present that is RSS_{r+1}, the ratio $R = \dfrac{RSS_r - RSS_{r+1}}{RSS_{r+1} \big/ (n-r-2)}$ (2.1.3.1)

is calculated and its value is as compared with an 'F–to enter' value, say F_e. If R is greater than $F_{e,}$ the variable is added to the selected set.

2.1.3.2 Criterion for deletion

If RSS_{r-1} is the smallest RSS which can be obtained after deleting any variable from the previously selected variables, the ratio

$$R = \frac{RSS_{r-1} - RSS_r}{RSS_r \big/ (n-r-1)} \qquad (2.1.3.2)$$

is calculated and its value compared with an 'F – to delete (or drop)' value, say F_d. If R is less than F_d, the variable is deleted from the selected set.

2.1.3.3 Convergence of the Algorithm

From (2.1.3.1) it follows that when the criterion for adding a variable is satisfied we have $\text{RSS}_{r+1} \leq \text{RSS}_r \,/\, \{1+ F_e \big/ (n-r-2)\}$ and from (2.1.3.2) when the criterion for deletion of a variable is satisfied we have

$RSS_r \leq RSS_{r+1}\{1 + F_d \big/ (n-r-2)\}$ .Consequently when an addition is followed by a deletion, the new RSS, say RSS^*_r, is such that

$$RSS_r^* \le RSS_r \times \frac{1+ {F_d}/{(n-r-2)}}{1+ {F_e}/{(n-r-2)}} \qquad (2.1.3.3)$$

The procedure stops when no further additions and deletions satisfying the criteria are possible. Since each RSS $_r$ is bounded below by the smallest RSS for any subset of r variables, by ensuring that the RSS is reduced each time that a new subset of r variables is found, convergence is guaranteed. From (2.1.3.3) it follows that a sufficient condition for convergence is that $F_d < F_e$.

However, there is no guarantee that this algorithm will find the best fitting subsets, though it often performs better than forward selection when some of the predictors are highly correlated.

2.1.4 Press

According to Draper and Smith (1981), the Press selection procedure proposed by D.M Allen in Technical Report No 23, Dept of Statistics, University of Kentucky, 1971, the procedure is a combination of all possible regressions, residual analysis and validation techniques.

If r is the number of parameters including β_o in a regression equation and there are n observations in all, the basic calculations entail:

1. Deleting the first observation on the response and predictor variables.
2. Fitting all possible regressions to the remaining n-1 data points
3. Using each fitted model to predict Y_1 by $\hat{Y}_{1r}$ (say) and so obtain a predictive discrepancy $(Y_1 - \hat{Y}_{1r})$ for all the possible regression models.
4. Repeating steps 1, 2 and 3, but deleting the second observation to give $(Y_2 - \hat{Y}_{2r})$ values, the third to give $(Y_3 - \hat{Y}_{3r})$ values, and so on, to n deletions.
5. Calculating the predictive discrepancy sum of squares $\sum_{i=1}^{n}(Y_i - \hat{Y}_{ir})^2$ for each subset regression model.

6. Choosing the “best” subset regression. This will have a comparatively small predictive sum of squares but not involve many predictors.

2.1.5 Principal Component Regression

This is a procedure which analyses the collaboration structures in some detail and was first proposed by Harold Hotelling (Draper and Smith (1981)).

Let **Z** represent the appropriate centred and scaled **X** matrix. Then the correlation matrix **Z'Z**, and the eigenvalues of this correlation matrix are the k solutions $\lambda_1, \lambda_2, ...\lambda_k$ of the determinantal equation

$$|\mathbf{Z'Z} - \lambda\mathbf{I}| = 0 \qquad (2.1.5.1)$$

for the model with all possible predictors Z1$_{,}$ Z_2,…,Z_k. By making each new variable column

$$Z_{ji} = (Z_{ji} - \overline{Z_j} \Big/ S_{jj}^{\frac{1}{2}} \qquad (2.1.5.2)$$

$$\text{where } n\overline{Z_j} = \sum_{i=1}^{n} Z_{ji}, S_{jj} = \sum_{i=1}^{n} (Z_{ji} - \overline{Z}_j)^2 \qquad (2.1.5.3)$$

with zero mean and unit sum of squares, we have orthogonalised out a new $\beta_0^{'}$ term, and cast the predictors into ‘correlation form’. The rank of the non-singular correlation matrix is k= r – 1. The total of all sums of the squares of the Z_j is clearly k (Draper & Smith (1981)). We call this the total variance of the Z’s.

For each eigenvalue, λ_j, there is a eigenvector γ satisfying

$$(\mathbf{Z}^{'}\mathbf{Z} - \lambda_j \mathbf{I})\gamma_j = 0 \qquad (2.1.5.4)$$

with $\gamma^{'}{}_{j}\gamma_j = \mathbf{1}$. The vectors r_j are used to re-express the Z’s in terms of principal components W’s, in the form

$$W_j = \gamma_{1j}z_1 + \gamma_{2j}z_2 + ... + \gamma_{kj}z_k \qquad (2.1.5.5)$$

and the sum of the squares of the new W_j column with elements W_{ji}, i=1, 2,………,n, is λ_j i.e. W_j picks up an amount of λ_j of the total variance. We note that $\sum_j \lambda_j = k$ and $\sum_j \sum_i W_{ji}^2 = k$ •

The W $_j$ corresponding to the largest λ_j value is called the principal component and accounts for the largest proportion of the variation in the standardised data set. Also $W_j^{'}$ s explain smaller and smaller proportions until all variation is explained i.e. $\sum_{j=1}^{r} \lambda_j = k$. The W_j's are not all used but a selection procedure of some sort is used, however, there is no universally agreed upon procedure.

2.1.6 Latent Root Regression

This is an extension of the principal component regression for examining alternative predictive equations and elimination of predictor variables by Webster and his co-workers (Draper and Smith (1981)). The data matrix of the centered and scaled predictor variables is augmented with the centered and scaled responsible variable to provide **Z***= (**y**,**Z**) where **Z** is the centered and scaled '**X** matrix' **y**=(**Y** – **1**$\overline{Y}$)/ $S_{YY}^{\frac{1}{2}}$ where **1** is an nx1 vector of 1's and $S_{YY} = \sum (Y_i - \overline{Y})^2$. It follows that $\mathbf{Z}^{*'}\mathbf{Z}^{*}$ is the augmented correlation matrix. The eigen values and their corresponding eigen vectors are calculated and the first element of each of the eigen vectors is used as a measure of predictability of the response by that eigen vector. The larger the size of the first element of the eigen value the more useful is that eigen vector in predicting the response variable and vice versa. The presence of small eigen values indicates potential linear dependence among predictor variables. Eigen vectors whose eigen values and corresponding first element of the eigen vectors are small are dropped and modified least squares estimation equation is obtained. The backward elimination procedure is then employed to remove predictor variables from the equation.

The vector of a modified least square (MLS) equation coefficients are given by:

$$\mathbf{b}^* = \begin{bmatrix} b_1^* \\ b_2^* \\ b_k^* \end{bmatrix} = c\sum_j {}^*\gamma_{0j}\lambda_j^{-1}\begin{bmatrix} \gamma_{1j} \\ \gamma_{2j} \\ \gamma_{kj} \end{bmatrix} \qquad (2.1.6.1)$$

where c = - $\{\sum_j {}^*\gamma_{oj}^{2}\lambda_j^{-1}\}^{-1}\{\sum_{i=1}^{n}(Y_i-\bar{Y})^2\}^{\frac{1}{2}}$ (2.1.6.2)

and $\sum {}^*$ denotes a summation over only those values of j whose vectors have been retained. Also $b_0^* = \bar{Y}$ for the model. The residual sum of the squares for any modified least squares (MLS) equation can be written as

$$RSS = \{\sum_{l=1}^{n}(Y_i-\bar{Y})^2\}\{\sum_j {}^*\gamma_{oj}^{2}\lambda^{-1}{}_j\}^{-1}$$

$$= -c\{\sum(Y_i-\bar{Y})^2\}^{\frac{1}{2}} \qquad (2.1.6.3)$$

the residual sum of squares that results from deletion of X$_l$, $l = 1,2....k$ from the MLS equation can be evaluated as

$$\{\sum_{i=1}^{n}(Y_i-\bar{Y})^2\}\{t_{00}-\frac{t_{l0}^{2}}{t_{ll}}\}^{-1} \qquad (2.1.6.4)$$

where $t_{rq} = \sum_j {}^*\frac{\gamma_{rj}\gamma_{qj}}{\lambda_j}$ (2.1.6.5)

The main advantage of this method is that by removing the effect of the non-predictive near singularities, the true influences of the independent variables on the dependent variable are more clearly represented.

2.1.7 Branch –and bound Techniques

Suppose that we are looking for the subset of r variables out of p variables which yields the smallest RSS. We begin by dividing all possible subsets into two branches, those which contain X_1, and those which do not. Within each branch we can have sub-branches including and excluding variable X_2, etc. Suppose at some stage we found a subset of r variables containing X_1

or X_2 or both giving RSS=100 say. Suppose we are about to start examining the sub-branch which excludes both $X_1$and X_2. A lower bound on the smallest RSS which can be obtained from this sub-branch is the RSS for all of the p-2 variables. If this is say, 108 then no subset of r variables can do better than this, and since we have already found a smaller RSS, this whole sub-branch can be skipped.

The technique is useful when there are ‘dominant’ variables which good-fitting subsets must include. It is of no value when there are more variables than observations, as the lower bounds are nearly always zero.

2.1.8 Variable Selection via the Elastic net

According to Zou and Hastie (2005), the elastic net encourages a grouping effect where strongly correlated predictors tend to be in or out of the model together. It is particularly useful when the number of predictors (p) is much bigger than the number of observations (n).

2.1.8.1 Naive Elastic net

Let $\mathbf{y} = (\mathrm{y}_{1,}..., y_n)'$ be the response and $\mathbf{X} = (\mathbf{x}_1 \,|\, ... \,|\, \mathbf{x}_p)$ the model matrix,

where $\mathbf{x}_j = (x_{1j},..., x_{nj})'$, j = 1,…,p, are the predictors. After a location and scale transformation, we can assume that the response is centered and the predictors are standardised, and hence

$$\sum_{i=1}^{n} y_i = 0 \quad \sum_{i=1}^{n} x_{ij} = 0 \text{ and } \sum_{i=1}^{n} x_{ij}^2 = 1, \text{ for j =1,…,p} \qquad (2.1.8.1.1)$$

For any fixed non-negative λ_1 and λ_2, we define the naïve elastic net criterion as:

$$L(\lambda_1, \lambda_{2,} \boldsymbol{\beta}) = |\, \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \,|^2 + \lambda_2 \,|\, \boldsymbol{\beta} \,|^2 + \lambda_1 \,|\, \boldsymbol{\beta} \,|_1 \qquad (2.1.8.1.2)$$

where

$$|\boldsymbol{\beta}|^2 = \sum_{j=1}^{p} \beta_j^2$$

$$|\boldsymbol{\beta}|_1 = \sum_{j=1}^{p} |\beta_j|$$

The naïve elastic net estimator $\hat{\boldsymbol{\beta}}$ is the minimiser of (2.1.8.1.2)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{L(\lambda_1, \lambda_2, \boldsymbol{\beta}\}. \qquad (2.1.8.1.3)$$

Let $\alpha = \dfrac{\lambda_2}{\lambda_1 + \lambda_2}$, then solving $\hat{\boldsymbol{\beta}}$ in (2.1.8.2) is equivalent to the optimisation problem

$\hat{\boldsymbol{\beta}}$=argmin$|\mathbf{y}$-$\mathbf{X}\boldsymbol{\beta}|^2$ subject to $(1\text{-}\alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2 \le t$ for some t.

The function $(1\text{-}\alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2$ is the elastic net penalty. In this discussion we consider the case where $\alpha<1$. For all $\alpha \in [0,1)$, the elastic net penalty function is singular (without first derivative) at 0 and it is strictly convex for all $\alpha>0$.

Lemma 1. Given data set (**y**,**X**) and (λ_1, λ_2), define an artificial data set $((\mathbf{y}^*, \mathbf{X}^*)$ by

$$\mathbf{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I} \end{pmatrix}, \qquad \mathbf{y}^*_{(n+p)} = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix}.$$

Let $\gamma = \dfrac{\lambda_1}{\sqrt{(1+\lambda_2)}}$ and $\boldsymbol{\beta}^* = \sqrt{(1+\lambda_2)}\boldsymbol{\beta}$. Then the naïve elastic net criterion can be written as

$L(\gamma, \boldsymbol{\beta}) = L(\gamma, \boldsymbol{\beta}^*) = |\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*|^2 + \gamma|\boldsymbol{\beta}^*|_1$.

Let $\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} L\{(\gamma, \boldsymbol{\beta}^*)\}$;

then $\hat{\boldsymbol{\beta}} = \dfrac{1}{\sqrt{(1+\lambda_2)}}\hat{\boldsymbol{\beta}}^*$.

2.1.8.2 Elastic net

Zou and Hastie (2005) point out that empirical evidence shows that the naïve elastic net does not perform satisfactorily unless it is very close to the lasso method discussed in section (2.2.2). This is why it is called naïve. The elastic net improves the prediction performance of the naïve elastic net.

Given (**y, X**), penalty parameter (λ_1, λ_2) and let $(\mathbf{y}^*, \mathbf{X}^*)$ be the artificial data, the naive elastic net solves a lasso-type problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}^*} |\,\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*\,|^2 + \frac{\lambda_1}{\sqrt{(1+\lambda_2)}}\,|\,\boldsymbol{\beta}^*\,|_1 \qquad (2.1.8.2.1)$$

The elastic net (corrected) estimates $\hat{\boldsymbol{\beta}}$ are defined by

$$\hat{\boldsymbol{\beta}}\,(\text{elastic net}) = \sqrt{(1+\lambda_2)}\,\hat{\boldsymbol{\beta}}^* \qquad (2.1.8.2.2)$$

We recall that $\hat{\boldsymbol{\beta}}$ (naïve elastic)={ $\frac{1}{\sqrt{(1+\lambda_2)}}$}$\hat{\boldsymbol{\beta}}^*$; thus

$$\hat{\boldsymbol{\beta}}\,(\text{elastic net}) = (1+\lambda_2)\hat{\boldsymbol{\beta}}\ \ (\text{naïve elastic net}). \qquad (2.1.8.2.3)$$

Hence the elastic net coefficient is a rescaled naïve elastic net coefficient.

An algorithm called LARS-EN (Zou and Hastie (2005)) is recommended to solve the elastic net efficiently. Algorithm LARS-EN sequentially updates the elastic net fits. In the p>n case, such as with micro array data, it is not necessary to run the algorithm to the end. Real data and simulated computational experiments show that the optimal results are achieved at an early stage of algorithm LARS-EN. If we stop the algorithm after m steps, then it requires $\mathbf{0}(m^3 + pm^3)$ operations.

2.1.9 Generating all Subsets

It is feasible to generate all subsets of variables if the number of predictor variables is not too large, say less than 20 and if only the RSS is calculated for each set. When the complete search has been carried out, a small number of the more promising subsets can be examined in more detail.

The disadvantage of generating all subsets is cost. The computational cost roughly doubles with each additional variable. Hence the availability of high-speed computing becomes imperative for this rather cumbersome procedure.

2.2 VARIABLE SELECTION IN THE COX REGRESSION MODEL

The Cox regression model or proportional hazards model for survival data assumes that

$$\mathrm{h}(\mathrm{t},\mathbf{x},\boldsymbol{\beta}) = h_0(t)\exp(\sum_j x_j\beta_j) \qquad (2.2.1)$$

where $h_0(t)$ is the hazard at time t given predictor values x = $(x_1..., x_p)$ and $h_0(t)$ is an arbitrary baseline function. We usually estimate the parameter $\boldsymbol{\beta} = (\beta_1,..., \beta_p)'$ here in the proportional hazards model without specifying $h_0(t)$ through maximization of the partial likelihood :

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{r\in D}\frac{\exp(\boldsymbol{\beta}'\mathbf{x}^{j_r})}{\{\sum_{j\in R_r}\exp(\boldsymbol{\beta}'\mathbf{x}^{j})\}} \qquad (2.2.2)$$

Performing a proportional hazards regression analysis requires a number of critical decisions. When selecting a subset of covariates, we must consider issues such as clinical importance and adjustment for confounding, as well as statistical significance. Once a subset is selected, we must determine whether the model is ‘linear’ in continuous covariates and, if not, what transformations are suggested by data and clinical considerations. Another important decision is the question of interactions, if any, to be included in the model.

Regardless of which method is used for covariate selection, any survival analysis should begin with a thorough bivariate analysis of association between survival time and all important covariates. For categorical covariates the logrank test must be employed whilst quartiles are used for continuous covariates to make them nominal for the logrank test to be employed.

Stepwise methods for the Cox regression are similar to those that will be discussed in Logistic regression in Chapter 3 and hence will not be considered in this section.

2.2.1 Purposeful Selection of variables

This is a method that is completely controlled by the data analyst. It begins with a multivariable model containing all variables significant in the bivariate analysis at the 20-25 percent level, as well as any other variable not selected with this criterion, but which are judged to be of clinical importance. The use of the above level of significance should lead to the inclusion of any variable that has the potential to be either an important confounder, or statistically significant in the preliminary multivariable model.

Following the fit of the initial multivariable model, we use the P-values from the Wald tests of the individual coefficients to identify covariates that might be deleted from the model. The P-value of the partial likelihood ratio test should confirm that the deleted covariate is not significant.

After fitting the reduced model, we assess whether or not removal of the covariate has produced an "important" change in the coefficients of the variables remaining in the model. We use a value of about 20 percent as an indicator of an important change in the coefficients. If the variable excluded is an important confounder, it is recommended that any variable excluded from the initial multivariable model be added back into the model to confirm that it is neither statistically significant nor an important confounder.

The next step is to examine the scale of continuous covariates in the preliminary main effects model. There are methods that can be employed to assess whether the effect of the covariate is linear in the log hazard and if not, which transformation is linear in the log hazard.
One of the methods involves replacing the continuous covariate with design variables such as quartiles or other purposeful cut-points that may have been used in the bivariate analysis. The estimated coefficients for the design variables are plotted against the midpoints of the groups and, at the midpoint of the first group, a point is plotted at zero. If the correct scale is linear in the log hazard, then the polygon connecting the points should be nearly a straight line. If there is a substantial departure from the linear trend, its form may be used to suggest a transformation of the covariate. The quartile method does not require any special software. However, it is not powerful enough to detect subtle, but often important, deviations from a linear trend.

Another approach is the method of fractional polynomials which we shall not discuss in this study. The only software that has fully implemented this method is STATA (Hosmer & Lemeshow (1998)).

In the final step we determine whether interactions are needed in the model. Special considerations may dictate the inclusion of certain interaction terms irrespective of whether the coefficients are statistically significant or not. In most settings there will be insufficient clinical theory to justify automatic inclusion of interactions.

Biologically plausible interactions are formed and those that are individually significant at the 5 percent level are included simultaneously in the main effects model. The inclusion of non-significant interactions will increase standard error estimates, resulting in wide confidence intervals. The inclusion of an interaction term will change the coefficients of the relevant main effects. When there is statistically significant interaction, we include the corresponding main effect terms in the model regardless of their statistical significance.

2.2.2 The Lasso Method (Tibshirani (1997))

We denote the log partial likelihood by $\lambda(\boldsymbol{\beta})=\log L(\boldsymbol{\beta})$, and assume that the x_{ij} are standardised so that $\sum_i x_{ij} / N = 0, \sum_i x_{ij}^2 / N = 1$.

We estimate $\boldsymbol{\beta}$ via the criterion

$$\hat{\boldsymbol{\beta}} = \arg\min \lambda(\boldsymbol{\beta}), \text{ subject to } \Sigma \,|\, \beta_j \,| \leq s \qquad (2.2.2.1)$$

where s > 0 is a user specified parameter. Suppose $\hat{\beta}^0$ are maximisers of the partial likelihood (2.2.2). Then if $s \geq \sum | \hat{\beta}_j^0 |$, the solution to (2.2.2.1) are the usual partial likelihood estimates. If s<$\sum | \beta_j^2 |$, the solutions to (2.2.2.1) are shrunken towards zero. An attractive feature of the particular constraint $\sum | \beta_j | \leq s$ is that quite often some of the solution coefficients are exactly zero and hence this makes for a more interpretable final model.

The strategy for solving (2.2.2.1) is to express the usual Newton-Raphson update as an iterative reweighted least squares (IRLS) step, and then replace the weighted least squares step by a constrained weighted least squares procedure. If X denotes the design matrix of regressor variables and $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, define $\mathbf{u} = \partial\ell/\partial\boldsymbol{\eta}$, $\mathbf{A}= \partial^2 \ell/\partial\boldsymbol{\eta}\boldsymbol{\eta}'$ and $\mathbf{z} = \boldsymbol{\eta}+\mathbf{A}^{-1}\mathbf{u}$**.** Then a one-term Taylor series expansion for $\ell(\boldsymbol{\beta})$ has the form

$$(\mathbf{z}-\boldsymbol{\eta})'\ \mathbf{A}(\mathbf{z}-\boldsymbol{\eta}) \qquad (2.2.2.2)$$

Hence to solve the original problem (2.2.2.1), we use the following procedure:

i) Fix s and initialise $\hat{\boldsymbol{\beta}}$=0.

ii) Compute $\boldsymbol{\eta}$, $\mathbf{u}$, $\mathbf{A}$ and $\mathbf{z}$ based on the current value $\boldsymbol{\beta}$**.**

iii) Minimise $(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})'\ \mathbf{A}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})$ subject to $\Sigma\,|\,\beta_j\,| \le s$.

iv) Repeat steps 2 and 3 until $\hat{\boldsymbol{\beta}}$ does not change.

