

**GROWTH CURVE ANALYSIS FOR CARDIO-PULMONARY RESPONSES TO
PHYSICAL EXERCISE**

**A DISSERTATION SUBMITTED TO THE FACULTY OF SCIENCE, UNIVERSITY
OF WITWATERSRAND, JOHANNESBURG, FOR THE DEGREE OF MASTER OF
SCIENCE**

BY

MARK JOEL PAIKER

JULY 1990

I, THE UNDERSIGNED, HEREWITH DECLARE THAT THIS DISSERTATION IS ENTIRELY MY OWN WORK AND HAS NOT BEEN PRESENTED FOR ANY DEGREE AT ANOTHER UNIVERSITY



MARK JOEL PAIKER

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation for the support and assistance I received from my family, friends and colleagues who made this work possible.

I am particularly indebted to my supervisor Professor P Fatti, whose enthusiasm, patient guidance and encouragement assured the completion of this work.

I am also very grateful for the assistance received from the staff at the Computer Centre of the University of the Witwatersrand.

Finally, I wish to thank Mrs M Manuel for her extreme patience, devotion and efficiency in typing the manuscript.

ABSTRACT

This dissertation investigates the fitting of growth curves to cardio-pulmonary responses on a treadmill.

The practical application is based on two approaches in the growth curve literature, namely the Richards family of growth curves and the polynomial of Potthoff and Roy, and Rao and Khatri.

The first part of this dissertation deals with the theoretical background to the two approaches.

Derivations of the Richards family of growth curves are considered, distinguishing characteristics discussed, and the various estimation techniques employed in the fitting of these nonlinear growth curves.

The different types of polynomials in the Potthoff and Roy model as well as the Rao-Khatri model are considered. The fitting of these growth curves as well as the hypothesis tests involved in choosing the correct degree of polynomial is also examined.

The second part of this dissertation deals with the data taken from an exercise test examining cardio-pulmonary responses. The two approaches in the first part of the dissertation are used to fit the data to a specific function. The best fitting functions characterizing the data are chosen, and physiological interpretations of these results are provided. Experimental design difficulties are also discussed.

CONTENTS

Introduction	Page
CHAPTER 1	
Introduction.....	4
1.1 Curve fitting—why?	5
1.2 Linear curves or curves that are transformable to the linear form	6
The straight line.....	8
The exponential curve.....	9
Power curves.....	9
The Hyperbolic curve.....	10
The parabola.....	11
Orthogonal polynomials.....	13
1.3 Nonlinear growth models.....	15
The Monomolecular model.....	15
The Logistic model.....	20
Inversesymmetry of logistic model.....	22
The Gompertz model.....	26
The Richard family of growth curves.....	30
1.4 Reparametrization.....	32
1.5 Estimation.....	33
Maximum likelihood estimation.....	33
Least squares.....	35
Weighted least squares.....	36
Comparison between Maximum likelihood and least-squares.....	37
Estimating initial parameters.....	38
1.6 Optimization.....	40
Newton Raphson procedure.....	40
Gauss Newton Method.....	41
1.7 Normal Theory maximum likelihood estimation in the presence of correlated residuals.....	44
1.8 Bayes' estimation when data are incomplete.....	46

CHAPTER 2	Page
Introduction	49
2.1 The Potthoff and Roy Model.....	49
2.2 Hypothesis testing.....	56
A Generalized linear hypothesis.....	56
Definition of a Wishart Distribution.....	57
Wilks likelihood ratio test.....	58
2.3 Choice for the matrix G	60
2.4 The Rao-Khatri model	61
General theory for analysis of covariance.....	62
Partitioning of the data matrix.....	63
The Model.....	64
2.5 Comparisons of the two methods.....	67
CHAPTER 3	
3.1 The Application.....	69
3.2 The Exercise Test.....	69
3.3 The Data.....	70
3.4 Equipment used.....	71
3.5 Physiological background.....	72
3.6 Computer programs used.....	73
3.7 The Sample.....	75
3.8 Results using SCORE.....	75
CHAPTER 4	
4.1 Applying the Logistic Function to heartrate.....	77
4.2 Applying the Logistic Function to ventilation.....	92
4.3 Applying the Logistic Function to oxygen consumption.....	98
4.4 Summary.....	104

CHAPTER 5	Page
5.1 Application of the Rao-Khatri model.....	105
Introduction	
Heart rate results.....	105
Ventilation results.....	110
Oxygen consumption results.....	112
5.2 Graphical comparisons between the Rao-Khatri model and Richards' logistic function.....	114
CONCLUSION	118
APPENDIX A	119
APPENDIX B	122
APPENDIX C	124
APPENDIX D	129
REFERENCES	130

INTRODUCTION

In many fields of application, such as biology, psychology, agriculture, botany, engineering and medicine, experiments are conducted in which a response is measured p times on each of N experimental units. The concept "response on an experimental unit" is used here to cover a wide range of situations.

Examples of these are: the height of a child measured annually from one to twelve years, the width of a leaf measured daily over a period of two weeks, the output of an electronic filter measured at ten time intervals, the weight of a mouse measured weekly over a period of three months, the measurement of heart rate of a person at ten second intervals on a treadmill and the reaction time of a person to a specific stimulus measured six times at equal intervals. Experimental units may therefore refer to individuals, animals, plants or other subjects. Models for such data are called growth curve models and in recent years considerable attention has been given to their analysis. Growth curve analysis usually refers to the fitting of models and testing of hypotheses concerning series of measurements in time on the same experimental unit of some quantity of interest.

There have been two distinct, dominant strands in the growth curve literature. The first is the "biological" strand, which seeks a model with a biological basis and "biologically meaningful" parameters. The other is the "statistical" strand which deals with a class of growth models, usually polynomials, which are analysed allowing for the dependence structure but not necessarily according to biological characteristics.

Richards (1969) reviews current growth processes in biological data. It becomes apparent how complex these data are to model due to a lack of knowledge in this vast field.

At the current state of knowledge about the vastly complex subjects of biology, biochemistry, physiology and environment in which growth occurs, the biologically based

models which Richards uses are merely approximations of the growth processes.

Biologically based models are acceptable if and only if they pass the empirical test.

Likewise all statistical models only approximate the truth and stand or fall on their empirical performance.

In the statistical strand there have been various approaches by Potthoff and Roy (1964), Rao (1965), Khatri (1966) and Grizzle and Allen (1969). These techniques enable us to fit polynomial growth curves to sets of data with arbitrary internal correlation to test the usual sorts of linear hypotheses about the true regression coefficients and to derive confidence intervals for the growth curves.

In very short series where no alternative model could be fitted, a polynomial may be adequate, for it may be the only model available.

However the choice of a polynomial to represent a growth curve leaves a biologist unsatisfied since there is no biological interpretation of the parameters.

Richards (1969) states "... the polynomial curves commonly adopted in statistical work are usually quite inappropriate for growth studies".

Sandland and McGilchrist (1979) list three reasons why polynomials are unattractive models for long structured growth series

- "(i) Growth processes undergo changes of phase, which are not appropriately handled by fitting a global polynomial.
- (ii) If we fit a polynomial to data for which it is not appropriate, the correlation structure in the observed residuals will be distorted.
- (iii) Polynomials are poor at fitting asymptotic approaches to limiting values."

This dissertation will consider the fitting of growth curve models to cardio-pulmonary responses on a graded exercise test. The first approach employed uses the non-polynomial, nonlinear Richards family of growth curves (the "biological" strand). Factors such as estimation processes, iterative procedures, the estimation of initial parameters, all associated with the fitting of such nonlinear growth curves are considered. The second approach used in this dissertation employs the polynomials of Potthoff and Roy and Rao and Khatri (the "statistical" strand). The fitting of such polynomials is considered. The use of concomitant variables, the various types of polynomials employed, and the hypothesis testing of such polynomials, are considered. The best fitting functions of both approaches are examined and physiological interpretation is provided.

CHAPTER 1

INTRODUCTION

In Section 1.1 of this chapter the reasons for curve fitting are considered.

Section 1.2 deals with curves that are linear or transformable to the linear form. In Section 1.3 the characteristics of the three growth curves constituting the Richards family of growth curves are considered, namely: the monomolecular model, the logistic model, and the Gompertz model. A derivation of each of these curves is provided, and brief backgrounds to the curves are discussed. Since we are considering nonlinear growth curves Section 1.4 investigate a suitable transformation of the parameters to simplify the optimising problem to one that has no restrictions. In Section 1.5 maximum likelihood estimation, and least squares estimation are investigated, and comparisons between the two processes are discussed. A method for initializing parameters and the derivation of confidence intervals are also examined. In 1.6 optimization techniques such as the Newton Raphson procedure and the Gauss Newton method are investigated. Section 1.7 considers estimation where residuals are correlated, and the maximum likelihood estimation of growth data. Section 1.8 examines the situation where the data are incomplete and censored estimation is used.

1.1 CURVE FITTING — WHY?

Before mentioning the various curves and their fitting to data, it is necessary to answer the question 'Why must a curve be fitted?'

Steyn, Du Toit and Smit (1981) consider the following reasons.

(i) Extrapolation

An important aim in statistical research is to make forecasts about observed data for a given data set.

If a function $\hat{y} = f(x)$ describes the relationship between the variables x and y well, then it is possible to forecast for a given x -value outside the interval of observed x values, only if the relationship still holds for this particular x value.

This process is known as extrapolation. Consider the example where we fit a curve to n pairs of observations (x,y) , where x is the work rate expressed as a percentage of maximum oxygen uptake, and y is the corresponding blood lactic acid concentration of an individual while exercising. If the lactic acid concentration is determined by the following x -values 20 30 40 50 60 (% $\dot{V}O_2$ max), then

$\hat{y} = f(70)$ is an estimate of the blood lactic acid concentration at a work rate of 70.

It is customary to represent the variable to be extrapolated by y . This variable is the dependent variable whilst x -value is the independent or predictor variable.

In general, extrapolation must be applied with caution, because the relationship might only be applicable within the interval of observed x -values.

(ii) Interpolation

Often it is necessary to estimate a y -value for a x -value within an interval of x -values. This process is known as interpolation.

Consider the example where the rate of oxygen uptake (y) for a specific individual is known if the work rate (x) is equal to 25 50 75 and 100% respectively. If a curve is fitted to these data, then $\hat{y} = f(35)$ is the estimated oxygen uptake for the individual if the work rate is 35%.

Note that 35% is a value in the interval 25 to 50% and that we therefore interpolated.

(iii) Postulating an empirical law

The data might be of such a nature that it might not be theoretically justified to use the process of extrapolation or interpolation.

We can however use the curve that is fitted to the data, to postulate an empirical law for the data set.

1.2 LINEAR CURVES OR CURVES THAT ARE TRANSFORMABLE TO THE LINEAR FORM.

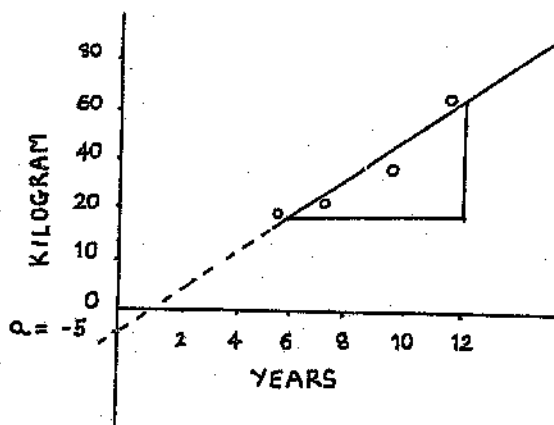
(i) The straight line $\hat{y} = a + bx$.

A plot of a bivariate data set (x_i, y_i) $i = 1, 2, \dots, n$ often indicates that there exists a linear relationship of the form $\hat{y} = a + bx$ between x and y .

In the presentation of results it is important to consider whether the value of the intercept a , has any practical importance or not.

To illustrate this point consider the following plot that represents the relationship between age(x) and mass(y) and that a curve $\hat{y} = a + bx$ was fitted to the data.

Figure 1.1 Straight line $y = a + bx$



If we are prepared to extrapolate outside the interval of the observed x -values for the value $x = 0$, then $\hat{y} = -5$.

From a mathematical point of view it means that the specific person weighed -5 kilograms at birth.

It has already been stressed that one must approach extrapolation with caution, because the assumed relationship between x and y is not necessarily linear outside the boundaries of the observed x -values.

The value of b is however of practical interest. Practically it measures the average change in y for a unit change in x . In the example above, b indicates that there is an average increase of 7 kg for every year that this specific person got older. The interpretation of b is usually only valid within the interval of observed x -values.

(ii) The exponential curve (Single log transformation)

In the case of a linear relationship between x and y where a straight line is fitted to the data, the rate of change of $\hat{y}$ remained constant.

In economical and demographic data where the x values are often equally spaced, it will result in the corresponding $\hat{y}$ -values forming an arithmetic sequence.

This type of data however often forms approximately a geometric, rather than an arithmetic sequence. e.g. When the growth is of a percentage increase per time interval (for example compound interest). When this is the case, the exponential curve $\hat{y} = ab^x$ may describe the relationship between x and y better than a straight line.

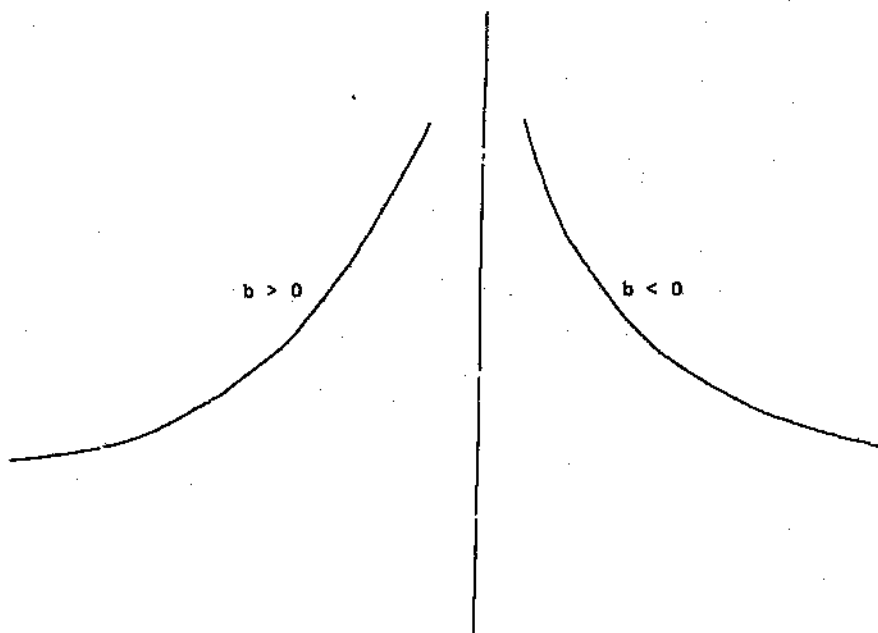
This function is often known as a growth curve. The connection is characterised by a constant ratio b between the y -values that correspond with the x -values 0,1,2,....

Through application of the rules for logarithms it follows that the exponential curve can be written as follows

$$\log \hat{y} = \log a + x \log b.$$

If $\log y$, $\log a$ and $\log b$ can be represented respectively by y' , a' and b' then it follows that $y' = a' + b'x$ is the equation of a straight line with gradient b' and intercept a' . The original form of the equation $\hat{y} = ab^x$ would then be obtained by calculating a and b to be the antilogarithms of a' and b' respectively. See Figure 1.2 for a typical plot of the exponential curve.

Figure 1.2 Typical shapes of the exponential curve $y = ab^x$



(iii) Power Curves (Double log transformation)

If the observed (x, y) points are plotted and no linear relation is indicated, then the $(x, \log y)$ points can be graphically represented and inspected for a linear relationship.

In the case where there still is no linear relationship between x and $\log y$, there might very well be a linear tendency between $\log x$ and $\log y$.

When this is the case $\hat{y} = ax^b$ is a suitable curve to describe the relationship between x and y .

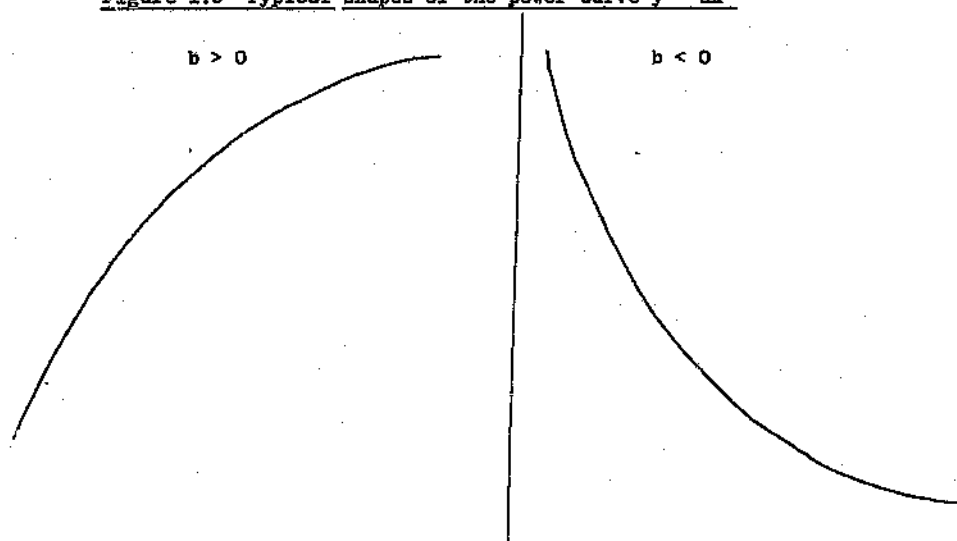
In logarithmic form, this equation can be written as

$$\log \hat{y} = \log a + b \log x.$$

Let y' , a' and x' respectively represent $\log y$, $\log a$ and $\log x$, then we will obtain the equation of a straight line $y' = a' + bx'$.

See Figure 1.3 for a typical plot of the Power Curve.

Figure 1.3 Typical shapes of the power curve $y = ax^b$



(iv) The Hyperbolic curve (Inverse transformation)

A curve with equation

$$\hat{y} = a + b/x, \quad x \neq 0$$

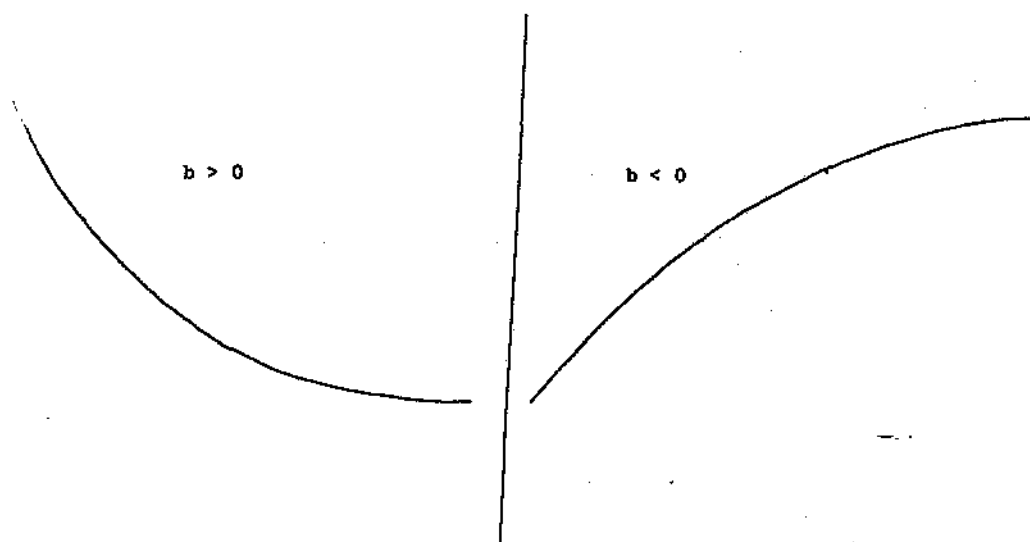
is known as a hyperbolic curve.

Let $z = 1/x$, it follows

$$\hat{y} = a + bz$$

Two possible forms of this curve are given in Figure 1.4.

Figure 1.4 Typical shapes of the hyperbolic curve $y = a + b/x$



(v) The parabola

An expansion of the relationship $\hat{y} = a + bx$ between y and x is a polynomial of degree q namely

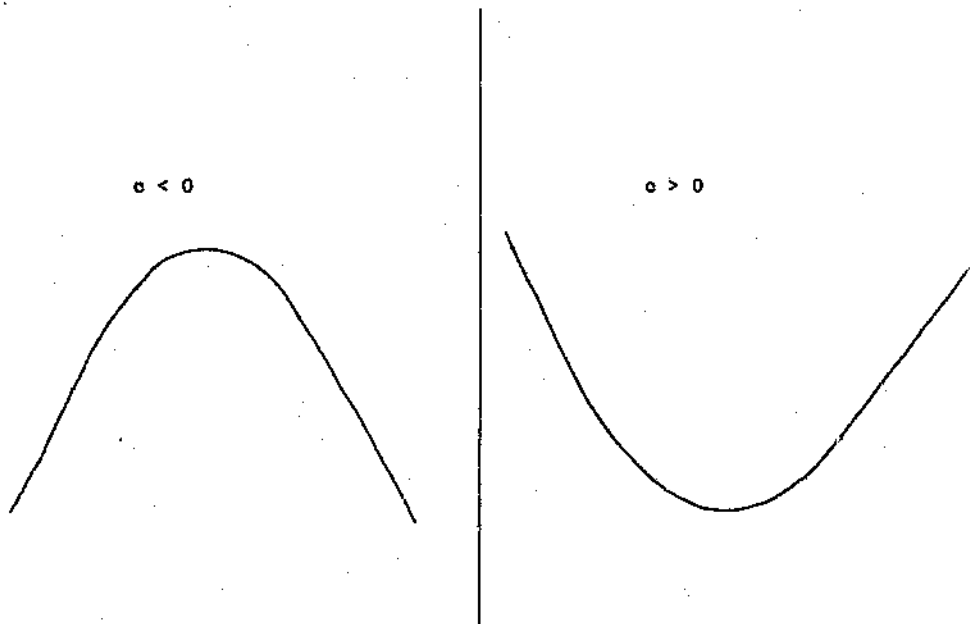
$$\hat{y} = a + bx + cx^2 + \dots + lx^q$$

where the degree q can be any values from 2, 3, ..., $n - 1$.

Although the relationship between x and y may be better described the larger q becomes, there are practical and theoretical considerations for keeping a polynomial's degree as low as possible. One such reason is the principle of parsimony which advises us to use models with the minimum number of parameters.

In polynomial regression, the assumed polynomial form of the model is usually chosen for convenience, so that a polynomial of given degree is not assumed to be the true regression model. Because of this, the testing of individual polynomial coefficients is little more than a guide in deciding on the degree of polynomial to be fitted. Of far more importance is a decision on what degree of variability about the regression model is acceptable, and this decision will be based on practical rather than merely statistical considerations. (William H. Kruskal and Judith M. Tanur, *International Encyclopaedia of Statistics*.) Figure 1.5 represents a typical plot of the parabola.

Figure 1.5 Typical shapes of the parabola $y = a + bx + cx^2$



(vi) Orthogonal polynomials

Draper and Smith (1981) discuss the properties of orthogonal polynomials.

Orthogonal polynomials are used to fit a polynomial of any order in one variable to a set of n data points (x_i, y_i) $i = 1, 2, \dots, n$.

Suppose we wish to fit the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_r X^r + \epsilon$$

where Y is the response variable and X is the predictor variable. Generally the columns of the resulting design matrix X will not be orthogonal.

Therefore we can construct polynomials of the form

$$\begin{aligned} \phi_0(X_i) &= 1 && \text{zero-order polynomial} \\ \phi_1(X_i) &= R_1 X_i + S_1 && \text{first order polynomial} \\ \phi_2(X_i) &= R_2 X_i^2 + S_2 X_i + T_2 && \text{second order polynomial} \\ &\vdots && \\ \phi_m(X_i) &= R_m X_i^m + S_m X_i^{m-1} + \dots + V_m && \text{mth order polynomial} \\ &\vdots && \end{aligned}$$

with the property that they are orthogonal polynomials, that is

$$\sum_{i=1}^n \phi_j(X_i) \phi_\ell(X_i) = 0 \quad (j \neq \ell) \quad (1.2.1)$$

for all $j, \ell < n-1$. We can thus rewrite the model as

$$Y = a_0 \phi_0(X) + a_1 \phi_1(X) + \dots + a_r \phi_r(X) + \epsilon \quad (1.2.2)$$

In this case we have

$$X = \begin{bmatrix} 1 & \phi_1(X_1) & \phi_2(X_1) & \dots & \phi_r(X_1) \\ 1 & \phi_1(X_2) & \phi_2(X_2) & \dots & \phi_r(X_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(X_n) & \phi_2(X_n) & \dots & \phi_r(X_n) \end{bmatrix}$$

so that $X'X = \begin{bmatrix} a_{00} & & & 0 \\ & a_{11} & & \\ & & a_{22} & \\ 0 & & & \ddots & \\ & & & & a_{nn} \end{bmatrix}$

where $a_{jj} = \sum_{i=1}^n \{\phi_j(X_i)\}^2$ and $a_{00} = n$ since all the off diagonal terms disappear (from (1.2.1)).

Since $(X'X)^{-1}$ also consists of diagonal elements the least squares procedure can be used to calculate estimates for a_j , $j = 0, \dots, r$ in (1.2.2), simplifies to:

$$\text{i.e. } \hat{a}_j = \frac{\sum_{i=1}^n Y_i \phi_j(X_i)}{\sum_{i=1}^n [\phi_j(X_i)]^2} \quad j = 0, 1, 2, \dots, r$$

$$= \frac{P_{jY}}{P_{jj}}$$

Pearson and Harley (1966) provide formulas for orthogonal polynomials $\phi_j(X)$ up to order six for equally spaced X 's and for any value of n :

$$\begin{aligned} \phi_0(X) &= 1 \\ \phi_1(X) &= \lambda_1 X \\ \phi_2(X) &= \lambda_2 \left\{ X^2 - \frac{1}{12}(n^2 - 1) \right\} \\ \phi_3(X) &= \lambda_3 \left\{ X^3 - \frac{1}{12}(3n^2 - 7)X \right\} \\ \phi_4(X) &= \lambda_4 \left\{ X^4 - \frac{1}{14}(3n^2 - 13)X^2 + \frac{3}{560}(n^2 - 1)(n^2 - 9) \right\} \\ \phi_5(X) &= \lambda_5 \left\{ X^5 - \frac{5}{18}(n^2 - 7)X^3 + \frac{1}{1008}(15n^4 - 230n^2 + 407)X \right\} \\ \phi_6(X) &= \lambda_6 \left\{ X^6 - \frac{5}{44}(3n^2 - 31)X^4 + \frac{1}{176}(5n^4 - 110n^2 + 329)X^2 \right. \\ &\quad \left. - \frac{5}{14784}(n^2 - 1)(n^2 - 9)(n^2 - 25) \right\} \end{aligned} \quad (1.2.3)$$

1.3 NONLINEAR GROWTH MODELS

The class of curves that have been considered so far can be classified as linear in the parameters or transformable to a linear form.

In this section we will consider a family of growth curves that are not transformable to the linear form. These curves are usually applied when it is necessary to describe the relationship between growth (y) and time (t) — from whence the name growth curve, arises.

The type of growth model used in a specific area and in a specific problem depends on the type of growth that occurs. In general, growth models are mechanistic rather than empirical. A mechanistic model usually arises as a result of assumptions being made about the type of growth, resulting differential or difference equations that represent these assumptions, and then solving these equations to obtain a growth model.

An empirical model, on the other hand, is a model chosen to approximate an unknown mechanistic model empirically.