Since **A** is a full matrix, it requires computation of $0(\mathrm{N}^2)$ elements. However, this difficulty can be avoided by replacing **A** with diagonal matrix **D** that has the same diagonal elements as **A.**
If the log partial likelihood is bounded in $\boldsymbol{\beta}$ for the given data set, then for fixed s a solution to (2.2.2.1) exists since the region $\Sigma\,|\,\beta_j\,| \le s$ is compact. But the solution may not be unique.

In some situations it is desirable to have an automatic method for choosing the parameter s based on the data. Tibshirani's proposal is to minimise an approximate Generalised Cross Validation (GCV) statistic. We write the constraint $\Sigma\,|\,\beta_j\,| \le s$ as $\sum \beta_j^2 / |\,\beta_j\,| \le s$. This latter constraint is equivalent to adding a Lagrangian penalty $\lambda \sum \beta_j^2 / |\,\beta_j\,|$ to the log partial likelihood, with λ≥0 depending on s. We may write the constrained solution $\hat{\boldsymbol{\beta}}$ step 3 in the form

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{D}\mathbf{X} + \boldsymbol{\lambda}\mathbf{W})^{-1}\mathbf{X}'\mathbf{D}\mathbf{z} \qquad (2.2.2.3)$$

$\mathbf{W} = \mathrm{diag}\ (\mathbf{W}_j)$, $\mathbf{W}_i = 1/|\,\tilde{\beta}_j\,|$ if $|\,\tilde{\beta}_j\,|$>0 and 0 otherwise. Therefore we may approximate the number of effective parameters in the constrained fit $\hat{\boldsymbol{\beta}}$ by

$$\mathrm{p(s)} = \mathrm{tr}[\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X} + \boldsymbol{\lambda}\mathbf{W}^{-})^{-1}\mathbf{X}'\mathbf{D}].$$

Letting λ_s be the log partial likelihood for the constrained fit with constraint s, we construct the GCV-style statistic

$$\text{GCV(s)} = \frac{1}{N}\frac{-\lambda_s}{N[1-p(s)/N]^2} \ .$$

The GCV criterion inflates the negative log partial likelihood by a factor that involves p(s), the effective number of parameters and larger values of p(s) cause more inflation of the negative log partial likelihood.

The simulation study by Tibshirani revealed that the lasso clearly outperforms stepwise selection and picked the correct number of zero coefficients. It is less variable than the stepwise approach and still yields interpretable models.

2.3 VARIABLE SELECTION FOR TIME SERIES DATA

Marriot and Pettitt (1997) proposed a model that takes the form:
Y_t= Filter + Covariates + noise
where the filter is a "time series filter" and is designed to capture stochastic and deterministic trends and seasonality and also to correct for possible auto correlated noise terms. We simply seek to remove the "time series behaviour" from the dependent variable to prevent it from hiding the effects that any exogenous explanatory variable or covariable might have.

The trend components take a lagged dependent variable and linear time trend, and the seasonal component is also a lagged dependent variable.
The proposed time series filter is given by

Filter $=\alpha + \beta\frac{t}{T} + vY_{t-1} + \partial Y_{t-s} + \sum_{i=1}^{p}\phi_i \Delta Y_{t-i}$ where T observations are available, Δ is the difference operator, $\Delta Y_t = Y_t - Y_{t-1}$ and s is the period of the seasonality. The exogenous explanatory variables or covariates are given as

Covariates = $\sum_{l=1}^{k}\psi_i$ $X_{t,i}$

where $X_{t,i}$,…,$X_{t,k}$, are observations on the covariates, and the complete model for observed data is

$$Y_t = \alpha + \beta\frac{t}{T} + rY_{t-1} + \partial Y_{t-s} + \sum_{l=1}^{p}\phi_i \Delta Y_{t-i} + \sum_{l=1}^{k}\psi_i x_{t,j} + \varepsilon_t \qquad \text{t = 1,2,…..T}$$

where $\varepsilon_t \sim iid \;\; \mathrm{N}\,(0,\sigma^2)$

The model is given in vector form as $\mathbf{Y} = \mathbf{Z\theta} + \varepsilon$ (2.3.2)

where **Z** = (**F, X**), the columns of **F** and **X** being sample values of the filter and covariates respectively, and $\boldsymbol{\theta}' = (\alpha, \beta, \nu, \partial, \phi_1, ..., \phi_p, \psi_p, ..., \psi_k)$.

From (2.3.2), Marriot and Pettitt (1997) point out that Zellner (1971) shows that using a non informative joint prior for parameters, and writing ***D*** to represent the past history of both Y_T and $X_{T,i}$ the marginal posterior density for $\boldsymbol{\theta}$ is:

$$f(\boldsymbol{\theta} \mid \mathbf{D}) \propto \{\nu s^2 + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathbf{Z}'\mathbf{Z}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\}^{-\frac{T}{2}}$$

where $\nu = T - p - k - 4$,

$$\mathrm{s}^2 = \frac{(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\theta}})'(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\theta}})}{\nu}$$

and $\hat{\boldsymbol{\theta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$

This is a multivariate *t*-density. The marginal posterior density for σ is

$$f(\sigma \mid \mathbf{D}) \propto \frac{1}{\sigma^{\upsilon+1}}\exp(-\frac{\nu s^2}{2\sigma^2})$$

which is the inverse gamma type distribution.

$$f(Y_F \mid \mathbf{D}, \tilde{\mathbf{z}}) = \int f(Y_F \mid \boldsymbol{\theta}, \sigma, \tilde{\mathbf{z}})\, f(\boldsymbol{\theta}, \sigma \mid \mathbf{D})\, d\boldsymbol{\theta} d\sigma \propto \{\upsilon + (Y_F - \tilde{\mathbf{z}}\hat{\boldsymbol{\theta}})'\mathbf{H}(Y_F - \tilde{\mathbf{z}}\hat{\boldsymbol{\theta}})\}^{-(\nu+1)/2}$$

where $\mathbf{H} = \frac{1}{s^2}\{1 - \tilde{\mathbf{z}}(\mathbf{Z}'\mathbf{Z} + \mathbf{z}\tilde{\mathbf{z}}')^{-1}\tilde{\mathbf{z}}'\}$

which is a *t*- density. The mean and variance of Y_F are $\mathrm{E}[Y_F / \mathbf{D}, \tilde{\mathbf{z}}] = \tilde{\mathbf{z}}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$

and $E[(Y_F - E[Y_F \mid \mathbf{D}, \tilde{\mathbf{z}}])^2 \mid \mathbf{D}, \tilde{\mathbf{z}}] = \frac{\nu}{\nu - 2}s^2\{1 + \tilde{\mathbf{z}}(\mathbf{Z}'\mathbf{Z})^{-1}\tilde{\mathbf{z}}'\}$

If we delete the i^{th} row of the **Z**-matrix to get $\mathbf{Z}_{-\text{i}}$, the complete Bayesian analysis using $\mathbf{Z}_{-\text{i}}$ in place of **Z** is undertaken to obtain the posterior densities. The deleted row $\mathbf{z}_\text{i}$ are used to obtain the predictive densities for observed Y value, $\text{Y}_{i.}$ The predictive mean $\text{E}[Y_i/\mathbf{D},\mathbf{z}_i]$ and standard deviation $\text{S}[Y_i \mid \mathbf{D} \mid \mathbf{z}_i]$ are then used in the construction of diagnostic plots.

The plots are designed to help to answer the questions of whether or not an exogenous explanatory variable makes a significant additional contribution to the model or not, where we consider any additional contribution to be significant if it appears to improve the predictive power of the model.

The order of including explanatory variables is given by backward elimination, the variable corresponding to the smallest value of

$$\left[\frac{E[\psi_i|\mathbf{D}}{S[\psi_i|\mathbf{D}}\right]$$

at each step being removed.

We plot the absolute value of the deviation (AD) of the observation from the predictive mean $\left|Y_i - E[Y_i|\mathbf{D},\mathbf{z}_i\right|$ against the predictive standard deviation (SD),

$\sqrt{\text{var}[Y_i \mid \mathbf{D},\mathbf{z}_i]}$ for each model.

We then plot the convex hull of the scatter. For a clearer picture of the data, all points on the convex hull are ‘peeled’ away and the set of points that form the convex hull of the remaining scatter is identified. The process is repeated until the central 50% of the scatter is reached, and the convex hull of the central 50% is then superimposed on the picture. Plots arising from different models are superimposed, suppressing the original scatter, and the resulting pictures make the relative performance of competing models easy to assess. The better model is the model that combines low predictive dispersion with few extreme values, graphically, the plot of its convex hull is closest to the origin.

If a graphical choice of a model is not clear cut, the sample means of the absolute mean deviations, MAD, and the standard deviations, MSD are used to select the optimal model. The use of sum of these two, MAD + MSD, provides a simple but useful numerical summary of the absolute deviation-standard deviation, ADSP, plot.

2.4 HYPOTHESIS TESTING

Suppose that by some method we have already selected *r* variables, where *r* may be zero, out of *p* variables available to include in our predictor subset. If the remaining variables contain no further information which is useful for predicting the response variable then we should certainly not make any further selection. But we need to know whether the remaining variables containing further information or not. The following hypothesis can be tested

H$_O$: $\beta_{r+1}, \beta_{r+2}, ..., \beta_p = 0$ where these $\beta's$ are the regression coefficients of the variables which have not been selected.

2.4.1 The lack-of-fit Test

If we have n observations and have fitted a linear model containing *r* out of *p* variables plus a constant, then the difference in RSS between fitting the r variables and fitting all the *p* variables, RSS$_r$ – RSS$_p$, can be compared with RSS$_r$ giving the lack-of-fit statistics:

$$\text{Lack of fit F} = \frac{\left. RSS_r - RSS_p \middle/ p - r \right.}{\left. RSS_P \middle/ (n - p - 1) \right.} \qquad (2.4.1.1)$$

If the usual conditions of independence, constant variance and normality are satisfied, then the lack-of-fit statistic is sampled from an F-distribution with (*p-r*) and (*n-p-1*) degrees of freedom.

2.4.2 The Coefficient of Determination, R^2

According to Miller (1990), the distribution of R^2 for a random subset of the Y-variable which is uncorrelated with the X-variables is a beta distribution with

$$\text{prob }(R^2<z) = \frac{1}{B(a,b)}\int_0^z t^{a-1}(1-t)^{b-1}dt$$

where a = $r/2$, b = $(n-r-1)/2$ if a constant has been included in the model but not counted in the *r* variables. Using the beta distribution and fitting constants to their tables, as Miller (1990) points out, Rencher and Pun obtained the following formula for the upper 100(*1-γ*) % point of the distribution of the maximum R^2 using the Efroymson's algorithm as

$R^2_\gamma = [[1 + \log_e \gamma / (\log_e N)^{1.8N^{0.4}}]F^{-1}(\gamma)]$ where (2.4.2.1)

N = ${}^pC_r$ and $F^{-1}(\gamma)$ is the value of z such that prob ($R^2<z$) = γ

Values of $F^{-1}(\gamma)$ can be obtained from the tables of the incomplete beta function or from tables of the F-distribution by writing Reg_r to denote the regression sum of squares on r variables, we have

$$R^2 = \mathrm{Re}\,g_r / (\mathrm{Re}\,g_r + RSS_r)$$

Write F = $\dfrac{\mathrm{Re}\,g_r / r}{RSS_r / (n-r-1)}$

as the usual variance ratio for testing the significance of the subset of *r* variables, if had been chosen a priori, then $R^2 = r/[r+(n-r-1)F]$. (2.4.2.2)

Thus the value of R^2 such that the prob ($R^2<z$) = γ is the value of *F* with prob($R^2<z$) = γ which is the value of F with r and (*n-r-1*) degrees of freedom for the numerator and denominator respectively so that the upper tail area is γ. The reciprocal of a variance ratio also has an *F* distribution but with the degrees of freedom interchanged, and use the tables with (*n-r-1*) and *r* degrees of freedom for numerator and denominator respectively and then take the reciprocal of the *F*-value read from the tables. The upper limit of R^2 is then obtained by substitution in (2.4.2.2) and finally into (2.4.2.1).

2.4.3 Minimum Adequate Sets

Miller (1990) points out that Aitkin advances the following argument:

If we decide on a prior for the comparison of subset X_2 with the full model, containing all the variables in X, then we should use the likelihood-ratio test which gives the variance ratio statistic:

$$F = \frac{(RSS_r - RSS_p) \big/ (p-r)}{RSS_p \big/ (n-p)} \qquad (2.4.3.1)$$

where the counts of variables (r and p) include one degree of freedom for a constant if it is included in the models. Under the null hypothesis that none of the (p-r) variables excluded from X_2 is in the 'true' model, this quantity is distributed as F(p-r,n-p), subject to assumption of independence, normality and homoscedacity of the residuals from the model. Aitkin then considers the statistic:

$$U(X_2) = (p\text{-}r)F \qquad (2.4.3.2)$$

The maximum value of *U* for all possible subsets including a constant is then

$$U_{\max} = \frac{RSS_1 - RSS_p}{RSS_p \big/ (n-p)}$$

where RSS_1 is the sum of squares of *Y* about the mean.

A simultaneous $100\alpha\%$ test for all the hypotheses $\beta_2 = 0$ for all subsets X_2 is obtained by testing that:

$$U(X_2) = (p\text{-}1)\ F\,(\alpha, p-1, n-p\,). \qquad (2.4.3.3)$$

Subsets which satisfy (2.4.3.3) are referred to as 'minimal adequate sets' and are such that if any variable is removed from the subset, it fails to satisfy the condition.

2.5 COMPARISON OF MODELS: SOLUTION CRITERIA

Once a manageable set of models is reached, criteria are needed to select or decide on appropriate subset among contending subsets .The accuracy of any model is measured by a discrepancy, a measure of lack of fit of the model at hand. The model which minimises the expected discrepancy is the 'best' model selected. The overall discrepancy consists of two components: discrepancy due to the approximation (bias) and discrepancy due to estimation (variance). The discrepancy due to

approximation decreases as the number of parameters increases; the discrepancy due to estimation increases as the number of parameters increases.
A consistent estimator of the expected discrepancy is called a criterion and is used for model selection.

2.5.1 Akaike's Information Criterion (AIC) and the Bayes Information Criterion (BIC).

According to George (2000) these two criteria are among the most popular criteria, motivated from very different view points.

Letting $\hat{l}_\gamma$ denote the maximum log likelihood of the γ^{th} model, AIC selects the model which maximises $A = l_\gamma - q_\gamma$ (2.5.1.1)

where q_γ is defined in paragraph (1.2) of Chapter1. Miller (1990) points out that the *AIC* has often been used as the stopping rule for selecting *ARIMA*(auto-regressive, integrated, moving average) models where selection is not only between models with different numbers of parameters but also between many models of the same size. He further suggests that the *AIC*, with various modifications of it, can be applied in situations in which normality is not assumed.
The *BIC* selects the model which maximises

$$B = \left(\hat{l}_\gamma - \frac{1}{2}(\log n) q_\gamma \right)$$

George (2000) mentions Haughton as saying that BIC is consistent when the model is fixed and Shibata saying that AIC is consistent if the dimensionality of the true model increases with n, the number of observations, (at an appropriate rate).

2.5.2 C$_p$ – Statistics (C$_r$ – Criterion)

According to Hocking & Leslie (1967), C L Mallows suggests that the standardised total squared error be used as a criterion and he developed an estimate C_p of this quantity given by:

$$\mathrm{C_p} = \frac{RSS_r}{\hat{\sigma}^2} - (n-2r)\,, \qquad (2.5.2.1)$$

where r is the number of variables in the regression, RSS_r is as defined in (2.1.3.1) and $\hat{\sigma}^2$ is an estimate of σ^2.

Now, if an equation with r parameters is adequate, that is, does not suffer from lack of fit, then $E(RSS_r) = (n-r)\sigma^2$ so that

$$E(C_p) \approx \frac{(n-r)\sigma^2}{\sigma^2} - (n-2r) \qquad (2.5.2.2)$$

$$\approx \mathrm{r}$$

for an adequate model. It follows that a plot of C_p versus r will show up the 'adequate models' as points fairly close to the line $C_p = r$. Thus subsets with small C_p and C_p close to r will be considered to be good.

Certainly, of the $\binom{p}{r}$ possible regressions of size r, only few will be considered to be good. We are interested in that subset of size r for which the residual sum of squares and thus the C_p is minimal.

Hocking & Leslie (1967) further describe a method that allows the subset of size r to be identified after having compared the residual sum of squares for only a small fraction of the possible $\binom{p}{r}$ subsets. This computation will mostly yield those regressions with small $\mathrm{C_p}$. Reference is made to the $k = p - r$ variables which are to be removed from the regression rather than the variables which are to be retained. Reference shall also be made to the "reduction in regression sum of squares" due to removing a set of k variables. Now the set of k variables for which this reduction is minimum determines that set of r variables to be retained for which the residual sum of squares is minimum.

If σ^2 is determined by the residual mean square for the complete regression, and Red_r denotes the reduction, the $\mathrm{C_p}$ statistic can also be computed from this reduction:

$$C_p = \frac{\mathrm{Re}\,d_r}{\hat{\sigma}^2} - (2r - p) \qquad (2.5.2.3)$$

If a single variable, say the i^{th} is removed from the regression, the reduction is given by $\sigma^2 t_i^2$ where

$$t_i^2 = \frac{(b_i)^2}{\hat{\sigma}_{b_i}^{\ 2}} \qquad (2.5.2.4)$$

is the square of the usual *t*- statistic associated with the i^{th} regression coefficient. The b_i are defined by $b_i = D_i' X' X D_i)^{-1} D_i' X' Y$. Let

$$\theta_i = \hat{\sigma}^2 t_i^2 \qquad (2.5.2.5)$$

= reduction due to eliminating i^{th} variable where $i = 1,\ldots,p$.

First, we compute the full regression by solving the normal equations:

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y} \qquad (2.5.2.6)$$

and then evaluate the r univariable reductions, θ_i . We assume that the variables are labelled according on the θ_i . That is

$$\theta_1 \le \theta_2 \le .....\theta_p \,. \qquad (2.5.2.7)$$

With this labelling, the subset of size *p-1* with minimum residual sum of the squares is obtained by deleting the first variable.

This approach is based on the fundamental property of quadratic forms which states that if the reduction in the regression sum of squares due to eliminating any set of variables for which the maximum subscripts *j* is not greater than θ_{i+1} , then no subset including any variable with subscripts greater than i can result in a smaller reduction.

We now describe a sequential method consisting of at most *r+1* stages for each value of *r =1,2……..,p-2*. The first stage consists of computing the reduction due to eliminating variables 1,2,…,*k* for *k*=*p*-*r* under labelling indicated in expression (2.5.2.7). If this reduction does not exceed θ_{k+1} , then, according the above property, the process is terminated and the regression consisting of the *r* variables *k+1,…,p* is to be the 'best' subset of size *r* in the sense of minimum residual sum of squares.

If the reduction computed in the first stage exceeds θ_{k+1}, then no decision can be made and we proceed to the second stage and variable *k+1* is included among the candidates for elimination. The $\binom{k}{1}$ reductions due to eliminating any set of *k* variables selected from the first θ_{k+1} but containing the (*k+1*)st variable are then computed. If the smallest of the *1* + $\binom{k}{1}$ reductions computed to this point does not exceed θ_{k+2} the process terminates and the corresponding subset is 'best'. If not, no decision is taken at this second stage and we proceed to the third stage.

In the third stage the reductions are computed for all subsets of the size *k* selected from the first *k+2* variates which contain variable *k+2,* a total of $\binom{k+1}{2}$ computations. The minimum of the *1* + $\binom{k}{1}+\binom{k+1}{2}$ reductions from the first three stages is now compared with θ_{k+3} and the iteration either terminates or continues to the next stage.

In general, at any stage, say the q^{th}, a total of $\binom{k+q-2}{q-1}$ reductions must be computed and checked to see if the 'best' subset can be identified. At this stage the largest subscript on any variable being considered is *k+q-1* and hence the search can be terminated if the minimum of the $\sum_{j=1}^{q}\binom{k+j-2}{j-1}$ reductions computed in the first *q* stages does not exceed θ_{k+q} and the corresponding subset is 'best'. If not, we proceed to stage *q+1* where subsets of size *k* containing variable *k+q* are considered. However, it has been observed that it rarely happens that all *r*+1 stages are completed except for very small values for *r*.

2.5.3 The S_p – Statistics (S_r – Statistics)

According to Thomson (1978) this method is regarded as being amongst the most suitable for variable selection in multivariable regression analysis where dependent variable y and the *p* independent variables have a (*p+1*)-dimensional normal distribution. The criterion used minimises the expected squared distribution between the true and predictable values of the dependent variable *y*.

The value of *y*, conditionally given some predictor set xD_r ,$r \le p$ may be expressed as follows :

$$y = \beta_0 + (xD_r - \overline{X}_r\beta_r) + \varepsilon_r \qquad (2.5.3.1)$$

where $\overline{X}_r$ (1xr) vector of means obtained from a regression sample for the *r* variables being used and $\varepsilon_r \sim$ N(0;$\sigma_r{}^2$). For some particular predictor set *x*, a future value of $y, \hat{y}_r$ is predicted by:

$$\hat{y}_r = b_0 + (xD_r - \overline{X}_r)b_r \qquad (2.5.3.2)$$

where $b_r = [D_r'(X - 1_n\overline{X}_r)'(x - 1_n\overline{X})D_r]^{-1}D_r'(X - 1_n\overline{X})'Y$ and n the regression sample size.

The method involves calculating the statistic:

$$S_p = \frac{MSE_r}{n - r - 2} \qquad \text{or} \qquad (2.5.3.3)$$

$$S_p = \frac{RED_r + SSE_p}{(n - r)(n - r - 2)} \qquad (2.5.3.4)$$

For subsets of the independent variable where RED_r, is the reduction in regression sums of squares between the full *p*-variable regression and the r variable regression, *r=1,2,…,p* and SSE_p is the error sums of squares. Equation (2.5.3.4) as opposed to (2.5.2.3) provides an efficient computational procedure for the use of this statistic.