The Monomolecular Model

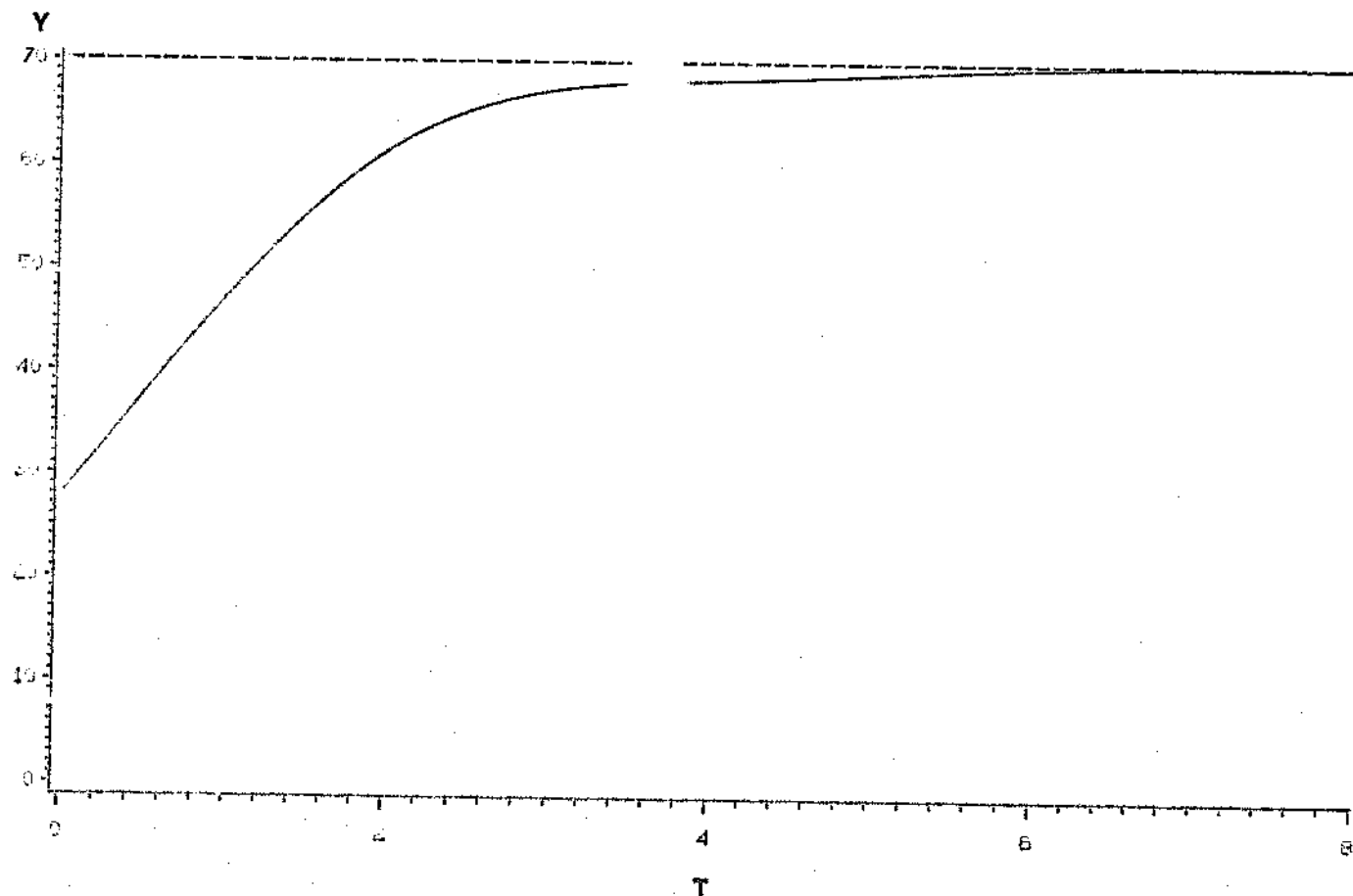
One assumes that for a specific growth situation the rate of growth is directly proportional to the amount of growth yet to be attained. This growth rate can be represented by the following formula.

$$\frac{dw}{dt} = k(a-w) \quad (1.3.1)$$

where a is the maximum growth size, w is the size of growth at time t and k is the constant of the growth rate pattern.

Figure 1.6

MONOMOLECULAR FUNCTION
 $Y=70*(1-0.8*(0.45)^{-T})$



Through simplification and integration of (1.3.1), one attains the desired growth function.

$$\text{i.e. } \frac{dw}{a-w} = kdt$$

$$\int \frac{dw}{a-w} = k \int dt$$

$$\int \frac{dw}{a-w} = kt + c$$

$$\Rightarrow -\ln(a-w) = kt + c$$

$$\Rightarrow w = a - e^{-kt} e^{-c}$$

$$\text{Letting } e^{-c} = \beta a$$

$$w = a(1 - \beta e^{-kt}) \quad (1.3.2)$$

which is known as the monomolecular function. See Figure 1.6 for an illustration of this curve where $a = 70$, $\beta = 0.6$, $e^{-k} = 0.45$. The reason for this name is that (1.3.2) is the equation that is used for a "monomolecular" chemical reaction of the first order, in which context a represents the initial amount of a substance, and w the quantity of it that has disappeared by the time t .

Letting $t = 0$ in this monomolecular function one can see that $\beta = (a - w_0)/a$ where w_0 is the value of w when $t = 0$, so that the constant β represents the potential growth relative to a from $t = 0$ to the limiting value a . The function therefore rises from $t = 0$ where $w = a(1 - \beta)$ to the limiting value a . The monomolecular function has no point of inflection since there is no change of sign in the second derivative for any t .

Different values of a will have differing effects on the vertical scale of the curve, whilst different values for k will have a contracting or stretching effect on the horizontal scale of the curve.

One can see the effect of varying values of β on the monomolecular function from Figure 1.7, where $w_0 = 70(1-\beta)$ is the starting growth value. (i.e. at $t = 0$.) and $a = 70$

F.G. Gregory (1928) made use of this function to fit the expansion in area of cotyledons in plants growing at a high temperature (90°F). The argument developed in support of its application is as follows. If the size of a fully expanded cell is large compared with that of a meristematic, the area of the leaf at any time will depend very largely on the number (x) of expanded cells present. The rate of increase in their numbers will be a function of the number of cells in the process of expansion, say $\frac{dx}{dt} = k(a-x)$, where a represents the total cell number. By integration $x = a(1-be^{-kt})$ which is equivalent to (1.3.2) the monomolecular function.

Figure 1.7

MONOMOLECULAR FUNCTION
 $W=70*(1-B*EXP(-T))$

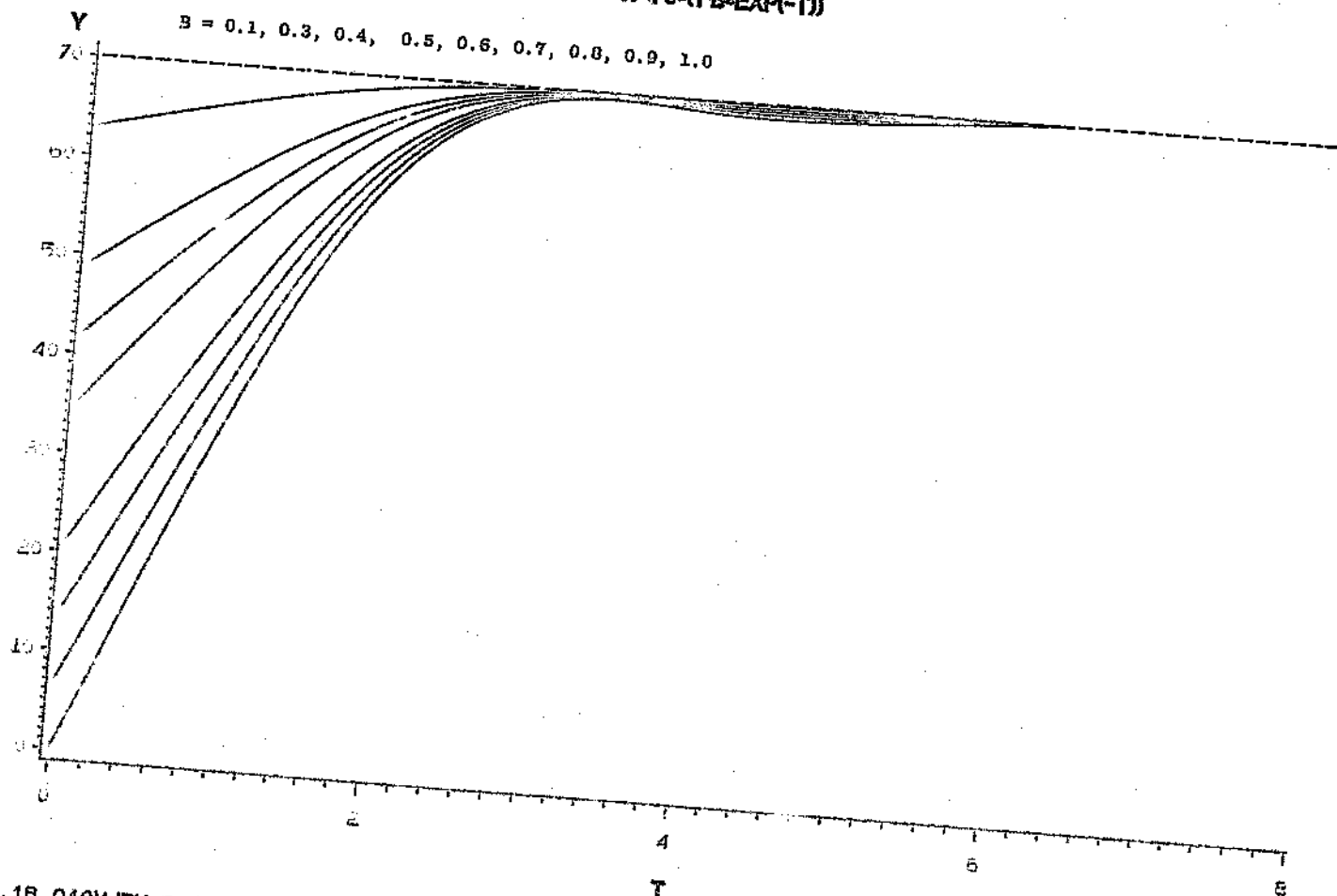
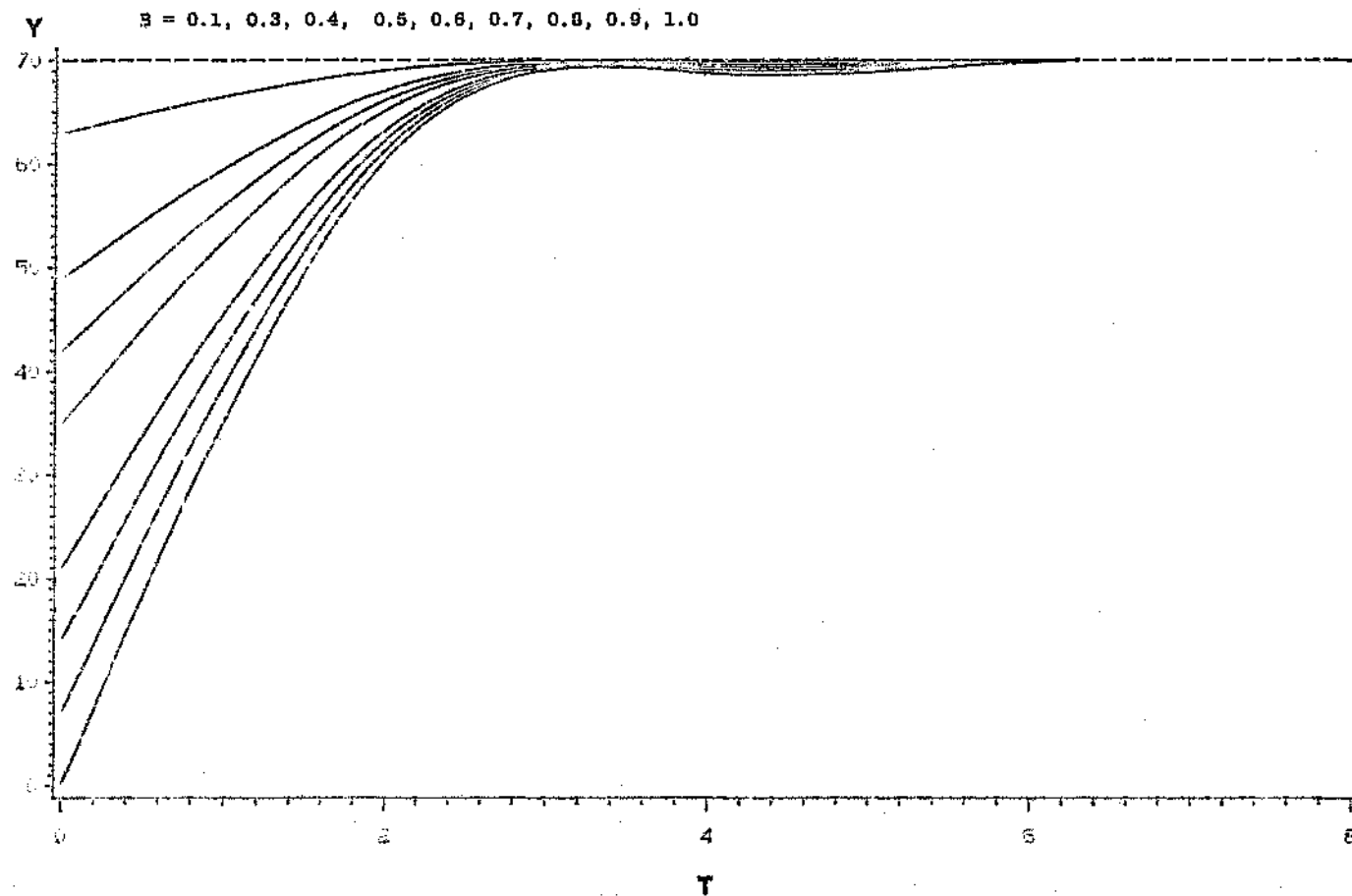


Figure 1.7

MONOMOLECULAR FUNCTION
 $W=70(1-B\cdot\text{EXP}(-T))$



The Logistic Model

Assuming that the growth rate is proportional to the product of the present size and the amount of growth yet to be achieved, with a being the limiting growth value then

$$\frac{dw}{dt} = kw(a-w) \quad (1.3.3)$$

Integrating (1.3.3)

$$\int \frac{dw}{w(a-w)} = k \int dt + c$$

$$\text{Now } \frac{1}{w(a-w)} = \frac{1}{aw} + \frac{1}{a(a-w)}$$

$$\begin{aligned} \text{Therefore } \int \frac{dw}{w(a-w)} &= \frac{1}{a} \ln w - \frac{1}{a} \ln(a-w) \\ &= kt + c \end{aligned}$$

$$\Rightarrow \frac{1}{a} \ln \frac{w}{a-w} = kt + c$$

$$\Rightarrow a = w(1 + e^{-\frac{k}{a}t} \cdot e^{-ac})$$

$$\text{Let } \beta = e^{-ac}, K = \frac{k}{a}$$

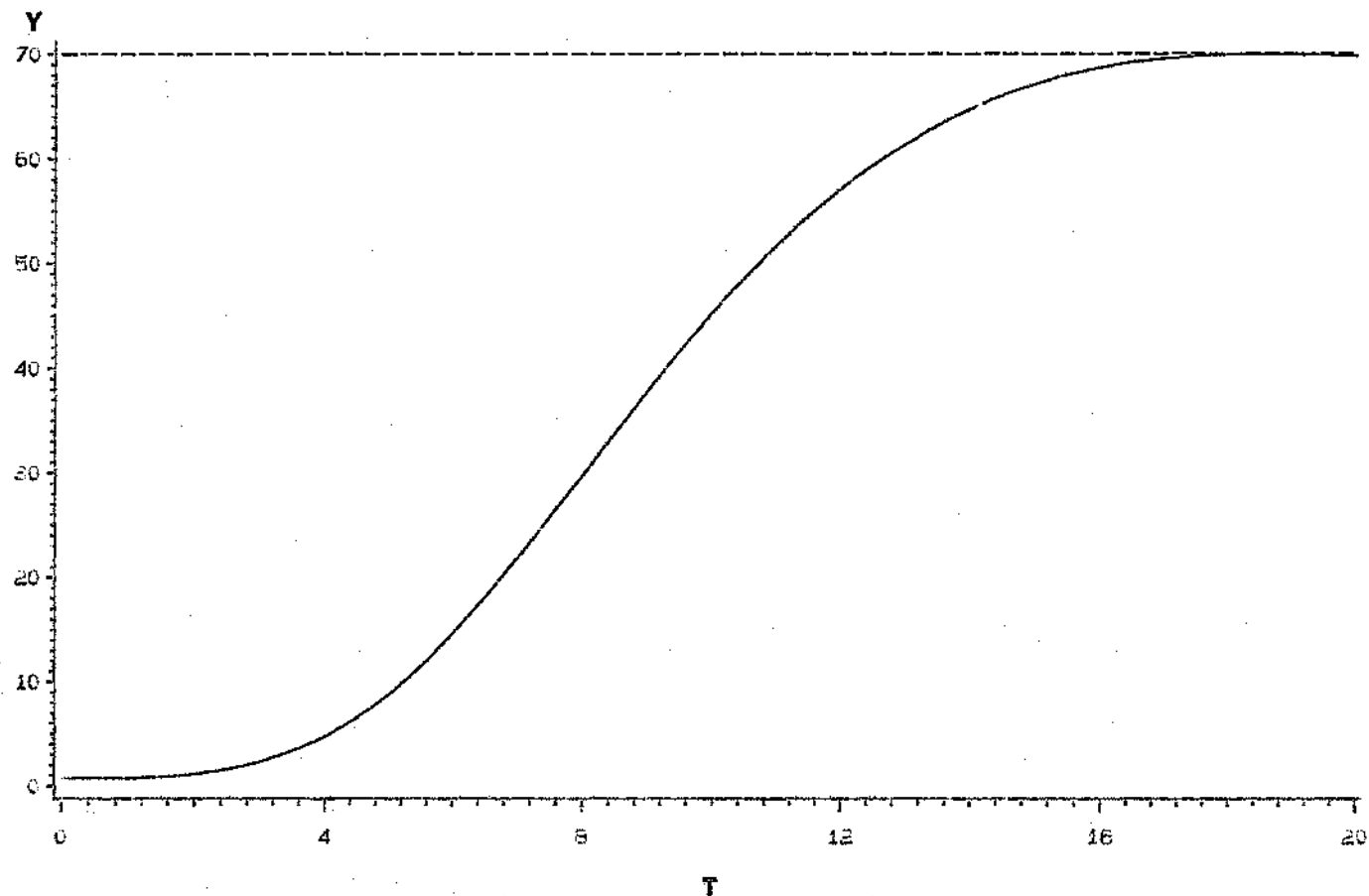
$$\text{so that } w = a/(1 + \beta e^{-Kt}) \quad (1.3.4)$$

which is known as the logistic or autocatalytic function.

See Figure 1.8 for a plot of this graph where $a = 10$, $\beta = 90$, $e^{-k} = 0.6$.

Figure 1.8

LOGISTIC FUNCTION
 $Y = 70 / (1 + 90 * (0.8)^{-T})$



The name "autocatalytic" comes from chemical theory where this equation (1.3.4) represents the progress of an autocatalyzed monomolecular reaction, a being the amount of some substance originally present and w the amount of the autocatalyst present at time t . Plotted against time this function yields a S-shaped curve, being asymptotic at its two ends to the lines $w = 0$ and $w = a$, and having its point of inflection at $w = a/2$ where

$$\frac{d^2w}{dt^2} = \frac{k}{a}(a-2w).$$

This point divides the curve into two halves identical in form but oriented in opposite directions. The resemblance between the typical S-curve of growth, whether plant or animal, and the progress of an autocatalyzed monomolecular reaction led Robertson (1927) to the conclusion that growth is indeed determined by some such chemical means.

Under the name "logistic" function, this equation has been used to express the law of human population increase (Pearl and Reed (1920)). The logistic function has also been applied to the growth of higher plants which appear in physiological literature. e.g. to plant height (Reed and Holland (1919)), and to embryo length (Rijven A.H.G. (1952)).

Inverse Symmetry of Logistic Model

Let t_I be the time point of inflection i.e. where $\frac{a}{2} = a/(1+\beta e^{-kt})$

$$\therefore t_I = \frac{(\ln \beta)}{k}.$$

When $t = t_I + u$ (1.3.4) can be written as

$$w = a/(1+e^{-ku})$$

To prove that w has inverse symmetry about its point of inflection let $g(u) = w - w_I$

$$\therefore w - w_I = \frac{a}{2} \left[\frac{2}{1+e^{-ku}} - 1 \right]$$

$$= \frac{a}{2} \left[\frac{1 - e^{-ku}}{1 + e^{-ku}} \right]$$

$$= g(u)$$

$$g(-u) = \frac{a}{2} \left[\frac{1 - e^{ku}}{1 + e^{ku}} \right] \times \frac{e^{-ku}}{e^{-ku}}$$

$$= \frac{a}{2} \left[\frac{e^{-ku} - 1}{e^{-ku} + 1} \right]$$

$$= -\frac{a}{2} \left[\frac{1 - e^{-ku}}{1 + e^{-ku}} \right]$$

$$= -g(u)$$

Since $g(-u) = -g(u)$ it is clear that the curve shape has inverse symmetry about its point of inflection.

The starting growth value when $t = 0$ is $w = a/(1+\beta)$ and for $t = \infty$ $w = a$, the limiting growth value.

One can see from the following two graphs the effects of changing k and β on the function. The varying values of k alters the steepness in the curves, whilst varying value of β , e.g. for $\beta = 1$ in Figure 1.9 and $\beta = 5$, in Figure 1.10, alters the starting point on the vertical axis when $t = 0$.

Figure 1.9

LOGISTIC FUNCTION
 $W=70/(1+\exp(-KT))$

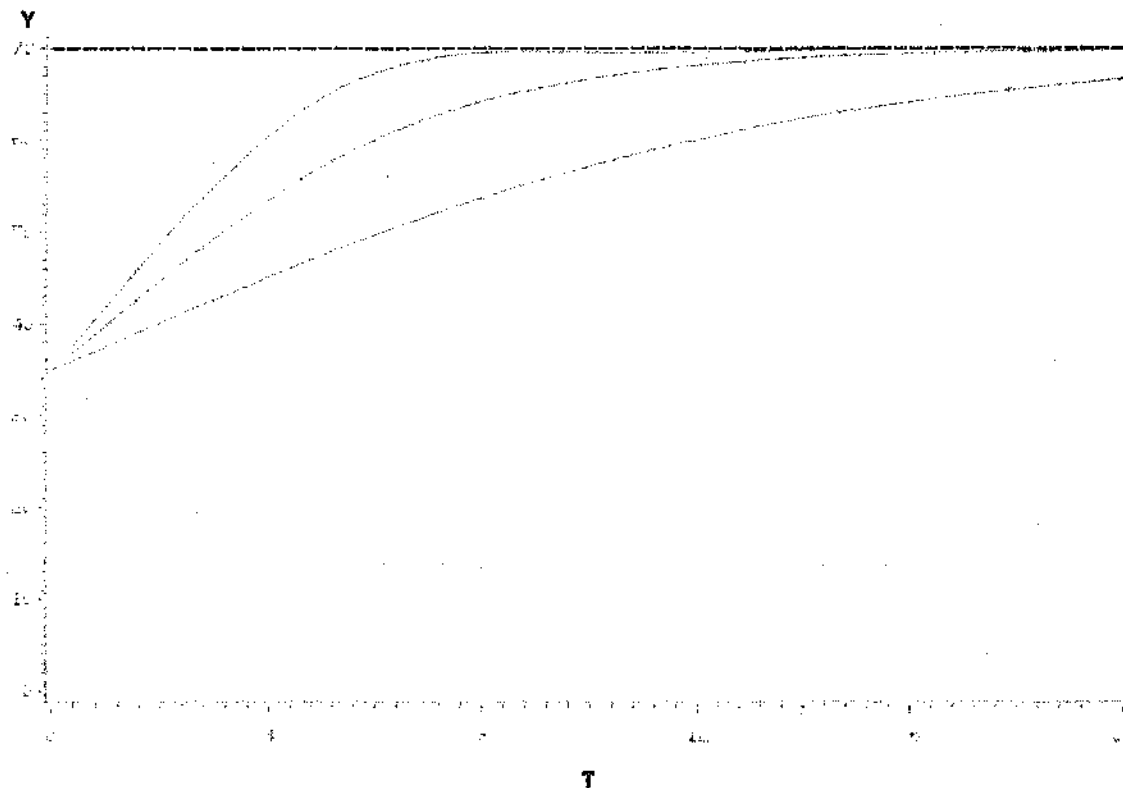
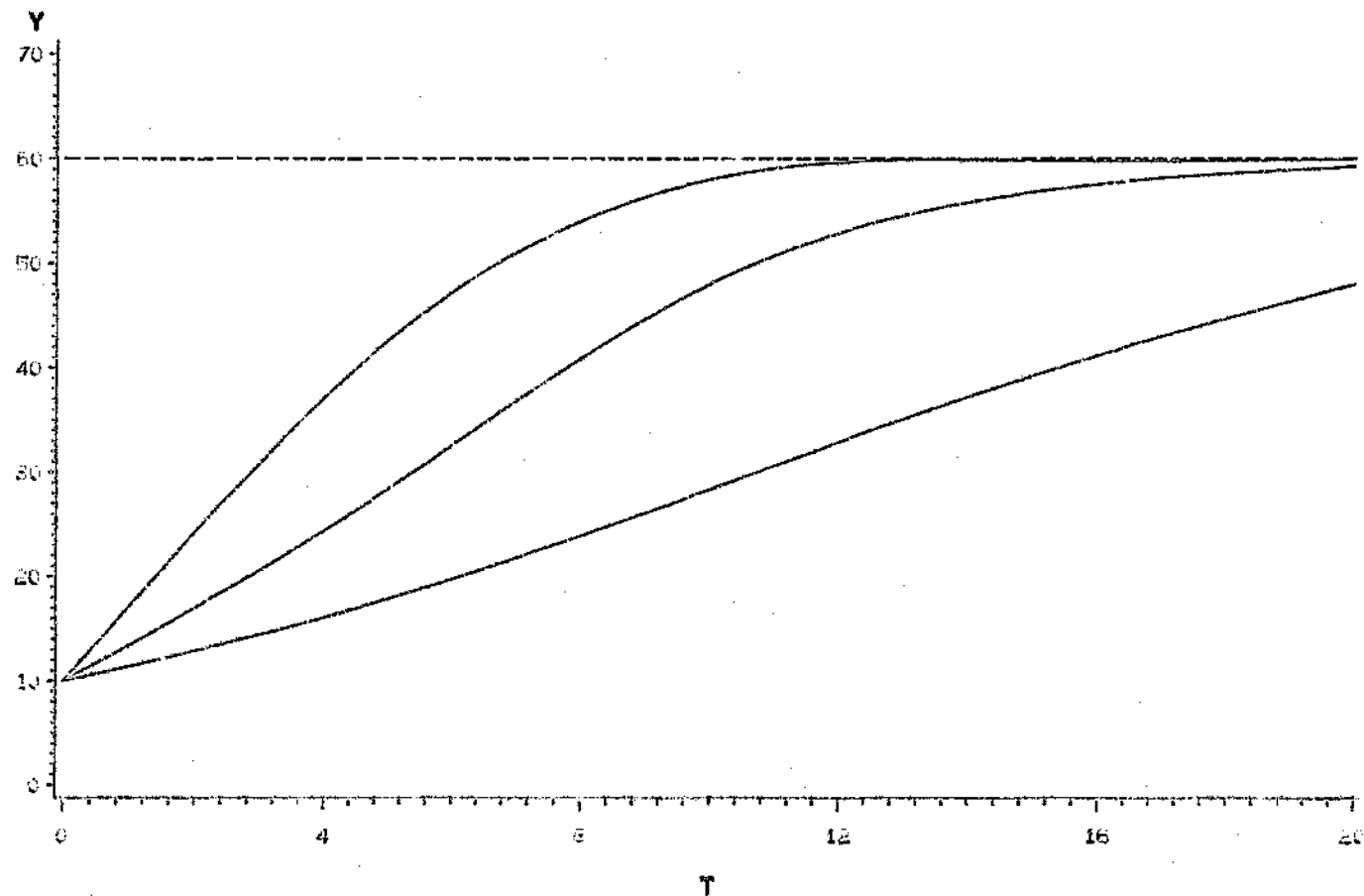


Figure 1.10

LOGISTIC FUNCTION
 $W = 60 / (1 + 5 \cdot \exp(-K(T)))$



Gompertz Model

If there is a negative linear relationship between the relative growth rate and $\log w$ then

$$\frac{1}{w} \cdot \frac{dw}{dt} = k(\log a - \log w) \quad (1.3.5)$$

so that $\frac{dw}{dt} = kw \log(a/w)$

i.e. $\frac{dw}{w \log(a/w)} = k dt$

then integrating both sides

$$\int \frac{dw}{w \log(a/w)} = k \int dt + c$$

$$\Rightarrow \int \frac{dw}{w \log(a/w)} = kt + c \quad (1.3.6)$$

Making the transformation $z = \log(a/w)$ in the integral on the left hand side of (1.3.6) yields

$$-\int \frac{dz}{z} = kt + c$$

$$\Rightarrow \ln z = -kt - c$$

Using the transformation for z

$$\Rightarrow \ln\left(\frac{w}{a}\right) = -\beta e^{-kt} \text{ where } \beta = e^{-c}$$

$$\Rightarrow w = a \exp(-\beta \exp(-kt)) \quad (1.3.7)$$

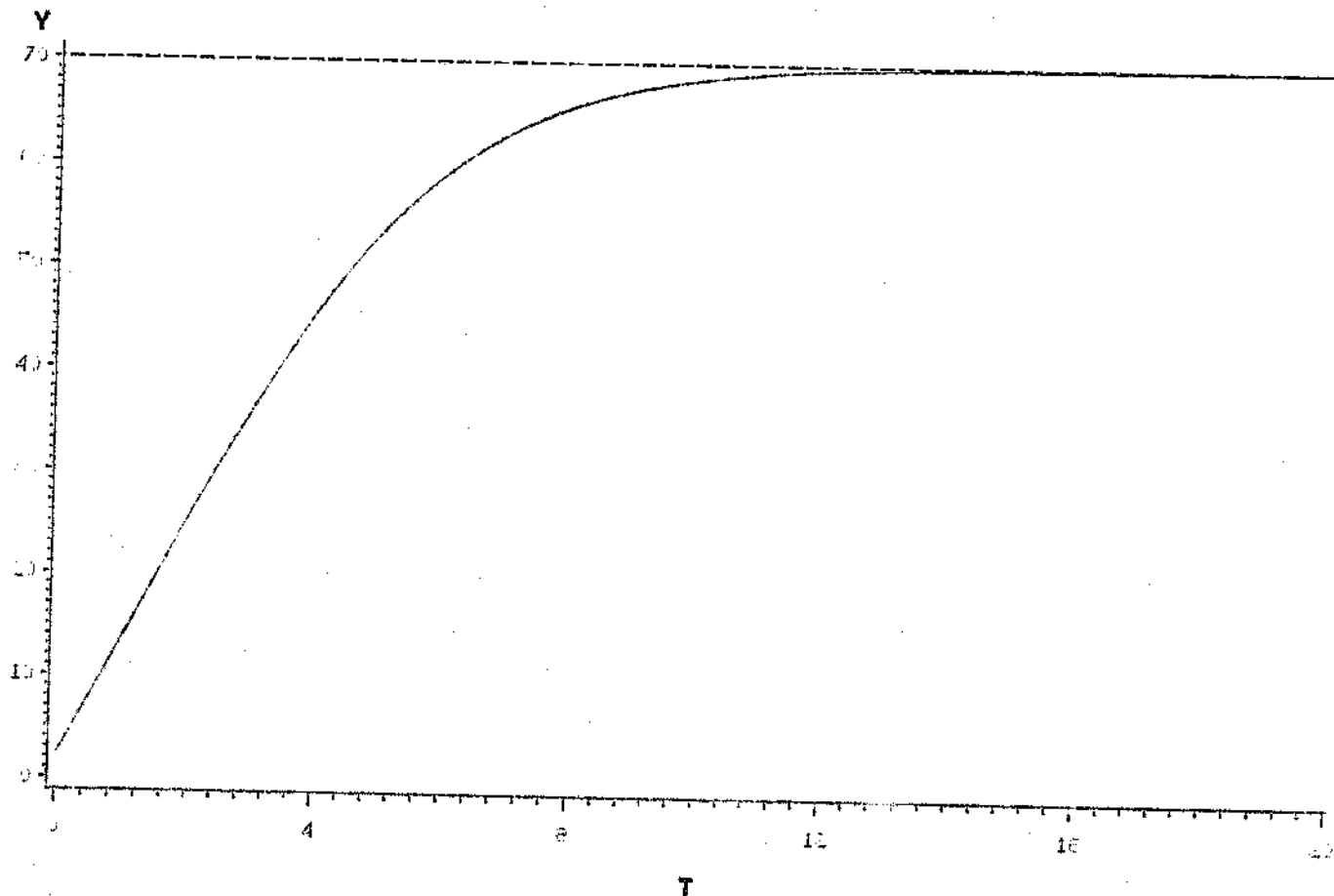
See Figure 1.11 for a plot of this function.

This function has a point of inflection at a/e , however it is not symmetric about this point.

Manipulation of (1.3.7) yields the following relationship

Figure 1.11

GOMPERTZ FUNCTION
 $Y=70 \cdot \text{EXP}(-3.6 \cdot (0.6)^{**}T)$



$$\frac{1}{w} \cdot \frac{dw}{dt} = k\beta e^{-kt}$$

$$\Rightarrow \log\left\{\frac{\frac{dw}{dt}}{w}\right\} = \log k\beta - kt$$

Which indicates that there is a linear relationship between the log relative growth rate and time.

From the plot in Figure 1.12, one can see the effect of a varying β on the Gompertz model. The starting point for this model i.e. when $t = 0$, $w = ae^{-\beta}$ and when $t \rightarrow \infty$ $w \rightarrow a$, the limiting growth value.

This function was introduced by Gompertz (1825) to assist actuaries in describing skew S-curves of population increase. It has also been applied to animal physiology as well as to botanical research.

P.B. Medawar (1940) first deduced and then confirmed, that the growth of a chicken's heart follows the Gompertz form. The change with age in "growth energy" were estimated by determining the concentration of an inhibitor which just prevented growth in explants taken at various ages; the values so determined proved to be proportional to the log relative growth rates of the heart at the time the explants were made, declining exponentially with the age of the organ.

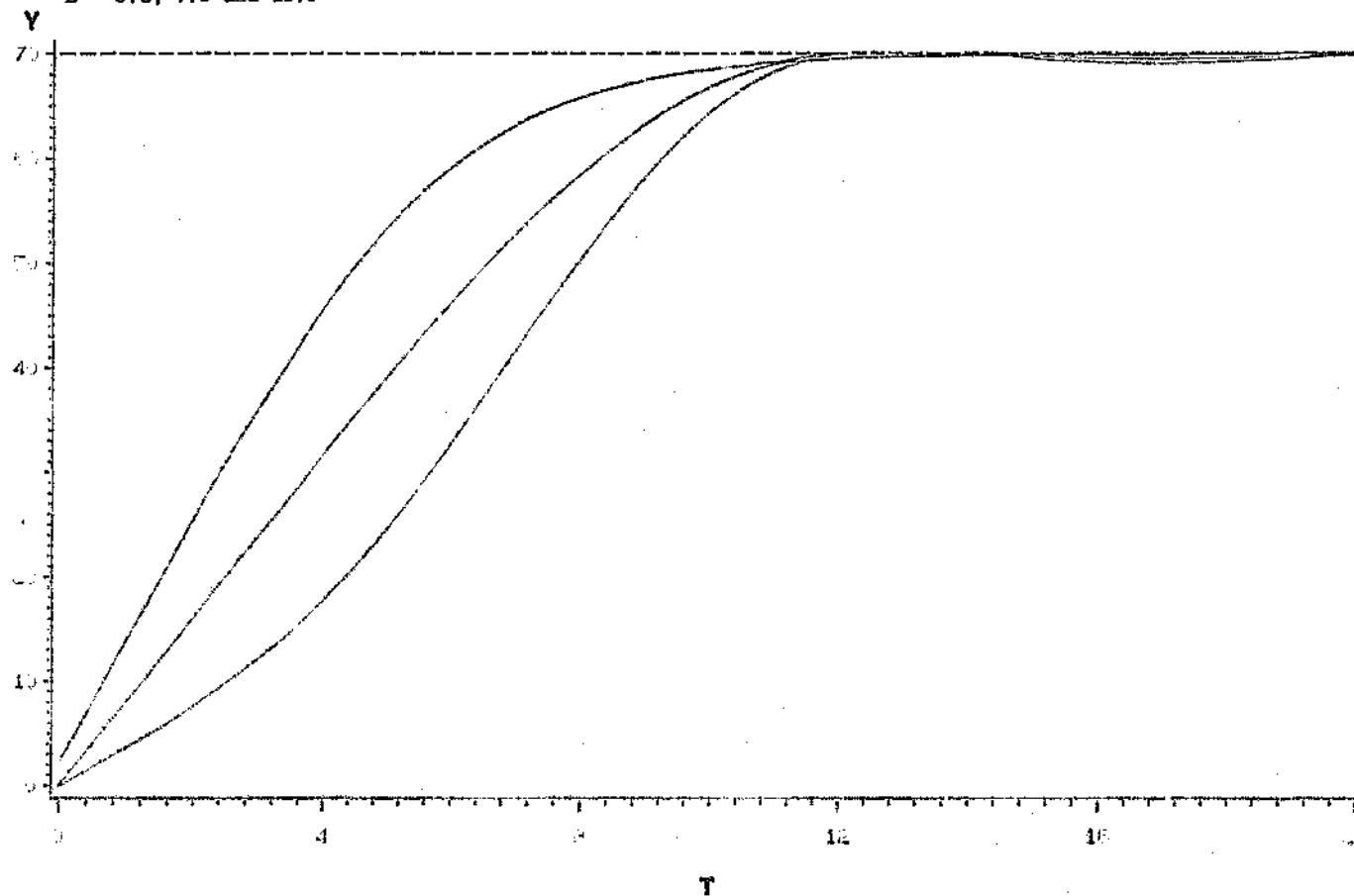
As regards botanical work, Amer and Williams (1957) used the Gompertz function more empirically, in a study of the growth in area of single *Pelargonium* leaves. The value of k proved to be constant under varying conditions e.g. water supply.

Due to the fact that k remained constant, β has no physiological significance and a is concerned only with the final size, the effects of treatment differences could be calculated on the value of k .

Figure 1.13

GOMPERTZ FUNCTION
 $W=70 \cdot \exp(-B \cdot (0.8)^{0.5 \cdot T})$

$B = 3.5, 7.5 \text{ and } 13.5$



The Richards Family of Growth Curves

The last three growth curves, i.e. the monomolecular, logistic, and Gompertz curves are known as the Richards family of Growth curves.

Du Toit (1979) has written this family as a five-parameter function

$$f(\underline{a}, t_i) = a(1+s\beta_{(s)}\rho^{t_i})^\lambda, \quad i = 1, \dots, n \dots \quad (1.3.8)$$

where

$$\underline{a}' = (a, \beta_{(s)}, \rho), \quad s = 1, -1$$

If $f(\underline{a}, t_i)$ is to be an increasing function in t_i , then $s = 1$ for $\lambda < 0$ and $s = -1$ for $\lambda > 0$.

For a decreasing function the signs of s are the opposite.

The parameters have the following restrictions.

$$a \geq 0$$

$$\beta_{(1)} \geq 0; \quad 0 \leq \beta_{(-1)} \leq 1$$

$$0 \leq \rho \leq 1 \quad \text{where } \rho = e^{-k}.$$

This implies $0 \leq e^{-k} \leq 1$

$$\text{i.e. } -\infty \leq -k \leq 0$$

$$\text{i.e. } k \geq 0.$$

$$(1.3.9)$$

Where k is the rate constant of the growth pattern, altering the steepness of the curve.

Note that s is just a constant that determines the sign of the parameter $\beta_{(s)}$, and it

prevents the possibility of $(1+s\beta_{(s)}\rho^{t_i})^\lambda$ having complex roots.

The three growth curves follow as special cases of (1.3.8) n1.

$$(a) \quad f(\underline{a}, t_i) = a(1+s\beta_{(s)}\rho^{t_i}), \quad \lambda = 1, \rho = e^{-k} \dots \text{monomolecular}$$

$$(b) \quad f(\underline{a}, t_i) = a/(1+s\beta_{(s)}\rho^{t_i}), \quad \lambda = -1, \rho = e^{-k} \dots \text{logistic}$$

(c) $f(\underline{a}, t_i) = ae^{-\beta^* t_i}, \quad |\lambda| \rightarrow \infty \quad \beta \rightarrow 0 \dots \text{Gompertz}$
 where $\beta^* = s\beta(s)^{\lambda}$.

Although polynomials may present a good fit to a set of growth data, the Richards family of curves are usually preferred because each parameter has a specific physical meaning.

As mentioned before, a is the limiting value of the observed function e.g. final size of a plant or an organ. β represents the potential increase (or decrease) in the value of the function during the time interval t_1 to t_n , and the parameter ρ represents the growth rate.

1.4 REPARAMETRIZATION

Due to the function (1.3.8) being non-linear in its parameters, it is not possible by using least squares to acquire a set of estimation equations where the solution is in closed form. This implies that an iterative procedure is necessary in the estimation process.

Note that the parameters of the growth curve have certain restrictions (see (1.3.9)).

A suitable transformation of the parameters can be implemented to simplify the optimising problem to one that has no restrictions i.e. unrestricted least squares.

Consider now the following set of transformations (Du Toit 1979).

$$\begin{aligned}\theta_1 &= \ln \alpha \\ \theta_2 &= \ln \beta_{(s)}, \quad s = 1 \\ &= \ln \left(\frac{1 - \beta_{(s)}}{\beta_{(s)}} \right), \quad s = -1 \\ \theta_3 &= \ln \left(\frac{1 - \rho}{\rho} \right)\end{aligned}\tag{1.4.1}$$

The original parameters can thus be expressed in the following unrestricted form

$$\begin{aligned}\alpha &= e^{\theta_1} \quad -\infty < \theta_1 < \infty \\ \beta_{(s)} &= [\delta_s + e^{\theta_2}]^s, \quad -\infty < \theta_2 < \infty \\ \rho &= \frac{1}{1 + e^{\theta_3}}, \quad -\infty < \theta_3 < \infty\end{aligned}$$

where

$$\begin{aligned}\delta_s &= 1 \quad \text{if } s = -1 \\ &= 0 \quad \text{if } s = 1.\end{aligned}\tag{1.4.2}$$

Let y_i denote a set of n responses which have an assumed structure

$$y_i = f(\underline{\theta}, t_i) + e_i, \quad i = 1, \dots, n\tag{1.4.3}$$

where it is assumed that the error terms e_i are uncorrelated and have an expected value of zero and variance σ^2 i.e.

$$E(e_i) = 0$$

$$\text{var}(e_i) = \sigma^2 > 0$$

The response function $f(\underline{\theta}, t_i)$ is the reparametrised function (1.3.8) and is defined as follows.

$$f(\underline{\theta}, t_i) = e^{\theta_1} \left[1 + \frac{s(\theta_{s-1} + e^{\theta_2})^s}{(1 + e^{\theta_3})^{t_i}} \right] \quad (1.4.4)$$

The least squares estimator $\underline{\theta}$ can be defined as the value $\underline{\hat{\theta}}$ which minimizes

$$F(\underline{\theta}) = \sum_{i=1}^n \{y_i - f(\underline{\theta}, t_i)\}^2 \quad (1.4.5)$$

It can be shown that (See section 1.6)

$$\frac{\partial F(\underline{\theta})}{\partial \theta_j} = -2 \sum_{i=1}^n \{y_i - f(\underline{\theta}, t_i)\} \frac{\partial f(\underline{\theta}, t_i)}{\partial \theta_j} \quad (1.4.6)$$

$$\frac{\partial^2 F(\underline{\theta})}{\partial \theta_k \partial \theta_j} = 2 \sum_{i=1}^n \left\{ \frac{\partial f(\underline{\theta}, t_i)}{\partial \theta_k} \cdot \frac{\partial f(\underline{\theta}, t_i)}{\partial \theta_j} - [y_i - f(\underline{\theta}, t_i)] \frac{\partial^2 f(\underline{\theta}, t_i)}{\partial \theta_k \partial \theta_j} \right\} \quad (1.4.7)$$

These equations will be used in the estimation process in Section 1.5.

1.5 ESTIMATION

There are techniques which provide good estimators, and which are widely used in applications. Two techniques to be considered are maximum likelihood estimation and the method of least squares.

Maximum likelihood estimation

Essentially maximum likelihood estimation involves estimating a parameter (or several parameters) by maximizing the likelihood function, this being simply the joint probability (density) of the observations regarded as a function of the parameter(s).

In other words the maximum likelihood estimate of a parameter θ is that value which makes the observed sample most likely to have occurred. So given a sample of n independent values, $x_1, \dots, x_n$ from the density function of $f(x, \theta)$ dependent on the single parameter θ , the likelihood function is given by

$$L(x_1, \dots, x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta).$$

The maximum likelihood principle implies that we take as our estimator of θ , that value $\hat{\theta}$, say, within the admissible range of θ which makes the likelihood function as large as possible.

If the likelihood function is a twice differentiable function of θ then $\hat{\theta}$ will be given by the root of

$$\frac{dL}{d\theta} = 0 \quad (1.5.1)$$

for which

$$\frac{d^2L}{d\theta^2} < 0 \quad (1.5.2)$$

If (1.5.1) and (1.5.2) are satisfied $\hat{\theta}$ will correspond to a possibly local maximum of the likelihood function. The value of $\hat{\theta}$ corresponding to the largest of these local maxima (if there are in fact more than one) will be the required estimate of θ .

In practice it is often simpler to work with the logarithm of the likelihood function which, since it has its maxima at the same positions as L , will lead to the same estimated value of θ .

Maximum likelihood estimators have desirable statistical qualities such as consistency and asymptotic normality. See for example B.S. Everitt (1987).

Least squares

Least squares estimation is used for the following sort of model. Suppose there are n observations of a variable Y , and corresponding values of variables $X_1, X_2, \dots, X_k$ which in some way are meant to predict the behaviour in Y . To be precise suppose that

$$E(Y) = f(X_1, X_2, \dots, X_k; \theta_1, \theta_2, \dots, \theta_r)$$

where f is a known function of the X_i 's and of the r unknown parameters $\theta_1, \theta_2, \dots, \theta_r$. Since Y will fluctuate around $E(Y)$, one plausible way to estimate $\theta_1, \theta_2, \dots, \theta_r$ is to use those values of $\theta_1, \theta_2, \dots, \theta_r$ for which the sum of the squared deviations of the observed values of Y from their corresponding $E(Y)$'s is a minimum. In other words if the n sets of observations are $(Y_i, X_{1i}, X_{2i}, \dots, X_{ki})$ $i = 1, 2, \dots, n$ the least squares estimates for $\theta_1, \dots, \theta_r$ will be the values of $\theta_1, \dots, \theta_r$ for which

$$S = \sum_{i=1}^n (Y_i - f(X_{1i}, X_{2i}, \dots, X_{ki}; \theta_1, \dots, \theta_r))^2 \quad (1.5.3)$$

is a minimum.

If it so happens that the $Y_i - f(X_{1i}, X_{2i}, \dots, X_{ki}; \theta_1, \dots, \theta_r) = \epsilon_i$ are independent normal errors, all with the same variance, then the method of least squares estimation is the same as that of maximum likelihood estimation.

This is true in particular for the general linear model which assumes that $k = r$ and

$$f(X_1, X_2, \dots, X_r; \theta_1, \theta_2, \dots, \theta_r) = \sum_{j=1}^r \theta_j x_j \quad (1.5.4)$$

According to the Gauss-Markov theorem, if the ε_i 's are uncorrelated, and the same variance, then no matter what their distribution is, no unbiased estimator of θ_j the

form $\sum_{i=1}^n a_i Y_i$ has a smaller variance than the least squares estimator of θ_j

(H. Scheffe (1959)).

Weighted Least Squares

In practice there may be additional information about the Y_i 's, for example that some observations are less reliable (i.e. have larger variance) than others. In such a case it may be necessary to weight the terms in (1.5.3) accordingly and minimize the sum

$$s_w = \sum_{i=1}^N w_i (y_i - \mu_i(\theta))^2 \quad (1.5.5)$$

where $\mu_i(\theta) = f(X_{1i}, X_{2i}, \dots, X_{ri}, \theta_1, \dots, \theta_r)$, and the terms w_i represent weights, e.g. $w_i = [\text{var}(Y_i)]^{-1}$.

More generally the Y_i 's may be correlated. Let V denote their variance-covariance matrix. In matrix notation (1.5.5) can be written as

$$s_w = (y - \mu)^T V^{-1} (y - \mu)$$

where $y = [y_1, \dots, y_N]^T$ and $\mu = [\mu_1, \dots, \mu_N]^T$.

In particular if the terms μ_i are linear combinations of parameters θ_j ($j=1, \dots, r < n$)

i.e. $\mu = X\theta$ for some $n \times p$ matrix X , then

$$s_w = (Y - X\theta)^T V^{-1} (Y - X\theta) \quad (1.5.6)$$

For (1.5.6) the derivatives of s_w with respect to the elements θ_j of θ are

$$\frac{\partial s_w}{\partial \theta} = -2X^T V^{-1} (Y - X\theta)$$

so the weighted least squares estimator of β is the solution of the normal equations

$$X^T V^{-1} X \hat{\theta} = X^T V^{-1} Y \quad (1.5.7)$$

When $V = \sigma^2 I$, (1.5.7) becomes $X^T X \hat{\theta} = X^T Y$ which is the normal equations for unweighted least squares.

Comparison between Maximum Likelihood estimation and Least Squares.

An important distinction between the methods of least squares and maximum likelihood is that the former can be used without making assumptions about the distributions of the response variables Y_i beyond specifying their expectations and possibly their variance-covariance structure.

In contrast, to obtain maximum likelihood estimators it is necessary to specify the joint probability distribution of the Y_i 's.

However, to obtain the sampling distribution of the least squares estimators $\hat{\theta}$ additional assumptions about the Y_i 's are generally required. Thus in practice there is little advantage in using the method of least squares unless the estimation equations are computationally simpler. (A. Dobson 1983)

Since for nonlinear models the set of equations

$$\frac{\partial F(\theta)}{\partial \theta_j} = 0, \quad j = 1, 2, 3$$

cannot be solved explicitly for θ (see (1.4.6)), an optimization procedure is necessary to estimate $\hat{\theta}$.

This optimization procedure usually requires the estimation of initial parameters.

Estimation theory

It can be shown (Malinvaud, 1970) that the least squares estimate $\hat{\theta}$ is asymptotically distributed $N(\underline{\theta}, \phi)$ (Goldfeldt and Quandt, 1972)

$$(1.5.12)$$

whereby $\phi \approx 2\hat{\sigma}^2 H^{-1}$

The (3x3) matrix H has typical elements (see(1.4.7))

$$[H]_{i,j} = \frac{\partial^2 F(\hat{\theta})}{\partial \theta_i \partial \theta_j}, \quad i, j = 1, 2, 3$$

and

$$\hat{\sigma}^2 = \frac{F(\hat{\theta})}{n-3}$$

Estimating initial parameters

Draper and Smith (1981) proposed the following method for estimating initial parameters.

Consider the logistic function (1.3.4) which can be written as

$$y = \delta - \ln(1 + \beta e^{-kt}) \quad (1.5.8)$$

where $y = \ln w$ and $\delta = \ln a$.

Since $\lim_{t \rightarrow \infty} y = \delta$ an initial estimate for δ is $\delta_0 = y_{\max}$, the maximum observed y .

Choosing any two observation points, i and j to obtain initial estimates β_0 and k_0

$$y_i = \delta_0 - \ln(1 + \beta_0 e^{-k_0 t_i}) \quad (1.5.9)$$

$$y_j = \delta_0 - \ln(1 + \beta_0 e^{-k_0 t_j}) \quad (1.5.10)$$

Rewriting (1.5.9) and (1.5.10) in the exponential form

$$\exp(\delta_0 - y_i) - 1 = \beta_0 \exp(-k_0 t_i)$$

$$\exp(\delta_0 - y_j) - 1 = \beta_0 \exp(-k_0 t_j)$$

By division and simplification

$$\frac{\exp(\delta_0 - y_i) - 1}{\exp(\delta_0 - y_j) - 1} = \frac{\exp(-k_0 t_i)}{\exp(-k_0 t_j)}$$

$$\therefore \ln \left[\frac{\exp(\delta_0 - y_j) - 1}{\exp(\delta_0 - y_i) - 1} \right] = k_0 t_j - k_0 t_i$$

$$\therefore k_0 = \frac{1}{t_j - t_i} \ln \left[\frac{\exp(\delta_0 - y_j) - 1}{\exp(\delta_0 - y_i) - 1} \right] \quad (1.5.11)$$

which gives an initial estimate for k . From the i th equation

$$\beta_0 = \exp(k_0 t_i) [\exp(\delta_0 - y_i) - 1]$$

giving a corresponding initial estimate of β .

The i th and j th observation should be widely distanced from one another in order to lead to more stable estimates.

Confidence Intervals for the Transformed Parameters e^{θ_1} and $1/(1+e^{\theta_1})$

An approximate $100(1-\alpha)\%$ confidence interval can be found for the unknown parameters α , $\beta_{(g)}$ and ρ , due to the fact that the transformations are one to one and that asymptotically $\hat{\theta}_i \sim N(\theta, \phi)$ $i = 1, 2, 3$ (see 1.5.12) so that

$$P(\hat{\theta}_i - z_{\alpha/2} \sqrt{\text{var } \hat{\theta}_i} < \theta_i < \hat{\theta}_i + z_{\alpha/2} \sqrt{\text{var } \hat{\theta}_i}) \approx 1 - \alpha.$$

where $z_{\alpha/2}$ is the critical value of the z score for probability level α .

It follows that

$$P[\eta_L \leq e^{\theta_1} \leq \eta_U] \approx 1 - \alpha$$

$$P[(1+\eta_U)^{-1} \leq \frac{1}{1+e^{\theta_1}} \leq (1+\eta_L)^{-1}] \approx 1 - \alpha \quad (1.5.13)$$

where

$$\eta_L = \exp \left\{ \hat{\theta}_1 - z_{\alpha/2} \sqrt{\text{var } \hat{\theta}_1} \right\}$$

$$\eta_U = \exp \left\{ \hat{\theta}_1 + z_{\alpha/2} \sqrt{\text{var } \hat{\theta}_1} \right\}$$

1.6 OPTIMIZATION

In the case where least squares estimation or maximum likelihood estimation do not have explicit solutions for the parameters being estimated, it is necessary to make use of an iterative process to estimate these parameters.

Two such processes to be considered in the context of growth curve analysis are the Newton Raphson method and the Gauss-Newton method.

Newton Raphson Procedure

Let the statistical model for subject i at time t_j be

$$y_{ij} = f(\underline{\theta}_i, t_j) + \varepsilon_{ij}, j = 1, 2, \dots, n \quad (1.6.1)$$

where f represents a twice-differentiable function of time and the individual growth parameters $\underline{\theta}_i = (\theta_{i1}, \theta_{i2}, \dots, \theta_{im})$, and ε_{ij} is an independent residual.

Consider the case where not all of the observations y_{ij} have the same variance. To obtain the weighted least squares estimator of $\underline{\theta}_i$, we minimize the weighted sum of squared residuals.

$$q_i = \sum_j^n w_j [y_{ij} - f(\underline{\theta}_i, t_j)]^2 \quad (1.6.2)$$

where the w_j are positive-valued weights.

A necessary condition on the minimum may be obtained by differentiating (1.6.2) and equating to zero to obtain the so-called "extremal" equations,

$$\begin{aligned} \frac{\partial q_i}{\partial \theta_{ig}} &= -2 \sum_{j=1}^n w_j [y_{ij} - f(\underline{\theta}_i, t_j)] \left[\frac{\partial f(\underline{\theta}_i, t_j)}{\partial \theta_{ig}} \right]_{\underline{\theta} = \hat{\theta}_i} \\ &= 0, g = 1, 2, \dots, m \end{aligned}$$

This system of implicit nonlinear equations in the m unknowns, $\underline{\hat{\theta}} = [\hat{\theta}_{i1}, \hat{\theta}_{i2}, \dots, \hat{\theta}_{im}]$ usually can be solved by a conditioned Newton-Raphson process based on the iterative corrections:

$$\hat{\theta}_{ig}^{(\ell+1)} = \hat{\theta}_{ig}^{(\ell)} + \gamma_{ig}^{(\ell)} \quad (1.6.3)$$

where $\gamma_{ig}^{(\ell)}$ is obtained at iteration ℓ by solving the linear system,

$$\sum_{h=1}^m \gamma_{ih}^{(\ell)} \left[\frac{\partial^2 q_i}{\partial \theta_{ig} \partial \theta_{ih}} + c_\ell \delta_{gh} \right] \theta_i = \hat{\theta}_i^{(\ell)} = \left[\frac{\partial q_i}{\partial \theta_{ig}} \right] \theta_i = \hat{\theta}_i^{(\ell)} \quad (1.6.4)$$

In (1.6.4), $\delta_{gh} = 1$ for $g = h$ and 0 otherwise (Kronecker's delta), and c_ℓ is a constant chosen to accelerate convergence of the iterations (see Chambers 1977).

To start the process, the elements of $\underline{\hat{\theta}}^{(\ell)}$ are assigned initial estimates, $\underline{\hat{\theta}}^{(0)}$. A good choice for $\underline{\hat{\theta}}^{(0)}$ in the present instance is the set of mean values of the growth parameter in the population from which the subject was drawn.

Provided the initial values are not too far removed from the solution point, the corrections will quickly become small, and the process will converge to the estimated parameters. If $\varepsilon_{ij} = y_{ij} - f(\underline{\hat{\theta}}_i, t_j)$, $j = 1, 2, \dots, n$ are independent normal errors, then the method of ordinary least squares or the method of maximum likelihood estimation can be used in the above process.

Gauss-Newton Method

The Gauss-Newton method consists in linearly approximating the response function $f(\underline{\hat{\theta}}_i, t_j)$ and then solving the resulting linear least squares problem by means of linear regression.

Due to the fact that the set of least squares equations $\frac{\partial F(\underline{\theta})}{\partial \theta_j} = 0$, $j = 1, 2, 3$ (see (1.4.6)) cannot be solved explicitly for growth curves, a minimizing program must be used to calculate $\hat{\underline{\theta}}$. Applying the Gauss-Newton method to equation (1.4.3) with response function (1.4.4) it is simpler to make use of vector notation.

Define $\underline{f}(\underline{\theta})$ as an $(n \times 1)$ vector with typical element

$$\underline{f}(\underline{\theta}) = f(\underline{\theta}, t_i), \quad i = 1, \dots, n$$

Using function (1.4.4) we can now further define an $(n \times 3)$ matrix X with typical element

$$[X]_{ij} = \frac{\partial f(\underline{\theta}, t_i)}{\partial \theta_j} \quad i = 1, \dots, n; \quad j = 1, 2, 3 \quad (1.6.5)$$

The set of regression equations can then be written as

$$\underline{y} = \underline{f}(\underline{\theta}) + \underline{e} \quad (1.6.6)$$

The sum of squares of residuals become

$$F(\underline{\theta}) = (\underline{y} - \underline{f}(\underline{\theta}))'(\underline{y} - \underline{f}(\underline{\theta})) \quad (1.6.7)$$

and the first order partial derivatives of $F(\underline{\theta})$ in relation to the elements of $\underline{\theta}$ is the (3×1) vector

$$\frac{\partial F(\underline{\theta})}{\partial \underline{\theta}} = -2X'(\underline{y} - \underline{f}(\underline{\theta})) \quad (1.6.8)$$

Starting at an initial value $\underline{\theta}^0$ of $\underline{\theta}$, the Gauss-Newton algorithm can be used in order to minimize $F(\underline{\theta})$ with respect to $\underline{\theta}$.

The Gauss-Newton method can be described as follows.