The subset of variables chosen is the one which yields the smallest value of S_p. However, if the independent variables cannot be regarded as randomly and normally distributed, the use of C_p is suggested.

2.5.4 RMS, R^2 and Adjusted R^2 Statistics

These are common criteria functions which are simple functions of the residual sum of the squares for the r-term equation denoted by RSS_r

2.5.4.1 The Residual Mean Square

The residual mean square is given by: $RMS_r = \dfrac{RSS_r}{n-r}$ (2.5.4.1)

Hocking (1976), points out that many statisticians voice preference for the residual mean square, RMS_r, as a criterion function. RMS_r, is plotted against r and the choice of r is based on

I. The minimum *RMS*.
II. The value of r such that $RMS_r = RMS$ for the full equation or
III. The value of r such that the locus of the smallest RMS_r turns sharply upwards.

2.5.4.2 The Squared Multiple Correlation Coefficients (SMCC)

The SMCC is given by: $R^2{}_r = 1 - \dfrac{RSS_r}{TSS}$. (2.5.4.2.1)

The plot of $R^2{}_r$ versus r may yield a locus of the minimum $R^2{}_r$ which remains quite flat as r is decreased and then turns sharply down. The value of r at which this 'knee' in the $R^2{}_r$ plot occurs is frequently used to indicate the number of terms in the model. However, it has been observed that R^2 is a measure of the residual sum of the squares proportional to the total sum of squares and, hence, would appear to be a reasonable measure of model adequacy. The relation of R^2 to C_p is given by

$C_p = (n-t-1)(1-R_r{}^2 / (1-R^2) + 2p - n$ (2.5.4.2.2)

It follows from this relation that, while the $R^2{}_r$ plot may be quite flat for a given range of r, the coefficient (n-t-1) can magnify small differences causing C_p to increase dramatically as r is decreased. Therefore, the R_r^2 criterion may suggest the deletion of more variables than the minimum C_p criterion. Simulation studies by some authors as described by Hocking (1976) indicate that essential variables may be deleted using the R_r^2 criterion. Also, lacking a precise definition of the knee, the qualitative inspection of the R_r^2 plot is dependent on the scale.

2.5.4.3 The Adjusted R^2 or Fisher's A-statistics

The adjusted R^2-statistic (adjusted for degrees of freedom) is usually defined as:

$$\overline{R}^2{}_r = 1-(1-R^2{}_r)\frac{(n-1)}{n-r} \qquad (2.5.4.3)$$

as an alternative to R^2. Some users recommend the adjusted squared multiple correlation coefficient $\overline{R}$ and suggest using the value of *r* for which $\overline{R}_r{}^2$ is maximum. Following the simple relation of $\overline{R}_r{}^2$ to C$_{p,}$ the adjusted R^2 -statistic is given by:

$$\overline{R}^2{}_r = 1-\frac{n-1}{TSS}RMS_r\ .$$

The $\overline{R}_r{}^2$ procedure is exactly equivalent to minimising *RMS$_r$*. There appears to be no advantage in using $\overline{R}_r{}^2$ over *RMS$_r$* in view of the above relation.

Chapter 3

THE LOGISTIC MODEL AND VARIABLE SELECTION FOR A BINARY OUTCOME VARIABLE

Having discussed variable selection procedures with regard to continuous outcome variables in Chapter 2, we now in this chapter, consider situations where the response variable is a categorical random variable, attaining only two possible outcomes. In the first place a model and estimation of its parameters is discussed in detail. Then variable selection for this model is presented.

In the next discussions, use was made of the following references :(Czepiel, S, Guyon, I and Elisseeff, A (2002). Joubert, G (1994). Hosmer, D W and Lemeshow, S (1989). Larson, P V (2001). Menard S, (2001)).

3.1 BINARY DATA

When the response variable is dichotomous, it is convenient to denote one of the outcomes as 'success' and the other as 'failure'. For example, if a patient is cured of a disease, the response is 'success', if not, then the response is 'failure'. If a mouse dies from toxic exposure, the response is 'success', if not (i.e. if it survives) the response is 'failure'. It is standard to let the response variable Z be the **binary** variable, which attains the value 1, if the outcome is 'success', and 0 if the outcome is 'failure'.

Let π = P(Z=1) so that P(Z=0) = 1 – π, then Z~ B(1, π). Suppose that data on p predictor variables are available for each patient or mouse, $x_1,\ldots,x_p$. The objective is to investigate the relationship between π and the predictor variables. In a regression situation, each response variable is associated with a given set of values of a set of explanatory variables $x_1,\ldots,x_k$. For example whether or not a patient is cured of a disease may depend on the particular medical treatment the patient is given, the patient's general state of health, age, gender, etc.; whether or not an item in a manufacturing process passes the quality control may depend on various conditions regarding the production process, such as temperature, quality of raw material, time since last service of the machinery, etc. It is often possible to group the observations in such a way that all observations within a group have the same values of predictor variables. For instance, we may group the patients in the disease example according to type of medical treatment, gender and age group, etc such that there are several patients in each grouping. When the data can be grouped it is easier to record the number of successes and failures for each group, rather than recording a long series of 0s and 1s.

Example 3.1 (Larsen 2005)
The link between the use of oral contraceptives and the incidence of myocardial infarction was investigated. The table below gives the number of women in the study, using the contraceptive pill, who suffered a myocardial infarction, and the number using the pill who did not suffer a myocardial infarction. The corresponding numbers for women not using the pill are also given.

		Infarction	
		Yes	No
Pill	Yes	23	34
	No	35	132

Example 3.1

3.2 LOGISTIC REGRESSION

Binomial logistic regression is a form of regression which is used when the response variable is a dichotomy and the predictor variable(s) is/are of any type (i.e. discrete or continuous). It can be used to predict a response variable on the basis of values of predictors and to determine the percentage of variance in the response variable explained by the predictors; to rank the relative importance of predictors; to assess interaction effects; and to understand the impact of covariate control variables. Logistic regression has proven to be one of the most versatile techniques in the class of generalised linear models (Czepiel, S).

Whereas linear regression models equate the expected value of the dependent variable to a linear combination of predictor variables and their corresponding parameters, generalised linear models equate the combination to some function of the probability of a given outcome on the dependent variable. In logistic regression, that function is the logit transform: the natural logarithm of the odds that some event will occur. In linear regression, parameters are estimated using the method of least squares by minimising the sum of squared deviations of predicted values from observed values. However, logistic regression is not capable of producing minimum variance unbiased (minvu) estimators of the actual parameters. In place of the minvu estimators maximum likelihood estimation is used to solve for the parameters.

3.2.1 Assumptions

Logistic regression is popular in part because it enables the researcher to overcome many of the restrictive assumptions of ordinary least square (OLS) regression:

i) Logistic regression does not require linear relationships between predictors and the response variable but assumes a linear relationship between the predictors and the logit of the response variable.
ii) The response need not be normally distributed (we need to assume its distribution is within the range of the exponential family of distributions, such as normal, Poisson, binomial, gamma).
iii) The response variable need not be homoscedastic for each combination of levels of the predictors; that is, there is no homogeneity of variance assumption.
iv) Normally distributed errors are not assumed. However, errors are assumed to be independent.
v) Logistic regression does not require that the predictors be measured on interval scale.
vi) Logistic regression does not require the dependents to be unbounded.

3.2.2 The Multiple linear Logistic Regression Model

Let Z be a dichotomous (termed 'success' and 'failure') random variable denoting the outcome of some experiment and let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ be a collection of predictor variables. Given a data set with a total sample size of M, where each observation is independent from all the others, **Z** can be considered as a column vector of M binomial random variables Z_i. The data is aggregated such that each row represents one distinct combination of values of the predictor variables. The rows are often referred to as 'populations'. Let N represent the total number of populations and let **n** be a column vector with elements n_i representing the number of observations in each population for i =1 to N where $\sum_{i=1}^{N} n_i$ =M, the total sample size.

Let **Y** be a column vector of length N where each element Y_i is a random variable representing the observed counts of the number of successes of Z for population i. Let the column vector **y** contain elements y_i representing the observed counts of the number of successes for each population. Let $\boldsymbol{\pi}$ be a column vector also of length N with elements π_i = P(Z_i=1|i), i.e., the probability of success for any given observation in the ith population.

The linear component of the model contains the design matrix and the vector of parameters to be estimated. The design matrix of predictor variables, **X**, is composed of N rows and p+1 columns, where p is the number of predictor variables specified in the model. For each design matrix, the first element x_{i0} = 1 for all i. This is the intercept. The parameter vector, $\boldsymbol{\beta}$**,** is a column vector of length p+1. There is one parameter corresponding to each of the p columns of predictor variables settings in **X,** plus one, β_0, for the intercept.

The logistic regression model equates the logit transform, the log-odds of the probability of a success, to the linear component:

$$\text{Logit}\,(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{k=0}^{p} x_{ik}\beta_k \qquad i = 1,2, \ldots, \text{N} \qquad (3.2.2.1)$$

$$= \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

If some of the independent variables are discreet, (nominal scaled variables such as race, sex, treatment group, and so forth), it is inappropriate to include them in the model as if they were interval scaled. In fact the numbers used to represent the various levels are simply identifiers, and have no numeric significance. The method of choice is to use a collection of **design variables** (or **dummy variables**). For example, if one of the predictor variables is race, say, coded as ‘‘white”, “black” or “other” then two design variables are necessary. Table 3.1 illustrates coding of the design variables, D1 and D2.

	Design Variable	
RACE	D1	D2
White	0	0
Black	1	0
Other	0	1

Table3.1. An example of the coding of Design Variable Race coded at three levels.

(In general, if a nominal scaled variable has k possible values, then k-1 design variables are needed).

3.3 PARAMETER ESTIMATION

The goal of logistic regression is to estimate the p+1 unknown parameters in equation (3.2.1.1). This is done with maximum likelihood estimation which entails the finding of a set of parameters for which the probability of the observed data is greatest.

3.3.1 Maximum likelihood Estimation

The maximum likelihood estimation equation is derived from the probability distribution of the dependent variable. Since each y_i represents a binomial count in the ith population, the joint density function of **Y** is:

$$\mathrm{f}(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} \ \pi_i^{y_i} (1\text{-}\pi_i)^{n_i - y_i} \qquad (3.3.1.1)$$

For each population, there are $\begin{pmatrix} n_i \\ y_i \end{pmatrix}$ different ways to arrange y_i success from n_i trials. Since the probability of a success for any one of the n_i trials is π_i, the probability of y_i successes is $\pi_i^{y_i}$. Likewise, the probability of $n_i - y_i$ failures is $(1\text{-}\pi_i)^{n_i - y_i}$.

The joint probability function in equation (3.3.1.1) expresses the values of **y** as function of known, fixed values for **β.** The likelihood function has the same form as the probability function, except that the parameters of the function are reversed: the likelihood function expresses the values of **β** in terms of the known values for **y**. Thus,

$$\mathrm{L}(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} \ \pi_i^{y_i} (1\text{-}\pi_i)^{n_i - y_i} \qquad (3.3.1.2)$$

The maximum likelihood estimates are the values for **β** that maximize the likelihood function in equation (3.3.1.2). The critical points of a function (maxima and minima) occur when the first derivative equals 0. Attempting to take the derivative of equation (3.3.1.2) with respect to **β** is a difficult task due to the complexity of multiplicative terms. However, the likelihood equation can be considerably simplified. We ignore the factorial terms since they do not contain π_i and their exclusion will come to the same results. After rearranging equation (3.3.1.2) we obtain:

$$\mathrm{L}(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N} \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i)^{n_i} \qquad (3.3.1.3)$$

Taking e to both sides of (3.2.2.1) gives,

$$\left(\frac{\pi_i}{1-\pi_i}\right) = \mathrm{e}^{\sum_{k=0}^{p} x_{ik}\beta_k} \qquad (3.3.1.4)$$

which after solving for π_i becomes,

$$\pi_i = \left(\frac{\mathrm{e}^{\sum_{\mathrm{k=o}}^{p} \mathrm{x}_{\mathrm{ik}}\beta_{pk}}}{1+e^{\sum_{k=0}^{p} x_{ik}\beta_k}}\right) \qquad (3.3.1.5)$$

Substituting equation (3.3.1.4) for (3.3.1.1) and equation (3.3.1.5) for (3.3.1.2), equation (3.3.1.3) becomes:

$$\mathrm{L}(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N}\left(e^{\sum_{k=0}^{p} x_{ik}\beta_k}\right)^{y_i}\left(1-\frac{e^{\sum_{k=0}^{p} x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{p} x_{ik}\beta_k}}\right)^{n_i} \qquad (3.3.1.6)$$

which can be written as:

$$\mathrm{L}(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N}\left(e^{y_i\sum_{k=0}^{p} x_{ik}\beta_k}\right)\left(1+e^{\sum_{k=0}^{p} x_{ik}\beta_k}\right)^{-n_i} \qquad (3.3.1.7)$$

This is the kernel of the likelihood to maximize. We simplify by taking its log and equation (3.3.1.7) becomes:

$$\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i\left(\sum_{k=0}^{p} x_{ik}\beta_k\right) - n_i \log(1+e^{\sum_{k=0}^{p} x_{ik}\beta_k}) \qquad (3.3.1.8)$$

We now find the critical points of the log likelihood function by differentiating it and obtain:

$$\frac{\partial \lambda(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{N} y_i x_{ik} - n_i \pi_i x_{ik} \qquad (3.3.1.9)$$

The critical point will be a maximum if the matrix of second partial derivatives is negative definite; that is, if every element on the diagonal of the matrix is less than zero. It is formed by differentiating each of the p+1 equations in equation (3.1.1.9) a second time with respect to each element of $\boldsymbol{\beta}$. The general form of the matrix of second partial derivatives is

$$\frac{\partial}{\partial \beta_k}\left(\frac{\partial \lambda(\boldsymbol{\beta})}{\partial \beta_k}\right) = \frac{\partial}{\partial \beta_k}\sum_{i=1}^{N} y_i x_{ik} - n_i x_{ik} \pi_i$$

$$= \frac{\partial}{\partial \beta_k}\sum_{i=1}^{N} - n_i x_{ik} \pi_i$$

$$= -\sum_{i=1}^{N} n_i x_{ik} \frac{\partial}{\partial \beta_k}\left(\frac{e^{\sum_{k=0}^{P} x_{ik}\beta_k}}{1 + e^{\sum_{k=0}^{P} x_{ik}\beta_k}}\right)$$

$$= -\sum_{i=1}^{N} n_i x_{ik} \pi_i (1 - \pi_i) x_{ik} \qquad (3.3.1.10)$$

Thus the critical point will be a maximum since the matrix of second partial derivatives is negative definite following the result obtained in equation (3.3.1.10).

3.3.2 The Newton-Raphson Method

Setting the equations in equation (3.3.1.9) equal to zero results in a system of p+1 nonlinear equations each with k+1 unknown variables. The solution to the system is vector $\hat{\boldsymbol{\beta}}_k$. However, solving a system of nonlinear equations is not easy since the solution cannot be derived algebraically as it can be done in the case of linear equations. The solution must be found using an iterative process. The most popular method for solving systems of nonlinear equations is Newton's method, also known as the Newton-Raphson method.

It is more convenient to use matrix notation to express each step of the Newton-Raphson method. We can write equation (3.3.1.10) as $\lambda^{/}(\boldsymbol{\beta}) = -\sum_{i=1}^{N} n_i x_{ik} \pi_i (1-\pi_i) x_{ik}$.

Let $\boldsymbol{\beta}^{(0)}$ represent the vector of initial approximations for each β_k , then the first step of Newton-Raphson can be expressed as:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + [\, -\lambda^{//}(\boldsymbol{\beta}^{(0)})]^{-1} \; \lambda^{/}(\boldsymbol{\beta}^{(0)}) \qquad (3.3.2.1)$$

Let **μ** be a column vector of length N with elements $\mu_i = n_i \pi_i$. Each element of **μ** can be expressed as μ_i = E(y_i), the expected value y_i . Using matrix multiplication, we can show that:

$$\lambda^{/}(\boldsymbol{\beta}) = -\, \mathbf{X}^{'}(\mathbf{y}\text{-}\boldsymbol{\mu}) \qquad (3.3.2.2)$$

is a column vector of length P+1 whose elements are $\frac{\partial(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k}$, as derived in equation (3.3.1.9). Let **W** be a square matrix of order N, with elements $n_i \pi_i (1-\pi_i)$ on the diagonal and zeros everywhere else. Again, using matrix multiplication, we can verify that

$$\lambda^{//}(\boldsymbol{\beta}) = \mathbf{X}^{'}\mathbf{W}\mathbf{X} \qquad (3.3.2.3)$$

is a p+1 × p+1 square matrix with elements $\frac{\partial^2 \lambda(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k^2}$. Now equation (3.3.2.1) can be written as

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + [\mathbf{X}^{'}\mathbf{W}\mathbf{X}]^{-1} \; \mathbf{X}^{'}(\mathbf{y}\text{-}\boldsymbol{\mu}) \qquad (3.3.2.4)$$

We continue to apply equation (3.3.2.4) until there is essentially no change between the elements of **β** from one iteration to the next. At this point, the maximum likelihood estimates are said to have converged, and equation (3.3.2.3) will hold the variance-covariance matrix of the estimates.

3.4 ODDS AND ODDS RATIO

The odds of some event happening (e.g. the event Y = 1) is defined as the ratio of probability that the event will occur divided by the probability that the event will not occur. That is, the **odds** of the event E is given by

$$\text{Odds (E)} = \frac{P(E)}{P(notE)} = \frac{P(E)}{1-P(E)}$$

Example 3.1 (continued from page 34)

An estimate of the probability of having a myocardial infarction for women in the study using the pill is given by P(E_{pill}) = 23/57 = 0.4035. Hence, the odds, amongst these women, of having a myocardial infarction when using the pill, is given by

$$\text{Odds } (E_{pill}) = \frac{0.4035}{1-0.4035} = 0.6764.$$

That is, the probability of having a myocardial infarction is around 2/3rds the probability of not having a myocardial infarction, for women using the pill.

Similarly, for women who are not using the pill, an estimate of the probability of having a myocardial infarction is given by $P(E_{no-pill}) = 35/167 = 0.2096$. The odds of having a myocardial infarction, when not using the pill, is given by

$$\text{Odds }(E_{no-pill}) = \frac{0.2035}{1-0.2096} = 0.2652.$$

Thus the odds are around 1 to 4 that a woman in the study not using the pill will have a myocardial infarction.

The **odds ratio** $R_{A,B}$ that compares the odds of events E_A and E_B (that is, Event E occurring in group A and B, respectively), is defined as the ratio between the two odds; that is

$$R_{A,B} = \frac{odds(E_A)}{odds(E_B)} = \frac{P(E_A)}{1-P(E_A)} \Big/ \frac{P(E_B)}{1-P(E_B)}.$$

Example 3.1 (continued from page 42)

The odds ratio comparing the odds of having a myocardial infarction for women using the pill with the odds of having a myocardial infarction for women not using the pill, is given by

$$R_{pill,no-pill} = \frac{odds(E_{pill})}{odds(E_{no-pill})} = 0.6764/0.2652 = 2.5505.$$

That is, the odds of having myocardial infarction are 2.55 times higher for women using the pill, than for women not using the pill. In particular, if an odds ratio is equal to one, the odds are the same for the two groups.

3.5 INTERPRETATION OF COEFFICIENTS

The interpretation of any fitted model requires that we be able to draw practical inferences from the estimated coefficients in the model. The estimated coefficients must be able to answer the questions that motivated the study. Interpretation involves determining the functional relationship between the response variable and the predictor variable, and appropriately defining the unit of change for the response variable.

3.5.1 Dichotomous Predictor Variables

The link function is the logit transformation g(x) = ln$\{\pi(x)/[1-\pi(x)]\} = \beta_0+\beta_1 x$ for one predictor variable x . We assume that x is coded either as 1 or 0.The log odds ratio (that is, the logarithm of the odds ratio) corresponding to the probability of success when the predictor variable has a value $x = 0$ and the probability of success when the predictor variable has the value $x = 1$, is given by

$$\ln(\psi) = \ln\{\pi(1)/[1-\pi(1)]\} - \ln\{\pi(0)/[1-\pi(0)]\}$$

where

$$\psi = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} = \frac{g(1)}{g(0)}.$$

Now

$$\begin{aligned}\ln(\psi) &= g(1)-g(0)\\ &= \beta_0+\beta_1.1 - (\beta_0+\beta_1.0)\\ &= \beta_1\end{aligned}$$

It follows that the odds ratio is given by $\psi = e^{\beta_1}$

In general, the estimate of the log odds for any predictor variable at two different levels, say x = a versus x = b, is given by

$$\begin{aligned}\ln[\hat{\psi}(a,b)] &= \hat{g}(x=a) - \hat{g}(x=b)\\ &= (\hat{\beta}_0 + \hat{\beta}_1 \times a) - (\hat{\beta}_0 + \hat{\beta}_1 \times b)\end{aligned}$$

$$= \hat{\beta}_1 \times (a - b) \tag{3.4.1.1}$$

and the estimated odds ratio is

$$\hat{\psi}(a,b) = \exp[\ \hat{\beta}_1 \times (a - b)\,] \tag{3.4.1.2}$$

where

$$\hat{\psi}(a,b) = \frac{\hat{\pi}(x = a)/(1 - \hat{\pi}(x = a)}{\hat{\pi}(x = b)/(1 - \hat{\pi}(x = b))}$$

is used to represent the odds ratio in equations (3.4.1.1) and (3.4.1.2).

The end points of the confidence interval for the odds ratio given in equation (3.4.1.2) are

$$\exp[\hat{\beta}_1(a - b) \pm z_{1-\frac{\alpha}{2}} \,|\, a - b \,|\, \times S\hat{E}(\hat{\beta}_1)\,]$$

3.5.2 Polytomous Predictor Variables

In paragraph 3.2.2 we mentioned that if a nominal scale variable has more than two levels, say k levels, we must model the variable using a collection of k-1 design variables as illustrated in Table 3.1. With this method, we choose one level of the variable to be the reference level usually the 0 level, against which all other levels are compared. We fit the model using design variables to obtain coefficients equal in number to the number of design variables.