Replace $\underline{f}(\underline{\theta})$ in (1.6.6) by a first order linear approximation

$$\underline{f}(\underline{\theta}^0 + \Delta \underline{\theta}) \approx \underline{f}(\underline{\theta}^0) + X' \Delta \underline{\theta} \quad (1.6.9)$$

$$\text{where } \underline{\theta} = \underline{\theta}^0 + \Delta \underline{\theta}$$

This can also be written as

$$f(\theta^0 + \Delta\theta) \approx f(\theta)^0 + \Delta\theta_1 f_1'(\theta) + \Delta\theta_2 f_2'(\theta) + \Delta\theta_3 f_3'(\theta)$$

It follows that

$$y - f(\theta^0) = X' \Delta\theta + e \quad (1.6.10)$$

which may be viewed as a standard linear regression equation with X playing the role of the vector of independent variable, the $\Delta\theta$ the role of regression coefficients and $y - f(\theta)$ the role of the dependent variable. $\Delta\theta$ is thus obtained by standard regression techniques.

$$\text{i.e. } \Delta\theta = (X'X)^{-1} X' (y - f(\theta^0))$$

$$= -\frac{1}{2} (X'X)^{-1} \frac{\partial F(\theta)}{\partial \theta} \quad \text{by (1.6.8)} \quad (1.6.11)$$

The initial value θ^0 is replaced by $\theta^0 + \Delta\theta$ and then the whole process is repeated until it converges.

The Gauss-Newton method is useful for the following three reasons.

1. The step $\Delta\theta$ will be distinct in both length and direction.
2. It does not require the computation of the $p(p+1)/2$ second partial derivatives of each of the functions $f(\theta, t_i)$ which is required by the Newton Raphson procedure.
3. When the response function happens to be linear in the parameters it converges in a single step.

The above minimizing process was considered under the assumption that the error terms e_i of equation (1.4.3) were uncorrelated with expected value of zero and variance σ^2 . However consideration must be made when this assumption is not true i.e. the residuals are correlated. The maximum likelihood method can be used as an estimation process in the presence of correlated residuals or generalized least squares.

1.7 NORMAL THEORY MAXIMUM LIKELIHOOD ESTIMATION IN THE PRESENCE OF CORRELATED RESIDUALS

In recent times dependent residual about trend lines has been universally encountered, (Box and Jenkins, 1975; Brillinger, 1975; Jenkins and Watts 1968) so that conventional methods such as least-squares for minimizing independent errors cannot be relied on for analysis of growth data, but must be replaced by the method of maximum likelihood, or the computationally equivalent method of generalized least squares, on the assumption of correlated normal errors.

Consider the following statistical model of subject i at time j

$$y_{ij} = f(t_j, \theta_i) + \epsilon_{ij}, j = 1, \dots, n \quad (1.7.1)$$

Assuming that the above n residuals have the multivariate normal distribution $N(0, \Sigma_\epsilon)$ where the $n(n+1)/2$ distinct elements in the $n \times n$ covariance matrix Σ_ϵ , are arbitrary (Subject to Σ_ϵ being positive definite). The j -th diagonal element of Σ_ϵ is σ_j^2 , the variance of the residual at time t_j ; the j,k off-diagonal element, $j \neq k$, is the covariance $\sigma_{jk} = \sigma_{kj} = \rho_{jk}\sigma_j\sigma_k$, where ρ_{jk} is the correlation between residuals at time t_j and t_k .

To obtain the maximum likelihood estimates of the parameters

Let

$$\lambda_i = \frac{1}{(2\pi)^{n/2} \sqrt{|\Sigma_{\epsilon i}|}} e^{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_i)' \Sigma_{\epsilon i}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)}$$

be the likelihood function and

$$L = \log \lambda_i = c - \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_i)' \Sigma_{\epsilon i}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \quad (1.7.2)$$

be the log likelihood function

$$\text{where } \mu_i = \begin{bmatrix} f(t_1, \theta_i) \\ f(t_2, \theta_i) \\ \vdots \\ f(t_n, \theta_i) \end{bmatrix} = f(\theta_i)$$

$$\Sigma_{\epsilon_i}^{-1} = \begin{bmatrix} \sigma_i^{11} & \sigma_i^{12} & \dots & \sigma_i^{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_i^{nn} \end{bmatrix}$$

and where $c = -\frac{n}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_{\epsilon_i}|$.

Given values for the σ_i^{jk} , the first derivatives of the likelihood function are set equal to zero to obtain the equations

$$\frac{\partial L}{\partial \theta_i} = \Sigma_{\epsilon_i}^{-1} (y_i - f(\theta_i)) \left[\frac{\partial f(\theta_i)}{\partial \theta_i} \right]_{\theta_i = \hat{\theta}_i} = 0 \quad (1.7.3)$$

The specification of the value of Σ_{ϵ_i} may present problems using the method of maximum likelihood, as it is assumed known but in fact is unknown.

Fortunately Σ_{ϵ_i} plays the role of a weight in (1.7.2) and thus does not need to be known with great accuracy.

Estimating Σ_{ϵ}

The following strategy for estimating an expected Σ_{ϵ} has been proposed by Fearn (1975).

First calculate ordinary least-squares estimates of the growth parameters for a number of subjects, calling these values $\hat{\theta}_i$, $i = 1, 2, \dots, N$ and calculate the provisional residuals,

$$\hat{e}_{ij}^* = y_{ij} - f(\hat{\theta}_i, t_j), \quad i = 1, 2, \dots, N; \quad j = 1, 2, \dots, n \quad (1.7.4)$$

From (1.7.4) calculate provisional covariance estimates

$$\Sigma_e^* = \frac{1}{N-1} \sum_{i=1}^N [y_{ij} - f(\hat{\theta}_i^*, t_j)][y_{ik} - f(\hat{\theta}_i^*, t_k)]' \quad (1.7.5)$$

thereafter go back to the data and obtain provisional maximum likelihood estimates using (1.7.2) and substituting Σ_e^* for Σ_e .

Once again repeat the process above by computing the residuals and then the corresponding covariance matrix. Due to the fact that only a rough approximation for Σ_e is required, the last estimates of the residuals and covariance estimates, should be sufficiently accurate to be used in calculating the final maximum likelihood estimates of the growth parameters.

1.8 BAYES ESTIMATION WHEN DATA ARE INCOMPLETE

Although least squares or maximum likelihood estimation may adequately fit a nonlinear growth model when complete measurements of the subject are available, these methods may fail or may produce implausible results when the record is truncated at either end or contains gaps. They have the weakness of noticing only the incomplete information in the observations and ignoring the subject's origin in a population about which, in favourable instances, we have some knowledge.

In particular, if the population has been the object of a longitudinal growth study, we may have information about the distribution of individual growth parameters in it.

For a specified growth model, this empirically derived information takes the form of a multivariate distribution of the model parameters in that population.

Given knowledge of this prior distribution, we are in a position to employ Bayes estimation to obtain, in combination with the available data, the posterior distribution of the particular subject's growth parameters.

The mean of the posterior distribution is called the Bayes estimate of the growth parameter. It optimally combines the information that we have observed in this subject with what we expect from our knowledge that the subject is drawn from a population specified by sex, ethnic background, age, etc. assuming the relevant normative data are available.

Fearn (1975) has discussed results for the Bayes estimate that are already known for the case where the distribution of the residuals for the given subject, as well as the distribution of growth parameters in the population, are multivariate normal and the statistical model is linear in its parameters and in the errors (see Smith 1973). In the present application, the normal distribution assumptions seem reasonable, but as we have seen the logistic, Gompertz and monomolecular models are nonlinear in their parameters and so the known results do not apply.

For nonlinear models, the simplest approach appears to be to follow the example of Lindley and Smith (1972) and estimate the mode of the posterior distribution rather than the mean. The use of modes in Bayes' formula instead of means, overcomes the problem of integrals which cannot all be expressed in closed form, and which are most complex to evaluate. As the number of data points increase, the posterior distribution can be expected to approach a multivariate normal distribution and the mode will approach the mean.

Let the distribution of the residuals be $N(0, \Sigma_\epsilon)$ and the prior distribution of growth parameters be $N(\mu_\theta, \Sigma_\theta)$. Then the logarithm of the posterior density of the parameters of individual i is

$$\phi_i = D - \frac{1}{2}(\mathbf{x}_i - \mathbf{f}(\boldsymbol{\theta}_i))\Sigma_\epsilon^{-1}(\mathbf{x}_i - \mathbf{f}(\boldsymbol{\theta}_i)) - \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_\theta)\Sigma_\theta^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_\theta) \quad (1.8.1)$$

$$\text{where } \mathbf{f}(\boldsymbol{\theta}_i) = \begin{bmatrix} f(t_1, \boldsymbol{\theta}_i) \\ f(t_2, \boldsymbol{\theta}_i) \\ \vdots \\ f(t_n, \boldsymbol{\theta}_i) \end{bmatrix}$$

$$\boldsymbol{\theta}_i = \begin{bmatrix} \theta_{i1} \\ \theta_{i2} \\ \vdots \\ \theta_{im} \end{bmatrix}$$

$$\boldsymbol{\mu}_\theta = \begin{bmatrix} \mu_{\theta_1} \\ \mu_{\theta_2} \\ \vdots \\ \mu_{\theta_m} \end{bmatrix}$$

and where D is a constant not involving $\boldsymbol{\theta}_i$.

The maximum of the posterior density with respect to variation in $\boldsymbol{\theta}_i$ is, therefore

$$\frac{\partial \phi_i}{\partial \boldsymbol{\theta}_i} = \Sigma_\epsilon^{-1}[\mathbf{x}_i - \mathbf{f}(\boldsymbol{\theta}_i)] \left[\frac{\partial \mathbf{f}(\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i} \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \Sigma_\theta^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_\theta) = 0 \quad (1.8.2)$$

Note that (1.8.2) differs from the likelihood equation (1.7.3) only in the assumption that Σ_ϵ is common for all subjects in (1.8.2) and the appearance of an additional term on the right derived from the prior density of $\boldsymbol{\theta}_i$. This term acts as a "penalty" function to prevent the estimated $\hat{\boldsymbol{\theta}}$ from straying too far from its population mean $\boldsymbol{\mu}_\theta$.

CHAPTER 2

INTRODUCTION

This chapter considers the use of polynomials in analyzing growth curves. Two approaches from the growth curve literature are examined. The first being Potthoff and Roy's model (Potthoff and Roy (1964)). In Section 2.1 a general model is formulated by appending a polynomial matrix to the usual MANOVA model. The various growth curve situations that may arise and their effect on the general model are examined. Section 2.2 deals with the hypothesis testing of such polynomials. The test statistics involved are adapted from a general linear hypothesis, and those similarities in deriving the formulas are made evident. Included in this section is the estimation of the parameters of the polynomial function. Potthoff and Roy model makes use of a weight matrix G . Section 2.3 examines the various options for choosing this G matrix.

The second approach considered is the Rao-Khatri model in Section 2.4 (see Rao (1965,1966,1967) and Khatri (1966)). A reduction of the growth curve model using an analysis of covariance model as well as the partitioning of the data matrix was proposed. Test statistics based on the general theory for analysis of covariance are derived, as well as the parameter estimates. In Section 2.5 comparisons of the two approaches are considered.

2.1 THE POTHOFF AND ROY MODEL

Pothoff and Roy (1964) developed an alternative method to handle growth curve analysis. The usual MANOVA (multivariate analysis of variance) model was generalized by appending a post-matrix to the expectation equation.

This generalized MANOVA model was applied to polynomials as a representation of the growth curve.

The following reasons were given for developing this model (Potthoff and Roy (1964)):

"The different previous approaches to the growth curve problem deal mainly with rather specialized aspects of it, and so a major purpose of the present paper is to furnish new tools for growth curve analysis which are of sufficient general applicability. These tools which are derived from the theory of multivariate normal analysis of variance and which are based on polynomial models for the growth curves, provide both confidence bands for the growth curves and over-all tests of significance for various kinds of compound hypotheses, even under the more complex types of experimental designs".

A hypothesis of the form

$$M_{(s \times m)} \xi_{(m \times p)} V_{(p \times u)} = 0_{(s \times u)} \quad (2.1.1)$$

where M and V are matrices of known constants, 0 is the null matrix and ξ is a matrix of unknown parameters, can be used to test the usual MANOVA model

$$E[Y_{(n \times p)}] = A_{(n \times m)} \xi_{(m \times p)} \quad (2.1.2)$$

where A is a matrix of known constants and where the different rows of Y are distributed mutually independently and the p elements in any row follow a multivariate normal distribution with unknown variance matrix $\Sigma_{(p \times p)}$.

The model (2.1.2) can be generalized by appending to it a post-matrix P .

The general model to be considered now is

$$E(Y_{0(n \times q)}) = A_{(n \times m)} \xi_{(m \times p)} P_{(p \times q)} \quad (2.1.3)$$

in which the different rows of $Y_{0(n \times q)}$ are distributed mutually independently and the q

elements in any row follow a multivariate normal distribution with unknown variance matrix $\Sigma_0(q \times q)$ and where A and P are matrices of known constants and ξ is a matrix of unknown parameters. Comparing the growth curve model (2.1.3) with the standard MANOVA model (2.1.2), we see that only the post matrix P has been added to the model (2.1.2). This implies that each response variate can be expressed as a linear regression model of the form $E(y) = P'\beta_i$ where $y_i(q \times 1)$ is the observation vector for the i th subject, and β_i is a vector of unknown parameters. It is therefore necessary to find a technique to test the hypothesis of the form (2.1.1) under the general model. (2.1.3)

Let us now consider the various types of growth curve situations that may arise, and apply them to the general model (2.1.3).

- (a) Consider the simplest situation where n subjects' responses, all subject to the same conditions, are observed at q time points, $t_1, t_2 \dots t_q$. The q observations on a given subject are not independent, but rather assumed to be multivariate normal with unknown variance matrix Σ_0 .

The growth curve being considered is assumed to be a polynomial in time of degree $p-1$. The expected value of any subject at time t_i will be $\xi_0 + \xi_1 t_i + \xi_2 t_i^2 + \dots + \xi_{p-1} t_i^{p-1}$.

The vector A is $n \times 1$ containing all 1's.

The vector ξ is $1 \times p$ consisting of $\xi_0, \xi_1, \xi_2, \dots, \xi_{p-1}$ and the matrix P consists of time points e.g. the s th row and i th column is t_i^{s-1} .

In matrix notation we can write (2.1.3) as

$$\begin{matrix} A \\ \left[\begin{array}{c} 1 \\ 1 \\ \vdots \\ 1 \end{array} \right] \end{matrix} (\xi_0 \ \xi_1 \dots \xi_{p-1}) \begin{matrix} P \\ \left[\begin{array}{cccc} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_q \\ \vdots & \vdots & & \vdots \\ t_1^{p-1} & t_2^{p-1} & \dots & t_q^{p-1} \end{array} \right] \end{matrix} \quad (2.1.4)$$

- (b) Consider now the situation where there is no longer only one group of subjects but rather m groups. Under the general model the following changes take place. The matrix A will consist of $n = n_1 + n_2 + \dots + n_m$ rows and m columns where there will be n_1 rows $(1, 0, \dots, 0)$, n_2 rows $(0, 1, \dots, 0)$, and n_m rows $(0, 0, \dots, 1)$. The (j, k) th element of ξ will be $\xi_{j, k-1}$, and P will be the same matrix as in (a) above.

In matrix notation the general model will take the following form

$$\begin{matrix} A \\ \left[\begin{array}{cccc} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{array} \right] \end{matrix} \begin{matrix} \xi \\ \left[\begin{array}{ccc} \xi_{10} & \xi_{11} & \dots & \xi_{1p-1} \\ \xi_{20} & \xi_{21} & \dots & \xi_{2p-1} \\ \vdots & \vdots & & \vdots \\ \xi_{m0} & \xi_{m1} & \dots & \xi_{mp-1} \end{array} \right] \end{matrix} \begin{matrix} P \\ \left[\begin{array}{cccc} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_q \\ \vdots & \vdots & & \vdots \\ t_1^{p-1} & t_2^{p-1} & \dots & t_q^{p-1} \end{array} \right] \end{matrix} \quad (2.1.5)$$

where the growth curve associated with the j th group at the i th time point is

$$\xi_{j0} + \xi_{j1}t_i + \xi_{j2}t_i^2 + \dots + \xi_{j,p-1}t_i^{p-1} \quad i = 1, \dots, q.$$

- (c) Consider the example of n individuals being tested on a treadmill at various time points. If the two effects, namely treadmill speed and treatment dosage were being

simultaneously tested on each subject and an additive model with no interaction between treadmill speed and treatment dosage is assumed, then a growth curve for an individual subjected to the m th treadmill speed and the s th treatment dosage can be represented by an equation of the form

$$(\xi_{m0} + \xi_{m1}t + \xi_{m2}t^2 + \dots + \xi_{m,p-1}t^{p-1}) + (\mu_{s0} + \mu_{s1}t + \mu_{s2}t^2 + \dots + \mu_{s,p-1}t^{p-1}) \quad (2.1.6)$$

where $\xi_{m0}, \xi_{m1} \dots \xi_{m,p-1}$ are the parameters for treadmill speed and $\mu_{s0}, \mu_{s1}, \dots, \mu_{s,p-1}$ are the parameters for treatment dosage.

Suppose that there were four treadmill speeds and three treatment dosages e.g. an individual was subjected to a treadmill speed of 2.7, 3.3, 4.4 or 6.5 km/h and a treatment dosage of 100, 200, or 300 mg.

The assumption is made that each individual is measured at the same q points, is subjected to the combination of the m th treadmill speed and the s th treatment dosage, $m = 1, 2, 3, 4$ $s = 1, 2, 3$, and has the same unknown variance matrix Σ_0 .

ξ will be a $7 \times p$ matrix whose (j,k) element is $\xi_{j,k-1}$ if $j = 1, 2, 3$ or 4 and is $\mu_{j-4,k-1}$ if $j = 5, 6$, or 7 . A will contain n rows consisting of 1's in the m th and $(4+s)$ th positions and 0's in the other five positions, for $m = 1, 2, \dots, 4$ and $s = 1, 2, 3$. p will be as before.

$$\xi = \begin{bmatrix} \xi_{10} & \xi_{11} & \xi_{12} & \dots & \xi_{1,p-1} \\ \xi_{20} & \xi_{21} & \xi_{22} & \dots & \xi_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ \xi_{40} & \xi_{41} & \xi_{42} & \dots & \xi_{4,p-1} \\ \mu_{10} & \mu_{11} & \mu_{12} & \dots & \mu_{1,p-1} \\ \mu_{20} & \mu_{21} & \mu_{22} & \dots & \mu_{2,p-1} \\ \mu_{30} & \mu_{31} & \mu_{32} & \dots & \mu_{3,p-1} \end{bmatrix} \quad (2.1.7)$$

- (d) Consider now the situation where more than one response is being measured. For example if heartrate (HR) as well as breathing rate (BR) is measured over time. There are m groups of individuals as in case (b).

There will now be $2q$ observations on each subject: q observations on (HR) and q observations on (BR). One can expect the (HR) observations to be correlated among themselves and the (BR) observations to be correlated amongst themselves. One can also expect the (HR) observations to be correlated with the (BR) observations.

Let us assume that the (HR) growth curve for the j th group of individuals is

$$\xi_{j0} + \xi_{j1}t + \xi_{j2}t^2 + \dots + \xi_{j,p_1-1}t^{p_1-1}$$

and that the BR growth curve for the j th group

$$\gamma_{j0} + \gamma_{j1}t + \gamma_{j2}t^2 + \dots + \gamma_{j,p_2-1}t^{p_2-1}$$

Note that the degree of the polynomials need not be the same.

The A matrix will be the same as in (b), ξ will be $m \times p$, where $p = p_1 + p_2$, and the (j,k) element of ξ will be $\xi_{j,k-1}$ if $k \leq p_1$ and $\gamma_{j,k-p_1-1}$ if $k > p_1$.

$$\text{i.e. } \xi = \begin{bmatrix} \xi_{10} & \xi_{11} \dots \xi_{1,p_1-1} & \gamma_{10} & \gamma_{11} & \dots \gamma_{1,p_2-1} \\ \xi_{20} & \xi_{21} \dots \xi_{2,p_1-1} & \gamma_{20} & \gamma_{21} & \dots \gamma_{2,p_2-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \xi_{m0} & \xi_{m1} \dots \xi_{m,p_1-1} & \gamma_{m0} & \gamma_{m1} & \dots \gamma_{m,p_2-1} \end{bmatrix} \quad (2.1.8)$$

Y_0 is assumed to be arranged so that the first q columns contain the HR measurements and the last q columns the BR measurements.

Then $P_{[(p_1+p_2) \times 2q]}$ will have a submatrix $P_1(p_1 \times q)$ in its upper left-hand corner a submatrix $P_2(p_2 \times q)$ in its lower right-hand corner, and 0's elsewhere.

The P matrix will take the following form in matrix notation

$$P = \begin{bmatrix} 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \\ t_1 & t_2 & \dots & t_q & 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ t_{p_1-1} & t_{p_1-1} & \dots & t_{p_1-1} & 0 & 0 & \dots & 0 \\ t_1 & t_2 & \dots & t_q & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & t_1 & \dots & t_q \end{bmatrix} \quad (2.1.9)$$

If (2.1.1) is to be the hypothesis that the BR growth curves are the same for all the m groups we set $s = m - 1$ and $u = p_2$, take M to be a matrix whose last column contains all minus ones and whose first $(m-1)$ columns constitute the identity matrix, and take V to be a matrix whose first p_1 rows contain all 0's and whose last p_2 rows constitute the identity matrix, $p = p_1 + p_2$.

$$\text{i.e. } M = \begin{bmatrix} 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & & 0 & -1 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 0 & -1 \end{bmatrix} \quad (2.1.10)$$

and

$$V = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

One can also test whether both the HR growth curves and the BR growth curves are the same for all m groups. In this case M is the same as in (2.1.10) and $V = I$ where $u = p$.

2.2 HYPOTHESIS TESTING

A Generalized Linear Hypothesis.

Consider the model $Y = XB + U$, where Y is $n \times p$, X is $n \times m$ of rank m , and the rows of U are independently identically distributed $N_p(0, \Sigma)$. We wish to test

$$H_0: CBD = 0 \quad (2.2.1)$$

where: C is $s \times m$ of rank s ($s \leq m$), B is $m \times p$ of rank p , and D is $p \times u$ of rank u ($u \leq p$). Since H_0 reduces to the hypothesis $CB = 0$ when $D = I_p$, a natural procedure for handling H_0 is to try and reduce the hypothesis $CBD = 0$ to $CB = 0$ by using a suitable transformation. This can be achieved by setting $Y_D = YD$

so that $Y_D = XBD + UD$

$$= X\tilde{B} + U_0 \quad (2.2.2)$$

say, where the rows of

$$U_0 = \begin{bmatrix} U_1' \\ U_2' \\ \vdots \\ U_n' \end{bmatrix} D = \begin{bmatrix} (D'U_1')' \\ (D'U_2')' \\ \vdots \\ (D'U_n')' \end{bmatrix} \quad (2.2.3)$$

are i.i.d $N_u(0, D'\Sigma D)$ by Theorem 2.1(i) G.A.F. Seber (1984) p.13.

Since H_0 is now $C\beta = 0$ and using the general theory for multivariate linear models (see Theorem 8.5 G.A.F. Seber (1984) p. 409.)

The Hypothesis sum of squares deviation matrix is

$$\begin{aligned} H &= (C\hat{\beta})'[C(X'X)^{-1}C']^{-1}C\hat{\beta} \\ &= Y_D'X(X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}C(X'X)^{-1}X'Y_D \\ &= (C\hat{B}D)'[C(X'X)^{-1}C']C\hat{B}D \text{ where } \hat{B} = (X'X)^{-1}X'Y \end{aligned} \quad (2.2.4)$$

The error sum of squares deviation matrix is

$$E = Y_D'(I_n - P_\beta)Y_D = D'Y'(I_n - P_\beta)YD \quad (2.2.5)$$

where $P_\beta = X(X'X)^{-1}X'$

so that $E \sim W_n(n-m, D'ED)$ and when H_0 is true $H \sim W_n(s, D'ED)$ (Seber (1984) p. 32-33 see (2.2.6)).

Definition of a Wishart Distribution

(Seber (1984) p. 21) defined a Wishart distribution as follows:

Suppose that $X_1, X_2, \dots, X_m$ are independently and identically distributed as $N_d(0, \Sigma)$; then

$$W = \sum_{i=1}^m X_i X_i'$$

is said to have a Wishart distribution with m degrees of freedom, so that

$$W \sim W_d(m, \Sigma) \quad (2.2.6)$$

A Test Statistic for testing $H_0: \text{CBD} = 0$ (see 2.2.1)

There are four test statistics that have been proposed to test the above hypothesis. They are Roy's $\lambda_{\max}$, Wilks' Λ , Hotelling T_g^2 , and Pillai's $V^{(s)}$. Schatzoff (1966), using the concept of expected significance level proposed by Dempster and Schatzoff (1965) compared the powers of the four test statistics and came to the conclusion that not one of the test statistics were uniformly better than any of the others.

For the purposes of this dissertation Wilks' likelihood ratio test has been used.

Wilks likelihood ratio test

Seber (1984) considers this test statistic.

Suppose H and E have independent Wishart distributions $W_u(s, D'ED)$ and $W_u(n-m, D'ED)$ respectively (see (2.2.6)), and $\Lambda = \frac{|E|}{|E+H|}$ then we say that Λ has a U -distribution with degrees of freedom u, s , and $n-m$, and write

$$\Lambda \sim U_{u,s,n-m} \text{ so that asymptotically } -\left(\frac{f}{C_\alpha}\right) \log_e \Lambda \sim \chi_{us}^2 \text{ where } f = n - m - \frac{1}{2}(u-s+1) \quad (2.2.7)$$

Given α and $\chi_{us}^2(\alpha)$ where $P(\chi_{us}^2 > \chi_{us}^2(\alpha)) = \alpha$, the upper α quantile value for $-\log \Lambda$ is $C_\alpha \chi_{us}^2(\alpha)$. Values of C_α and $\chi_{us}^2(\alpha)$ are given in the tables of Seber(1984) for the parameters $u, s (u \leq s, n-m)$ and $M = n - m - u + 1$.

For $u > s$ the u and s are interchanged. C_α is a conversion factor and from the tables one can observe that $C_\alpha \geq 1$ and $C_\alpha \rightarrow 1$ as $n-m$, and therefore M , tends to infinity.

If $s = \min(u, s)$ is 1 or 2 the upper tail of an appropriate F -distribution is used. If $s = 1$

$$\frac{1-\Lambda}{\Lambda} \cdot \frac{v_2+1}{v_1+1} \sim F_{2v_1+2, 2v_2+2}, \quad (2.2.8)$$

while if $s = 2$

$$\frac{1-\Lambda^{\frac{1}{2}}}{\Lambda^{\frac{1}{2}}} \cdot \frac{2v_2+2}{2v_1+3} \sim F_{4v_1+6, 4v_2+4}, \quad (2.2.9)$$

where $v_1 = \frac{1}{2}(|s-u|-1)$ and $v_2 = \frac{1}{2}(n-m-u-1)$

Since Σ_1 is not known and $\Sigma \neq \sigma^2 I$, Potthoff and Roy suggested the choice of a nonsingular $q \times q$ weight matrix G as an estimate of Σ_0 . An iterative procedure using a generalized least squares estimation process was suggested for estimating Σ_0 (see Section 2.3). Potthoff and Roy based this argument on Σ_0 representing the optimal choice of G , based on the criterion of minimum variance unbiased estimation.

Therefore to analyze (2.1.1) the $p \times p$ matrix $PG^{-1}P'$ is nonsingular and we transform Y_0 to $Y_0 C_1$, where

$$C_1 = G^{-1}P'(PG^{-1}P')^{-1} \quad (2.2.10)$$

is $q \times p$ of rank p ,

$$\text{Then } PC_1 = I_p \text{ and } Y_1 = Y_0 C_1 \quad (2.2.11)$$

and the $p \times 1$ transformed observation vectors are independently normally distributed with mean

$$\begin{aligned} E(Y_1) &= E(Y_0 C_1) \\ &= A\xi PC_1 \\ &= A\xi \quad \text{from (2.2.11)} \end{aligned}$$

with dispersion matrix $\Sigma_1 = C_1' \Sigma_0 C_1$ by theorem 2.1(i) Seber(1984) p.18

The original model (2.1.3) has thus been reduced to $Y_1 = A\xi + u_1$, where the rows of u_1 are i.i.d. $N_p(0, \Sigma_1)$ and (2.1.1) can be tested using the theory for a generalized linear hypothesis section 2.2 with $Y_0 \rightarrow Y_1$, $B \rightarrow \xi$, $Y_D \rightarrow VY_1$ where

$$H_1 = (M\xi'V)'[M(A'A)^{-1}M']^{-1}M\xi'V \quad (2.2.12)$$

$$\text{and} \quad E_1 = V'Y_1'(I_n - A(A'A)^{-1}A')Y_1V \quad (2.2.13)$$

$$\text{and where} \quad \xi = (A'A)^{-1}A'Y_1$$

$$\hat{\xi} = (A'A)^{-1}A'YG^{-1}P'(PG^{-1}P')^{-1} \text{ from (2.2.10)} \quad (2.2.14)$$

Therefore H_1 and E_1 are independently distributed as $W_u(s, V'\Sigma_1 V)$ and $W_u(n-m, V'\Sigma_1 V)$ respectively when (2.1.1) is true.

2.3 CHOICE FOR THE MATRIX G

(i) $G = I$

A simple choice of G is to set $G = I$.

Such a choice of G will certainly simplify one's calculations and make computing that much easier. Another advantage of using $G = I$ is for those experimenters who regard arbitrariness as something to be seriously avoided and are as a result prepared to pay a high price for avoiding it.

(ii) $G = \Sigma_0$

Potthoff and Roy (1964) considered the choice of $G = \Sigma_0$. This choice cannot be made since Σ_0 is unknown. However if before the experiment a fairly good guess of Σ_0 is available, then it might be preferable to set G equal to this guess of Σ_0 rather than to set $G = I$.

Even though G might differ radically from Σ_0 the analysis being made will still be valid so long as G is a symmetric positive definite matrix such that $PG^{-1}P'$ is of full rank. However the more G differs from Σ_0 , the worse is the power of the test, the wider the confidence bands, and the larger the variance of the unbiased estimators become.

Since the analysis of (2.1.3) is valid for any choice of the $(q \times q)$ matrix G , so long as G and $PG^{-1}P'$ are non-singular, an iterative procedure could be considered where one can start with $G = I$. Thereafter the model will be fitted and hence an estimate

of Σ_0 will be obtained. Now use this estimate as the value for G in the second iteration. This process could be continued until convergence is reached and the optimum choice of G is obtained.

(iii) **Stochastic or non-Stochastic G**

According to Potthoff and Roy the G that is chosen need not necessarily be non-stochastic. A stochastic G may be used so long as it is Stochastically independent of Y_0 , since the distributions relating to the confidence bounds and tests are mathematically independent of G .

(iv) **External information about G**

If an experiment was carried out which produced Y_0 , and experimental information available outside this experiment, then a stochastic G based on this information would be independent of Y_0 and legitimate for use in the experiment.

For example let E be the experiment which yields the observations Y_0 under model (2.1.3) with unknown variance matrix Σ_0 . Suppose another independent experiment E' was carried out, whose results can be utilized to estimate Σ_0 , then this estimate of Σ_0 can be used for G in experiment E .

E' does not have to have the same model as E nor does it have to be based on such a model (2.1.3).

2.4 THE RAO - KHATRI MODEL

Rao (1965, 1966, 1967) and Khatri (1966) independently proposed an alternative method to analyze growth curves. They proposed a reduction of the growth curve model, using an analysis of covariance model, which gets around the use of an arbitrary matrix G .

General Theory for Analysis of Covariance

A summary of the basic results for the analysis of covariance are given below. The general theory is given in detail in Seber. [1977: Sections 3.7, 3.8.3, and Chapter 10]

Consider the following covariance model

$$F: Y = A\xi + Z\Gamma + U \quad (2.4.1)$$

where A is $n \times m$ of rank m , Z is $n \times t$ of rank t consisting of values taken by concomitant variables, the columns of Z are linearly independent of the columns of A , and $U \sim N_p(0, E)$.

To test $H_0: M\xi = 0$

The error SSD matrix is

$$E = Y'RY - \hat{\Gamma}_F'Z'RY \quad (2.4.2)$$

where $R = I_n - A(A'A)^{-1}A'$

and $\hat{\Gamma}_F = (Z'RZ)^{-1}Z'RY \quad (2.4.3)$

The Hypothesis SSD matrix is

$$H = (M\hat{\xi}_F)'[M(A'A)^{-1}M' + MLNL'M']^{-1}M\hat{\xi}_F \quad (2.4.4)$$

where $\hat{\xi}_F = (A'A)^{-1}A'(Y - Z\hat{\Gamma}_F) \quad (2.4.5)$

and $L = (A'A)^{-1}A'Z$ and $N = (Z'RZ)^{-1}$.

When (2.4.1) is true, $H \sim W_p(s, E)$ and $E \sim W_p(n-m-t, E)$.

To test $H_0: M\xi^2V = 0 \quad (2.4.6)$

as in (2.1.1) where V is $p \times u$ of rank u H, E and p are replaced by $V'HV$, $V'EV$ and u respectively

so that

$$H = (M\hat{\xi}_F V)' [M(A'A)^{-1}M' + MLNL'M']^{-1} M\hat{\xi}_F V \quad (2.4.7)$$

$$E = V'Y'RYV - V'\hat{\Gamma}_F Z'RYV \quad (2.4.8)$$

Partitioning of the Data Matrix

Consider the usual linear model $Y = AX + U$ and hypothesis $H_0: M\xi = 0$ where A is $n \times m$ of rank m , M is $s \times m$ of rank s and the rows of U are i.i.d. $N_p(0, \Sigma)$. Suppose we partition the data matrix $Y = (Y_1, Y_2)$ where Y_1 represents n observations on p_1 y -variables and Y_2 represents n observations on a further p_2 ($= p - p_1$) variables.

Partitioning $\xi = (\xi_1, \xi_2)$ in the same way, we consider the model and hypothesis.

$$Y_2 = A\xi_2 + U_2 \quad (2.4.9)$$

with null hypothesis $H_0: M\xi_2 = 0$ where the rows of U_2 are i.i.d. $N_{p_2}(0, \Sigma_{22})$

where

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}$$

From Theorem 2.1 (viii) Seber (1984) it follows that the rows of Y_1 conditional on Y_2 are independently normally distributed with common dispersion matrix

$$\begin{aligned} \Sigma_{11.2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \text{ and that} \\ E(Y_1|Y_2) &= A\xi_1 + (Y_2 - A\xi_2)\Sigma_{12}\Sigma_{22}^{-1} \\ &= A(\xi_1 - \xi_2\Sigma_{12}\Sigma_{22}^{-1}) + Y_2\Sigma_{12}\Sigma_{22}^{-1} \\ &= A\delta + Y_2\Gamma \end{aligned} \quad (2.4.10)$$

where $\Gamma = \Sigma_{12}\Sigma_{22}^{-1}$

which is written in the form of model (2.4.1).

The Model

The Rao-Khatri reduction model is derived from the Pothoff and Roy model as follows:

First choose a $q \times p$ matrix C_1 of rank p such that $PC_1 = I_p$ as in (2.2.10) and a $q \times (q-p)$ matrix C_2 of rank $q-p$ such that $PC_2 = 0$ where $C = (C_1, C_2)$ is a $q \times q$ nonsingular matrix.

A transformation on the observation Y can now take place so that

$$Y_C = YC = (Y_1, Y_2)$$

$$\text{where } Y_i = YC_i \quad (i=1,2) \quad (2.4.11)$$

$$\text{Now } E(Y_1) = A\xi PC_1 \text{ from (2.1.3)}$$

$$= A\xi$$

$$E(Y_2) = A\xi PC_2$$

$$= 0.$$

The analysis of covariance model is as follows (see (2.4.10)).

$$\begin{aligned} E(Y_1 | Y_2) &= A\xi + Y_2 \Gamma \\ &= (A, Y_2) \begin{bmatrix} \xi \\ \Gamma \end{bmatrix} \end{aligned} \quad (2.4.12)$$

The matrix (A, Y_2) will have rank $m+q-p$, as the m columns of A are linearly independent, and Y_2 is an $n \times (q-p)$ unrestricted random matrix with linearly independent columns that are also linearly independent of the columns of A .

Since the rows of Y_2 are conditionally distributed as independent multivariate normals with a common covariance matrix, the analysis of covariance theory of Section 2.4 is applicable. (with $Z = Y_2$).

The least squares estimates of Γ and Λ for the conditional model (2.4.12) are (see (2.4.3) and (2.4.5)).

$$\begin{aligned}\hat{\Gamma} &= (Z' R Z)^{-1} Z' R Y_1 \\ &= (C_2' Y' R Y C_2)^{-1} C_2' Y' R Y C_1 \quad \text{from (2.4.11)} \\ &= (C_2' S C_2)^{-1} C_2' S C_1\end{aligned}\quad (2.4.13)$$

since
$$S = \frac{Y' R Y}{n-m}$$

$$= \frac{Y'(I_n - A(A' A)^{-1} A') Y}{n - m}$$

which is the error covariance matrix

and
$$\begin{aligned}\hat{\xi}_R &= (A' A)^{-1} A' (Y_1 - Z \hat{\Gamma}) \\ &= (A' A)^{-1} A' (Y C_1 - Y C_2 \hat{\Gamma}) \\ &= (A' A)^{-1} A' Y (C_1 - C_2 [C_2' S C_2]^{-1} C_2' S C_1) \quad \text{from (2.4.13)}\end{aligned}$$

From Seber (1984) Appendix B3.5 p. 536.

$$C_2 (C_2' S C_2)^{-1} C_2 = S^{-1} - S^{-1} P (P' S^{-1} P)^{-1} P' S^{-1} \quad (2.4.14)$$

so that $\hat{\xi}_R = (A' A)^{-1} A' Y S^{-1} P (P' S^{-1} P)^{-1} P' C_1$

(since $S S^{-1} = I_m$)

$$\therefore \hat{\xi}_R = (A' A)^{-1} A' Y S^{-1} P (P' S^{-1} P)^{-1} \quad (2.4.15)$$

since $P C_1 = I_p$.

In order to test $H_0: M\xi V = 0$ see (2.1.1) we first of all require the E and H matrices for testing $M\xi = 0$. From (2.4.2)

$$\begin{aligned}
E &= Y_1' R Y_1 - \hat{F}' Y_2' R Y_1 \\
&= C_1' Y' R Y C_1 - \hat{F}' C_2' Y' R Y C_1 \\
&= (n-m) \{ C_1' S C_1 - C_1' S C_2 (C_2' S C_2)^{-1} C_2' S C_1 \} \text{ by (2.4.13)} \\
&= (n-m) \{ C_1' P (P' S^{-1} P)^{-1} P' C_1 \} \text{ by (2.4.14)}
\end{aligned}$$

$$E = (n-m) (P' S^{-1} P)^{-1} \text{ by (2.2.11)} \quad (2.4.16)$$

Using (2.4.4)

$$H = (M \hat{\xi})' [M(A'A)^{-1} M' + M L N L' M']^{-1} M \hat{\xi} \quad (2.4.17)$$

where $L N L' = (A'A)^{-1} A' Y_2 (Y_2' R Y_2)^{-1} Y_2' A (A'A)^{-1}$

$$= \frac{1}{n-m} (A'A)^{-1} A' Y C_2 (C_2' S C_2)^{-1} C_2' Y' A (A'A)^{-1}$$

so that from (2.4.14)

$$L N L' = \frac{1}{n-m} (A'A)^{-1} A' Y [S^{-1} - S^{-1} P (P' S^{-1} P)^{-1} P' S^{-1}] Y' A (A'A)^{-1} \quad (2.4.18)$$

To test $H_0: M \hat{\xi} V = 0$ we merely replace the matrices H and E by

$$H_V = V' H V \text{ and } E_V = V' E V$$

so that

$$H_V = (M \hat{\xi} V)' [M(A'A)^{-1} M' + M L N L' M']^{-1} M \hat{\xi} V \quad (2.4.19)$$

$$\text{and } E_V = (n-m) V' (P' S^{-1} P)^{-1} V \quad (2.4.20)$$

When $H_0: M \hat{\xi} V = 0$ is true then H_V and E_V are independently distributed as u -dimensional Wishart distributions with a common dispersion matrix and degrees of freedom s and $n-(m+q-p)$ respectively.

2.5 TEST FOR INTERNAL ADEQUACY

The linear model which has been considered is $E(Y_0) = AB$, where $B = \xi P$ and since $PC_2 = 0$

$$E(Y_2) = ABC_2 = A\xi PC_2 = 0$$

Since A is $n \times m$ of rank m , $A\xi PC_2 = 0$ if and only if $BC_2 = 0$. The adequacy of the model $B = \xi P$ can be tested by testing $E(Y_2) = 0$, that is by testing $H_0: I_m BC_2 = 0$ for the model $E(Y) = AB$. This test is a special case of the linear hypothesis test described in Section 2.2 with $C = I_m$.

2.6 COMPARISONS OF THE TWO METHODS.

When $G = I_p$

Since (2.4.15), (2.4.19) and (2.4.20) do not depend on G , and if Potthoff and Roy's transformation $C_1 = G^{-1}P(P'G^{-1}P)^{-1}$ is used, then testing $H_0: M\xi'V = 0$ in the Rao-Khatri method is independent of G and so we can set $G = I_p$. If $G = I_p$ then $C_1 = P(P'P)^{-1}$, and if normalized orthogonal polynomials are used, then $P'P = I_q$, $C_1 = P$, and C_2 can be chosen so that $C = (C_1, C_2)$ is orthogonal.

If such a P is used, C_1 can be chosen to be the matrix of normalized orthogonal polynomials of degrees 0 to $p-1$, and C_2 to be a similar matrix for degrees p to $q-1$ (Grizzle and Allen (1969)).

The above method uses all $p-q$ concomitant variables in Y_2 . Grizzle and Allen (1969) discusses the suggestion of Rao(1965) that a better procedure might be to use only a subset of the concomitant variables. The choice of the concomitant variables was evaluated by the correlations between the columns of Y_1 and Y_2 where $Y_1 = YC_1$ and $Y_2 = YC_2$.

If the correlation of each column of Y_2 with each of Y_1 is zero, then Y_2 provides no information about Y_1 , and Y_2 should be discarded. However if there is a high correlation between the columns of Y_1 and Y_2 , those covariates are included in the model. If the correlation between the columns of Y_1 and Y_2 are all zero then the Potthoff and Roy method is equivalent to not using any covariates in the Rao-Khatri method. This approach is used in the analysis of the experimental data in Section 4.

When $G = S$

Potthoff and Roy did not develop the theory for allowing $G = S$, where S is proportional to the estimate of Σ , (the population covariance matrix) calculated from the data used to estimate ξ (2.2.14). Grizzle and Allen (1969) noted that when $G = S$, Potthoff and Roy's estimate $\hat{\xi}$ (2.2.14) is the same as Rao-Khatri's $\hat{\xi}_R$ (2.4.15). Baksalary et al. (1978) extended this observation made by Grizzle and Allen to the case where not all the concomitant variables are used.

For hypothesis testing the transformation $G = S$ is not appropriate in the Potthoff and Roy method, as S is not independent of Y_0 (2.1.3) (see Section 2.3 (iii)). As a result the two methods have different test statistics for testing $H_0: M\xi V = 0$ (2.1.1).

Timm (1975) raised two queries about the Rao-Khatri reduction method. Firstly he commented that different covariates may be appropriate for different groups within a sample. Secondly the covariates that are selected may vary from sample to sample. However he did point out that a major advantage of the Rao-Khatri method is that one can apply standard multivariate analysis of covariance programs to the data.

CHAPTER 3

3.1 THE APPLICATION

Stress testing, the assessment of a physiological function during physical exercise, has become an increasingly accepted concept in cardiac and pulmonary evaluation in the past 15 years.

The stress electrocardiogram, (ECG), is regularly used to assess the risk of cardiovascular disease. The increased metabolic and gas exchange demands of physical exercise, though are cooperatively supported by the cardiovascular and respiratory systems, so that the same conditions that provide important information about cardiovascular function under stress also yield critical information about pulmonary function.

The standard resting pulmonary function tests, (PFT's), as with the resting ECG, gives only a portion of the complete picture. Crucial data concerning performance are missing.

An abnormal resting PFT estimates the degree of pulmonary impairment, but cannot indicate the degree to which this damage translates into a reduced work or functional capacity.

Cardio-pulmonary exercise testing, (CPET), can quantify this deficit, predict performance capacity and identify the components of the cardio-pulmonary system responsible for limiting exercise capacity.

3.2 THE EXERCISE TEST

1000 subjects, male and female underwent the maximal graded exercise test, as part of a multifasic medical checkup at the Mor Institute for Medical Data, Ramat Gan, Israel. Out of these 1000 subjects, only those subjects without any previous serious illnesses or any chronic disorders and with resting systolic and diastolic blood pressure below 150 and 90 mm Hg respectively, served as subjects for this dissertation. Competitive athletes were also excluded from the test.

Subjects came to the laboratory at least two hours after their last meal. They were informed to refrain from drinking coffee 4 hours before they arrived.

Once they arrived, they were given a questionnaire to fill in concerning their occupation, smoking history, previous health, leisure activities and exercise related symptoms.

After this they underwent a physical examination by a cardiologist. Subjects' weight, height and body fat were determined.

Thereafter the exercise test began.

The exercise test comprised of the subjects walking on a treadmill at a speed between 5 – 6.5 km/h (modified Balke Protocol) predetermined by age, habitual physical activity and medical status. The exercise load was increased by elevating the treadmill slope by 2% every minute until one of the following conditions were reached:

- (a) Volitional exhaustion – where the subject feels he can no longer continue (maximal exercise).
- (b) Maximal age predicted heart rate.
- (c) Respiratory quotient, (RQ), of at least 1.3.
- (d) Symptoms such as chest pains, nausea or dizziness.
- (e) Clinical signs such as ECG abnormalities and others that would dictate an end to exercise.

3.3 THE DATA

For every subject in the exercise test, the following variables were measured every five seconds. ECG, Systolic and Diastolic blood pressure, VO_2 (Oxygen Consumption), VE (Ventilation) VE/VO_2 (Ventilatory Equivalent for O_2), VE/VCO_2 (Ventilatory Equivalent for CO_2), R (Respiratory Quotient), HR (Heart Rate), O_2 pulse (Indication of Stroke Volume), ETO_2 and ETCO_2 (O_2 and CO_2 concentration in the Alveoli), VD/VT (Respiratory Dead Space), VT (Volume of air in a single breath), BR (Breathing Rate).

There were five additional variables taken into account in the analysis of the experiment.

- (i) workload — after each minute the gradient of the slope was increased by 2%.
- (ii) age group — each subject was placed in the age group above 35 or below 35.
- (iii) smoke — a subject was either a smoker or a non-smoker where smoker = 1 and non-smoker = 0
- (iv) lifestyle — a subject either led an active lifestyle or a sedentary lifestyle.

Those subjects exercising at least twice a week for more than a half hour were classified as active, otherwise they were classified as leading a sedentary way of life, where active = 1 and sedentary = 0

- (v) weight — indicating the subject's weight in kg.

See Appendix D for a subset of the data taken from one subject whilst undergoing the exercise stress test.

3.4 EQUIPMENT USED

- (a) Treadmill (Quinton Q65, U.S.A.).
- (b) 12 Leads Electrocardiograph (Siemens, West Germany).
- (c) Blood Pressure Mercury Sphignometer.
- (d) Metabolic cart (Sensormedics 4400 U.S.A. — for analysis of expired gases).
- (e) Weight and Height measured on a Seca Scale (West Germany).
- (f) Skinfold Caliper (Skindex, U.S.A.).
- (g) There are two Hewlett-Packard disk drives built into the Sensormedics 4400 where the data was collected and saved on 3.5 inch floppy disks.

The author was involved in processing the raw data. This involved transferring the data from a Hewlett Packard computer (the Sensormedic 4400), to an IBM personal computer. The data was then transferred from floppy disks onto a magnetic tape, and finally from the magnetic tape onto the mainframe IBM computer at the University of the Witwatersrand.

3.5 PHYSIOLOGICAL BACKGROUND

Every cell in the body requires a continuous supply of energy to carry out its normal functions. This energy is obtained by simple combustion of hydrocarbons (oxidation). In principle this is identical to motor car engines which also utilize oxidation of hydrocarbons as a source of energy. In the body, however, the combustion is highly regulated by a series of enzymes and therefore does not take place in an explosive manner. The most important fuel used by the cell is the hydrocarbon glucose.

The rate at which oxidation of glucose occurs is dependent on the metabolic rate or amount of exercise performed. The oxygen is extracted from the atmosphere by the lungs and dissolved in the blood. The blood circulates through the body due to the pump action of the heart and this delivers the oxygen to the cells. The amount of oxygen in the blood is dependent on the ventilation occurring at the lungs i.e. the amount of atmospheric air intake per minute. On the other hand, oxygen delivery to the cells is dependent on both the content of the blood and the cardiac output i.e. the amount of blood pumped by the heart each minute.

The cardiac output is clearly equal to the product of heartrate and the amount of blood pumped by the heart with each beat (stroke volume).

Exercise increases energy demand or metabolic rate and thus the oxygen consumed to burn the glucose. It is thus clear that metabolism is dependent on the lungs and the circulatory system. For this reason one representative variable from each of the above systems was chosen for study viz: Oxygen consumption as the variable of metabolism, heartrate as the variable of circulation and ventilation as the variable of lung function.

3.8 COMPUTER PROGRAMS USED

Three programs were used to analyze the data; a FORTRAN program called SCURVE, and the SAS procedures PROC NLIN, and PROC IML. SCURVE is an interactive computer program that was developed by Prof. S.H.C. Du Toit of the University of Pretoria for the analysis of well known asymptotic growth curves belonging to the Richards family of growth curves (See Chapter 1).

Since this is an established program, only preliminary results were used for the purpose of this dissertation. To illustrate how this program works an example can be found in Appendix A.

The NLIN (Non Linear regression) procedure produces least-squares estimates of the parameters of a nonlinear model.

For a nonlinear model to be analyzed, one must specify:

- the names and starting values of the parameters to be estimated
- the model, using a single dependent variable
- partial derivatives of the model with respect to each parameter.

One can also:

- confine the estimation procedure to a certain range of values of the parameters by imposing bounds on the estimates
- produce new SAS data sets containing predicted values, residuals, estimates of parameters, and the residual sum of squares. (See Appendix B for an example program of how a model for heartrate was fitted using PROC NLIN).

The last procedure used was PROC IML which comes from the SAS/IML software, an Interactive Matrix Language. The IML software is a multilevel, interactive programming language with dynamic capabilities. A copy of the program using IML can be found in Appendix C.

The programs SCURVE and PROC NLIN were used to apply the theory of the Richards family of growth curves to the experimental data (see Section 1.3), whilst PROC IML was used to apply the Rao-Khatri method to the data. (see Section 2.4)

3.7 THE SAMPLE

A sample of 80 was taken from the thousand subjects who underwent the exercise test.

In order to acquire a representative sample of the population ensuring equal coverage of all subgroups, the following eight groups each comprising ten subjects were chosen.

	<u>Smoke/non-smoker</u>	<u>age</u>	<u>Lifestyle</u>
1)	smoker	> 35 years	active
2)	smoker	> 35 years	sedentary
3)	smoker	< 35 years	active
4)	smoker	< 35 years	sedentary
5)	non-smoker	> 35 years	active
6)	non-smoker	> 35 years	sedentary
7)	non-smoker	< 35 years	active
8)	non-smoker	< 35 years	sedentary

Every fifth subject from the population was considered until the sample size of eighty was attained.

Since only fifty females underwent the exercise test during the given timespan (i.e. 1st June 1986 to 1st June 1987) only males were considered for the sample. Table 3.1 provides the mean scores of weight and age for the various subgroups.

Table 3.1 Mean weight and age for subgroups

	<u>weight</u>	<u>age</u>
Active smokers	78.22	39.44
Active non-smokers	76.54	39.45
Sedentary smokers	79.11	39.76
Sedentary non-smokers	79.85	39.18

3.8 RESULTS USING SCURVE

The program SCURVE was used in a preliminary study to decide on the best model to be used from the Richards family of growth curves. This was achieved by fitting the three growth curve functions (i.e. Gompertz, Monomolecular and Logistic, ignoring smoking age and lifestyle) to all 80 subjects over time and comparing their mean square errors (MSE).

$$\text{MSE} = \left\{ \frac{1}{n} \sum_{i=1}^n e_i^2 \right\}^{\frac{1}{2}}$$

$$= \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right\}^{\frac{1}{2}} \text{ where at time point } i, e_i \text{ is the}$$

residual, y_i the observed value and $\hat{y}_i$ the predicted value.

From Table 3.2 below it is clear that the logistic function is the best fitting function for all three variables (i.e. heartrate, ventilation and oxygen consumption). This is evident since the mean square error for the logistic function is the smallest in each instance.

TABLE 3.2 **MEAN SQUARE ERRORS USING SCURVE**

	<u>Gompertz</u>	<u>Logistic</u>	<u>Monomolecular</u>
heartrate	276.25	124.84	310.5
ventilation	136.25	120.11	191.63
oxygen consumption	13.76	13.03	14.4

Therefore in all subsequent analyses using the Richards functions, the logistic function was used.

CHAPTER 4

4.1 APPLYING THE LOGISTIC FUNCTION FOR HEARTRATE

The following logistic function was fitted using PROC NLIN.

$$HR = A / (1 + \exp(B_0 + B_1 \text{ time} + B_2 \text{ wl} + B_3 \text{ sm} + B_4 \text{ lifestl} + B_5 \text{ wt} + B_6 \text{ time} \times \text{wl} + B_7 \text{ time} \times \text{sm} + B_8 \text{ time} \times \text{lifestl} + B_9 \text{ time} \times \text{wt} + B_{10} \text{ wl} \times \text{sm} + B_{11} \text{ wl} \times \text{lifestl} + B_{12} \text{ wl} \times \text{wt} + B_{13} \text{ sm} \times \text{lifestl} + B_{14} \text{ sm} \times \text{wt} + B_{15} \text{ lifestl} \times \text{wt})). \dots \quad (4.1.1)$$

where A is the asymptotic value reached for heartrate (assuming $B_1 < 0$).

wl = work load, sm = smoke, lifestl = lifestyle, and wt = weight

sm = 1... smoker, sm = 0...non-smoker, lifestl = 1...active, lifestl = 0...sedentary

TABLE 4.1 ESTIMATE AND CONFIDENCE INTERVALS FOR EQUATION (4.1.1)

Parameter	Effect	Estimate	Asymptotic	Asymptotic 95%	
			Standard Deviation	Confidence	Interval
A	asymptote	250	30.4162	190.366	300.634
B0	constant	-0.0412	0.2657	-0.562	-0.073
B1	time	-0.1402	0.1726	-0.4786	-0.103
B2	wl	-0.0076	0.0512	-0.1088	0.0935
B3	sm	0.2875	0.0737	0.1429	0.4319
B4	lifestl	-0.2897	0.0785	-0.4436	-0.1356
B5	wt	0.0057	0.0013	0.0032	0.0083
B6	time*wl	0.0039	0.0005	0.0029	0.0049
B7	time*sm	-0.0122	0.0442	-0.0996	0.0745
B8	time*lifestl	0.0284	0.0448	-0.0595	0.1164
B9	time*wt	-0.0019	0.0021	-0.0061	0.0022
B10	wl*sm	0.0027	0.0133	-0.0233	0.0288
B11	wl*lifestl	0.0010	0.0133	-0.0272	0.0252
B12	wl*wt	0.0001	0.0006	-0.0113	0.0013
B13	sm*lifestl	0.1939	0.0333	-0.1285	0.2594
B14	sm*wt	-0.0041	0.0009	-0.0059	0.0022
B15	lifestl*wt	0.0024	0.0008	-0.0008	0.0040

From the results in Table 4.1 it can be seen that only the main effects excluding workload, are significant. A parameter estimate is considered significant if zero does not fall in the 95% confidence interval.

Since workload was increased every minute, it was highly correlated with time. This high correlation with time could be the reason for workload not being significant.

The variable agegroup was not included in the model because of inconsistencies which may have arisen if it was included. The reason for this was that different agegroups were subjected to different treadmill speeds and thus different conditions. i.e. agegroup and treadmill speed were completely confounded in the experiment. Table 4.2 indicates the treadmill speeds for the different agegroups.

TABLE 4.2 TREADMILL SPEEDS FOR THE VARIOUS AGE GROUPS

Age(years)	Treadmill speed (km/h)
< 30	6.5
30 - 45	5.1
45 - 60	4.4
< 60	3.3

A new logistic function was fitted using only those effects that were significant in (4.1.1)

$$\text{i.e. } HR = A / (1 + \text{EXP}(B_0 + B_1 \text{ time} + B_2 \text{ smoke} + B_3 \text{ lifestyle} + B_4 \text{ weight})) \quad (4.1.2)$$

Table 4.3 gives a summary of results for function (4.1.2). It is evident from Table 4.3 that all the parameter estimates are significant. However the estimate for smoke which is 0.0969653 is not what would be expected, as a positive coefficient in the exponential section of the denominator implies a decreasing effect on heartrate as the exercise test progressed. One would expect a smoker (note, smoker = 1 and non-smoker = 0) to have a higher heartrate than a non-smoker the longer the exercise test progressed. No explanation could be found for this result.

Table 4.3

Non-linear least squares statistics for heart rate

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	6	78613250.676	
RESIDUAL	4005	888454.324	13102208.446
UNCORRECTED TOTAL	4011	79501705.000	221.836
(CORRECTED TOTAL)	4010	2437692.623	

PARAMETER	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95 % CONFIDENCE INTERVAL	
			LOWER	UPPER
A	191.1139815	3.0075452535	185.21721691	197.01074609
B0	-0.4160960	0.0622024330	-0.53804959	-0.29414243
B1	-0.3481678	0.0162314980	-0.37999113	-0.31634445
B2	0.0969653	0.0133047275	0.07088019	0.12305047
B3	0.0801572	0.0130671265	0.05453794	0.10577654
B4	0.0026258	0.0006102507	0.00142930	0.00382221

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

CORR	A	B0	B1	B2	B3	B4
A	1.0000					
B0	0.5986	1.0000				
B1	0.9567	0.5333	1.0000			
B2	-0.3008	-0.2605	-0.2917	1.0000		
B3	-0.2033	-0.3048	-0.2917	0.0465	1.0000	
B4	-0.1869	-0.8659	-0.1810	0.0349	0.1512	1.0000

It was decided for the purposes of this dissertation that the effects of smoking and lifestyle on the function would be examined, and therefore weight, even though it was significant in the overall function was not included in the subgroups (See Table 4.3 & 4.6).

Another worrying factor was the high correlation of 0.9567 between the asymptote A and B_1 , the parameter for time. In order to acquire more stable estimates with smaller correlation between them, the following function was fitted.

$$HR = \frac{1}{a + \exp(b_0 + b_1 \text{ time} + b_2 \text{ smoke} + b_3 \text{ lifestl} + b_4 \text{ weight})} \quad (4.1.3)$$

Here the asymptote is in the denominator. Equation (4.1.3) can also be written as

$$\begin{aligned} HR &= \frac{1}{a(1 + (\frac{e^{b_0}}{a})e^{b_1 \text{ time} + b_2 \text{ smoke} + \dots + b_4 \text{ weight}})} \\ &= \frac{A^*}{1 + e^{\beta_0^* + b_1 \text{ time} + \dots + b_4 \text{ weight}}} \end{aligned}$$

$$\text{where } A^* = \frac{1}{a}$$

$$e^{\beta_0^*} = \frac{1}{a} e^{b_0}$$

$$HR = \frac{A^*}{1 + e^{\beta_0^* + b_1 \text{ time} + \dots + b_4 \text{ weight}}}$$

which is an adjusted form of the logistic function.