Fitting the model using Table3.1 will give the following results with regard to coefficients: (Here the category 'white' is used as reference category)

Variable	Estimated Coefficient
Black Other	$\hat{\beta}_{11}$ $\hat{\beta}_{12}$

Table 3.2 An example showing coefficients that will be obtained when fitting the model using design variables in Table 3.1

Comparing Whites with Blacks we obtain

$$
\begin{aligned}
&\ln[\hat{\psi}(black, white)] \\
&= \hat{g}(black, white) \\
&= \hat{\beta}_0 + \hat{\beta}_{11} \times (D_1 = 1) + \hat{\beta}_{12} \times (D_2 = 0) - (\hat{\beta}_0 + \hat{\beta}_{11}(D_1 = 0) + \hat{\beta}_{12}(D_2 = 0) \\
&= \hat{\beta}_{11}
\end{aligned}
$$

Similarly, comparing others and with whites we obtain:

$$\ln[\hat{\psi}(other, white)] = \hat{\beta}_{12}$$

Thus the odds ratio of any level with the reference level will be the exponential of the coefficient of that level. If comparison is not with a reference level, the odds ratio will be the exponential of the difference between the coefficients in question.

The limits for a 100(1-α) percent CI for the coefficient are

$$\hat{\beta}_{ij} \pm z_{1-\frac{\alpha}{2}} \times S\hat{E}(\beta_{ij})$$

and the corresponding limits for the odds ratio are

$$\exp[\hat{\beta}_{ij} \pm z_{1-\frac{\alpha}{2}} \times S\hat{E}(\hat{\beta}_{ij})].$$

3.5.3 One Continuous Predictor Variable

We assume that the logit is linear in the continuous predictor, x, then the equation of the logit is

$$g(x) = \beta_0 + \beta_1 x.$$

The log odds for a change of c units in x is obtained from the logit difference $g(x+c) - g(x) = c\beta_1$ and the associated odds ratio is obtained by exponentiating this logit difference, $\psi(c) = \psi(x+c, x) = \exp(c\beta_1)$. An estimate may be obtained by replacing β_1 with its maximum likelihood estimate $\hat{\beta}_1$. The end points of the 100(1-α) percent CI estimate $\psi(c)$ are

$$\exp[c\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} cS\hat{E}(\hat{\beta}_1)]$$

3.5.4 Multivariable Case

We now face the situation in which the model contains two predictor variables, where one variable is dichotomous say, x_1 coded 0 and 1 and one continuous, x_2 with primary interest focused on the effect of the dichotomous variable. The equation of the logit will then be $g(x_{1,}x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. If x_1 changes from 0 to 1 with x_2 = a i.e. held constant, then the log odds ratio is:

$$\begin{aligned}&\ln(\psi) = g(x_1 = 1, x_2 = a) - g(x_1 = 0, x_2 = a)\\&= \beta_0 + \beta_1.1 + \beta_2.a - (\beta_0.0 + \beta_2.a)\\&= \beta_1\end{aligned}$$

and the odds ratio is ψ = e^{β_1}

Similarly, holding x_1 constant when x_2 changes from x to $x+c$ the odds ratio is $\psi = e^{c\beta_1}$. Confidence intervals are calculated as before.

3.5.5 One Dichotomous and one Continuous and their Interaction

If the primary interest is focused on the effect of the dichotomous variable x_1 coded 0 and 1 and x_2 is the continuous covariate, then the equation of the logistic interaction is

$$g(x_{1,}x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 .$$

If x_1 changes from 0 to 1 and $x_2 = a$ the log odds ratio is

$$\begin{aligned}&\ln(\psi) = g(x_1 = 1, x_2 = a) - g(x_1 = 0, x_2 = a)\\&= \beta_0 + \beta_1.1 + \beta_2.a + \beta_3.a - (\beta_0 + \beta_1.0 + \beta_2.a + \beta_3.0.a)\\&= \beta_0 + \beta_3.a\end{aligned}$$

The odds ratio is thus $\psi = e^{\beta_1 + a\beta_3}$ which does not depend on the variable of interest only. The 100(1-α) percent CI for the odds ratio is

$$\exp[\hat{\beta}_1 + \hat{\beta}_3.a \pm z_{1-\frac{\alpha}{2}} S\hat{E}(\hat{\beta}_1 + \hat{\beta}_3.a)]$$

where

$$S\hat{E}(\hat{\beta}_1 + \hat{\beta}_3.a) = \hat{\text{v}}\text{ar}(\hat{\beta}_1) + a^2\ \hat{\text{v}}\text{ar}(\hat{\beta}_2) + 2a\hat{C}ov(\hat{\beta}_1, \hat{\beta}_2)$$

3.6 TESTING FOR THE SIGNIFICANCE OF THE MODEL

3.6.1 The Likelihood Ratio Test

After fitting a particular multiple logistic regression model, we do an assessment of the model. We begin by assessing the significance of the p regression coefficients in the model. A likelihood ratio test for overall significance of the p coefficients for the predictor variables in the model is performed. This test is based on the statistic

$$G = 2[L_p(\boldsymbol{\beta}) - L_p(0)]$$

Under the null hypothesis that the coefficients for the predictors in the model are all equal to zero, the distribution of G will be a chi-square with p degrees of freedom. The exceedance probability value (P-value) for the test is P= Pr[$\chi^2(p)$ > G]. Rejection of the null hypothesis leads to the conclusion that at least one and perhaps all p coefficients are significantly different from zero.

3.6.2 Wald Test Statistics

Before we conclude that all of the coefficients are nonzero, we may wish to look at the univariate Wald test statistics:

$$W_j = \frac{\hat{\beta}_j}{S\hat{E}(\hat{\beta}_j)}\ .$$

This test is commonly used to test the significance of the individual logistic regression coefficients for each independent predictor variable (that is, to test the null hypothesis in logistic regression that a particular logit (effect) coefficient is zero). It is the ratio of the logit coefficient to its standard error and is approximated by the standard normal distribution under the said null hypothesis.

3.6.3 Using Deviances to Compare Likelihoods

Suppose that model one has t parameters while model two is a subset of model one with only r of the t parameters so that r < t. Model one will have a larger log-likelihood than model two. For large sample sizes, the difference between these two likelihoods, when multiplied by two, will behave like the chi-square distribution with t-r degrees of freedom. This fact can be used to test the null hypothesis that the t-r parameters that are not in model two (as above) are zero. The difference denoted by D is calculated using results from statistical packages, as follows:

$$D = -2[(\text{model } 2) - (\text{model } 1)]$$
$$= -2\log L\ (\text{model } 2) - -2\log L\ (\text{model } 1),$$

and D $\sim \chi^2(t-r)$, when the sample size is large.

3.7 INTERACTION AND CONFOUNDING

The term confounding is used by epidemiologists to describe a covariate that is associated with both the outcome variable of interest AND a primary predictor variable or risk factor. When both associations are present then the relationship between the risk factor and the outcome variable is said to be confounded.

Consider a model containing a dichotomous risk factor variable and a continuous covariate. If the association between the covariate and the outcome variable is the same within each level of risk factor, there is no interaction between the covariate and the risk factor. Graphically the absence of interaction yields a model with two parallel lines of outcome variable on covariate, one for each level of risk factor variable. In general, the absence of interaction is characterised by a model that contains no product terms involving two or more variables.

When interaction is present, the association between the risk factor and the outcome variable differs or depends in some way on the level of the covariate. That is, the covariate modifies the effect of the risk factor. The term 'effect modifier' is used by epidemiologists to describe a variable that interacts with a risk factor.

Determining if a covariate is an effect modifier and/or a confounder involves several issues. Determining effect modification status involves the parametric structure of the logit, while determination of confounder status involves two things. First, the covariate must be associated with the outcome variable. This implies the logit must have a nonzero slope in the covariate. Second, the covariate must be associated with the risk factor.

In practice, the confounder status of a covariate is ascertained by comparing the estimated coefficient for the risk factor variable from models containing and not containing the covariate. Any "biologically important" change in the estimated coefficient for the risk factor would dictate that the covariate is a confounder and should be included in the model, regardless of the statistical significance of the estimated coefficient for the covariate. On the other hand, we believe that a covariate is an effect modifier only when the interaction term added to the model is biologically meaningful and statistically significant. When a covariate is an effect modifier, its status as a confounder is of secondary importance and the estimate of the effect of the risk factor depends on the specific value of the covariate.

3.8 VARIABLE SELECTION FOR LOGISTIC REGRESSION

According to Hosmer and Lemeshow (1989), in logistic regression the errors are assumed to follow a binomial distribution and the significance of a variable is assessed via the likelihood ratio chi-square. At any step in the procedure the most important variable in statistical terms will be the one that produces the greatest change in the log-likelihood relative to the model not containing the variable.

3.8.1 Purposeful Selection of Variables

3.8.1.1 Screening of Variables

This method is almost similar to the one discussed in section (2.2.1) under the proportional hazards regression model. This method is also analyst driven.

Hosmer and Lemeshow (1989) suggest that the selection process should begin with a univariate analysis of each variable. Hence it is suggested that the selection process should begin with a careful univariate analysis of each variable. For nominal, ordinal, and continuous predictor variables with few integer values, it is suggested this be done with a contingency table of outcome (y= 0, 1) versus the k levels of the predictor variable. The likelihood chi-square test with k-1 degrees of freedom is exactly equal to the value of the likelihood ratio test for the significance of the coefficients for the k-1 design variables in a univariate logistic regression model that contains that single predictor variable.

Particular attention should be paid to any contingency table with a zero cell. Strategies for handling zero cells include: collapsing the categories of the predictor variable in some sensible way to eliminate the zero cells: eliminating the categories completely: or, if the variable is ordinally scaled, modelling the variable as if it is continuous.

For continuous predictor variables the most desirable univariate analysis involves fitting a univariate logistic regression with each predictor to obtain the estimated coefficient, the estimated standard error, the likelihood ratio test for the significance of the coefficient, and the univariate Wald statistic.

The completion of univariate analyses is followed by selection of variables for multivariate analysis. Any variable whose univariate test has a P-value<0.25 should be considered as a candidate for a multivariable model along with all other variables of known biologic importance.

The univariate approach has the disadvantage of excluding predictor variables which can collectively be important predictors of outcome, whilst individually weakly linked with the outcome. This problem can be overcome by choosing a significance level large enough to allow the suspect variables to be included.

After fitting the multivariable model, the importance of each variable included in the model should be verified. This should include (a) an examination of the Wald statistic for each variable and (b) a comparison of each estimated regression coefficient with the coefficient from the univariate model containing only that specific variable. Variables that do not contribute to the model based on these criteria should be eliminated and a new model should be fitted. Comparison of models is done through the likelihood ratio test. Also, estimated coefficients for any remaining variables should be compared to those of the full model. Marked change in magnitude would imply that one or more of the excluded variables were important in the sense of providing a necessary adjustment of the effect of variables that remained in the model. This process is done repeatedly until it appears that all of the important variables are included in the model and those excluded are either biologically or statistically unimportant.

3.8.1.2 Scale of Continuous Predictors

For continuous scaled predictor variables we must check the assumption of linearity in the logit. Since the concept of scale selection is the same for the multivariable models, we describe this approach using the univariable model. One method to ascertain linearity is to plot the fitted line on the scatter-plot of the logit versus the predictor variable and look for any obvious systematic deviations from the line. A modification of this approach is to break the range of the predictor variable into groups and, for each group, plot the average value of the logit versus the group midpoint. This approach in logistic regression requires that we transform the vertical axis to the logit. Thus we would plot, for each group, the logit of the group mean versus the midpoint of the group. The plot is examined with respect to the shape of the resulting “curve”.

An alternative to scale identification in logistic regression is the Box-Tidwell transformation for linear regression. According to Hosmer and Lemeshow (1989), the use of this transformation has been examined for use in logistic regression by Guero and Johnson (1982). This approach adds a term of the form $x\ln(x)$ to the model. If the coefficient for this variable is significant, we have evidence for non-linearity in the logit. This procedure, however, has low power in detecting small departures from linearity.

3.8.1.3 Inclusion of Interactions

Once continuous variables are on the correct scale, we begin to check for interactions in the model. An interaction between two variables implies that the effect of one of the variables is not constant over levels of the other. For example, an interaction between sex and age would imply that the regression coefficient for age is different for males and females. The need to include interaction terms in a model is assessed by first creating the appropriate product of the variables in question and then using a likelihood ratio test to assess their significance (that is their contributions to the model). (See paragraph (3.5.3)). In general, for an interaction term to alter both the point and interval estimates, the estimated coefficient must attain at least a moderate level of statistical significance. The final decision as to whether an interaction term should be included in a model should be based on statistical as well as practical considerations.

3.8.2 Stepwise Forward Selection

This procedure starts by fitting only the intercept term, then for each of the possible predictor variables, a univariate logistic regression containing the intercept and that predictor (say x_j) is fitted. The log- likelihood of the intercept model (L_0) is compared with the log-likelihood of each of the univariate model (L_j) by means of the ratio test statistic:
$G_j = 2(L_j - L_0)$.

Its P-value is determined by $P = \Pr(\chi^2(v) > G_j)$, where ν=1 if x_j is continuous and ν= k-1 if x_j has k categories. The most important predictor variable is the one with minimum P-value and this variable, denoted by $x_{e,}$ is entered into the model. The subscript "e" indicates that the variable is a candidate for entry. The choice of an "alpha"(significance level) level used to judge the importance of variables is a crucial aspect. Let α_E denote our choice where the "E" stands for entry and this choice for α_E will determine how many variables will eventually be included in the model. Choosing a value for α_E in the range 0.15 to 0.2 is highly recommended. Moreover, using

α_E in this range will provide assurance that the procedure selects variables whose coefficients are different from zero (Hosmer and Lemeshow (1989)).

After the variable x_e has been entered, the next step is to determine whether any of the remaining p-1 variables are important once x_e is in the model by fitting the p-1 logistic regression models containing x_e and x_j, j = 1,2,3 ….. p and j $\neq$ e. The log-likelihoods of these models are compared with that of the model containing the intercept and x_e. The variable with the smallest P-value at this step is entered, and the algorithm continues provided P-value$<\alpha_E$, otherwise it stops.

3.8.3 Stepwise Backward Selection

The process starts with a full model containing all variables. In the first step the log-likelihood of the model containing all variables (L_f) is compared to that of p-1 variables with x_j is removed denoted by (L_{-j}) by using the likelihood ratio test statistic

$G_{-j} = 2(L_f - L_{-j})$.

To ascertain which variable should be deleted from the model, we select that variable which, when removed, gives the maximum P-value. We denote the minimal level of continued contribution to the model by α_R where "R" stands for remove. The value we choose for α_R must exceed the value for α_E, to avoid the possibility of having to enter and remove the same variable at successive steps.

In the next step the log- likelihood of the model excluding the one removed at the previous step is compared to those of all p-1 models with one of the remaining variables removed. If P-value$>\alpha_R$, a variable is removed. Generally the choice of α_R is 0.2 or 0.25. However, important variables can be forced to remain in the model.

The algorithm stops when all variables have entered the model or when all variables in the model have P-values to which is less than α_R.

3.8.4 Stepwise Selection (Forward and backward)

This is a combination of forward and backward selection procedures discussed above. It is based on a statistical algorithm that allows moves in either direction, dropping or adding variables at various steps based on the 'importance' of variables. The 'importance' of a variable refers to the statistical significance of its coefficient. Since, in logistic regression the errors are assumed to follow a binomial distribution, the significance is assessed via the likelihood ratio chi-square test. Thus at any step in the procedure the most important variable will be the one that result in the largest likelihood ratio statistic, G.

Since the magnitude of G depends on its degrees of freedom, any procedure based on the likelihood ratio test statistic, G must account for possible differences of degrees of freedom of variables. This is achieved by assessing significance through the p-value for G.

3.8.5 Best Subset Selection

This is an alternative to stepwise selection. This model building approach has been available in linear regression. Typical software implementing this method for linear regression will identity a specified number of 'best' models containing one, two, three variables, and so on, up to the single model containing all p variables. According to Hosmer and Lemeshow (1989), we may use any best subsets linear regression program to execute the computations for best subsets logistic regression.

The subsets of variables selected for 'best' models depend on the criterion for 'best'. In logistic regression the Score and the C_p criteria are preferred. A model with high score- value will be preferred to a model with a smaller score-value whereas a model with a small C_p value or $C_p \approx$ r will be preferred where r is the number of predictor variables in the model. It is important to note that variables suggested by best subset strategy should not be accepted without considerable critical evaluation.

Though we discussed several selection procedures in Chapter 2, a few of them have been discussed, and others left out in this chapter. The reason is that such procedures do not apply to the logistic regression

3.8.6 General

From the information in this chapter, it is clear that selection methods for binary outcome variables are lacking. For this reason, we will be evaluating a new method, based on the ROC curve, briefly in Chapter 5. We will first discuss the concept of a ROC curve in Chapter 4.

Chapter 4

THE RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

4.1 BACKROUND

We discuss ROC curves as a separate chapter because we will be endeavouring (chapter 5) to utilise these curves as additional model (or variable) selection method. Specifically: the area under the curve (AUC) will be evaluated as a selection criterion. The AUC will be discussed in section 4.5.

Researchers and analysts allocate a great deal of effort to the development of prediction models to support decision making. However, too often insufficient attention is allocated to the tool(s) used to evaluate the model(s) in question. The issue is that accurate prediction models may be measured inappropriately based upon the information available regarding classification error rate and the context of application. In the end, poor decisions are made because of selecting wrong models, using an inappropriate evaluation method.

In the context of consumer risk prediction, understanding how to evaluate models which predict potential customers to be ‘good’ or ‘bad’ credit risks is critical to managing Customer Relationship Management (CRM). Since the dependent variable of concern is categorical, the issue is one of binary classification. For a binary classification problem (i.e. prediction of ‘good’ versus ‘bad’), logit analysis utilises a linear combination of the predictor variables and transforms the result to lie between 0 and 1, to equate to a probability.

One method of evaluation, which enables a comprehensive analysis of all possible error severities, is the Receiver Operating Characteristic (ROC) curve. According to Morrison & Michelle (2005), ROC curves were developed in the field of statistical decision theory, and later used in the field of signal detection during WW II. ROC curves enabled radar operators to distinguish between an enemy target, a friendly ship, or noise. They further point out that ROC curves assess the value of diagnostic tests by providing a standard measure of the ability of the test to correctly classify subjects. Mention is made of Metz (1978) stating that the biomedical field uses ROC curves extensively to assess the efficacy of diagnostic tests in discriminating between healthy and diseased individuals. ROC curves have since been used in fields ranging from electrical engineering and weather prediction to Psychology and are used almost everywhere in the literature on medical testing to determine the effectiveness of medications (Nargundkar and Priestly (2003)).

4.2 DEFINITION OF AN ROC CURVE

Consider diagnostic tests with dichotomous outcomes, with positive outcomes suggesting presence of disease. For dichotomous tests, there are two potential types of error. A false- positive error happens when a non-diseased individual has a positive test result. On the other hand, a false-negative error happens when a diseased individual has a negative test result. The rates of occurrence of these errors, termed false-positive and false negative rates, together constitute the operating characteristics of the dichotomous diagnostic test. These notions can be generalised to non-binary tests in this way: Let *D* be a binary (0/1) indicator of the disease status with $D = 1$ for diseased subjects. Let *Y* denote the test result with the convention that larger values of *Y* are more indicative of disease for some threshold value C. Now 1 minus the false-negative rate (or true positive rate) and 1 minus true negative rate (false-positive) associated with this decision criterion can be written as P*r* ($Y \geq C \left| D = 1 \right.$) and Pr(Y<C|D=0), respectively. An ROC curve is a plot of the true positive rate versus 1 minus true negative rate across all positive threshold values, *C*. When *Y* is continuous, a clear and brief way of writing the ROC curve is ROC(t) = $F_D \left\{ F_{\bar{D}}^{-1}(t) \right\}$ t $\in (0,1)$, where F_D and $F_{\bar{D}}$ are the survivor functions of Y in the diseased and non-diseased populations, respectively, and where *t* is the false positive rate which varies from 0 to 1 as the corresponding implicit threshold value, *C*, varies from ∞ to -∞ . When Y is discrete the ROC curve can also be written in the form $F_D \left\{ F_{\bar{D}}^{-1}(t) \right\}$ but the domain for ROC (t) is restricted to the range of $F_{\bar{D}}(.)$, that is, the set of all possible false positive rates associated with the test. By definition, the ROC curve is a monotone increasing function from $[0,0]$ to $[1,1]$

4.3 DIAGNOSTIC TEST INTERPRETATION

The basic idea of diagnostic test interpretation is to calculate, for example, the probability that a patient has a disease under the consideration given certain result. A 2 by 2 table is employed in this regard (See Table 4.3.1).

4.3.1 2 X 2 Table or Contingency Matrix

	Disease Present	Disease Absent	
Test Positive	True Positives (TP)	False Positives (FP)	Total Positive
Test Negative	False Negatives (FN)	True Negatives (TN)	Total Negative
	Total with Disease	Total without Disease	Grand Total

Table 4.1 An example of a Contingency Table

4.3.2 Basic Concepts

In this discussion we refer back to Table 4.1.

4.3.2.1 Sensitivity

Sensitivity is the proportion of patients with disease whose tests are positive.

$P(T+|D+)=TP/(TP+FN)$

High sensitivity is important when:

- The disease is serious and should not be missed.
- The disease is treatable.
- *FP* results do not lead to serious physic, psychological

or economic trauma to the patient.

4.3.2.2 Specificity

Specificity is the proportion of patient without disease whose tests are negative.

P(T-|D-) = TN/ (TN + FN)

High specificity is needed when:

- The disease is serious.
- The disease is not treatable or curable.
- *FP* results do not lead to serious physic, psychological or economic trauma to the patient.

4.3.2.3 Pre-test Probability

Pre-test probability is the prevalence of the disease in the population. It is also called efficiency of the test.

P(D+) = (TP+N)/(TP+FP+TN+FN)

Higher Efficiency is needed when:

- The disease is serious.
- The disease is curable
- FP and FN are essentially equally serious damages.

4.3.2.4 Predictive Value of a Positive Test

Predictive values of a positive test is the proportion of patients with positive tests who do have disease.

P(D+|T+) = TP/(TP+P)

These values measure:

- The same thing as posttest probability of disease given a positive test.
- Measures how well the test rules in disease.

4.3.3.5 Predictive Value of a Negative Test

Predictive value of a negative is the proportion of patients with negative tests who do not have disease.

P(D-|T-) = TN/(TN+N)

This value measures how well the test rules out the disease.

4.4 ROC REGRESSION MODEL

Let the false positive rate be denoted by t and let τ denote the set of possible values for t, namely the range of $F_{\bar{D}}$, which is a subset of [0, 1]. Let Z denote some factors which potentially influence test accuracy and let X be a corresponding vector of covariates. For example, if Z is a categorical variable, X might be the associated vector of dummy variables. The covariate vector X is a function of the factors Z. We write the ROC curve associated with Z as $ROC_z(t)$ and model it as

$$ROC_z(t) = g\{\alpha_0(t), \beta X\} \ (\mathrm{t} \in \tau_z),$$

where α_o(t) is a univariate baseline function of t, βX is a linear predictor which characterises the effect of the covariates X on the ROC curve, g is a known function and τ_z denotes the domain of the ROC function associated with Z. In general the covariate vector X may include interactions between factors in Z and t, in which case we write the covariate vector X(t). Since the ROC curve is a monotone increasing function by definition, g and α must be chosen such that monotonicity in ROC$_z$ is ensured.

4.5 AREA UNDER THE ROC CURVE (AUC)

4.5.1 Interpretation of the Area

The area under the ROC curve is commonly used as a summary measure of diagnostic accuracy. It takes values from 0.5 to 1.0. The AUC statistic can be interpreted as the probability that the test result from a randomly chosen diseased individual is more indicative of disease than that from a randomly chosen non-diseased individual or a measure of a model's ability to discriminate between those who experience the outcome of the interest versus those who do not. $AUC = P(X_i \geq X_j | D_i = 1, D_j = 0)$. An ROC curve summarises the possible set of 2 X 2 matrices that results when the cut-off value is varied continuously from its highest possible value down to its smallest possible value. An area of 1 represents a perfect discrimination. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the

discrimination. On the other hand an area of 0.5 represents a worthless discrimination. The closer the curve comes to the 45 degrees diagonal of the ROC space, the less accurate the test.
An area of

- 0.9 – 1.0 = excellent discrimination
- 0.80 -0.90 = good discrimination
- 0.70 -0.80 = fair discrimination
- 0.60 -0.70 = poor discrimination
- 0.50 – 0.60 = fail, i.e. no discrimination

However, in practice it is extremely unusual to observe areas under the curve greater than 0.9.

4.5.2 Comparison of Tests

When results from multiple tests have been obtained, the ROC plots can be graphed together. The relative positions of the plots indicate the relative accuracies of the tests. A plot lying above and to the left of another plot indicates greater observed accuracy. If the curves for two tests cross, a meaningful difference between the tests over the range of interest might not be picked up by the AUCs.

If we have two curves of similar area and we wish to decide whether the two curves differ significantly, we can use bivariate statistical analysis.

Where we have different areas derived from two tests applied to different sets of cases, it is appropriate to calculate the standard error of the difference between the two areas, thus:

$$SE_{(A_1-A_2)} = \sqrt{SE_{A1}{}^{2} + SE_{A2}{}^{2})}$$

This approach is not appropriate where two sets are applied to the same set of patients. Hanley and McNeil (1982) show that in these circumstances, the correct formula is:

$$SE_{(A_1-A_2)} = \sqrt{SE^{2}{}_{A_1} + SE^{2}{}_{A_2} - 2r.SE_{A_1}SE_{A_2}}$$

where r is the quantity that represents the correlation induced between the two areas by the study of the same set of cases.

Once we have the standard error of the difference in areas, we can then calculate the statistic:

$Z = (A_1 - A_2) / (SE_{(A_1 - A_2)})$

If Z is above a critical level, then we accept that the two areas are different. Commonly this critical value is set at 1.96, and we then have a 0.05 chance of making a type I error in rejecting the hypothesis that the two curves are similar.

Assuming we have two tests T1 and T2 that classify our cases into either normal (n) or abnormal (a), and we have already calculated the AUCs for each test, r is calculated as follows:

1. Look at (n), the non-diseased patients. We find how the two tests correlate for these patients and obtain a value r_n for this correlation.
2. Similarly we derive r_a, the correlation between the two tests for the patients
3. Average r_n and r_a.
4. Average out the areas A_1 and A_2 by calculating $(A_1 + A_2)/2$.
5. Look up the value of r in Hanley and McNeil's Table I (Hanley and McNeil (1982)) given the the average areas of r_n and r_a.

4.5.3 Advantages and Disadvantages of ROC

The ROC plot is a simple, graphical and easily appreciated visually. It is a comprehensive representation of pure accuracy, i.e. discriminating ability, over the entire range of a test. It provides a direct visual comparison between tests on a common scale and it requires no grouping and binning of data. With appropriate software, ROC plotting is quite readily done.

Actual decision thresholds are usually not displayed in the plot. The number of subjects is also not shown on the display and as the sample size decreases, the ROC plot tend to become increasingly jagged and bumpy. However, even with a large number of subjects, the plot may be bumpy.

CHAPTER 5

MODEL BUILDING USING REAL DATA

In this chapter we will look at the application of the procedures and methods outlined in chapters 3 and 4 with regard to selection of variables. Some of the criteria, discussed in Chapter 2, such as the Akaike Information Criterion may come into play since they also are applicable to logistic regression and needless to say, Cox regression as well.

The data set to be used was developed for a study of factors associated with success of first year students at the Tshwane University of Technology (TUT) from the year 1999 to 2002. Information on 18047 students was obtained.

Table 5.1 describes the response, predictor variables and their codes.

Variable	Description and code
Pass	pass=1, fail=0
Campuss	main campus=1, satellite campus =2
Genderr	female = 1, male=2
Agregate	aggregate mark for all subjects in matric exam for an individual student
Maritall	marital status (single=1, married=2)
Finaidd	Financial aid (aided=1, not aided)
Age	student age at first registration
English	Performance in English in matric exam (good=1,not good=2)
Race	(white=1, coloured=2, Asian=3 and black=4)
Faculty	(Engineering=1, Commerce =2, Social Science=3, Arts=4, Natural Science =5, Agricultural Science=6 and Health =7)

Table 5.1 Code Sheet of the Variables used in the Data set for the Study of Factors Associated with Success of First Year Students at TUT from 1999 to 2002

5.1 PURPOSEFUL SELECTION OF VARIABLES

We begin with a univariable description of all predictors; both categorical and continuous variables are shown in Tables 14 and 15 of the appendix respectively.

The univariable analysis does not reveal any variable for which there are illegal values. All binary variables are coded as 1; 2. Race and Faculty are the only non-binary categorical variables. We create indicator variables for the Faculty variable as shown in Table 5.2:

Faculty	Label	faculty_2	faculty_3	faculty_4	faculty_5	Faculty_6	Faculty_7
1	Engineering	0	0	0	0	0	0
2	Commerce	1	0	0	0	0	0
3	Social Sci	0	1	0	0	0	0
4	Arts	0	0	1	0	0	0
5	Natural Sci	0	0	0	1	0	0
6	Agric Sci	0	0	0	0	1	0
7	Health	0	0	0	0	0	1

Table 5.2 Indicator Variables for the Variable Faculty.

Since the numbers of Indians and Coloureds were quite small, each less than 2% of the total, a dichotomous variable Brace (black race for blacks) was created. Brace takes the value 1 if race is black and the value 0 for other races (White, Coloured and Indian). The dependent variable was the logit $\pi = (\log\pi/(1-\pi))$, where π is the probability that a student passed.

Univariable logistic regressions were fitted to the data and the results are given in Table 5.3.

Predictor Variable	Estimated Coefficient	Estimated Standard Error	Estimated Odds ratio	Wald Test 95% CI	P-value
Age	-0.0552	0.00726	0.759	(0.707,0.815)	<0.0001
Agregate	0.00287	0.000078	1.267	(0.673,1.287)	<0.0001
Campuss	-0.1645	0.0172	0.720	(0.673,0.770)	<0.0001
Maritall	0.2043	0.0762	1.505	(1.116,2.028)	0.0073
Finaidd	0.3483	0.0264	2.007	(1.810,2.225)	<0.0001
Genderr	0.1662	0.0167	1.394	(1.306,1.489)	<0.0001
English	0.3890	0.0199	2.177	(2.014,2.353)	<0.0001
Faculty_2	0.2447	0.0586	1.277	(1.139,1.433)	<0.0001
Faculty_3	0.5835	0.0620	1.792	(1.587,2.024)	<0.0001
Faculty_4	1.8045	0.0757	6.077	(5.239,7.048)	<0.0001
Faculty_5	0.7191	0.0744	2.053	(1.774,2.375)	<0.0001
Faculty_6	-0.0894	0.0866	0.914	(0.772,1.084)	0.3020
Faculty_7	1.2743	0.0288	3.576	(3.040,4.207)	<0.0001
Brace	-0.9388	0.0344	0.391	(0.366,0.418)	<0.0001

Table 5.3 Univariable Logistic Regression Models

For the variables Age and Agregate in Table 5.3 odds ratios are for an increase of 5 years and 100 marks respectively. A change of 1 mark or 1 year would not be meaningful.

With the exception of variables Faculty_6 and Agregate, there is evidence that each of the variables has some association with the outcome variable, pass. This is based on the observation that the confidence interval estimates do not contain 1. Furthermore, all variables are significant with P-value≤0.25 for the Wald test. We now, based on the univariable results, begin the multivariable model including all variables besides Faculty_6 which is not significant. The model is shown in Table 5.4.

The Wald statistics is now used to delete variables one by one that do not appear to be significant at the P-value≤0.05 level, starting with the least significant one.

```
                              Intercept   Intercepts and
              Criterion         Only         Covariates

              AIC             21460.178      19424.526
              SC              21467.979      19533.737
              -2 Log L        21458.178      19396.526

              Testing Global Null Hypothesis: BETA=0

      Test                   Chi-Square       DF      Pr > ChiSq

      Likelihood Ratio        2061.6516       13          <.0001
      Score                   2085.0164       13          <.0001
      Wald                    1798.0246       13          <.0001


             Analysis of Maximum Likelihood Estimates

Parameter       DF     Estimate    Std  Error  Wald Chi-Square   Pr > ChiSq

Intercept        1      -2.5845      0.2700        91.6291       <.0001
age              1      -0.0101     0.00855         1.3903       0.2384
agregate         1      0.00162    0.000094       298.6322       <.0001
Campuss          1       0.0662      0.0255         6.7383       0.0094
maritall         1       0.0637      0.0946         0.4535       0.5007
finaidd          1       0.3634      0.0282       165.6272       <.0001
genderr          1       0.1001      0.0186        29.0649       <.0001
english          1       0.0778      0.0237        10.7918       0.0010
faculty_2        1       0.6209      0.0564       121.0806       <.0001
faculty_3        1       0.5730      0.0581        97.1171       <.0001
faculty_4        1       1.6185      0.0878       340.1683       <.0001
faculty_5        1       0.7232      0.0862        70.4490       <.0001
faculty_7        1       1.1871      0.0815       211.9119       <.0001
Brace            1      -0.5585      0.0437       163.3471       <.0001

                          Odds Ratio Estimates
                                     Point          95% Wald
                Effect             Estimate      Confidence Limits

                age                   0.990        0.974        1.007
                agregate              1.267        0.673        1.287
                Campuss   1 vs 2      1.142        1.033        1.262
                maritall  1 vs 2      1.136        0.784        1.646
                finaidd   1 vs 2      2.068        1.852        2.311
                genderr   1 vs 2      1.222        1.136        1.314
                english   1 vs 2      1.168        1.065        1.282
                faculty_2             1.861        1.666        2.078
                faculty_3             1.774        1.583        1.988
                faculty_4             5.046        4.248        5.992
                faculty_5             2.061        1.741        2.440
                faculty_7             3.278        2.793        3.846
                Brace                 0.572        0.525        0.623
```

Table 5.4 Multivariable Model Containing Variables Identified in the Univariable Analysis.

The model at the end of the process of removing non-significant variables is shown in Table 5.6.

At this point, we allow each of the variables not in the model, the opportunity to re-enter the model one by one. As each variable enters the model, we evaluate its statistical significance using the Wald test and also ascertain whether the variable is a confounder or not of other variables in the model by calculating the extent of change of coefficients of variables in the model.

There is no significant change in the coefficients of other variables when Faculty_6 re-enters the model but according to the Wald test the variable is however, not statistically significant. The same argument holds for the variables Maritall and Age when they re-enter the model. Therefore, the preliminary main-effects model is as given in Table 5.6.

Before proceeding to determine interactions we need to examine the variables that have been modelled as continuous to obtain the correct scale in the logit. In this case the variable we need to check is Agregate.

We start by determining the quartiles of the distribution of Agregate from appendix 1 Table 14 and create three design variables using the lowest quartile as the reference group. The results of the quartile analysis are shown in Table 5.5.

Quartile	Midpoint	Coefficient	95%CI for Odds Ratios
1	775	0	
2	955	0.2898	(1.208,1.478)
3	1137	1.0672	(2.516,3.359)
4	1680	0.9989	(2.407,3.063)

Table 5.5 Results of Quartile Analyses of the Variable Agregate from the Multivariable Model Containing Variables shown in Table 5.6

```
                          Intercept     Intercept and
           Criterion      Only          Covariates
           AIC           21460.178      19423.992
           SC             21467.979      19517.601
           -2 Log L       21458.178      19399.992

           Testing Global Null Hypothesis: BETA=0
  Test                  Chi-Square       DF     Pr > ChiSq
  Likelihood Ratio       2058.1862       11         <.0001
  Score                  2083.3017       11         <.0001
  Wald                   1796.8154       11         <.0001

           Analysis of Maximum Likelihood Estimates
                                    Standard                  Wald
Parameter      DF     Estimate       Error     Chi-Square    Pr > ChiSq
Intercept       1      -2.7409      0.1250      480.9214        <.0001
agregate        1      0.00163    0.000093      304.3700        <.0001
Campuss    1    1       0.0730      0.0252        8.3651        0.0038
finaidd    1    1       0.3637      0.0282      165.9746        <.0001
genderr    1    1       0.1029      0.0184       31.1306        <.0001
english    1    1       0.0837      0.0234       12.8165        0.0003
faculty_2       1       0.6197      0.0564      120.6196        <.0001
faculty_3       1       0.5729      0.0582       97.0702        <.0001
faculty_4       1       1.6244      0.0877      342.9895        <.0001
faculty_5       1       0.7316      0.0861       72.2871        <.0001
faculty_7       1       1.1867      0.0815      211.7846        <.0001
Brace           1      -0.5567      0.0437      162.5179        <.0001

             Odds Ratio Estimates
                        Point          95% Wald
Effect                  Estimate      Confidence Limits
agregate                 1.267       0.673       1.287
Campuss    1 vs 2        1.157       1.048       1.277
finaidd    1 vs 2        2.070       1.853       2.312
genderr    1 vs 2        1.229       1.143       1.321
english    1 vs 2        1.182       1.079       1.296
faculty_2                1.858       1.664       2.076
faculty_3                1.773       1.582       1.988
faculty_4                5.075       4.274       6.027
faculty_5                2.078       1.756       2.460
faculty_7                3.276       2.792       3.844
Brace                    0.573       0.526       0.624

Association of Predicted Probabilities and Observed Responses

Percent Concordant          69.8    Somers' D    0.405
Percent Discordant          29.3    Gamma        0.409
Percent Tied                 0.9    Tau-a        0.164
Pairs                   65896012    c            0.703

Adjusted Odds Ratios
Effect            Unit      Estimate
agregate         100.0         1.177
agregate        -100.0         0.850
```

Table 5.6 Preliminary Main Effects Model

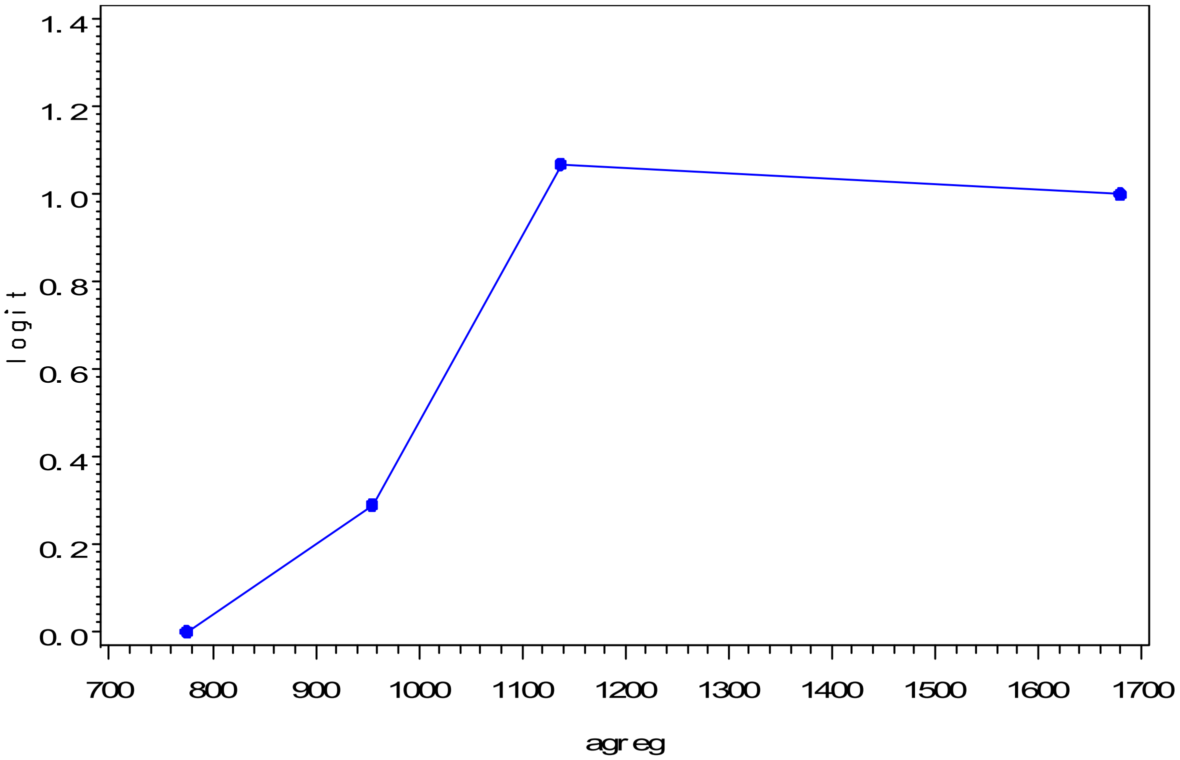
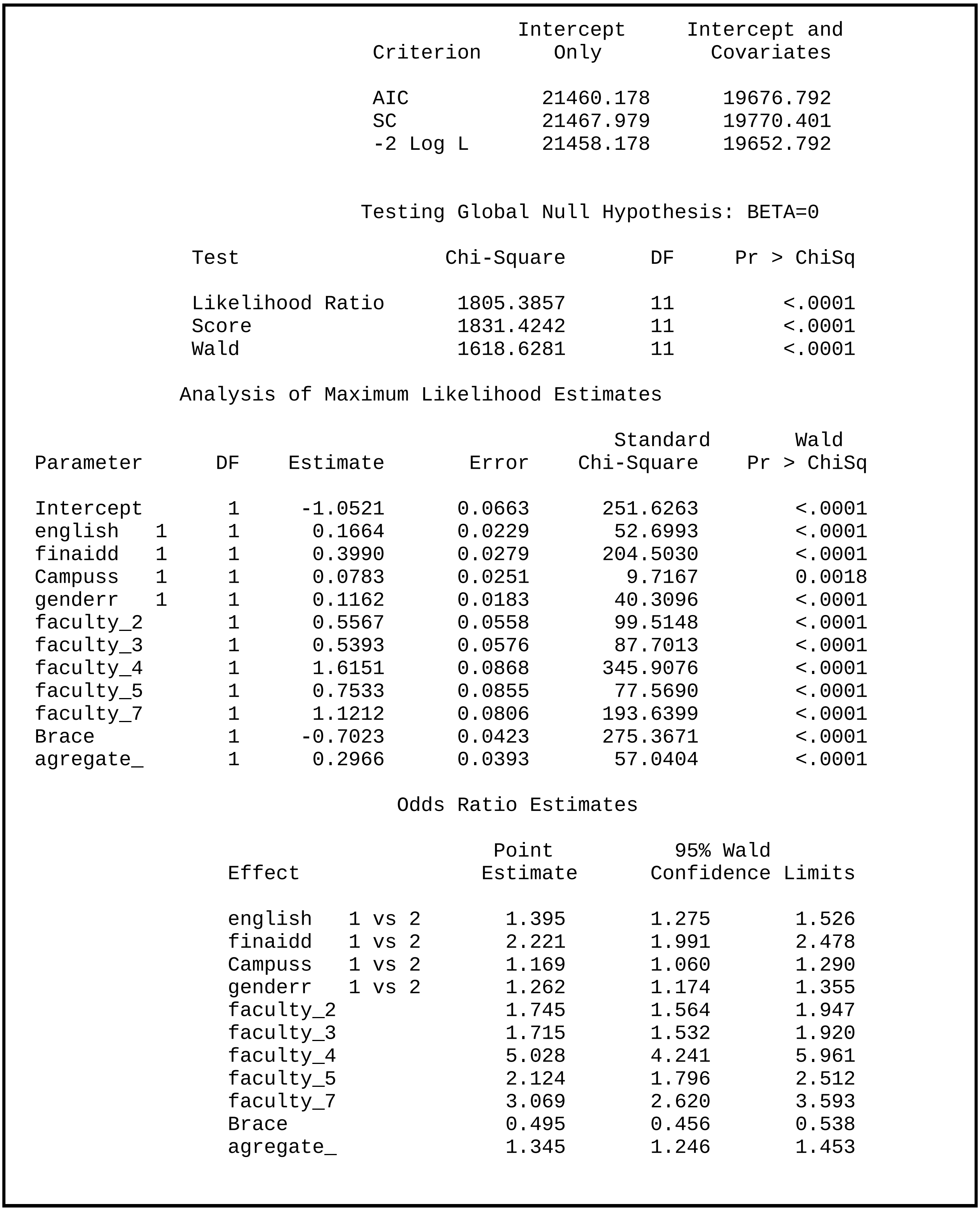


Figure 5.1 Plot of quartile midpoints against coefficients.