Fitting (4.1.3) provided the following results in Table 4.4.

TABLE 4.4 ESTIMATES AND CONFIDENCE INTERVALS FOR EQUATION (4.1.3)

<u>PARAMETER EFFECT</u>	<u>ESTIMATE</u>	<u>ASYMPTOTIC STD. ERROR</u>	<u>ASYMPTOTIC 95% CONFIDENCE INTERVAL</u>	
			<u>LOWER</u>	<u>UPPER</u>
B0 constant	- 5.5170	0.0512	-5.6175	-5.4165
B1 time	-33.3623	1.5282	-36.3585	-30.3661
B2 smoke	0.0933	0.0126	0.0686	0.1181
B3 lifestyle	0.0781	0.0124	0.0538	0.1024
B4 weight	0.0017	0.0006	0.0006	0.0029
a	0.0051	0.0001	0.0050	0.0053

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

<u>CORR</u>	<u>B0</u>	<u>B1</u>	<u>B2</u>	<u>B3</u>	<u>B4</u>	<u>A</u>
B0	1.000	0.2606	-0.1803	-0.2785	-0.9369	-0.3212
B1	0.2606	1.0000	-0.2914	-0.2187	-0.1298	-0.9575
B2	-0.1803	-0.2914	1.0000	0.0477	0.0105	0.3009
B3	-0.2785	-0.2187	0.0477	1.0000	0.1396	0.2116
B4	-0.9369	-0.1298	0.0105	0.1396	1.0000	0.1343
A	-0.3212	-0.9575	0.3009	0.2116	0.1343	1.0000

Workload was again not included in the model because of the high correlation between time and workload. Therefore the interaction of time and workload was not included either. From Table 4.4 it is evident that there is still a very high correlation of -0.9575 between the asymptote A and B1 the parameter for time.

Another form of the logistic function was attempted. The following model was fitted after following the same procedure that was used in acquiring function (4.1.2) (i.e. only those significant effects were taken into consideration).

$$Y_{HR} = \frac{1}{a_0 + a_1 \text{ smoke} + a_2 \text{ lifestyle} + e^{b_0 + b_1 \text{ time}}} \quad (4.1.4)$$

From (4.1.4) it is clear that the effects of smoking and lifestyle are in the linear section of the denominator whilst the effect of time is in the exponential section.

The variable weight was not significant and as a result it was excluded from the function.

Below is a summary of results for model (4.1.4).

TABLE 4.5 ESTIMATES AND CONFIDENCE INTERVALS FOR EQUATION (4.1.4)

<u>PARAMETER</u>	<u>ESTIMATE</u>	<u>ASYMPTOTIC ASYMPTOTIC 95%</u>		
		<u>STANDARD</u>	<u>CONFIDENCE INTERVAL</u>	
		<u>ERROR</u>	<u>LOWER</u>	<u>UPPER</u>
B0	- 5.3743	0.0143	-5.4023	-5.3463
B1	-34.8832	1.6521	-38.0635	-31.7029
A0	0.0052	0.0001	0.0049	0.0054
A1	0.0002	0.00002	0.00013	0.00022
A2	0.00015	0.00002	0.00010	0.00019

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

CORRELATION	B0	B1	A0	A1	A2
B0	1.0000	0.4294	-0.4205	-0.0335	0.1237
B1	0.4294	1.0000	-0.9384	-0.0233	0.1080
A0	-0.4205	-0.9384	1.0000	-0.0539	-0.2446
A1	-0.0335	-0.0233	-0.0539	1.0000	-0.0574
A2	0.1237	0.1080	-0.2446	-0.0574	1.0000

Once again it is evident that there is still a very high correlation of -0.9384 between the asymptote (which in (4.1.4) is the intercept in the linear section of the function) and B1 the parameter for time.

Convergence was not always reached and as a result estimates were not always available for specific functions.

Since (4.1.3) and (4.1.4) did not significantly decrease the high correlation between the parameters, and because convergence posed a problem in both these functions, it was decided to make use of (4.1.2) for further analysis of the experiment. This high correlation between parameters could undermine a parameter's contribution to the function value.

Figure 4.1 is a residual plot against time for function (4.1.2). It is evident from the plot that the residuals form a band around the horizontal zero line, indicating a constant variance. There is however a tendency for underestimation at low time and overestimation at high time (i.e. the longer the experiment carried on).

Figure 4.1

LOGISTIC RESIDUAL PLOT FOR VENTILATION
 $VE = A / (1 + \exp(B0 + B1 \cdot TIME + B4 \cdot SMOKE + B5 \cdot LFESTL + B7 \cdot WEIGHT))$

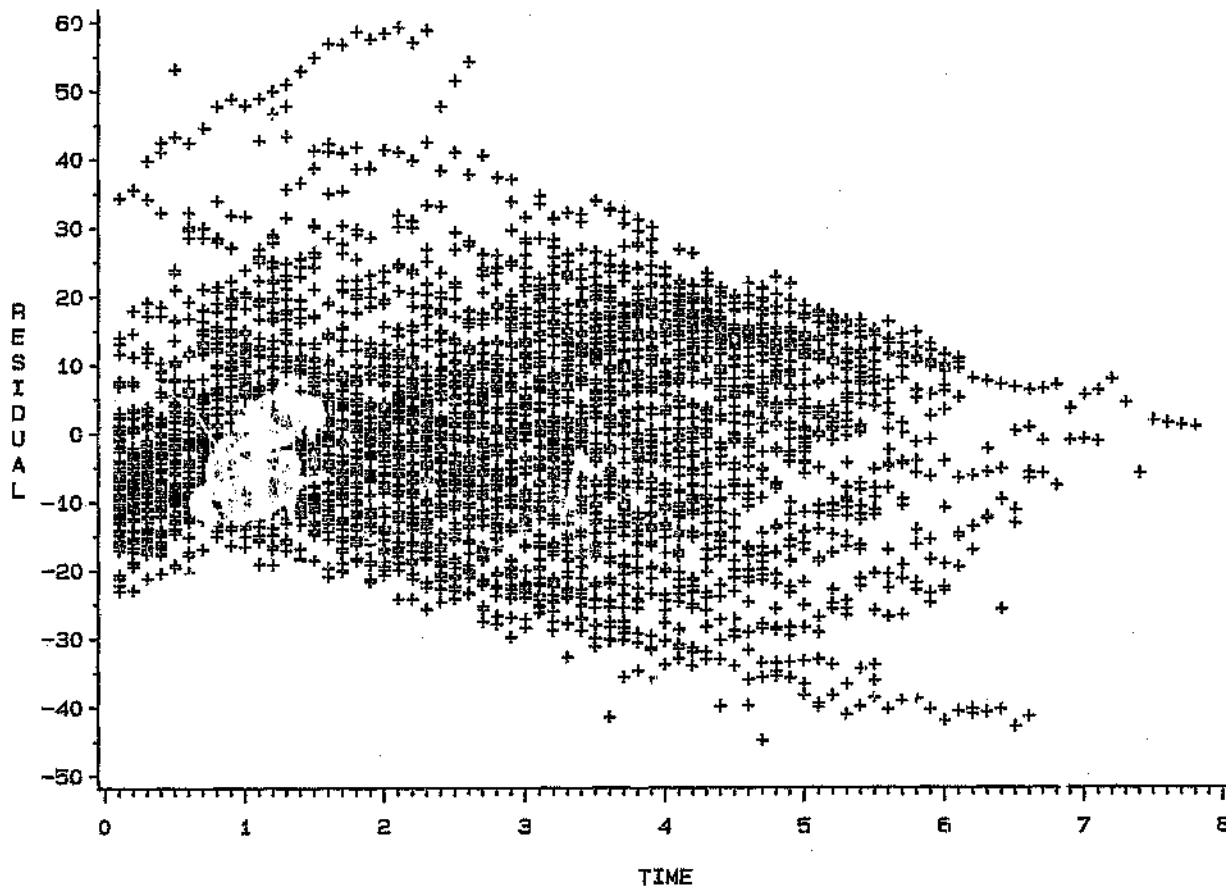
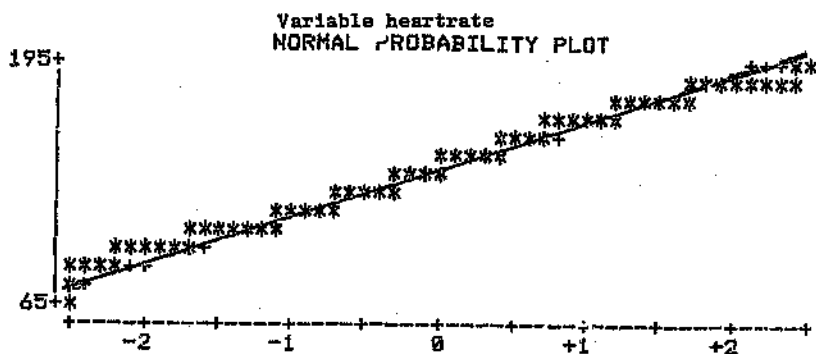


Figure 4.2 represents a normal probability plot for function (4.1.2). A normal probability plot is a quantile-quantile plot of the data. The empirical quantiles are plotted against the quantiles of a standard normal distribution. Asterisks (*) mark the data values. The vertical coordinate is the data value, and the horizontal coordinate is

$$\Phi^{-1}((r_i - 3/8)/(n + 1/4))$$

where r_i is the rank of the data value, Φ^{-1} is the inverse of the standard normal distribution function and n is the number of non-missing data values. The plus signs (+) provide a reference straight line that is drawn using the sample mean and standard deviation. If the data are from a normal distribution, they should tend to fall along the reference line. From Figure 4.2 it is evident that the data values do lie on the reference line and that normality can be assumed.

Figure 4.2



In order to investigate the effect of the various variables on the function, the following four groups were fitted: active smokers, active non-smokers, sedentary smokers, sedentary non-smokers. Of interest to a physiologist is the value of the asymptote 'A' (the final predicted heartrate value, reached by the function). Table 4.6 provides these estimates for the various subgroups. Four separate functions are fitted to the subgroups as a confirmation of the results of the overall function (see Table 4.3).

**TABLE 4.6 ESTIMATES AND CONFIDENCE INTERVALS FOR ASYMPTOTIC
HEARTRATE VALUE.**

	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95% CONFIDENCE INTERVAL	
			LOWER	UPPER
Active non-smokers	192.9217	6.9895	179.2062	206.6372
Active smokers	175.7577	3.8351	168.2325	183.2829
Sedentary smokers	212.9973	9.4269	194.4959	231.4986
Sedentary non-smokers	221.4755	15.3046	201.4235	261.4915

From Table 4.6 it is once again evident that smoking has the opposite effect to what would be expected.

The active non-smokers have a higher asymptotic heartrate estimate than active smokers. The effect of lifestyle however does produce the expected results (i.e. active smokers have a lower asymptotic heartrate than sedentary smokers).

The squared multiple correlation coefficient (R^2) for function (4.1.2) is

$$\begin{aligned}
 R^2 &= 1 - \frac{\text{RESIDUAL SUM OF SQUARES}}{\text{CORRECTED TOTAL SUM OF SQUARES}} \\
 &= 1 - \frac{888454.324}{2437692.823} \\
 &= 0.64
 \end{aligned}$$

which indicates that 64% of the (corrected) total variation is attributed to the fit rather than left to residual error, which indicates a reasonable predictive ability for the model.

Of interest to a physiologist is to compare the asymptotic values of the subgroups in Table 4.6 with each other. These tests can be performed by using Bonferonni bounds. (see Srivastava and Carter 1983).

If $a_1, a_2 \dots a_k$ are independent and approximately normally distributed with respective standard deviations $s_1, s_2, \dots, s_k$ then the test statistic to compare them all is

$$t_{ij} = \frac{a_i - a_j}{\sqrt{s_i^2 + s_j^2}} \sim N(0,1) \text{ if the true parameter values are equal} \quad (4.1.5)$$

There are altogether $\binom{k}{2}$ tests. If each test is done at the $100\alpha\%$ significance level, then the overall significance level is $\leq \binom{k}{2}\alpha$. For example if $k = 5$ and $\alpha = 0.01$ then $\binom{k}{2} = 10$ and the overall significance level for all 10 pairwise tests is $\leq 10 \times 0.01 = 0.10$.

Therefore if you require an overall significance level of α^* ($= 0.05$, say), then the individual tests must be performed at the $\alpha = \frac{\alpha^*}{\binom{k}{2}}$ level e.g. $\frac{0.05}{10} = 0.005$ level.

TABLE 4.7 COMPARISONS BETWEEN THE ASYMPTOTIC HEARTRATE
ESTIMATES OF THE SUBGROUPS

	<u>T - value</u>
Active non-smokers vs Active smokers	2.153
Active non-smokers vs Sedentary smokers	-1.711
Active non-smokers vs Sedentary non-smokers	-2.191
Active smokers vs Sedentary smokers	-3.659 **
Active smokers vs Sedentary non-smokers	-3.531 **
Sedentary smokers vs Sedentary non-smokers	-1.028

For an overall significance level of 0.05 the six individual tests were performed at the $\alpha = \frac{0.05}{\binom{4}{2}} = 0.008$ level using (4.1.5).

From Table 4.7 it is evident that the asymptotic heartrate estimate of active smokers differed significantly from that of sedentary smokers and sedentary non-smokers. (** after the t-value indicates where a significant difference occurred).

In order to illustrate the effect of the various variables on the function the graphs in Figures 4.3 and 4.4 were plotted.

From Figure 4.3, which plots the logistic function for heartrate of active smokers and sedentary smokers, it is evident that for any given duration of exercise the sedentary subjects tend to have a higher heartrate than active smokers. The reason for this can be interpreted as greater demands of oxygen are necessary in an unfit person due to a reduction in metabolic efficiency at the cellular level.

As both groups are smokers it can be assumed that the amount of lung damage due to smoking is the same in both groups and therefore also the ability to oxygenate the blood. Therefore to comply with higher oxygen demands an increase in cardiac output must occur. It can thus be seen that one of the compensatory mechanisms to increase cardiac output (and therefore oxygen supply) would be to increase the heartrate.

Even though smoking has the opposite effect to what is expected the curves in Figure 4.4 confirm the results in Table 4.3 the active smoker has a lower heartrate over time than the sedentary non-smoker. These results warrant further investigation from a physiologist.

LOGISTIC HEARTRATE FUNCTION ACTIVE SMOKERS VS SEDENTARY SMOKERS

Figure 4.3

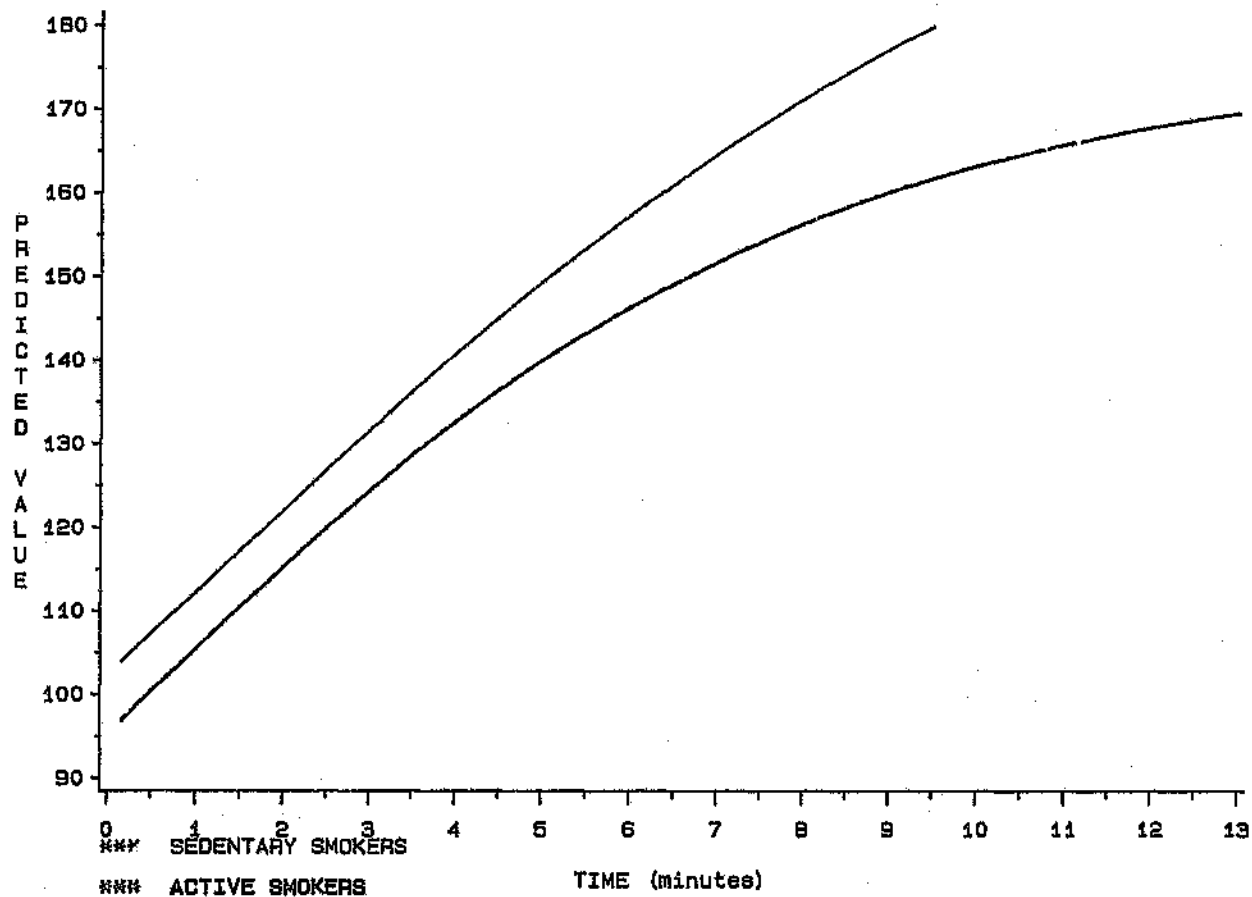
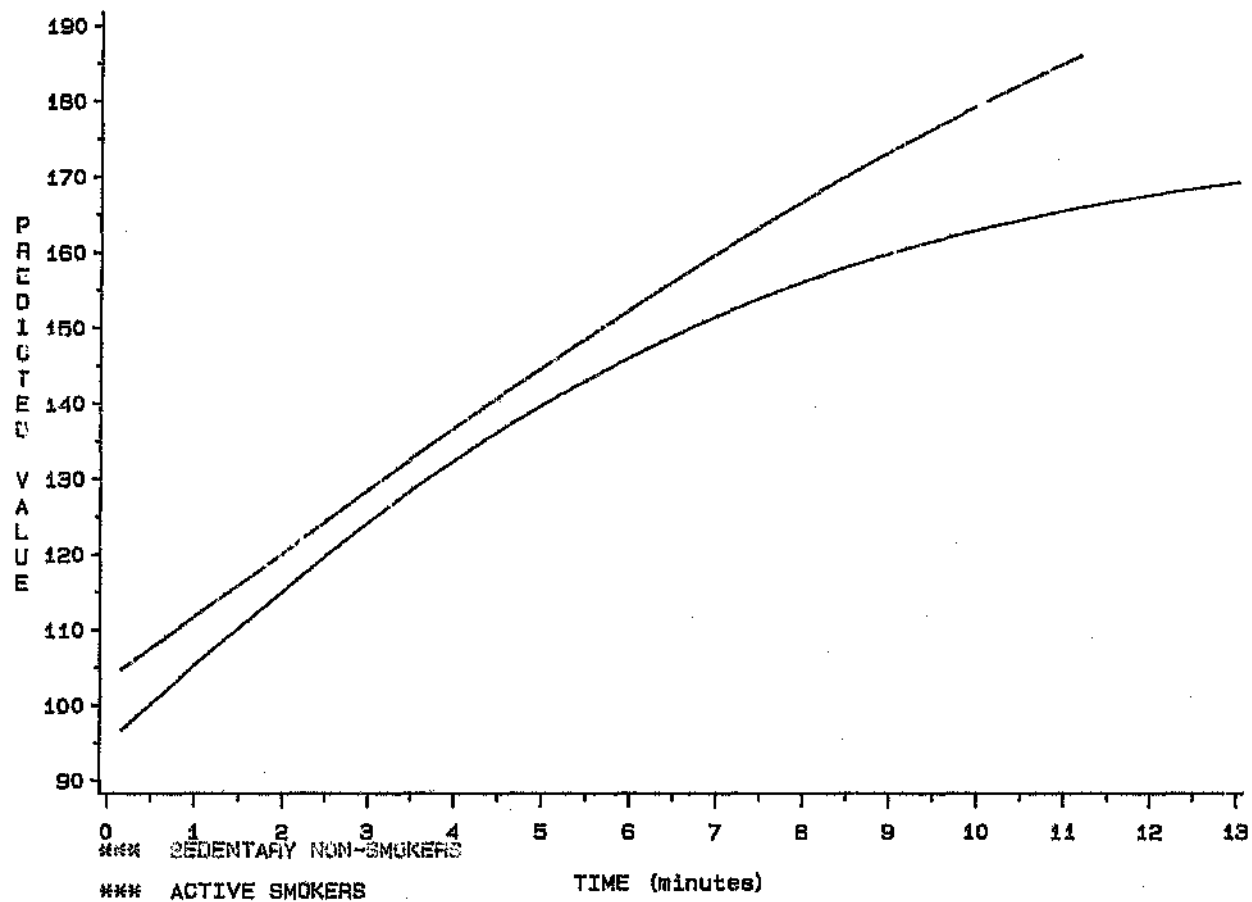


Fig. 4.4

LOGISTIC HEARTRATE FUNCTION
ACTIVE SMOKER VS SEDENTARY NON-SMOKER



4.2 APPLYING THE LOGISTIC FUNCTION FOR VENTILATION

Using the same procedure as in heartrate to arrive at the chosen function, resulted in the following:

$$VE = A / (1 + \text{EXP}(B_0 + B_1 \text{ time} + B_2 \text{ smoke} + B_3 \text{ Lifestyle} + B_4 \text{ weight})) \quad (4.2.1.)$$

From Table 4.8 one can observe that all the parameters are significant in the expected direction. The longer the exercise test the higher the ventilation; smoking increases ventilation; an active lifestyle results in a decreasing effect on ventilation and the heavier the subject the higher the ventilation tends to be. Once again there is a high correlation between the asymptote 'A' and the parameter 'B1' for time. For the same reasons provided in the results for heartrate, function (4.2.1) was chosen as the best fitting function for ventilation

$$\begin{aligned} R^2 &= 1 - \frac{\text{RESIDUAL SUM OF SQUARES}}{\text{CORRECTED TOTAL SUM OF SQUARES}} \\ &= 1 - \frac{82887.397}{2104035.981} \\ &= 0.61 \end{aligned}$$

indicating that 61% of the (corrected) total variation is attributed to fit rather than left to residual error. The residuals in Figure 4.5 form a band around the horizontal zero line, indicating a constant variance. There is however a tendency of underestimation for low time and of overestimation for high time. The normal plot for Fig. 4.6 is a straight line except for a slight deviation in the left-hand tail which confirms the assumption of normality. From Table 4.9 it would seem that the effect of smoking on ventilation is far greater than that of lifestyle. However Table 4.8 says the opposite since $|\hat{\beta}_3| > |\hat{\beta}_2|$. This could be ascribed to the fact that the variable weight was excluded from the subgroups. It would appear however from Table 3.1 that this is not the case since all the subgroups have more or less the same average weight.

Table 4.8

Non-linear least square statistics for ventilation

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	6	13292541.603	2215423.600
RESIDUAL	4005	826887.397	206.963
UNCORRECTED TOTAL	4011	14121429.000	
(CORRECTED TOTAL)	4010	2104035.981	

PARAMETER	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95 % CONFIDENCE INTERVAL	
			LOWER	UPPER
A	98.20196199	2.2360537837	93.817973300	102.58595069
B0	2.27203107	0.0910613001	2.093497032	2.45056511
B1	-0.51318701	0.0194914995	-0.551401885	-0.47497214
B2	-0.14436494	0.0218490485	-0.187402004	-0.10172787
B3	0.15722076	0.0218900588	0.114303311	0.20013825
B4	-0.01532111	0.0011433853	-0.017562817	-0.01307939

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

CORR	A	B0	B1	B2	B3	B4
A	1.0000					
B0	-0.2705	1.0000				
B1	0.9101	-0.3676	1.0000			
B2	0.2077	-0.1692	0.2098	1.0000		
B3	-0.1807	-0.1274	-0.1991	-0.0610	1.0000	
B4	0.4287	-0.9461	0.4273	0.0604	0.0009	1.0000

Figure 4.5

LOGISTIC RESIDUAL PLOT FOR HEARTHATE
 $HR = A / (1 + \exp(B_0 + B_1 \cdot TIME + B_2 \cdot SMOKE + B_3 \cdot LIFESTL + B_4 \cdot WEIGHT))$

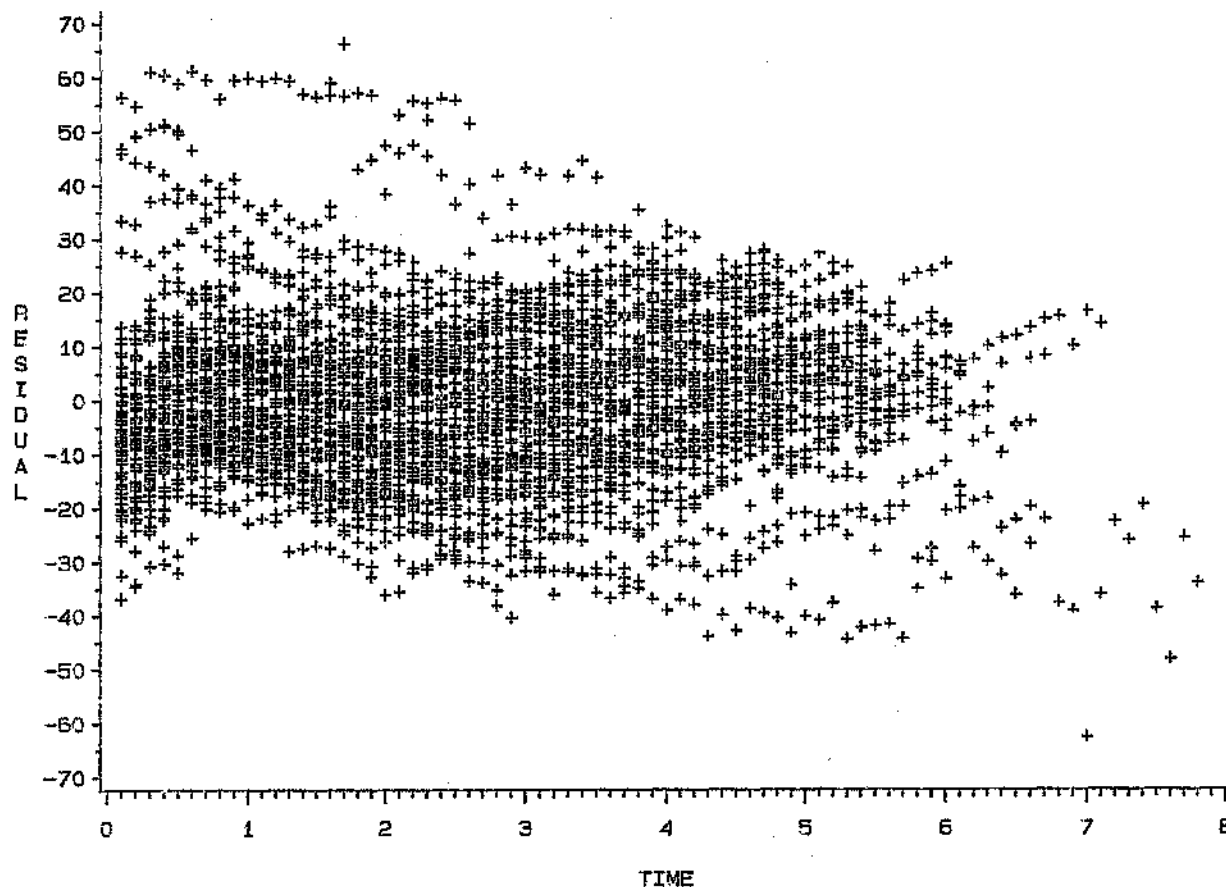
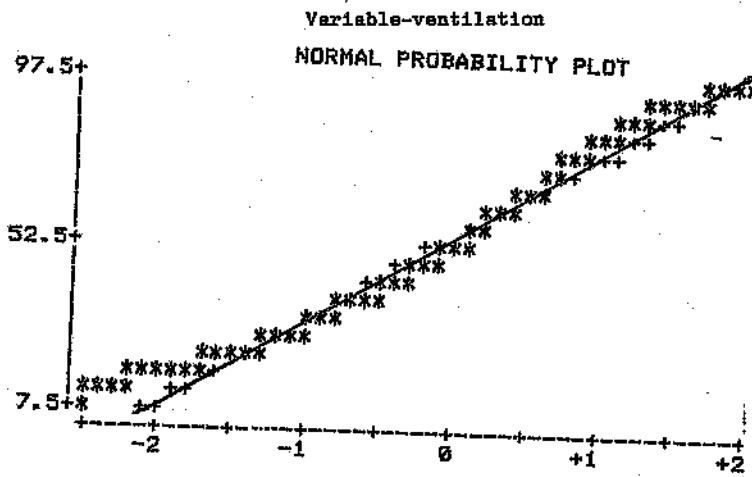


Figure 4.6

Therefore no apparent reason could be found for this anomalous result. From 4.10 it is evident that non-smokers had significantly lower ventilation than smokers irrespective of whether they were active or sedentary.