The results of plotting quartile midpoints against the coefficients are shown in figure 5.1. The plot of the coefficients supports an assumption of non linearity in the logit. Addition of the variable [Agregate*ln(Agregate)] to the model containing Agregate as a continuous variable yields a significant coefficient for the variable [Agregate*ln(Agregate)]. This confirms that agregate is not linear in the logit.

From Table 5.5 the two coefficients in the third and fourth quartiles are almost similar in magnitude and their confidence intervals have a great deal of overlap. These observations suggest the creation of a dichotomous variable taking on the value 1 if Agregate is in the third and fourth quartiles and the value of zero otherwise as also being supported by figure 5.1.

The results of including a dichotomous variable Agregate_ in the multivariable model are shown in Table 5.7.



```
                              Intercept     Intercept and
                  Criterion      Only         Covariates

                  AIC           21460.178      19676.792
                  SC            21467.979      19770.401
                  -2 Log L      21458.178      19652.792


                  Testing Global Null Hypothesis: BETA=0

           Test                 Chi-Square       DF     Pr > ChiSq

           Likelihood Ratio      1805.3857       11         <.0001
           Score                 1831.4242       11         <.0001
           Wald                  1618.6281       11         <.0001

          Analysis of Maximum Likelihood Estimates

                                                      Standard        Wald
Parameter        DF     Estimate        Error     Chi-Square    Pr > ChiSq

Intercept         1      -1.0521       0.0663       251.6263        <.0001
english    1      1       0.1664       0.0229        52.6993        <.0001
finaidd    1      1       0.3990       0.0279       204.5030        <.0001
Campuss    1      1       0.0783       0.0251         9.7167        0.0018
genderr    1      1       0.1162       0.0183        40.3096        <.0001
faculty_2         1       0.5567       0.0558        99.5148        <.0001
faculty_3         1       0.5393       0.0576        87.7013        <.0001
faculty_4         1       1.6151       0.0868       345.9076        <.0001
faculty_5         1       0.7533       0.0855        77.5690        <.0001
faculty_7         1       1.1212       0.0806       193.6399        <.0001
Brace             1      -0.7023       0.0423       275.3671        <.0001
agregate_         1       0.2966       0.0393        57.0404        <.0001

                            Odds Ratio Estimates

                                   Point          95% Wald
              Effect             Estimate      Confidence Limits

              english    1 vs 2     1.395       1.275        1.526
              finaidd    1 vs 2     2.221       1.991        2.478
              Campuss    1 vs 2     1.169       1.060        1.290
              genderr    1 vs 2     1.262       1.174        1.355
              faculty_2             1.745       1.564        1.947
              faculty_3             1.715       1.532        1.920
              faculty_4             5.028       4.241        5.961
              faculty_5             2.124       1.796        2.512
              faculty_7             3.069       2.620        3.593
              Brace                 0.495       0.456        0.538
              agregate_             1.345       1.246        1.453
```

Table 5.7 Multivariable Model With Dichotomous Variable Agregate_.

We now form all possible two way interaction using the variables in Table 5.7.

engagr=english*agregate_

engfin=english*finaidd

amfin=campuss*finaidd

finfac2=finaidd*faculty_2

finfac3=finaidd*faculty_3

finfac4=finaidd*faculty_4

finfac5=finaidd*faculty_5

finfac7=finaidd*faculty_7

racfin=brace*finaidd

agrbrac=agregate*brace

engbrac=english*brace

The interaction terms are added to the model containing main effects one by one. Table 5.8 shows those interactions that were significant when added one by one to the main effects model. Interactions which are not significant will be excluded from the model. A model with significant interactions is shown in Table 5.9. However, it should be noted that when there is statistically significant interaction, we include the corresponding main effects in the model regardless of their statistical significance.

Table 5.9 gives the final model containing main effects and interactions. From Table 5.10, we see that (12308+1046)=13354 or 73% of the 18047 observations in our data are correctly classified by the logistic regression model in Table 5.9. Of the 5083 observed passes, 1046 or 20.6% are correctly classified as predicted passes. 4037 of these observations are incorrectly classified as predicted fails. They are called false-negatives. Only 656 of the observed fails are incorrectly classified as predicted passes. These observations are called false-positives.

The c statistic in Table 5.11 gives the area under the ROC curve (the AUC) in figure 5.2. This c-value is 0.694 and indicates that the model has low predictive accuracy. But the low predictive accuracy does not imply the model does not fit.

```
                                        Intercept     Intercept and
                          Criterion     Only          Covariates

                          AIC           21460.178      19635.958
                          SC            21467.979      19791.972
                          -2 Log L      21458.178      19595.958


              Analysis of Maximum Likelihood Estimates

                                            Standard          Wald
Parameter      DF     Estimate       Error    Chi-Square    Pr > ChiSq

Intercept       1      -0.8944      0.3361        7.0820        0.0078
Campuss   1     1       0.0806      0.0252       10.2242        0.0014
genderr   1     1       0.1099      0.0184       35.7304        <.0001
finaidd   1     1       0.2188      0.0922        5.6253        0.0177
english   1     1       0.1100      0.1428        0.5936        0.4410
faculty_2       1       0.8804      0.2504       12.3660        0.0004
faculty_3       1       0.5433      0.0577       88.6276        <.0001
faculty_4       1       0.4241      0.4249        0.9962        0.3182
faculty_5       1      -0.0852      0.3634        0.0550        0.8146
faculty_7       1       1.1313      0.0808      195.8363        <.0001
Brace           1      -0.5948      0.3306        3.2371        0.0720
agregate_       1       0.6119      0.1228       24.8345        <.0001
engagr          1      -0.1501      0.0936        2.5717        0.1088
engfin          1      -0.1932      0.1293        2.2326        0.1351
finfac4         1       0.6224      0.2182        8.1332        0.0043
finfac2         1      -0.1775      0.1301        1.8627        0.1723
finfac5         1       0.4584      0.1878        5.9573        0.0147
racfin          1      -0.1765      0.1473        1.4345        0.2310
agrbrac         1      -0.2025      0.0875        5.3520        0.0207
engbrac         1       0.3499      0.1325        6.9747        0.0083
```

Table 5.8 A model containing Interactions which were Significant when Added One by One to the Main Effects Model.

```
                         Intercept    Interaction and
          Criterion      Only         Covariates

          AIC           21460.178      19636.828
          SC            21467.979      19769.441
          -2 Log L      21458.178      19602.828

       Analysis of Maximum Likelihood Estimates

                                                Standard         Wald
Parameter      DF     Estimate        Error    Chi-Square    Pr > ChiSq

Intercept       1      -1.3260       0.0944      197.1637        <.0001
Campuss    1    1       0.0802       0.0252       10.1358        0.0015
genderr    1    1       0.1097       0.0184       35.6720        <.0001
finaidd    1    1       0.3905       0.0441       78.5287        <.0001
english    1    1       0.3305       0.0609       29.4292        <.0001
faculty_2       1       1.0687       0.2342       20.8300        <.0001
faculty_3       1       0.5447       0.0577       89.1103        <.0001
faculty_4       1       0.3639       0.4263        0.7287        0.3933
faculty_5       1      -0.0899       0.3664        0.0602        0.8061
faculty_7       1       1.1313       0.0808      196.0118        <.0001
Brace           1      -0.9373       0.1737       29.1054        <.0001
agregate_       1       0.4453       0.0688       41.8797        <.0001
finfac4         1       0.6557       0.2187        8.9884        0.0027
finfac2         1      -0.2770       0.1212        5.2212        0.0223
finfac5         1       0.4603       0.1892        5.9179        0.0150
agrbrac         1      -0.2363       0.0838        7.9462        0.0048
engbrac         1       0.3694       0.1315        7.8870        0.0050
```

Table 5.9 Final Model with Interactions

		Predicted by Model		
		0	1	Total
Actual Classification	0	12308	656	12964
	1	4037	1046	5083
	Total	16345	1702	18047

Table 5.10 Contingency Matrix for model in Table 5.9

```
                          Odds Ratio Estimates

                                     Point          95% Wald
                Effect             Estimate      Confidence Limits

                Campuss   1 vs 2      1.174       1.064       1.296
                genderr   1 vs 2      1.245       1.159       1.338
                finaidd   1 vs 2      2.183       1.837       2.595
                english   1 vs 2      1.937       1.525       2.459
                faculty_2             2.912       1.840       4.607
                faculty_3             1.724       1.540       1.930
                faculty_4             1.439       0.624       3.318
                faculty_5             0.914       0.446       1.874
                faculty_7             3.100       2.646       3.632
                Brace                 0.392       0.279       0.551
                agregate_             1.561       1.364       1.786
                finfac4               1.926       1.255       2.958
                finfac2               0.758       0.598       0.961
                finfac5               1.585       1.094       2.296
                agrbrac               0.790       0.670       0.931
                engbrac               1.447       1.118       1.872



        Association of Predicted Probabilities and Observed Responses

                    Percent Concordant        68.6    Somers' D    0.388
                    Percent Discordant        29.8    Gamma        0.394
                    Percent Tied               1.7    Tau-a        0.157
                    Pairs                 65896012    c            0.694
```

Table 5.11 Odds Ratios and Association of Predicted Probabilities and Observed Responses for the Final Model in Table 5.9

5.2 OTHER LOGISTIC REGRESSION SELECTION PROCEDURES.

The results of applying Forward, Backward, Stepwise and Best-Subset selection procedures are given in appendices 2, 3, 4, and 5 respectively.

All the stepwise procedures except the Forward Selection produced eleven-variable models. The Forward Selection included two additional variables, Age and Faculty, which are non-significant at 5% significance level according to the Wald test. These variables satisfied the entry level of P=0.25 but could not leave the model since the Forward procedure does not provide room for non significant variables to leave the model.

The Best Subset procedure using the C_p - criterion pointed to a model with twelve variables from the two 'best' models requested for in the procedure. With regard to the Best Subset procedure using the Score-criterion we requested for 'best' two models as well, of each size (i.e. from a model containing one variable to a model with 13 variables). From the two 'best' models with twelve variables the Score- criterion selected the same model as the C_p - criterion.

The Purposeful Selection procedure like Backward and Stepwise procedures produced a model with eleven variables. However, Purposeful Selection warranted for the variable Agregate to enter the model as a binary variable following analysis of scale of continuity of this variable.

5.3 INVESTIGATION OF THE AUC AS A SELECTION TOOL

An attempt is now made to establish if the area under the ROC curve (AUC) can be used as a tool for selection of variables. In other words building a model by including variables that are increasing the AUC as they enter the model. A variable stays in the model provided it is significant in accordance with the Wald test. Like in the Forward stepwise selection, variables enter the model one at a time.

The process starts by building one-variable models and recording the AUC and the P-values as shown in Table 23. The one-variable model with the highest AUC provides the first variable to enter the model. In the next step all other variables will enter the model one by one and only the two-variable models with AUC greater than the highest AUC obtained in the first step will be considered. In the third step, a two-variable model with the highest AUC will be the basis for a three-variable model and only models with AUC higher than the largest obtained in the previous step will be considered. In any step, if there is more than one model with the same maximum, the model to be considered to the basis for next step will be selected using AIC. The process continues in this way until the AUC does increase further even when the number of variables in the model increases. However, only variables that are significant according to the Wald test will be allowed to stay in the model.

Tables 23 to 36 give the results of applying the above procedure to our data set. We note that in the last two steps (Tables 35 and 36) there are non significant variables. The final model is given in Table 34 with eleven variables, also the same as the other eleven-variable model obtained previously using Purposeful, Backward and Stepwise selection procedures.

The ROC curve for the model in Table 34 is given by figure 5.2. The area under this curve is 0.703 as shown in the table in question. This value of the area indicates a fair discrimination (predictive accuracy) by the model.

From Table 5.12 we see that (12240+1191) =13431 or 74% of the observations in our data are correctly classified by the logistic regression model in Table34. Out of 5083 observed passes, 1191 or 23% are correctly classified as predicted passes. 3892 or 77% of these observations are incorrectly classified as predicted fails (false negatives). Only 724 or 5.6% o the observed fails are incorrectly classified as predicted passes (false positives).

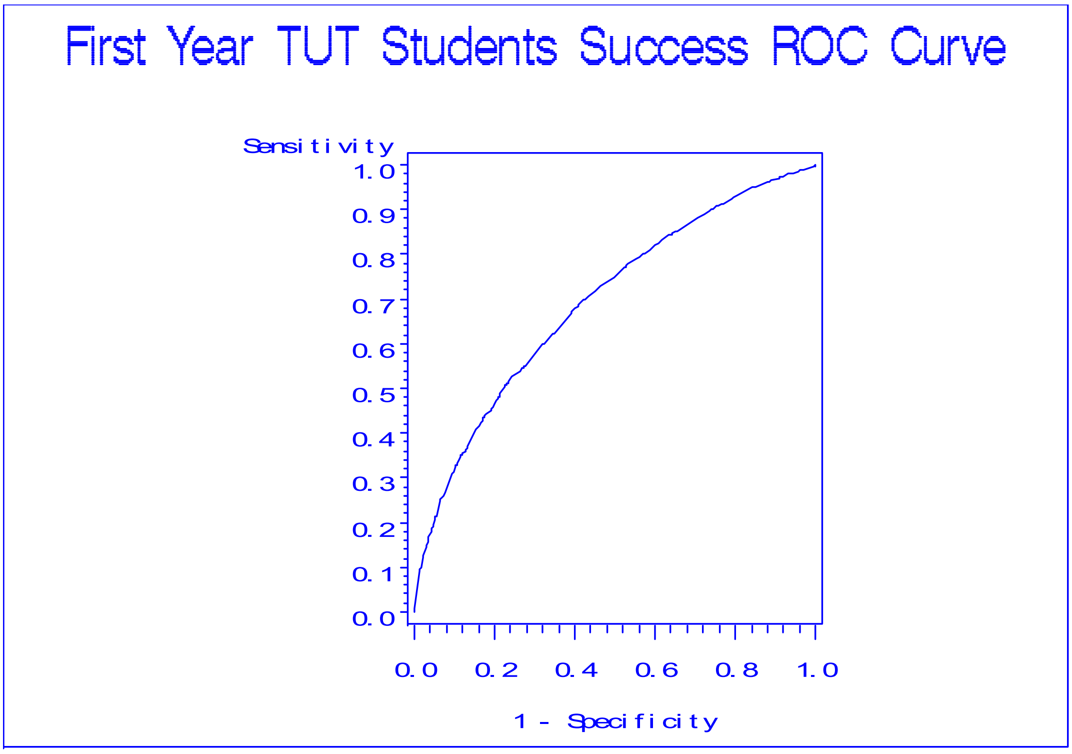


Figure 5.2 ROC curve for the model obtained using AUC procedure.

		Predicted by Model		
		0	1	Total
Actual Classification	0	12240	724	12964
	1	3892	1191	5083
	Total	16132	1915	18047

Table 5.12 Contingency Matrix for the Model in Table 34

5.4 THE AUC AND THE STEPWISE SELECTION PROCEDURES

These two selection procedures produced similar models. We note that these procedures involve 'picking' and 'dropping' of variables and we now investigate the sequence or the order of the variables entering and leaving the models. The comparison is shown in Table13.

Stepwise Procedure			AUC Procedure			
Step	Variable Entered/Removed	Wald P-value	Step	Variable Entered /Removed	Wald P-value	AUC
1	Agregate	0.0001	1	Agregate	0.0001	0.637
2	Faculty_4	0.0001	2	Brace	0.0001	0.656
3	Faculty_7	0.0001	3	Finaidd	0.0001	0.671
4	Finaidd	0.0001	4	Faculty_4	0.0001	0.681
5	Brace	0.0001	5	Faculty_7	0.0001	0.687
6	Genderr	0.0001	6	Genderr	0.0001	0.690
7	Faculty_6	0.0001	7	Faculty_6	0.0001	0.694
8	Faculty_2	0.0001	8	Faculty_2	0.0001	0.695
9	Faculty_3	0.0001	9	Faculty_3	0.0058	0.697
10	Faculty_5	0.0001	10	Faculty_4	0.0001	0.701
11	Faculty_6 Removed	0.6888	10	Faculty_6 Removed	0.6888	0.702
12	English	0.0002	11	English	0.0002	0.703
13	Campuss	0.0038	12	Campuss	0.0038	0.703
14	Age	0.862	13	Age Entered and Removed	0.0864	0.703
15	Age Removed	0.864	14	Maritall Entered & Removed	0.1584	0.703

Table 5.13 Comparison of the Stepwise and the AUC procedures

From Table 5.13 both procedures have Agregate as the first variable to enter the model. In step 2 up to step 4 the same variables entered the model though not in the same sequence. From step 5 up to the end, the two procedures yielded almost the same results. But the Stepwise procedure did not consider the variable Maritall for entry into the model.

The example used is perhaps not ideal for investigating the ROC curve as a variable selection technique. Here we have a lot of potential variables to be selected; all of them only make small contributions to the predicted probabilities. However, almost all of all of these contributions are statistically significant because of the huge sample size! Judging according to the AUC's, the increase in AUC from Table 32 to Table 36 (Appendix 6) is only 0.2% and from Table 29 to Table 36 only 0.8%. These are small increases and one may as well decide to use the model of Table 34 as the final model. It is clear that much more research on the use of the AUC's is needed.

CHAPTER 6

DISCUSSION AND CONCLUSION

The purpose of this study was to explore methods and procedures used to select predictor variables for binary response variables. However, as the point of departure selection procedures for a continuous response variable were also discussed in order to illuminate the whole question of variable selection.

We have seen that selection procedures for binary responses and continuous dependent variables are basically the same, for example, all methods used in Logistic regression are almost similar to those used for the Cox regression model. For both regressions, the 'Purposeful Selection of Variables' emerges as the most interesting and recommended procedure for selecting variables, since the method is completely controlled by the analyst. The stepwise and the best subset procedures are statistical algorithms which, to some extend, do the selection automatically. In situations where the number of variables is not large, Purposeful selection is recommended as the sole tool for selection. It can be coupled with Stepwise selection when the number of variables is too large, in which case stepwise selection will reduce the number of predictor variables to a reasonable number before Purposeful selection is used. Another advantage of Purposeful selection is the inclusion of variables that are scientifically relevant or known to interact with other variables regardless of their statistical significance. Thus the analyst, not the computer, becomes responsible for the review and evaluation of the model.

The results of a fitted logistic regression model can intuitively be summarised via classification tables. In this regard, the logistic regression model is a diagnostic test and the classification table measures the prediction accuracy. However, this measure is statistically insensitive. On the other-hand the area under the ROC Curve, another measure of the predictive accuracy, is not an extremely sensitive measure to compare two models. It is important to note that a model with high predictive accuracy does not necessarily provide evidence that the model fits well. We may have a situation where the logistic regression model is in-fact the correct model and thus fits the data but classification or discrimination is poor. These measures should, therefore, supplement more rigorous methods of assessment of fit.

The results in Tables 5.7, 5.13 and 34 suggest that to some extent, the AUC can be used as criterion for variable selection with the P-value of the Wald test used to remove insignificant variables. Perhaps even as an alternative to Purposeful and Stepwise selection procedures. However, further research is required to investigate this approach, especially for highly correlated variables.

It is further recommended that the data set used to fit the model should not be used to test for the predictive accuracy, otherwise the results become biased. A new set of observation should be used to avoid this bias, and the method called jack-knifing should be applied. The following are some of the major challenges for evaluating diagnostic tests and for applying ROC methodology in particular:

(1) Status, for example disease status, is often not a fixed entity, but rather may evolve over time. Now, how can the time aspect, be incorporated sensibly into ROC analysis?

(2) The statistical literature on diagnostic testing assumes that the test result is a simple numeric value. However, test results may be much more complicated, involving several components. Do ROC curves and the AUC have a role to play in determining how to combine different sources of information to optimise diagnostic accuracy?

The very brief investigation into the use of ROC curves and the AUC, in this thesis, yields, by no means, definitive answers to the question: How effective is the ROC curve as a tool for subset selection? Much more research is needed.

Finally, as the information revolution brings us larger data sets, with more and more variables, the demand for variable selection will strengthen and continue to be a basic strategy for data analysis. New problems will also appear as demand increases for data mining of massive data sets.

APPENDIX 1A

```
                           The UNIVARIATE Procedure
                           Variable:  age (age)

                                              Moments

N                       18047    Sum Weights              18047
Mean               20.0791821    Sum Observations        362369
Std Deviation      2.72214158    Variance            7.41005479
Skewness           4.36262118    Kurtosis            28.9348552
Uncorrected SS        7409795    Corrected SS        133721.849
Coeff Variation    13.5570342    Std Error Mean      0.02026321

                   Basic Statistical Measures

          Location                         Variability

Mean     20.07918     Std Deviation            2.72214
Median   19.00000     Variance                 7.41005
Mode     19.00000     Range                   38.00000
Interquartile Range      2.00000


                  Tests for Location: Mu0=0

           Test           -Statistic-           -----p Value------

           Student's t    t  990.9182    Pr > |t|    <.0001
           Sign           M    9023.5    Pr >= |M|   <.0001
           Signed Rank    S  81428064    Pr >= |S|   <.0001


                   Quantiles (Definition 5)

                   Quantile      Estimate
                   100% Max            54
                   99%                 33
                   95%                 24
                   90%                 22
                   75% Q3              21
                   50% Median          19
                   25% Q1              19
                   10%                 18
                   5%                  18
                   1%                  17
                   0% Min              16

                Extreme Observations

        ----Lowest----        ----Highest---

         Value      Obs        Value       Obs

            16    17517           51      2298
            16    17497           51     11190
            16    11294           52      7516
            16    10238           52     13182
            16     9455           54      3372
```

Table 14 Univariate Analysis of the Variable Age

```
                The UNIVARIATE Procedure
                Variable:  agregate (agregate)

                        Moments

    N                        18047    Sum Weights             18047
    Mean                1056.67779    Sum Observations     19069864
    Std Deviation       218.254459    Variance             47635.0089
    Skewness             0.5336167    Kurtosis             -0.1263512
    Uncorrected SS      2.10103E10    Corrected SS          859621371
    Coeff Variation     20.6547788    Std Error Mean         1.624653

                 Basic Statistical Measures

                Location                    Variability

   Mean     1056.678     Std Deviation          218.25446
   Median   1075.000     Variance                   47635
   Mode     1075.000     Range                       1440
   Interquartile Range    365.00000


           Tests for Location: Mu0=0

     Test           -Statistic-           -----p Value------

     Student's t    t  650.4021     Pr > |t|    <.0001
     Sign           M    9023.5     Pr >= |M|   <.0001
     Signed Rank    S  81428064     Pr >= |S|   <.0001

           Quantiles (Definition 5)

           Quantile      Estimate

           100% Max          2160
           99%               1612
           95%               1440
           90%               1320
           75% Q3            1200
           50% Median        1075
           25% Q1             835
           10%                835
           5%                 720
           1%                 720
           0% Min             720

       Variable:  agregate  (agregate)

       Extreme Observations

         ----Lowest----        ----Highest---

         Value    Obs         Value      Obs

         720    18044          1705    12262
         720    18041          1715     6901
         720    18038          1750     2105
         720    18032          1750     8313
         720    18022          2160     9123
```

Table 15 Univariate Analysis of the Variable Agregate

APPENDIX 1B

```
                             Faculty

                                                Cumulative  Cumulative
 Faculty     Frequency      Percent      Frequency   Percent
   2         6586        36.49          6586        36.49
   3         3771        20.90         10357        57.39
   1         2506        13.89         12863        71.28
   5         1540         8.53         14403        79.81
   6         1390         7.70         15793        87.51
   4         1313         7.28         17106        94.79
   7          941         5.21         18047       100.00

                            Race

                                             Cumulative    Cumulative
 Race     Frequency      Percent      Frequency       Percent

  4        12105        67.07          12105         67.07
  1         5334        29.56          17439         96.63
  3          341         1.89          17780         98.52
  2          267         1.48          18047        100.00

                           Campuss

                                               Cumulative    Cumulative
Campuss     Frequency      Percent      Frequency       Percent

  1        12004        66.52          12004         66.52
  2         6043        33.48          18047        100.00

                           english

                                               Cumulative    Cumulative
english     Frequency      Percent      Frequency       Percent
   1        12520        69.37         12520          69.37
   2         5527        30.63         18047         100.00


                           genderr

                                               Cumulative    Cumulative
genderr     Frequency      Percent      Frequency       Percent
   1         9207        51.02          9207          51.02
   2         8840        48.98         18047         100.00

                           maritall

                                                Cumulative    Cumulative
maritall     Frequency      Percent      Frequency       Percent

   1        17782        98.53         17782           98.53
   2          265         1.47         18047           100.00

                           finaidd

                                               Cumulative    Cumulative
finaidd     Frequency      Percent      Frequency       Percent

   2        16391        90.82         16391         90.82
   1         1656         9.18         18047         100.00
```

Table 16 Analysis of Categorical Variables

APPENDIX 2

```
                                  Intercept       Intercept and
                    Criterion     Only            Covariates

                    AIC           21460.178      19424.820
                    SC            21467.979      19534.030
                    -2 Log L      21458.178      19396.820

                    Testing Global Null Hypothesis: BETA=0
   Test                 Chi-Square       DF      Pr > ChiSq

   Likelihood Ratio      2061.3579       13          <.0001
   Score                 2085.9704       13          <.0001
   Wald                  1798.6426       13          <.0001

                     Analysis of Maximum Likelihood Estimates
                                              Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept       1     -2.4774      0.2060      144.5645        <.0001
faculty_2       1      0.6342      0.0653       94.3829        <.0001
faculty_3       1      0.5862      0.0660       78.8456        <.0001
faculty_4       1      1.6319      0.0922      312.9702        <.0001
faculty_5       1      0.7349      0.0908       65.4564        <.0001
faculty_6       1      0.0370      0.0909        0.1654        0.6842
faculty_7       1      1.2008      0.0877      187.6919        <.0001
age             1     -0.0130     0.00750        2.9863        0.0840
agregate        1     0.00162    0.000094      298.0360        <.0001
Campuss   1     1      0.0657      0.0256        6.5952        0.0102
genderr   1     1      0.0987      0.0187       27.8821        <.0001
finaidd   1     1      0.3636      0.0282      165.7910        <.0001
english   1     1      0.0779      0.0237       10.8080        0.0010
Brace           1     -0.5574      0.0437      162.8453        <.0001

                    Odds Ratio Estimates
                              Point          95% Wald
         Effect              Estimate      Confidence Limits

         faculty_2              1.886       1.659       2.143
         faculty_3              1.797       1.579       2.045
         faculty_4              5.113       4.268       6.127
         faculty_5              2.085       1.745       2.492
         faculty_6              1.038       0.868       1.240
         faculty_7              3.323       2.798       3.946
         age                    0.987       0.973       1.002
         agregate               1.002       1.001       1.002
         Campuss   1 vs 2       1.140       1.032       1.261
         genderr   1 vs 2       1.218       1.132       1.311
         finaidd   1 vs 2       2.069       1.852       2.311
         english   1 vs 2       1.169       1.065       1.282
         Brace                  0.573       0.526       0.624

         Association of Predicted Probabilities and Observed Responses

        Percent Concordant        70.0    Somers' D    0.405
        Percent Discordant        29.5    Gamma        0.408
        Percent Tied               0.5    Tau-a        0.164
        Pairs                 65896012    c            0.703

                    Adjusted Odds Ratios
                    Effect        Unit     Estimate

                    age           5.0000   0.937
                    age          -5.0000   1.067
                    agregate      100.0    1.176
                    agregate     -100.0    0.851
```

Table 17 The Results of Forward Selection Procedure

APPENDIX 3

```
                                   Intercept      Intercept and
                     Criterion     Only           Covariates
                     AIC           21460.178      19423.992
                     SC            21467.979      19517.601
                    -2 Log L      21458.178      19399.992

                    Testing Global Null Hypothesis: BETA=0

                    Test                    Chi-Square       DF      Pr > ChiSq

                    Likelihood Ratio         2058.1862       11          <.0001
                    Score                    2083.3017       11          <.0001
                    Wald                     1796.8154       11          <.0001

                    Analysis of Maximum Likelihood Estimates

                                                          Standard          Wald
  Parameter       DF     Estimate        Error      Chi-Square    Pr > ChiSq

  Intercept        1      -2.7409       0.1250      480.9214        <.0001
  faculty_2        1       0.6197       0.0564      120.6196        <.0001
  faculty_3        1       0.5729       0.0582       97.0702        <.0001
  faculty_4        1       1.6244       0.0877      342.9895        <.0001
  faculty_5        1       0.7316       0.0861       72.2871        <.0001
  faculty_7        1       1.1867       0.0815      211.7846        <.0001
  agregate         1      0.00163     0.000093      304.3700        <.0001
  Campuss   1      1       0.0730       0.0252        8.3651        0.0038
  genderr   1      1       0.1029       0.0184       31.1306        <.0001
  finaidd   1      1       0.3637       0.0282      165.9746        <.0001
  english   1      1       0.0837       0.0234       12.8165        0.0003
  brace            1      -0.5567       0.0437      162.5179        <.0001

                     Odds Ratio Estimates

                                       Point          95% Wald
                     Effect          Estimate      Confidence Limits
                     faculty_2          1.858        1.664       2.076
                     faculty_3          1.773        1.582       1.988
                     faculty_4          5.075        4.274       6.027
                     faculty_5          2.078        1.756       2.460
                     faculty_7          3.276        2.792       3.844
                     agregate           1.002        1.001       1.002
                     Campuss   1 vs 2   1.157        1.048       1.277
                     genderr   1 vs 2   1.229        1.143       1.321
                     finaidd   1 vs 2   2.070        1.853       2.312
                     english   1 vs 2   1.182        1.079       1.296
                     Brace              0.573        0.526       0.624

                 Association of Predicted Probabilities and Observed Responses

                     Percent Concordant         69.8    Somers' D    0.405
                     Percent Discordant         29.3    Gamma        0.409
                     Percent Tied                0.9    Tau-a        0.164
                     Pairs                  65896012    c            0.703

                 Adjusted Odds Ratios
                 Effect           Unit      Estimate
                 agregate        100.0         1.177
                 agregate       -100.0         0.850
```

Table 18 The Results of The Backward Selection Procedure

APPENDIX 4A

```
                           Intercept     Intercepts and
             Criterion     Only          Covariates
             AIC           21460.178      19423.992
             SC            21467.979      19517.601
             -2 Log L      21458.178      19399.992

             Testing Global Null Hypothesis: BETA=0

      Test                 Chi-Square       DF     Pr > ChiSq
      Likelihood Ratio      2058.1862       11         <.0001
      Score                 2083.3017       11         <.0001
      Wald                  1796.8154       11         <.0001

                    Analysis of Maximum Likelihood Estimates

                                    standard   wald
Parameter      DF    Estimate      Error    Chi-Square    Pr > ChiSq
Intercept       1     -2.7409      0.1250      480.9214        <.0001
faculty_2       1      0.6197      0.0564      120.6196        <.0001
faculty_3       1      0.5729      0.0582       97.0702        <.0001
faculty_4       1      1.6244      0.0877      342.9895        <.0001
faculty_5       1      0.7316      0.0861       72.2871        <.0001
faculty_7       1      1.1867      0.0815      211.7846        <.0001
agregate        1     0.00163    0.000093      304.3700        <.0001
Campuss   1     1      0.0730      0.0252        8.3651        0.0038
genderr   1     1      0.1029      0.0184       31.1306        <.0001
finaidd   1     1      0.3637      0.0282      165.9746        <.0001
english   1     1      0.0837      0.0234       12.8165        0.0003
Brace           1     -0.5567      0.0437      162.5179        <.0001

               Odds Ratio Estimates

                    Point          95% Wald
Effect              Estimate      Confidence Limits

faculty_2              1.858       1.664       2.076
faculty_3              1.773       1.582       1.988
faculty_4              5.075       4.274       6.027
faculty_5              2.078       1.756       2.460
faculty_7              3.276       2.792       3.844
agregate               1.002       1.001       1.002
Campuss   1 vs 2       1.157       1.048       1.277
genderr   1 vs 2       1.229       1.143       1.321
finaidd   1 vs 2       2.070       1.853       2.312
english   1 vs 2       1.182       1.079       1.296
Brace                  0.573       0.526       0.624

Association of Predicted Probabilities and Observed Responses

Percent Concordant        69.8    Somers' D    0.405
Percent Discordant        29.3    Gamma        0.409
Percent Tied               0.9    Tau-a        0.164
Pairs                 65896012    c            0.703

Adjusted Odds Ratios
Effect            Unit     Estimate
agregate         100.0        1.177
agregate        -100.0        0.850
```

Table 17 Results of The Stepwise Selection Procedure

APPENDIX4B

```
                                              Intercept
                               Intercept            and
                  Criterion          Only     Covariates

                  AIC           21460.178      19604.472
                  SC            21467.979      19744.885
                  -2 Log L      21458.178      19568.472


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio      1889.7059       17         <.0001
          Score                 1928.0421       17         <.0001
          Wald                  1700.7783       17         <.0001

                   Analysis of Maximum Likelihood Estimates

                                         Standard          Wald
   Parameter      DF    Estimate         Error      Chi-Square    Pr > ChiSq

   Intercept       1      0.8217        0.3777          4.7341        0.0296
   faculty_2       1      1.0533        0.2352         20.0490        <.0001
   faculty_3       1      0.5549        0.0578         92.2633        <.0001
   faculty_4       1     -1.0342        0.4876          4.4991        0.0339
   faculty_5       1     -1.4882        0.4363         11.6360        0.0006
   faculty_7       1      1.1323        0.0809        196.1369        <.0001
   agregate_       1      0.4487        0.0688         42.5332        <.0001
   Campuss    1    1     -0.7741        0.1477         27.4734        <.0001
   genderr    1    1      0.1077        0.0184         34.2445        <.0001
   finaidd    1    1     -0.1440        0.1016          2.0090        0.1564
   english    1    1      0.3295        0.0610         29.1972        <.0001
   Brace           1     -0.9424        0.1739         29.3771        <.0001
   camfin          1     -0.9071        0.1547         34.3832        <.0001
   finfac2         1     -0.2644        0.1217          4.7179        0.0298
   finfac4         1      1.4005        0.2526         30.7308        <.0001
   finfac5         1      1.2059        0.2277         28.0501        <.0001
   agrbrac         1     -0.2351        0.0839          7.8607        0.0051
   engbrac         1      0.3704        0.1317          7.9155        0.0049


                     Odds Ratio Estimates

                                 Point          95% Wald
           Effect              Estimate      Confidence Limits

           faculty_2              2.867       1.808       4.547
           faculty_3              1.742       1.555       1.951
           faculty_4              0.356       0.137       0.924
           faculty_5              0.226       0.096       0.531
           faculty_7              3.103       2.648       3.636
           agregate_              1.566       1.369       1.792
           Campuss    1 vs 2      0.213       0.119       0.379
           genderr    1 vs 2      1.240       1.154       1.333
           finaidd    1 vs 2      0.750       0.503       1.117
           english    1 vs 2      1.933       1.522       2.455
           Brace                  0.390       0.277       0.548
           camfin                 0.404       0.298       0.547
           finfac2                0.768       0.605       0.975
           finfac4                4.057       2.473       6.657
           finfac5                3.340       2.137       5.218
           agrbrac                0.790       0.671       0.932
           engbrac                1.448       1.119       1.875

       Association of Predicted Probabilities and Observed Responses

           Percent Concordant         68.7    Somers' D    0.392
           Percent Discordant         29.5    Gamma        0.399
           Percent Tied                1.9    Tau-a        0.159
           Pairs                  65896012    c            0.696
```

Table 20 The Results of The Stepwise Procedure with Interactions included.

APPENDIX 5

```
Number of  Score
Variables  Chi-Square Variables included in the model

      1   976.4420   agregate
      1   768.8932   Brace
      2   1431.5177  faculty_4 agregate
      2   1256.2088  agregate Brace
      3   1597.4285  faculty_4 faculty_7 agregate
      3   1580.0377  faculty_4 agregate Brace
      4   1749.6162 faculty_4 agregate finaidd1 Brace
      4   1734.2736  faculty_4 faculty_7 agregate finaidd1
      5   1869.5660  faculty_4 faculty_7 agregate finaidd1 Brace
      5   1838.9862  faculty_4 agregate genderr1 finaidd1 Brace
      6   1950.3123  faculty_4 faculty_7 agregate genderr1 finaidd1 Brace
      6   1905.0736  faculty_2 faculty_4 faculty_7 agregate finaidd1 Brace
      7   1976.8571  faculty_4 faculty_6 faculty_7 agregate genderr1 finaidd1 Brace
      7   1976.3923  faculty_2 faculty_4 faculty_7 agregate genderr1 finaidd1 Bra
      8   2036.9823  faculty_2 faculty_3 faculty_4 faculty_5 faculty_7 agregate finaidd1 Brace
      8   2017.8803  faculty_2 faculty_3 faculty_4 faculty_7 agregate genderr1 finaidd1 Brace
      9   2071.1645  faculty_2 faculty_3 faculty_4 faculty_5 faculty_7 agregate genderr1 finaidd1
                     Brace
      9   2044.7677  faculty_2 faculty_3 faculty_4 faculty_5 faculty_7 agregate finaidd1 english1
                     Brace
     10   2077.5397  faculty_2 faculty_3 faculty_4 faculty_5 faculty_7 agregate genderr1 finaidd1
                      English Brace
     10   2077.2280  faculty_2 faculty_3 faculty_4 faculty_5 faculty_7 agregate Campuss1 genderr1
                      finaidd1 Brace
     11   2083.3017  faculty_2 faculty_3 faculty_4 faculty_5 faculty_7 agregate Campuss1 genderr1
                      finaidd1 english1 Brace
     11   2080.0984  faculty_2 faculty_3 faculty_4 faculty_5 faculty_7 age agregate genderr1
                      finaidd1 english1 Brace
     12   2084.8214  faculty_2 faculty_3 faculty_4 faculty_5 faculty_7 age agregate Campuss1
                      genderr1 finaidd1 english1 Brace
     12   2084.3485  faculty_2 faculty_3 faculty_4 faculty_5 faculty_7 agregate Campuss1 genderr1
                      marital1 finaidd1 english1 Brace
     13   2085.9704  faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7 age agregate
                     campuss1 genderr1 finaidd1 english1 Brace
     13   2085.4222  faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7 agregate Campuss1
                      genderr1 maritall finaidd1 english1 Brace
     14   2086.1659  faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7 age agregate
                      campuss1 genderr1 maritall finaidd1 english1 Brace
```

Table 21 The Results of Best Subset Selection Procedure using Score Criterion.

```
                                        C(p) Selection Method

                               Number of Observations Read       18047
                               Number of Observations Used       18047

                                              Weight: v
  Number in
  Model         C(p)  R-Square Variables in Model

      12    11.6129    0.0902 faculty_2 faculty_3 faculty_4 faculty_5 faculty_7
                              age agregate Campuss genderr finaidd English Brace

      11    12.4942    0.0900 faculty_2 faculty_3 faculty_4 faculty_5 faculty_7
                              Campuss agregate gender finaidd English Brace
```

Table 22 The Results of Best Subset Selection Procedure using Cp Criterion.

APPENDIX 6

Variable	p-value	AUC
age	<0.0001	0.532
agregate	<0.0001	0.637 *
campuss	<0.0001	0.537
genderr	<0.0001	0.541
marital	0.0073	0.503
finaidd	<0.0001	0.532
english	<0.0001	0.576
brace	<0.0001	0.608
faculty_2	<0.0001	0.545
faculty_3	0.0148	0.508
faculty_4	<0.0001	0.555
faculty_5	<0.0001	0.508
faculty_6	<0.0001	0.520
faculty_7	<0.0001	0.523

Table 23 Step1 of the AUC procedure

Variables :	agregate	marital
p-value :	<0.0001	<0.0001
AUC :	0.637	
Variables :	agregate	faculty_5
p-value :	<0.0001	0.4389
AUC :	0.637	
Variable :	agregate	faculty_3
p-value :	<0.0001	0.5985
AUC :	0.638	
Variable :	agregate	gender
p-value :	<0.0001	<0.0001
AUC :	0.642	
Variable :	agregate	campuss
p-value :	<0.0001	<0.0001
AUC :	0.643	
Variable :	agregate	faculty_6
p-value :	<0.0001	<0.0001
AUC :	0643	
Variable :	agregate	english
p-value :	<0.0001	<0.0001
AUC :	0.647	
Variable :	agregate	finaidd
p-value :	<0.0001	<0.0001
AUC :	0.648	
Variable :	agregate	faculty_4
p-value :	<0.0001	<0.0001
AUC :	0.655	
Variable :	aggregate	brace
p-value :	<0.0001	<0.0001
AUC :	0.656 *	

Table 24 Step 2 of the AUC procedure

```
Variables : agregate   brace     faculty_2
p-value   :  <0.0001   <0.0001  <0.0132
AUC       :  0.658
Variables  : agregate   brace     english
p-value    :  <0.0001   <0.0001  <0.0001
AUC        : 0.658
Variable    : agregate   brace      campuss
p-value    : <0.0001   <0.0001   <0.0001
AUC        : 0.660
Variable    : agregate   brace      faculty_7
p-value     : <0.0001   <0.0001    <0.0001
AUC        : 0.660
Variable    :  agregate   brace     genderr
p-value    :  <0.0001    <0.0001   <0.0001
AUC        : 0.662
Variable    : agregate   brace      faculty_6
p-value    : <0.0001    <0.0001    <0.0001
AUC        :  0.663
Variable   : agregate   brace       faculty_4
p-value  :  <0.0001   <0.0001     <0.0001
AUC      :  0.666
Variable  : agregate    brace       finaidd
p-value  :  <0.0001   <0.0001     <0.0001
AUC      : 0.671*
```

Table 25 Step 3 of the AUC procedure

```
Variables: agregate   brace     finaidd   faculty_2
p-value  :  <0.0001  <0.0001   <0.0001  <0.0305
AUC      : 0.673
Variables: agregate   brace     finaidd    campuss
p-value   : <0.0001   <0.0001   <0.0001  <0.0001
AUC      : 0.674
Variables: agregate   brace     finaidd     english
p-value  :  <0.0001  <0.0001   <0.0001   <0.0001
AUC      : 0.672
Variables: agregate   brace     finaidd    genderr
p-value  :  <0.0001  <0.0001   <0.0001   <0.0001
AUC      : 0.677
Variables: agregate   brace     finaidd   faculty_6
p-value  :  <0.0001  <0.0001   <0.0001  <0.0001
AUC      : 0.677
Variables: agregate   brace     finaidd   faculty_4
p-value  :  <0.0001  <0.0001   <0.0001  <0.0001
AUC      : 0.681*
```