TABLE 4.9 ESTIMATES AND CONFIDENCE INTERVALS FOR ASYMPTOTIC VENTILATION VALUE

	ESTIMATE	ASYMPTOTIC STANDARD ERROR	ASYMPTOTIC 95% CONFIDENCE INTERVAL	
			LOWER	UPPER
Active smokers	97.6094	2.8473	92.0223	103.1965
Active non-smokers	79.9040	3.1805	73.6628	86.1452
Sedentary smokers	98.1095	4.0405	90.1796	106.0393
Sedentary non-smoker	81.2077	2.1225	77.0424	85.3729

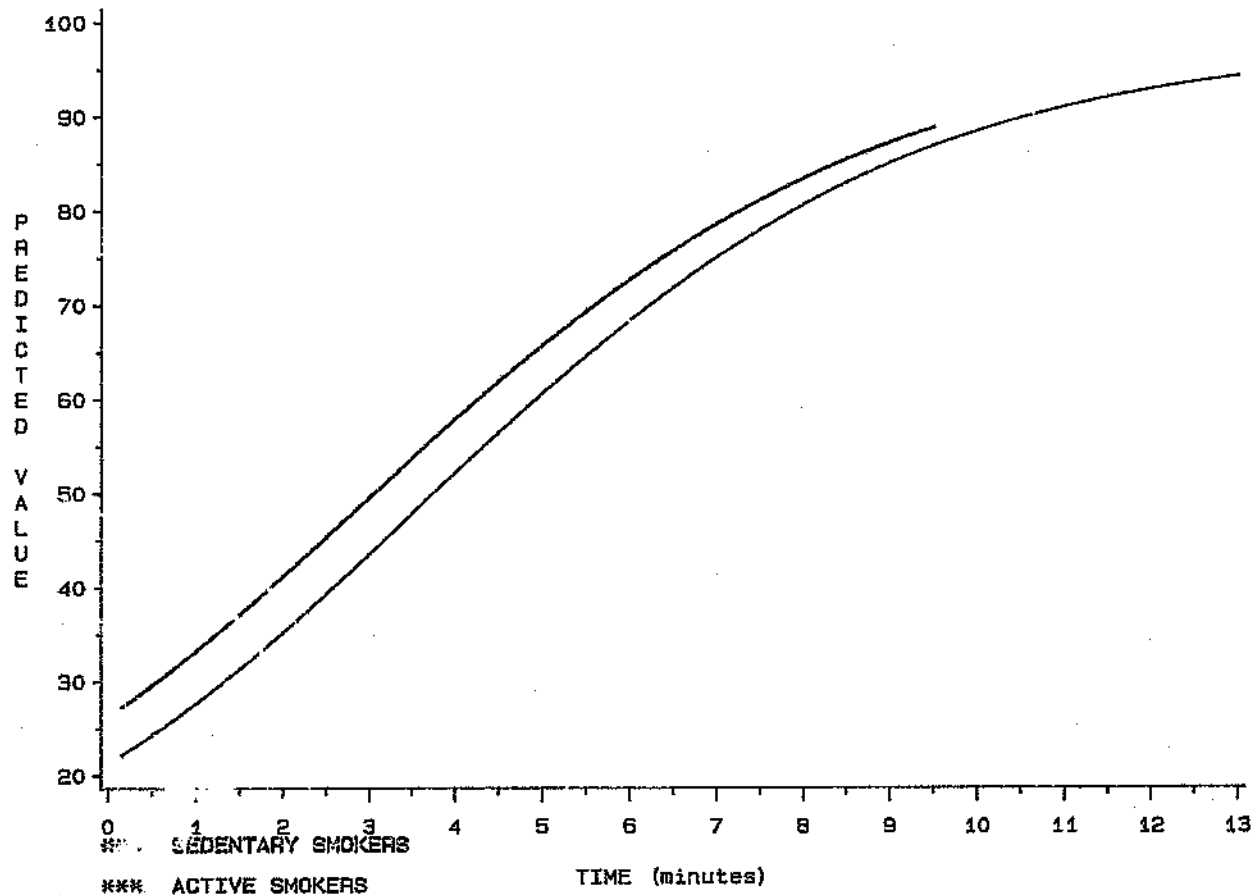
TABLE 4.10 COMPARISONS BETWEEN THE ASYMPTOTIC VENTILATION ESTIMATES OF THE SUBGROUPS

	<u>T-value</u>
Active non-smokers vs. Active smokers	4.148 * *
Active non-smokers vs sedentary smokers	-3.541 * *
Active non-smokers vs sedentary non-smokers	-0.341
Active smokers vs sedentary smokers	-0.101
Active smokers vs sedentary non-smokers	4.619 * *
Sedentary smokers vs sedentary non-smokers	3.706 * *

The curve in Fig. 4.7 reflect what is expected i.e. sedentary people require a greater ventilatory effort than active people to support prolonged exercise because their efficiency in utilising oxygen is lower.

LOGISTIC VENTILATION FUNCTION ACTIVE SMOKERS VS SEDENTARY SMOKERS

Figure 4.7



4.3 APPLYING THE LOGISTIC FUNCTION FOR OXYGEN CONSUMPTION

The best fitting model was:

$$VO_2\text{-kg} = A / (1 + \text{EXP}(B_0 + B_1 \text{ time} + B_2 \text{ smoke} + B_3 \text{ lifestyle} + B_4 \text{ weight} + B_5 \text{ smoke} \times \text{lifestyle})) \quad (4.3.1)$$

$VO_2\text{-kg}$ is a measure of the subjects metabolic rate and its unit of measurement is millilitres per minute per kilogram. (See Section 3.5)

Table 4.11 indicates that whilst the parameters for time and smoking are significant in the correct direction, the parameters for weight and lifestyle have the opposite effect on oxygen consumption during exercise to what is expected. There is also a significant interaction term, smoking $\times$ lifestyle which has an increasing effect on the heartrate. A contributing factor to this surprising result could be the self selection of the subjects that underwent the exercise stress test i.e. the subjects' volunteered for the test and were not scientifically selected. Further investigation from a physiologist is warranted in this case.

$$\begin{aligned} R^2 &= 1 - \frac{\text{RESIDUAL SUM OF SQUARES}}{\text{CORRECTED TOTAL SUM OF SQUARES}} \\ &= 1 - \frac{198200.8567}{372935.8456} \\ &= 0.47 \end{aligned}$$

which indicates that only 47% of the (corrected) total variation is attributed to the fit rather than left to residual error.

The residual plot in Fig. 4.8 illustrates that the observed values were generally close to the expected values for this specific model. There were two individuals whose oxygen consumption increased dramatically with time. Their measurements appear as outliers in Figure 4.8 which could account for the fact that R^2 is only 0.47 and why the variance increases with time. The normal plot of Figure 4.9 confirms the assumption that the data values are approximately normally distributed.

Table 4.11. Non-linear least squares statistics for oxygen consumption

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	7	2183030.4072	311851.4867
RESIDUAL	3937	160209.3928	40.6933
UNCORRECTED TOTAL	3944	2343239.8000	
(CORRECTED TOTAL)	3943	339464.4544	

PARAMETER	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95 % CONFIDENCE INTERVAL	
			LOWER	UPPER
A	37.00000000	0.78353246456	35.463804671	38.536195329
B0	0.59090485	0.10560168074	0.383461979	0.797947715
B1	-0.52443403	0.02283617704	-0.569206680	-0.479661371
B2	-0.11817732	0.03687714145	-0.190478716	-0.045875921
B3	-0.21629218	0.03774499139	-0.290295082	-0.142289269
B4	0.00391938	0.00124515009	0.003478133	0.008360621
B5	-0.10547037	0.05224316065	-0.207898415	-0.003042363

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

CORR	A	B0	B1	B2	B3	B4	B5
A	1.0000	0.2153	0.8943	0.1029	0.2232	-0.1776	0.0935
B0	0.2153	1.0000	0.1051	-0.1680	-0.2230	-0.9428	0.1900
B1	0.8943	0.1051	1.0000	0.1104	0.2147	-0.1705	0.0815
B2	0.1029	-0.1680	0.1104	1.0000	0.5095	-0.0097	-0.6882
B3	0.2232	-0.2230	0.2147	0.5095	1.0000	0.0610	-0.6623
B4	-0.1776	-0.9428	-0.1705	-0.0097	0.0610	1.0000	-0.0663
B5	0.0935	0.1900	0.0815	-0.6882	-0.6623	-0.0663	1.0000

LOGISTIC RESIDUAL PLOT FOR OXYGEN CONSUMPTION
 $VO2_KG = A / (1 + \exp(B0 + B1 \cdot TIME + B4 \cdot SMOKE + B5 \cdot LIFESTL + B7 \cdot WEIGHT))$

Figure 4.8

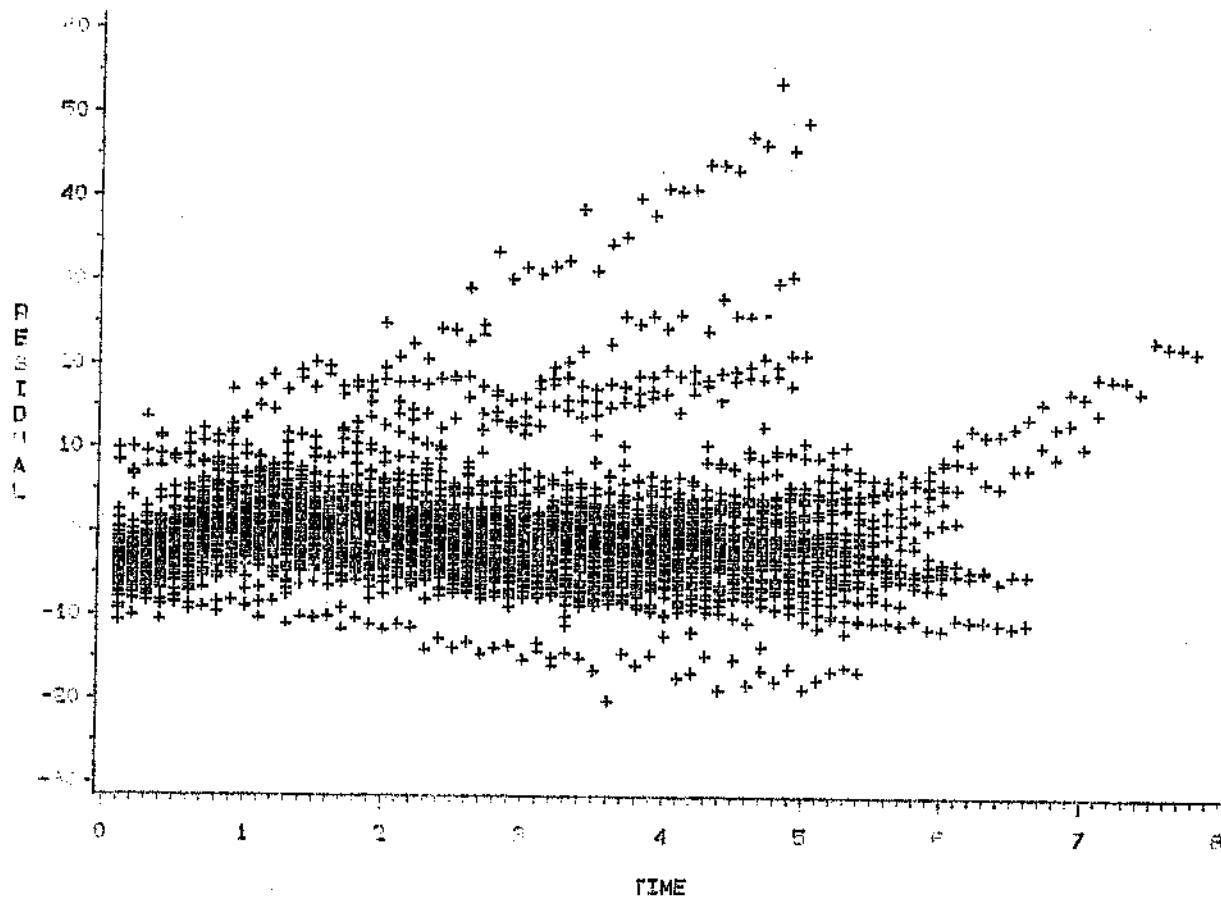


TABLE 4.12 ESTIMATES AND CONFIDENCE INTERVALS FOR ASYMPTOTIC
OXYGEN CONSUMPTION VALUES

	<u>ESTIMATE</u>	<u>ASYMPTOTIC STD ERROR</u>	<u>ASYMPTOTIC 95% CONFIDENCE INTERVAL</u>	
			<u>LOWER</u>	<u>UPPER</u>
Active smokers	37.0000	1.3139	34.4218	39.5781
Active non-smokers	34.2927	1.0701	32.1927	36.3927
Sedentary smokers	34.1925	1.2529	31.7334	36.6515
Sedentary non-smokers	23.6778	0.4437	22.8069	24.5486

From Table 4.12 it is evident that there is a dramatic increase in the oxygen consumption between sedentary non-smokers and sedentary smokers, but not such a great increase in active people.

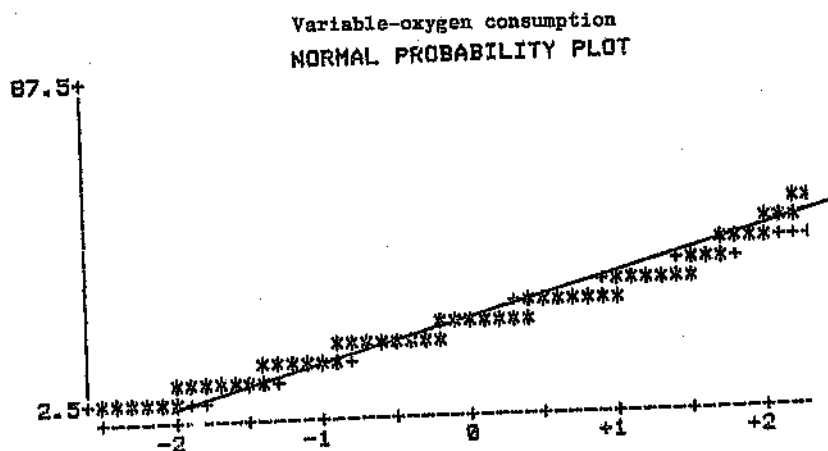
The curves in Figure 4.10 are exactly as would be expected. The smokers have a diminished capacity for gaseous exchange and therefore require greatly increased ventilation i.e. laboured breathing. This extra work in breathing results in an increased metabolic rate and therefore increased oxygen consumption.

TABLE 4.13 COMPARISONS BETWEEN THE ASYMPTOTIC OXYGEN
CONSUMPTION ESTIMATES OF THE SUBGROUPS

	<u>T - value</u>
Active smokers vs Active non-smokers	1.598
Active smokers vs Sedentary smokers	1.546
Active smokers vs Sedentary non-smokers	9.824 * *
Active non-smokers vs Sedentary smokers	0.061
Active non-smokers vs Sedentary non-smokers	9.167 * *
Sedentary smokers vs Sedentary non-smokers	7.912 * *

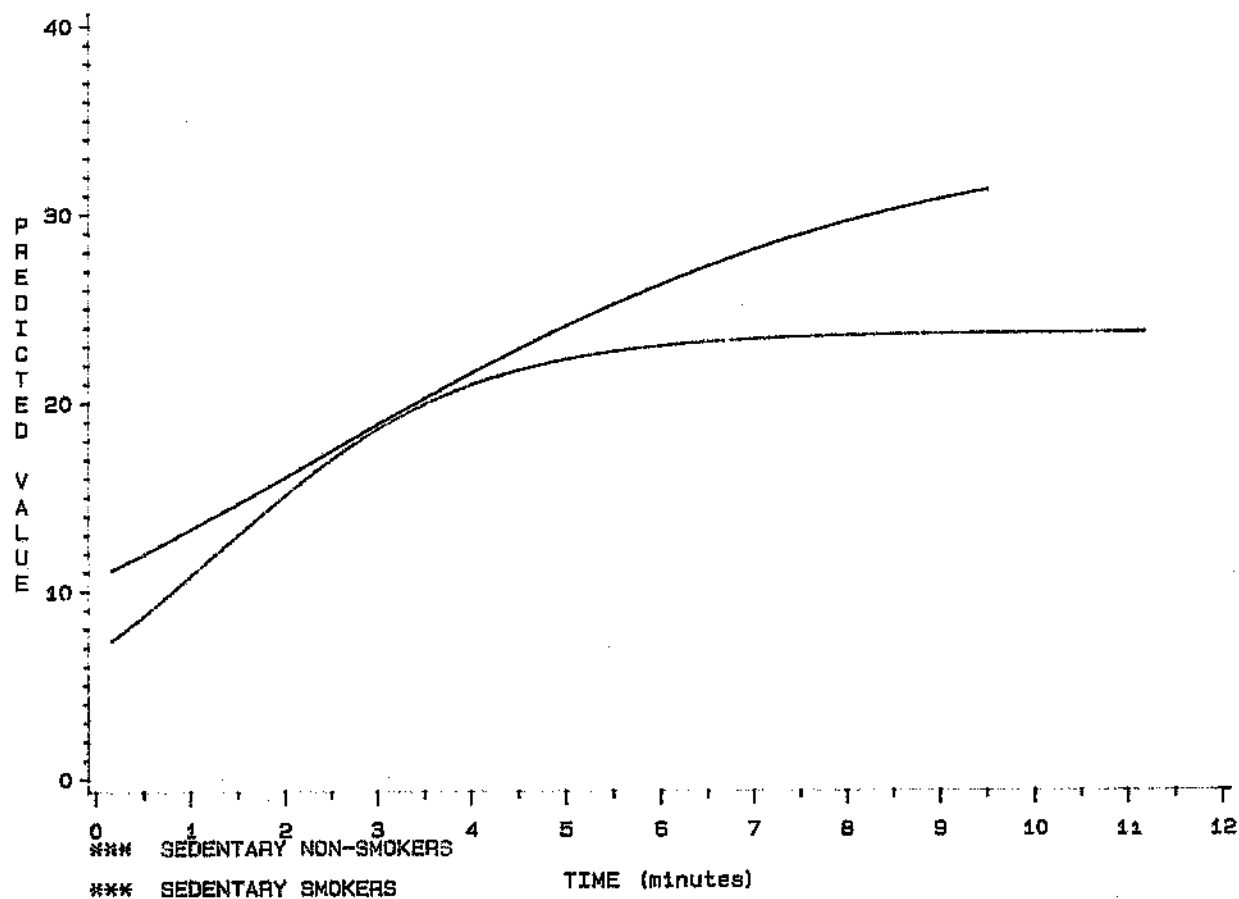
From Table 4.13 it is evident that there is a significant difference in the asymptotic oxygen consumption estimate between active non-smokers and sedentary non-smokers, and between sedentary smokers and sedentary non-smokers. This confirms the results in Table 4.11 that both lifestyle and smoking have an effect on the asymptotic oxygen consumption estimate.

Figure 4.9



LOGISTIC OXYGEN CONSUMPTION FUNCTION SEDENTARY SMOKERS VS SEDENTARY NON-SMOKERS

Figure 4.10



4.4 SUMMARY

It is evident from this section that the experimental design had a major influence on the results and the type of logistic function chosen. The fact that workload and time were completely confounded in the experiment, as well as treadmill speed and age, created difficulties in interpreting the results. Alternative forms of the logistic function were applied in order to reduce the high correlation between the asymptote and time. Unfortunately these new approaches did not have the desired effect, and the original function that was proposed, was used.

The parameter estimates for the variable ventilation were as expected i.e. smoking, time and weight had an increasing effect whilst an active lifestyle had a decreasing effect. The parameter estimate for smokers on the heartrate variable was contrary to what is expected i.e. smoking had a decreasing effect on the heartrate estimate. The effects of time, lifestyle and weight were however physiologically acceptable. For the variable oxygen consumption the parameters for time and lifestyle were significant in the correct direction whilst the parameters for weight and smoking were significant in the opposite direction to what is expected. The fact that the subjects were self selected could be the reason for some of these unexpected results. The squared multiple correlation coefficient was high enough to indicate a reasonable predictive ability for all three models.

CHAPTER 5

5.1 APPLICATION OF THE RAO-KHATRI MODEL

INTRODUCTION

In this section the Rao-Khatri model is applied to the data. A discussion of the similarities and differences between the Rao-Khatri model and the Potthoff and Roy model can be found in Section 2.6. There were two main reasons for applying the Rao-Khatri model instead of the Potthoff and Roy model. Firstly there was a lack of information concerning the matrix G for the Potthoff and Roy model. Since Σ_0 is unknown $G = \Sigma_0$ cannot be used. Grizzle and Allen (1989) proved that when $G = S$ the two models are actually equivalent. Another reason for applying the Rao-Khatri model is that standard multivariate analysis of covariance programs can be used for the modelling procedure, whereas the Potthoff and Roy model requires an involved program to handle the iterative procedure necessary to estimate Σ_0 . The three variables heartrate, ventilation, and oxygen consumption are fitted by this model. Firstly a decision is made on what degree polynomial to use, and thereafter the inclusion of covariables is considered. Comparisons between the Rao-Khatri model and Richards' logistic function are discussed and graphical representations are presented depicting these differences.

HEARTRATE RESULTS

Consider the model $E(Y_0) = A\xi P$ (See(2.1.1)) where A is the design matrix, ξ the parameters being estimated and P a third degree orthogonal polynomial. Since there was equal spacing between the data observations, orthogonal polynomial coefficients could be used. To test the adequacy of the third degree polynomial we compare it with the matrix C_2 , the coefficients corresponding to the orthogonal polynomials of degrees 4,5 and 6. (Pearson and Hartley (1966)). The matrices A , P , and C_2 can be found in Appendix C.

Since $PC_2 = 0$ the test for adequacy given in Section 2.5 is equivalent to testing

$H_0: BC_2 = 0$ for the model $E(Y_0) = AB$.

Using the theory of Section 2.4 we have $M = I_4$ since there are four groups, $V = C_2$, $u = 3$ (columns in C_2) $s = 4$ (rank of M) and $n - m = 51$. Applying the theory leading up to (2.2.4) and (2.2.5) we have

$$H = (\hat{B}C_2)'(A'A)\hat{B}C_2 \quad (5.1.1)$$

and
$$E = C_2'Y'(I_{51} - A(A'A)^{-1}A')YC_2 \quad (5.1.2)$$

so that Wilk's lambda is

$$\Lambda = \frac{|E|}{|E + H|} \sim U_{3,4,51} \quad (\text{See (2.2.6)}).$$

$$= 0.410.$$

It was decided for the application of this method to only use those subjects who underwent the experiment for 7.5 minutes and longer.

This restriction ensured that there were no missing observations. However it resulted in there being a reduction of 25 subjects from the original sample. i.e. there were 25 subjects that lasted on the treadmill for less than 7.5 minutes.

Since $-f \log \Lambda = 19.748$ (where $f = n - m - \frac{1}{2}(u - s + 1)$ see (2.2.7)) is less than $\chi^2_{12}(0.05) = 21.03$, we do not reject H_0 . i.e. a third degree polynomial describes the data well, at the 5% significance level.

Testing the adequacy of a 2nd degree polynomial produced the following:

$$\Lambda = \frac{|E|}{|E+H|} = 0.3803.$$

In this case the exact F-test is used as opposed to the approximate χ^2 test. (See (2.2.8)).

$$\frac{1 - \lambda^{\frac{1}{2}}}{\lambda^{\frac{1}{2}}} \cdot \frac{2v_2 + 2}{2v_1 + 3} \sim F_{4v_1+6, 4v_2+4}$$

$$= 8.324 \text{ with } V_1 = \frac{1}{2} \text{ and } V_2 = 24.$$

Since $24 > F_{8,100,0.05} (\approx 2.08)$ the hypothesis that a second degree polynomial fits the data is rejected at a 5% significance level. Therefore a 3rd degree polynomial was used.

In order to estimate the parameters for the four groups (Active smokers, active non-smokers, sedentary smokers and sedentary non-smokers) we first consider which of the three covariates corresponding to the fourth, fifth and sixth degree orthogonal polynomial shall be used. These covariates are used if there is a high correlation with the variables corresponding to the third degree polynomial. Covariates are included in the estimation process but their coefficients are not estimated. (See Section 2.6).

First choose C_1 and C_2 such that $P'C_1 = I_4$ and $P'C_2 = 0$ and transform the data matrix to YC where $C = (C_1, C_2)$.

From Table 5.1 it is evident that none of the columns in the (1,2) part of the table has all its correlations small, so that all three covariates are appropriate. i.e. all three covariates are to be included in the estimation process.

TABLE 5.1 CORRELATION MATRIX FOR THE COLUMNS OF THE TRANSFORMED DATA (YC_1 , YC_2)

	YC_1				YC_2		
YC_1	1.000	0.2289	-0.2310	-0.0681	0.0556	-0.0771	-0.0573
	-	1.000	-0.3464	-0.0650	0.1367	-0.0462	0.2178
	-	-	1.000	-0.2981	-0.4360	0.3856	-0.1881
	-	-	-	1.000	-0.1339	-0.4460	0.0763
YC_2	-	-	-	-	1.000	-0.4499	0.2520
	-	-	-	-	-	1.000	-0.5545
	-	-	-	-	-	-	1.0000

Therefore in estimating ξ , the parameters for the four groups, the coefficients corresponding to the fourth, fifth and sixth-degree orthogonal polynomials are also included in the P matrix as covariates when estimating ξ in (5.1.3). i.e. they provide information about the third degree polynomial coefficients being estimated.

$$\xi = (A'A)^{-1}A'YS^{-1}P(P'S^{-1}P)^{-1} \dots \text{from (2.4.15)} \quad (5.1.3)$$

$$\text{where } S = \frac{Y'(I_{51} - A(A'A)^{-1}A')Y}{51}$$

	Intercept	Linear	Quadratic	Cubic	
Therefore $\hat{\xi} =$	131.8	1.1072	.0018474	.0010366	group 1
	131.5	1.1039	.0038426	3.9E-04	group 2
	123.8	0.9781	.0051974	-1.0E-04	group 3
	126.6	1.4180	.0016696	8.2E-04	group 4

Using Section 1.2 (VI) on orthogonal polynomials, the fitted curves for heartrate for the four groups are in Table 5.2.

TABLE 5.2 Orthogonal Polynomials For Heartrate

Active smokers

$$\begin{aligned} Y &= 131.8 + (1.1072)(1)X + 0.00185(3)[X^2 - \frac{1}{12}(45^2 - 1)] \\ &\quad + 0.0010366(\frac{5}{8})[X^3 - \frac{1}{20}(3(45)^2 - 7X)] \\ &= 131.8 + 1.1072X + 0.0055(X^2 - 168.67) + 0.0008(X^3 - 303.4X) \\ &= 130.87 + 0.8645X + 0.0055X^2 + 0.0008X^3 \end{aligned}$$

Active non-smokers

$$\begin{aligned} Y &= 131.5 + 1.1039X + 0.01152(X^2 - 168.67) + 0.000325(X^3 - 303.4X) \\ &= 129.5569 + 1.0052X + 0.0115X^2 + 0.000325X^3 \end{aligned}$$

Sedentary smokers

$$\begin{aligned} Y &= 123.8 + 0.9781X + 0.01557(X^2 - 168.67) - 0.0008(X^3 - 303.4X) \\ &= 121.174 + 1.2208X + 0.01557X^2 - 0.0008X^3 \end{aligned}$$

Sedentary non-smokers

$$\begin{aligned} Y &= 126.6 + 1.418X + 0.0049(X^2 - 168.67) + 0.0007(X^3 - 303.4X) \\ &= 125.7735 + 1.2056X + 0.0049X^2 + 0.0007X^3 \end{aligned}$$

time and X correspond as follows:

$$\phi_1(X_i) = X_i: -22 -21 -20 \dots 0 \dots 20 21 22$$

time: 10 20 30 ... 230 ... 430 440 450 seconds

Clearly the required coding is

$$X = \frac{1}{10}(\text{time} - 230)$$

The same procedure for fitting polynomial curves to the heartrate data was used for the ventilation and the oxygen consumption data.

VENTILATION RESULTS

A third degree polynomial proved to be the best fitting function to describe the data with

$$\lambda = 0.476$$

so that $-f \log \lambda = 16.44$, which is less than $\chi^2_{12}(0.05) = 21.03$.

Table 5.3 indicates that none of the columns in the (1,2) part of the table has all its correlations small, so that all three covariates should be included in the estimation process.

TABLE 5.3 Correlation Matrix For The Columns Of The Transformed Data (YC_1, YC_2)

	YC_1				YC_2		
	1.000	0.7282	-0.3195	-0.0794	0.4361	-0.1220	-0.0471
	-	1.000	-0.0117	0.4336	0.3388	-0.1776	0.0572
YC_1	-	-	1.000	0.0090	0.0992	0.1390	-0.3057
	-	-	-	1.000	-0.2183	-0.0934	-0.0220
	-	-	-	-	1.000	-0.1568	-0.0374
YC_2	-	-	-	-	-	1.000	-0.1120
	-	-	-	-	-	-	1.0000

The parameter estimates are therefore

$$\hat{\xi} = \begin{bmatrix} 45.9877 & 1.0913 & .0032915 & .0015517 \\ 38.6033 & 0.8188 & 7.4E-04 & .0018633 \\ 50.7020 & 1.1706 & .0037024 & .0017374 \\ 50.4352 & 1.4819 & .0012755 & -1.2E-04 \end{bmatrix} \begin{matrix} \text{Active smokers} \\ \text{Active non-smokers} \\ \text{Sedentary smokers} \\ \text{Sedentary non-smokers} \end{matrix}$$

Using Section 1.2 (VI) we have the following four functions with $X = \frac{1}{10}(\text{time} - 230)$ in Table 5.4.