Table 26 Step 4 of the AUC Procedure

```
Variables: agregate  brace    finaidd  faculty_4  faculty_3
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    0.0226
AUC      : 0.682
Variables: agregate  brace    finaidd  faculty_4  english
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    <0.0001
AUC      : 0.683
Variables: agregate  brace    finaidd  faculty_4  faculty_2
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    <0.0001
AUC      : 0.683
Variables: agregate  brace    finaidd  faculty_4  faculty_6
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    <0.0001
AUC      : 0.685
Variables: agregate  brace    finaidd  faculty_4  genderr
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    <0.0001
AUC      : 0.686
Variables: agregate  brace    finaidd  faculty_4  faculty_7
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    <0.0001
AUC      : 0.687*
```

Table 27 Step 5 of the AUC procedure

```
Variables: agregate  brace    finaidd  faculty_4  faculty_7   faculty_5
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    <0.0001     <0.0001
AUC      : 0.688
Variables: agregate  brace    finaidd  faculty_4  faculty_7  faculty_3
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    <0.0001    <0.0001
AUC      : 0.688
Variables: agregate  brace    finaidd  faculty_4  faculty_7  english
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    <0.0001    <0.0001
AUC      : 0.688
Variables: agregate  brace    finaidd  faculty_4  faculty_7  faculty_2
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    <0.0001    <0.0001
AUC      : 0.690*
Variables: agregate  brace    finaidd  faculty_4  faculty_7 genderr
p-value  :  <0.0001  <0.0001  <0.0001 <0.0001    <0.0001   <0.0001
AUC      : 0.690*
```

Table 28 Step 6 of the AUC procedure

```
Variables: agregate  brace     finaidd   faculty_4   faculty_7  genderr  faculty_5
p-value  :  <0.0001  <0.0001   <0.0001 <0.0001     <0.0001    <0.0001  0.0065
AUC      : 0.691
Variables: agregate  brace     finaidd   faculty_4   faculty_7  genderr  faculty_3
p-value  :  <0.0001  <0.0001   <0.0001 <0.0001     <0.0001    <0.0001  0.0002
AUC      : 0.692
Variables: agregate  brace     finaidd   faculty_4   faculty_7  genderr  english
p-value  :  <0.0001  <0.0001   <0.0001 <0.0001     <0.0001    <0.0001  0.0003
AUC      : 0.692
Variables: agregate  brace     finaidd   faculty_4   faculty_7  genderr  faculty_2
p-value  :  <0.0001  <0.0001   <0.0001 <0.0001     <0.0001    <0.0001  <0.0001
AUC      : 0.693
Variables: agregate  brace     finaidd   faculty_4   faculty_7  genderr   faculty_6
p-value  :  <0.0001  <0.0001   <0.0001 <0.0001     <0.0001    <0.0001   <0.0001
AUC      : 0.694*
```

Table 29 Step 7 of the AUC procedure

```
Variables: agregate  brace     finaidd   faculty_4   faculty_7  genderr   faculty_6   english
p-value  :  <0.0001  <0.0001   <0.0001 <0.0001     <0.0001    <0.0001   <0.0001     0.0010
AUC      : 0.695*
Variables: agregate  brace     finaidd   faculty_4   faculty_7  genderr   faculty_6  faculty_2
p-value  :  <0.0001  <0.0001   <0.0001 <0.0001     <0.0001    <0.0001   <0.0001     <0.0001
AUC      : 0.695*
```

Table 30 Step 8 of the AUC procedure

```
Variables: agregate  brace     finaidd   faculty_4   faculty_7  genderr   faculty_6  faculty_2   english
p-value  :  <0.0001  <0.0001   <0.0001 <0.0001     <0.0001    <0.0001   <0.0001     <0.0001    <0.0001
AUC      : 0.696
Variables: agregate  brace     finaidd   faculty_4   faculty_7  genderr   faculty_6  faculty_2   faculty_3
p-value  :  <0.0001  <0.0001   <0.0001 <0.0001     <0.0001    <0.0001   <0.0001    <0.0001     0.0058
AUC      : 0.697*
```

Table 31 Step 9 of the AUC procedure

```
Variables : agregate   brace     finaidd     faculty_4  faculty_7   genderr    faculty_6  faculty_2  faculty_3  english
 p-value   : <0.0001   <0.0001  <0.0001    <0.0001     <0.0001     <0.0001   0.0091      <0.0001    <0.0001    0.0002
 AUC       : 0.698
Variables : agregate   brace     finaidd     faculty_4  faculty_7   genderr    faculty_6  faculty_2  faculty_3  faculty_5
 p-value   : <0.0001   <0.0001  <0.0001    <0.0001     <0.0001     <0.0001   0.6888      <0.0001    <0.0001    0.0001
 AUC       : 0.701*
Variables : agregate   brace     finaidd     faculty_4  faculty_7   genderr    faculty_2  faculty_3   faculty_5
 p-value   : <0.0001   <0.0001  <0.0001    <0.0001     <0.0001     <0.0001  < 0.0001    <0.0001     0.0001
 AUC       : 0.701*
```

Table 32 Step 10 of the AUC procedure

Variables:	agregate	brace	finaidd	faculty_4	faculty_7	genderr	faculty_2	faculty_3	faculty_5	english
p-value :	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0002
AUC : 0.702 *										

Table 33 Step 11 of the AUC procedure

Variables:	agregate	brace	finaidd	faculty_4	faculty_7	genderr	faculty_2	faculty_3	faculty_5	english	campuss
p-value :	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	< 0.0001	0.0038
AUC : 0.703*											

Table 34 Step 12 of the AUC procedure

Variables:	agregate	brace	finaidd	faculty_4	faculty_7	genderr	faculty_2	faculty_3	faculty_5	english	campuss	age
p-value :	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0090	<0.0001	0.0864
AUC : 0.703*												

Table 35 Step 13 of the AUC procedure

Variables:	agregate	brace	finaidd	faculty_4	faculty_7	genderr	faculty_2	faculty_3	faculty_5	english	campuss	maritall
p-value :	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0005	<0.0060	0.1584
AUC : 0.703*												

Table 36 Step 14 of the AUC procedure

APPENDIX 7

SAS PROGRAMME

```
data jimmy;
set sasuser.osiame;
proc freq order=freq;
tables faculty race campuss english genderr maritall finaidd;
run;
data jimmy2;
set sasuser.osiame;
proc univariate;
var age agregate;
title;
run;
data joseph1;
set sasuser.osiame;
if score ='1' then pass = 1;
else if score ='2' then pass=0;
if faculty = '2' then faculty_2=1;
else faculty_2=0;
if faculty = '3' then faculty_3 =1;
else faculty_3 =0;
if faculty ='4' then faculty_4 =1;
else faculty_4 = 0;
if faculty ='5' then faculty_5 =1;
else faculty_5=0;
if faculty = '6' then faculty_6 =1;
else faculty_6 = 0;
if faculty ='7' then faculty_7=1;
else faculty_7 =0;
if race >3 then Brace=1;
else Brace=0;
Keep faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7
      pass age agregate campuss genderr maritall finaidd english brace;
       run;


proc logistic descending;
class campuss maritall finaidd genderr english;
model pass =age;
units age=5 -5;
run;

proc logistic descending;
class campuss maritall finaidd genderr english;
model pass=agregate;
units agregate=100 -100;
run;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass=campuss;
run;
proc logistic descending;
class campuss maritall finaidd genderr english;
```

```
model pass=maritall;
run;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass=finaidd ;
run;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass=genderr;
run;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass=english;
run;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass=faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7 ;
run;

*/ The model without the variable Faculty_6 insignificant in
   Univariate Logistic regression;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass=age agregate campuss maritall  finaidd genderr english faculty_2
      faculty_3 faculty_4 faculty_5  faculty_7 brace;
units age=5 -5 agregate= 100 -100;
run;
*/ The model without the variables faculty_6,maritall;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass=age agregate campuss  finaidd genderr english faculty_2
      faculty_3 faculty_4 faculty_5  faculty_7 brace ;
units age=5 -5 agregate= 100 -100;
run;
*/ The model without the variables faculty_6,maritall and age;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass= agregate campuss  finaidd genderr english faculty_2
      faculty_3 faculty_4 faculty_5  faculty_7  brace   ;

units  agregate= 100 -100;
run;

*/ Variable faculty_6 re-enters the model
proc logistic descending;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass= agregate campuss  finaidd genderr english faculty_2
      faculty_3 faculty_4 faculty_5  faculty_7   brace faculty_6   ;

units  agregate= 100 -100;
run;
*/ Variable maritall re-enters the model;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass= agregate campuss  finaidd genderr english faculty_2
      faculty_3 faculty_4 faculty_5  faculty_7 brace maritall  ;
```

```
units    agregate= 100 -100;
run;
*/ The variable age re-enters the model;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass= agregate campuss  finaidd genderr english faculty_2
      faculty_3 faculty_4 faculty_5  faculty_7 brace faculty_6  age ;

units age=5 -5   agregate= 100 -100;
run;
*/ Variables thet give the pleliminary Main eefects model;
proc logistic descending;
class campuss maritall finaidd genderr english;
model pass= agregate campuss  finaidd genderr english faculty_2
      faculty_3 faculty_4 faculty_5  faculty_7 brace ;

units    agregate= 100 -100;
run;
*/ Examining the scale of continuous covariate;
*/ The variable agregate is analysed using quartiles;
data joseph3;
set joseph1;
if 720 <= agregate <= 835 then agregroup =1;
else if 835 < agregate <=1075 then agregroup=2;
else if 1075 < agregate <=1200 then agregroup=3;
else if agregate > 1200 then agregroup=4;
if agregroup='2' then agre_2=1;
else agre_2=0;
if agregroup ='3' then agre_3=1;
else agre_3=0;
if agregroup ='4' then agre_4 = 1;
else agre_4=0;
run;
proc logistic descending;
class campuss maritall finaidd genderr english;;
model pass= agre_2 agre_3 agre_4  campuss  finaidd genderr english faculty_2
            faculty_3 faculty_4 faculty_5 faculty_7 brace;

run;

data midpoints;
input    agreg    coef;
cards;
         775.5       0
         955        .2898
           1137.5     1.0672
           1680       .9989
        ;
        run;
        goptions  reset =all;
        symbol c=blue v=dot h=.8 i=j;
        axis order=(0 to 1.5 by .2) label=(a=90 'logit');
        proc gplot data=midpoints;
        plot coef*agreg / vaxis=axis;
        run;
        quit;
```

```
     data joseph6;
       set joseph3;
       proc chart;
       vbar agregate / midpoints=100 to 2200 by 100
       GROUP=pass;
       run;
       data scale;
       set joseph3;
       exlinex=agregate*log(agregate);
       run;
     proc logistic descending;
     class campuss maritall finaidd genderr english;;
     model pass=agregate exlinex  campuss  finaidd genderr english faculty_2
           faculty_3 faculty_4 faculty_5 faculty_7 brace;

     run;

    */ The variable agregate is dichotomised;
       data joseph4;
       set joseph3;
       if agregate >= '1075' then agregate_=1;
       else agregate_=0;
       run;
       */ Fitting a dichotomous variable agregate_;
       proc logistic descending;
       class english finaidd campuss genderr;
       model pass = english finaidd campuss genderr faculty_2 faculty_3
                    faculty_4 faculty_5 faculty_7 brace agregate_;
       run;


data interaction;
set joseph4;
engagr=english*agregate_;
 engfin=english*finaidd;
camfin=campuss*finaidd;
finfac2=finaidd*faculty_2;
finfac3=finaidd*faculty_3;
finfac4=finaidd*faculty_4;
finfac5=finaidd*faculty_5;
finfac7=finaidd*faculty_7;
racfin=brace*finaidd;
 agrbrac=agregate_*brace;
 engbrac=english*brace;
 run;

proc logistic data=interaction;
 class  campuss genderr finaidd english ;
model pass= campuss genderr  finaidd english faculty_2 faculty_3
            faculty_4 faculty_5 faculty_7 brace
            agregate_ engagr;
run;

proc logistic data=interaction;
 class  campuss genderr finaidd english ;
model pass= campuss genderr  finaidd english faculty_2 faculty_3
            faculty_4 faculty_5 faculty_7 brace
```

```
               agregate_ engfin;
 run;
proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_ camfin;
 run;


 proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_ finfac2;
 run;
 proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_ finfac3;
               run;
proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_ finfac4;
 run;
 proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_ finfac7;
 run;

proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_ finfac5;
 run;
proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_ racfin;
 run;
proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_  agrbrac;
 run;
 proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
```

```
               agregate_ engbrac;
 run;
proc logistic data=interaction descending;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_ engagr  engfin  finfac4 finfac2
               finfac5 racfin agrbrac  engbrac ;
 run;


proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_ engagr  engfin  finfac4 finfac2
               finfac5 racfin agrbrac  engbrac ;
 run;
proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_ engagr  engfin  finfac4 finfac2
               finfac5  agrbrac  engbrac ;
 run;
 proc logistic data=interaction;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_   engfin  finfac4 finfac2
               finfac5  agrbrac  engbrac ;
 run;
 proc logistic data=interaction descending noprint;
  class  campuss genderr finaidd english ;
 model pass= campuss genderr  finaidd english faculty_2 faculty_3
             faculty_4 faculty_5 faculty_7 brace
               agregate_   finfac4 finfac2
               finfac5  agrbrac  engbrac ;
            output out=probability predicted=phat;
 run;
data probability1;
set probability;
predicts=(phat>=.5);
run;
proc freq data=probability1;
tables pass*predicts / norow nocol nopercent;
run;

proc logistic data=interaction descending;
class campuss genderr finaidd english;
model pass=campuss genderr finaidd english faculty_2 faculty_3 faculty_4
      faculty_5 faculty_7 brace agregate_ finfac4 finfac2 finfac5 agrbrac
        engbrac / outroc=rocl;
        goptions cback=white
                 colors=(blue)
                     border;
                     axis1 length=2.5in;
                     axis2 order =(0 to 1 by .1) length=2.5in;
proc gplot data=rocl;
symbol1 i=join v=none;
title 'First Year TUT Students Success ROC Curve';
plot _sensit_*_1mspec_ / haxis=axis1 vaxis=axis2;
run;
quit;
*/ Forward selection procedure;
data foward;
set joseph1;
proc logistic descending;
class campuss genderr finaidd english;
model pass=faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7
       age agregate campuss genderr maritall finaidd english brace
        / selection=forward slentry=.25 details ;
            units age=5 -5 agregate=100 -100;
            run;
*/ Backward Selection procedure;
proc logistic descending;
class campuss genderr finaidd english;
model pass=faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7
       age agregate campuss genderr maritall finaidd english brace
        / selection=backward details slstay=.05;
            units age=5 -5 agregate=100 -100;
run;
*/ Stepwise Selection procedure;
proc logistic descending;
class campuss genderr finaidd english;
model pass=faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7
        agregate campuss genderr maritall age  finaidd english brace
        / selection=stepwise slentry=.25;
            units  agregate=100 -100;
            run;

*/ Stepwise Selection procedure used to select Interactions;
proc logistic descending;
class campuss genderr finaidd english;
model pass=faculty_2 faculty_3 faculty_4 faculty_5  faculty_7
        agregate_ campuss genderr  finaidd english brace engagr
         engfin camfin finfac2 finfac3 finfac4 finfac4 finfac5 finfac7 racfin
         agrbrac engbrac / selection=stepwise slentry=.25 include=11;
                        run;


*/ Best Subset Selection procedure using Score criterion;
      proc logistic descending;
class campuss genderr finaidd english;
model pass=faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7
       age agregate campuss genderr maritall finaidd english brace
        / selection=score best=2;
            units age=5 -5 agregate=100 -100;
            run;
*/ Best Subset procedure using Cp criterion;
proc logistic descending;
class campuss genderr finaidd english;
model pass=faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7
       age agregate campuss genderr maritall finaidd english brace;
```

```
output out=best2 prob=pihat;
     run;
     data best3;                                          set best2 ;
     z=log(pihat/(1-pihat))+((pass-pihat)/(pihat*(1-pihat)));
     v=pihat*(1-pihat);
     run;
     proc reg;
     model z=faculty_2 faculty_3 faculty_4 faculty_5 faculty_6 faculty_7
        age agregate campuss genderr maritall finaidd
        english brace/selection=cp best=3;
              weight v;
     run;
     quit;
```

Reference

Beale, E M L (1970). Note on procedures or variable selection in multiple regression. Technometrics, 12, 909-14.

Bergerud, W A (1996). Introduction to Regression Models: With worked forestry examples. Biom. Imf.Hand. Res.Br., B.C. Min. For., Victoria, B.C. work. Pap. 26/1996.

Cody, R P and Smith, JK (1997). Applied Statistics and the SAS programming Language. London. Prentice and Hall.

Cook, E D (2001). Solutions Manual to Accompany Applied Logistic Regression 2^{nd} Edition by Hosmer, D W and Lemeshow, S.

Czepiel S, http:// www.czep.net/contact.html.

Dallal, G E (2001). Logistic regression. http:// www.tufts.edu/~gdallal/logistic.htm

Delwiche, D L and Slaughter, S J (1995). A premier, Cary, NC: SAS Institute Inc

Draper, N R and Smith, H (1981). Applied Regression Analysis, Second Edition. New York. Wiley.

ERTAþ, G. Evaluation of Diagnostic Test Accuracy by Receiver Operation Characteristic (ROC) Analysis. Boðazici University, Biomedical Engineering Institute, 80815, Bebek, Ýstanbul, e-mail: erstasg@boun.edu.tr.

George, E I (2000). The variable selection problem. Journal of the American Statistical Association, vol 95, No 452, Vignettes.

Gorman, J W and Toman, R J (1966). Selection of variables for fitting equations to data. Technometrics, 12, No.1.

Guyon, I and Elisseeff, A (2002). Special Issue on Variable and Feature selection. Journal of Machine Learning Research.

Hanley, J A and McNeil, B J (1982). The meaning and use of the Area under a Receiver Operating Characteristic (ROC) curve. Radiology, 143, 29-36.

Hocking, R R and Leslie, R N (1967). Selection of best subset in Regression Analysis.Technometrics, 9, 531-540.

Hocking, R R (1972). Criteria for Selection of a subset Regression: Which one should be used? Technometrics, 14, No.4.

Hocking, R R (1976). The Analysis and Selection of Variables in linear regression, Biometrics, 32, 1-49.

Hosmer, D W and Lemeshow, S (1989). Applied Logistic Regression. New York. Wiley and Sons.

Hosmer, D W and Lemeshow, S (1998). Applied Survival Analysis: Regression Modeling of Time to Event Data. New York.Wiley and Sons.

Hosmer, D W and Lemeshow, S (2002). Applied Logistic Regression 2nd Edition.
New York. Wiley and Sons.

jo@anaestethetist.com (2001). The magnificent ROC. Google's cache of http://www.anaestethetist.com/mnm/stats/roc/

Joubert, G (1994). Variable Selection in Logistic Regression, with Special Application to Medical data.

Karp, A H. Using logistic regression to predict customer retention. http://www.Sierrainformation.com

Larsen, P V(2001). Module 14: Logistic regression. http:// www.statmaster.sdu.dk/courses/st111/module14/.

Mallows, C L (1973). More comments on C_p . Technometrics, 15, 661-676.

Mantel N (1970). Why stepwise selection in multiple regression. Technometrics, 12 621-25.

Marzban, C (2004). A comment on the ROC Curve measures. http://www.nhn.ou.edu/marzban.

Marriott, J M and Pettitt, A N (1997). Graphical Techniques for selecting explanatory variables for the time series data. Journal of Applied Statistics, 46, 253-264.

McClish, D K (1989). Analysing a portion of the ROC curve. Medical Decision Making, 9, 190-195.

McCullagh, P and Nelder, J A (1989). Applied Regression Analysis. New York. Wiley and Sons.

Menard, S (2001). Applied Logistic Regression Analysis. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-106. Thousand Oaks, CA: Sage.

Metz, C E, Herman, B A and Shen, J (1998). Maximum likelihood estimation of ROC from continuously distributed data. Statistics in Medicine, 17, 1033-1053.

Miller, A J (1984). Selection of subsets and regression variables. Journal of the Royal Statistical Society, A, 147, 389-425.

Miller, A J (1990). Subset Selection in Regression, London. Chapman and Hall.

Morrison, Ann Michelle (2005). Receiver Operating Characteristic (ROC) curve Analysis of Antecedent Rainfall and the Alewife/Mystic River Receiving water. Water Resource Authority, Report ENQUAD 2005-23.26p.

Nargundkar, S and Priestly, J L (2003). Assessment of Evaluation Methods for Prediction and Classification of Consumer Risk in the Credit Industry. Federal Reserve System Report. http://www.federalreserve.gov/rnd.htm.

Pepe, M S (1997). A regression modelling framework for receiver operating characteristic curves in the medical diagnostic testing. Biometrika, 84/3, 595-608.

Raftery et al. Statistics in the 21st century, Monographs on Statistics and Applied 93, 60- . London. Chapman and Hall/CRC.

Tosteson, A and Begg, C B (1988). A General Regression Methodology for ROC Curve Estimation. Medical Decision Making, 8, 204-15.

Thomson, M L (1978). Selection of Variables in multiple regression: part II. Chosen Procedures, Computations and Examples. Internal Statistics Review, 46, 129-146.

Tibshirani, R (1997). The Lasso Method for variable selection in the Cox model. Statistics in Medicine, 16, 385-395.
Walters, S J (2001). What is a Cox Model? http://www.evidence-ased.medicine.co.uk.

Zou, H and Hastie, T (2005). Regularisation and Variable Selection via elastic net. Journal of the Royal Statistics Society, 67, 301-320.

Zweig, M H and Campell, G (1993). Receiver-Operating Characteristic (ROC). A Fundamental Evaluation Tool in Clinical Medicine. Clinical Chemistry, 39/4, 561-577.