TABLE 5.4 Orthogonal Polynomials For Ventilation

Active smokers

$$\begin{aligned} Y &= 45.9877 + 1.0913X + 0.0099(X^2 - 168.67) + 0.0013(X^3 - 303.4X) \\ &= 44.3178 + 0.6969X + 0.0099X^2 + 0.0013X^3 \end{aligned}$$

Active non-smokers

$$\begin{aligned} Y &= 38.6033 + 0.8188X + 0.0022(X^2 - 168.67) + 0.0015(X^3 - 303.4X) \\ &= 38.2322 + 0.3637X + 0.0022X^2 + 0.0015X^3 \end{aligned}$$

Sedentary smokers

$$\begin{aligned} Y &= 50.702 + 1.1706X + 0.0111(X^2 - 168.67) + 0.0014(X^3 - 303.4X) \\ &= 50.702 + 0.7456X + 0.0111X^2 + 0.0014X^3 \end{aligned}$$

Sedentary non-smokers

$$\begin{aligned} Y &= 50.4352 + 1.4819X + 0.00381(X^2 - 168.67) - 0.0001(X^3 - 303.4X) \\ &= 49.8097 + 1.5122X + 0.0038X^2 - 0.0001X^3 \end{aligned}$$

OXYGEN CONSUMPTION RESULTS

A third degree polynomial once again proved to be the best fitting function to describe the data with $\lambda = 0.465$ so that $-f \log \lambda = 16.96$ which is less than $21.03 (= \chi^2_{12}(0.05))$.

All the covariates corresponding to the 4th, 5th and 6th degree polynomial coefficients should be included in the estimation process since none of the columns in the (1,2) part of the table has all its correlations small.

TABLE 5.5 Correlation Matrix For The Columns Of The Transformed DATA (YC_1, YC_2)

	YC_1				YC_2		
YC_1	1.000	0.8765	-0.1703	-0.0794	0.1559	-0.1033	-0.0983
	0.8765	1.000	-0.1054	-0.2930	-0.1074	-0.1618	0.1046
	-0.1703	0.1054	1.0000	-0.4569	-0.1090	0.2721	-0.1002
	-0.0769	-0.2930	-0.4569	1.0000	-0.3322	-0.3559	-0.2202
YC_2	-0.1559	-0.1074	-0.1090	-0.3322	1.0000	-0.2671	-0.7304
	0.1033	0.1618	0.2721	-0.3559	-0.1671	1.0000	-0.0602
	0.0983	0.1046	0.1002	0.2202	-0.7304	-0.0602	1.0000

The parameter estimate for the four groups are therefore

$$\hat{\xi} = \begin{bmatrix} 14.8591 & 0.3505 & -0.001143 & 1.3E-04 \\ 16.6081 & 0.2974 & -2.4E-04 & 5.1E-04 \\ 19.4863 & 0.3907 & -3.6E-04 & 2.1E-04 \\ 21.4493 & 0.4951 & -0.015995 & 5.03E-04 \end{bmatrix}$$

with the final polynomial functions given in Table 5.6.

TABLE 5.6 Orthogonal Polynomials For Oxygen Consumption

Active smokers

$$Y = 14.8591 + 0.3505X - 0.0034(X^2 - 168.67) + 0.0001(X^3 - 303.4X) \\ = 15.4325 + 0.3202X - 0.0034X^2 + 0.0001X^3$$

Active non-smokers

$$Y = 16.6081 + 0.2974X - 0.0007(X^2 - 168.67) + 0.0004(X^3 - 303.4X) \\ = 16.7261 + 0.1760X - 0.0007X^2 + 0.0004X^3$$

Sedentary smokers

$$Y = 19.4863 + 0.3907X - 0.0011(X^2 - 168.67) + 0.0002(X^3 - 303.4X) \\ = 19.6718 + 0.3300X - 0.0011X^2 + 0.002X^3$$

Sedentary non-smokers

$$Y = 21.4193 + 0.4951X - 0.0048(X^2 - 168.67) + 0.0004(X^3 - 303.4X) \\ = 22.2589 + 0.3737X - 0.0048X^2 + 0.0004X^3$$

where $X = \frac{1}{10}(\text{time} - 230)$.

5.2 GRAPHICAL COMPARISONS BETWEEN THE RAO-KHATRI MODEL AND RICHARDS' LOGISTIC FUNCTION

Since the application of the Rao-Khatri model had 25 subjects less than the sample used for the application of the logistic function, only graphical plots were used to compare the two approaches. In order to illustrate these differences an example was used for each variable: i.e. for oxygen consumption sedentary non-smokers were used, for heartrate, active non-smokers and for ventilation, sedentary smokers. It is evident from Figures 5.1, 5.2 and 5.3 that the two functions display similar tendencies in the first half of the experiment. In Figure 5.3 the predicted Rao-Khatri polynomial for example indicates an acceptable heartrate of 110 after 7.5 minutes whilst the predicted logistic function indicates a heartrate of 120. However the Rao-Khatri polynomial provides unacceptable values in the second half of the experiment, for example a heartrate of 400 after 11 minutes. This is due to the square and cubic effects in the polynomial. The logistic function on the other hand reaches a more acceptable asymptotic value, for example a heartrate of 180 after 11 minutes.

It therefore indicates that the Rao-Khatri model is suitable for interpolation but not for extrapolation. The logistic function provided more physiologically acceptable estimates the longer the exercise test endured. The major advantage however of the Rao-Khatri method is that one can apply a standard multivariate analysis of covariance program to the data.

OXYGEN CONSUMPTION FUNCTIONS FOR SEDENTARY NON-SMOKERS RAO-KHATRI POLYNOMIAL vs LOGISTIC FUNCTION

Figure 5.1

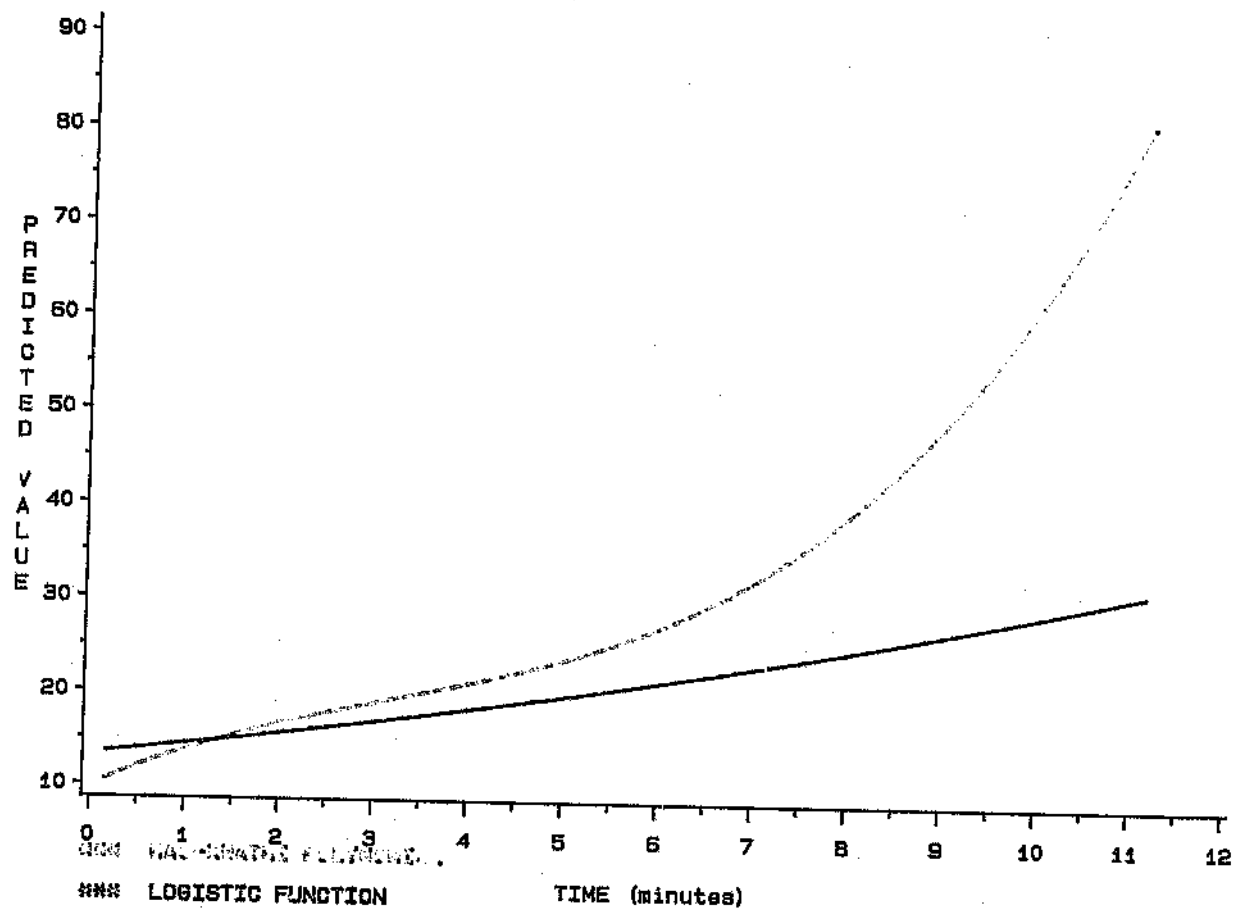


Figure 5.2

VENTILATION FUNCTIONS FOR SEDENTARY SMOKERS
RAO-KHATRI POLYNOMIAL vs LOGISTIC FUNCTION

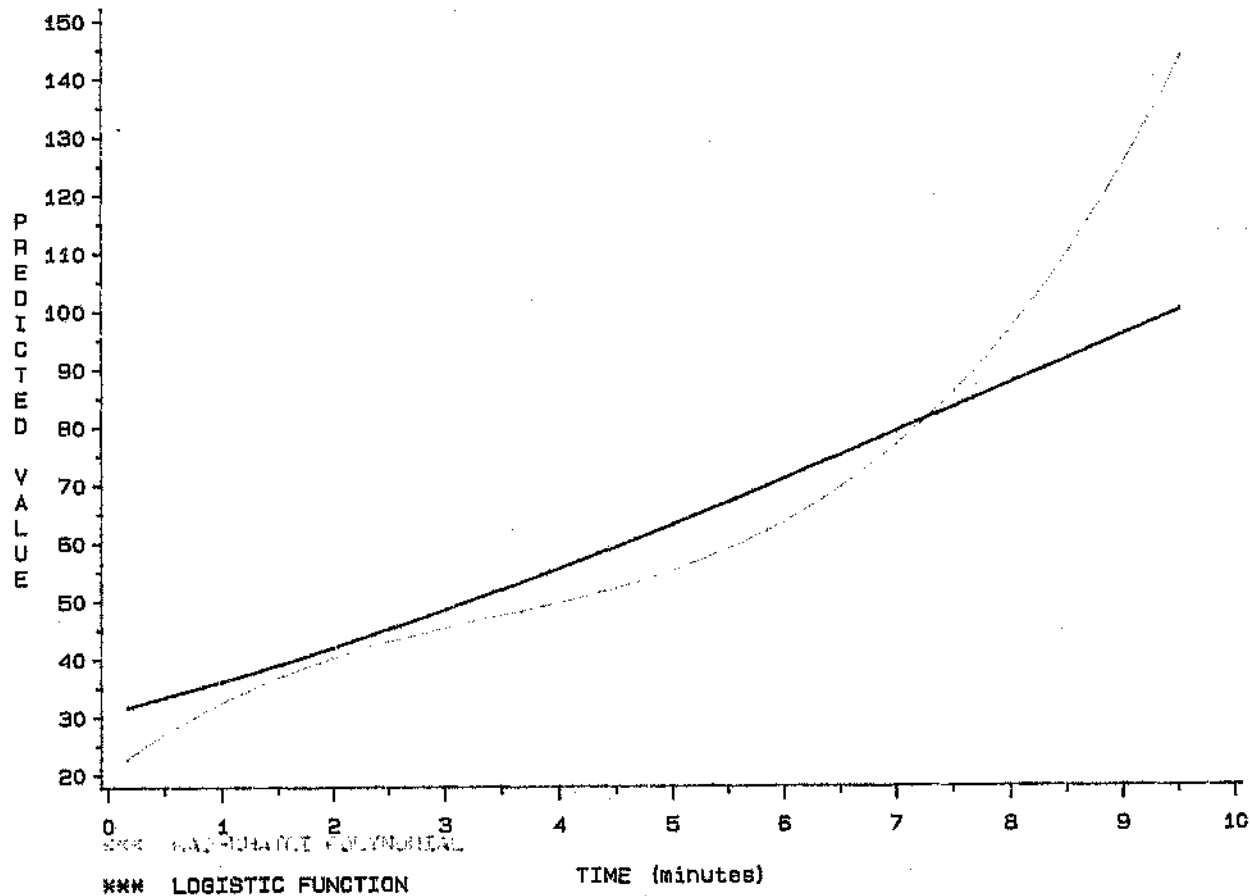
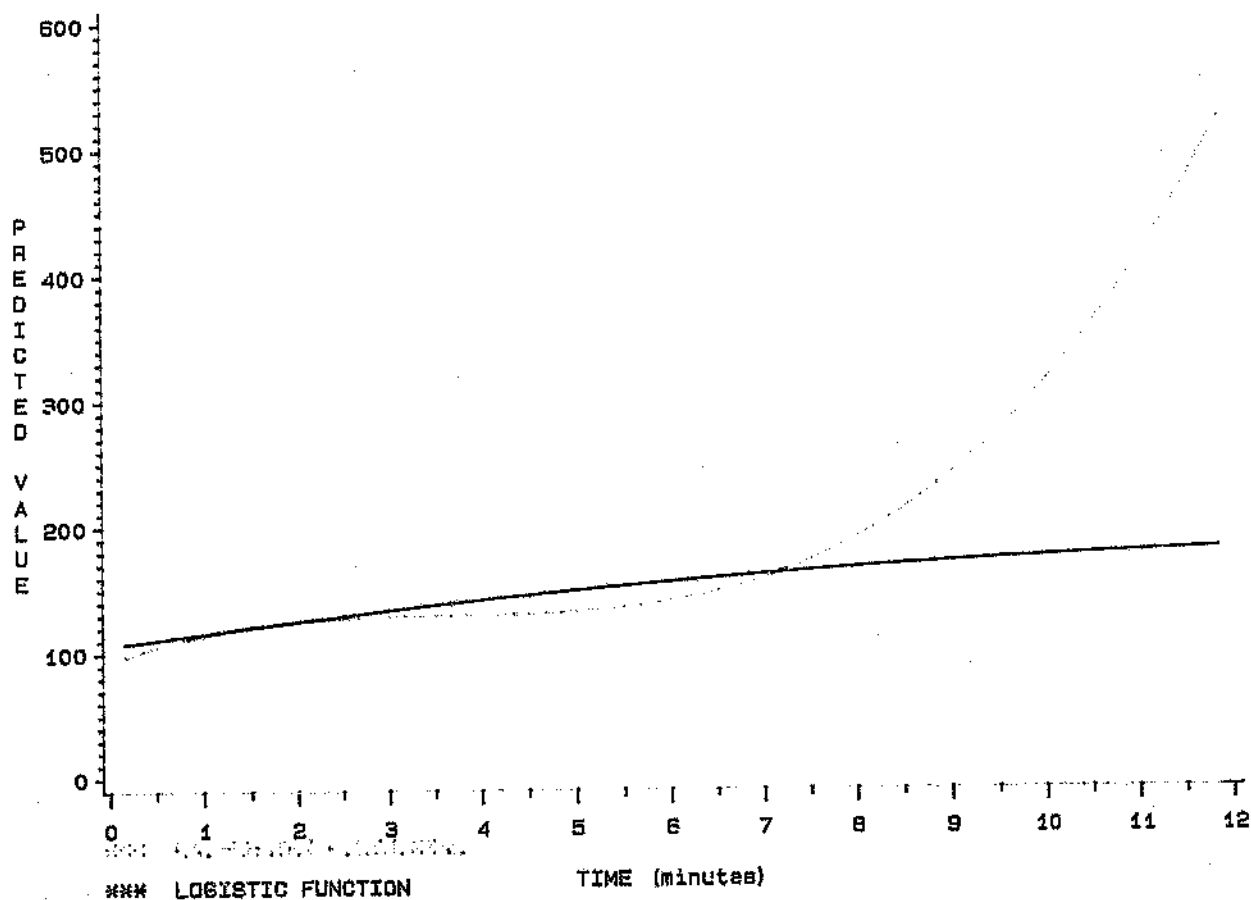


Figure 5.3

HEARTRATE FUNCTIONS FOR ACTIVE NON-SMOKERS
RAO-KHATRI POLYNOMIAL vs LOGISTIC FUNCTION



CONCLUSION

The question now should be "which approach should be used to analyze cardio-pulmonary responses?" The Rao-Khatrı polynomial has the obvious shortcoming that the parameters do not have any physiological meaning. The polynomial also provided unacceptable values, the longer the experiment endured. Richards' logistic function on the other hand does provide a physical meaning for each of the parameters. From a physiologist's point of view, the Richards' growth curves provided more acceptable results. The calculation of the asymptote for each variable, along with its confidence interval, could be useful in creating norms for similar exercise tests. The main advantage of using the Rao-Khatrı model is that one can apply a standard multivariate analysis of covariance program to the data.

The design of the experiment did however have shortcomings. The reason for this was that the data from the Mor Institute in Israel were not collected exclusively for the purposes of this dissertation. Nevertheless, some of the design aspects were bad.

The variable work load was highly correlated with time which made it impossible to distinguish between the effects of work load and time on the measurements. The observed effect is clearly a composite of the two effects. The various age groups were subjected to different treadmill speeds and thus different conditions and thus the effect of age could not be measured again, what is observed is a composite of the age and workload effects. The subjects volunteered and were not scientifically selected.

Factors such as the amount of cigarettes smoked by a 'smoker' each day, or the intensity of exercise an 'active' subject underwent each week, were not examined.

However with the correct experimental design, Richards' logistic function could possibly provide meaningful results in analyzing cardio-pulmonary responses to a graded exercise test.

APPENDIX A

OUTPUT OF SCURVE PROGRAM

GIVEN N OBSERVATIONS $Y(T)$, $T = 1, \dots, N$ MADE AT TIMEPOINTS T, THEN
 ANY OF THE FOLLOWING 3 CURVES CAN BE FITTED

- 1) MONOMOLECULAR: $Y(T) = A * (1 + S * B * (R ** T))$
- 2) LOGISTIC CURVE: $Y(T) = A / (1 + S * B * (R ** T))$
- 3) GOMPERTZ CURVE: $Y(T) = A * \text{EXP}(-B * (R ** T))$

REMARK: S IS =1 OR -1 AND THE VALUE IS CALCULATED BY SCURVE

IF YOU REQUIRE THE MONOMOLECULAR FUNCTION, TYPE '1'

IF YOU REQUIRE THE LOGISTIC FUNCTION, TYPE '2' AND

IF YOU REQUIRE THE GOMPERTZ FUNCTION, TYPE '3'

?

1

ENTER THE AMOUNT OF OBSERVATIONS (N) IN

?

7

IF YOUR DATA IS ALREADY IN "CURVE DATA", ENTER "S" ELSE THE
 NUMBER "4"

?

4

ENTER FIRST THE N OBSERVATIONS AND THEN THE TIME POINTS

?

24.53 28.61 31.26 31.75 33.26 33.28 33.53

?

0 1 2 3 4 5 6

- MONOMOLECULAR

- THE CALCULATED F-STATISTIC FOR TESTING H_0 :

"THE CORRELATION COEFFICIENT = 0" IS 298.129

THE DEGREES OF FREEDOM = 2 AND 4

IF THE CALCULATE $F > F$ -TABLE VALUE THEN H_0 IS REJECTED AND THE FUNCTION CAN BE REGARDED AS SIGNIFICANT.

- RESIDUAL SUM OF SQUARES = 0.43197

- THE VALUE OF THE ESTIMATED PARAMETERS IS THE FOLLOWING

"A" = 33.8089

THE SIGN "S" OF THE PARAMETER "B" = -1

"B" = 0.27452

"R" = 0.55319

- THE VARIANCE OF THE OBSERVATIONS IS 10.80385

- THE CORRELATION COEFFICIENT IS 0.990

- 95% C.I. FOR PARAMETER "A" 26.0935 43.8057

- 95% C.I. FOR PARAMETER "B" 0.0813 0.6180

- 95% C.I. FOR PARAMETER "R" 0.0173 0.9886

IF YOU REQUIRE A 95% C.I. FOR A SPECIFIC Y-VALUE TYPE "1" ELSE "0"

?

1

TYPE THE T-VALUE FOR A 95% C.I. FOR $\hat{Y}(T)$

?

7

-THE 95% C.I. IS: 10.49390 57.88417

- THIS IS THE 1ST DATASET ANALYZED. IF YOU REQUIRE ANOTHER
DATASET TYPE "1" ELSE "0"

?

0

ESTIMATION OF THE LOGISTIC FUNCTION

APPENDIX B

```
/* INSTRUCTIONS FOR GRAPHICS PLOT */
```

```
GOPTIONS HANDSHAKE=NONE AUTOFEED NOPROMPT;
GOPTIONS GACCESS=GSASFILE ;
CMS FILEDEF GSASFILE DISK STATA44A HPGL O;
GOPTIONS DEVICE=HP7530A PENMOUNTS=4;
GOPTIONS COLORS=(BLACK,RED,GREEN,BLUE);
*GOPTIONS DEVICE=IBM3279;
```

```
/* ARRANGEMENT OF THE DATA FILE */
```

```
DATA NEW;
  SET MERGE2.DATA;
  IF AGE <= 35 AND AGE > 10 THEN AGE = 1;
  IF AGE > 35 THEN AGE = 0;
  OBJ1 = 'AC SM';
  OBJ2 = 'SED SM';
  OBJ3 = 'SM';
  OBJ4 = 'AC NSM';
  OBJ5 = 'SED NSM';
  OBJ6 = 'NSM';
  T1 = SUBSTR(OBJECTV,1,6);
  T2 = SUBSTR(OBJECTV,1,3);
  T3 = SUBSTR(OBJECTV,1,7);
  IF T1 = OBJ1 OR T1 = OBJ2 OR T2 = OBJ3 THEN SMOKE = 1;
  IF T1 = OBJ4 OR T2 = OBJ6 OR T3 = OBJ5 THEN SMOKE = 0;
  IF T1 = OBJ1 OR T1 = OBJ4 THEN LIFESTL = 1;
  IF T1 = OBJ2 OR T3 = OBJ5 THEN LIFESTL = 0;
  IF N = 1 THEN OLDNUM = NUMBER;
  IF NUMBER = OLDNUM THEN DO;
    COUNT = COUNT + 1;
    IF COUNT = 6 THEN DO;
      WL = WL + 2;
      COUNT = 0;
    END;
  END;
  ELSE DO;
    COUNT = 0;
    WL = 0;
  END;
  OLDNUM = NUMBER;
  RETAIN WL 0 COUNT -1 OLDNUM;
  KEEP HR WEIGHT TIME WL SMOKE NUMBER LIFESTL AGE;
PROC SORT; BY TIME;
```

```
/* USING PROCEDURE NLIN TO CALCULATE PARAMETER ESTIMATES
FOR THE LOGISTIC MODEL */
```

```
PROC NLIN DATA = NEW METHOD = GAUSS;
  PARMs A=1.62 B0=0 B1=0 B2=0 B3=0 B4=0;
  BOUNDS 170<=A<=250;
  MJ= EXP(B0 + B1*TIME + B2*SMOKE +
  B3*LIFESTL + B4*WEIGHT);
```

```
/* SPECIFICATION OF THE MODEL BEING PLOTTED */
```

MODEL HR = A/(1 + MJ);

/* PARTIAL DERIVATIVE OF THE EXPRESSION IN THE MODEL STATEMENT
WITH RESPECT TO THE PARAMETER BEING ESTIMATED */

DER.B0 = (A*-1*MJ)/(1 + MJ)**2;
DER.B1 = (A*-1*MJ*TIME)/(1+MJ)**2;
DER.B2 = (A*-1*MJ*SMOKE)/(1+MJ)**2;
DER.B3 = (A*-1*MJ*LIFESTL)/(1+MJ)**2;
DER.B4 = (a*-1*mj*weight)/(1+mj)**2;
DER.A = 1/(1+MJ);
OUTPUT PREDICTED = BN1 RESIDUAL = BN2;
/* */

/* SPECIFICATION OF THE OUTLAY OF THE GRAPHICS */

PROC GPLOT;
TITLE1 F=XSWISS C=RED J=C H=.95 'LOGISTIC RESIDUAL PLOT FOR HEARTRATE';
TITLE2 F=XSWISS C=RED J=C H=.95 'HR=A/(1 + EXP(B0 + B1*TIME'
' + B2*SMOKE + B3*LIFESTL + B4*WEIGHT)';
SYMBOL1 I=NONE C=BLUE V=STAR;
SYMBOL3 I=NONE C=BLUE V=STAR;
PLOT BN2*TIME=3
/ OVERLAY ;

TESTING THE ADEQUACY OF THE RAO-KHATRI POLYNOMIAL APPENDIX C

```

*/      Arrangement of the data file      /*

DATA NEW;
  SET MERGE2.DATA;
  IF NUMBER = '0126680' OR NUMBER = '2443677' OR NUMBER = '0338930'
  OR NUMBER = '5071815' THEN DELETE;
  if time > 4.5 then delete;
  IF AGE <= 35 AND AGE > 10 THEN AGE = 1;
  IF AGE > 35 THEN AGE = 0;
  OBJ1 = 'AC SM';
  OBJ2 = 'SED SM';
  OBJ3 = 'SM';
  OBJ4 = 'AC NSM';
  OBJ5 = 'SED NSM';
  OBJ6 = 'NSM';
  T1 = SUBSTR(OBJECTV,1,6);
  T2 = SUBSTR(OBJECTV,1,3);
  T3 = SUBSTR(OBJECTV,1,7);
  IF T1 = OBJ1 OR T1 = OBJ2 OR T2 = OBJ3 THEN SMOKE = 1;
  IF T1 = OBJ4 OR T2 = OBJ6 OR T3 = OBJ5 THEN SMOKE = 0;
  IF T1 = OBJ1 OR T1 = OBJ4 THEN LIFESTL = 1;
  IF T1 = OBJ2 OR T3 = OBJ5 THEN LIFESTL = 0;
  IF _N_ = 1 THEN OLDNUM = NUMBER;
  IF NUMBER = OLDNUM THEN DO;
    COUNT = COUNT + 1;
    IF COUNT = 6 THEN DO;
      WL = WL + 2;
      COUNT = 0;
    END;
  END;
  ELSE DO;
    COUNT = 0;
    WL = 0;
  END;
  OLDNUM = NUMBER;
  RETAIN WL 0 COUNT -1 OLDNUM;
  KEEP HR vs vo2_kg WEIGHT TIME WL SMOKE NUMBER LIFESTL ;
  proc sort ;
  by number smoke lifestl;
  proc transpose data=new out=yd;
  var vo2_kg;
  by number smoke lifestl;
  proc sort;
  by smoke lifestl;
  data mj;
  set yd;

  /*      Remove missing observations      /*

  array h 45 col1 - col45;
  do i = 1 to 45 ;
    if h i = . then delete;
  end;

  /*      Design matrix A      /*

```



```

*/   Third degree  orthogonal polynomial coefficients.  /*
*/   Matrix P                                           /*

```

```

data kd;
input k1 - k4;
cards;
1 -22 946 -3311
1 -21 817 -2408
1 -20 694 -1610
1 -19 577 -912
1 -18 466 -309
1 -17 361 204
1 -16 262 632
1 -15 169 980
1 -14 82 1253
1 -13 1 1456
1 -12 -74 1594
1 -11 -143 1672
1 -10 -206 1695
1 -9 -263 1668
1 -8 -314 1596
1 -7 -359 1484
1 -6 -398 1337
1 -5 -431 1160
1 -4 -458 958
1 -3 -479 736
1 -2 -494 499
1 -1 -503 252
1 0 -506 0
1 1 -503 -252
1 2 -494 -499
1 3 -479 -736
1 4 -458 -958
1 5 -431 -1160
1 6 -398 -1337
1 7 -359 -1484
1 8 -314 -1596
1 9 -263 -1668
1 10 -206 -1695
1 11 -143 -1672
1 12 -74 -1594
1 13 1 -1456
1 14 82 -1253
1 15 169 -980
1 16 262 -632
1 17 361 -204
1 18 466 309
1 19 577 912
1 20 694 1610
1 21 817 2408
1 22 946 3311
;
*/   Orthogonal polynomial coefficients corresponding to the fourth, /*
*/   fifth and sixth degree.                                         /*
*/   Matrix G2                                                         /*

```

```

data cd;
input c1 - c3;
cards;

```

```

19393 -38786 504218
10578 -12341 22919
3608 5494 -234520
-1722 16359 -332059
-5607 21714 -321958
-8232 22848 -246064
-9772 20888 -137032
-10392 16808 -19480
-10247 11438 88922
-9482 5473 176429
-8232 -518 236264
-6622 -6083 265727
-4767 -10878 265370
-2772 -14658 238238
-732 -17268 189176
1268 -18634 124202
3153 -18754 49946
4858 -17689 -26845
6328 -15554 -99736
7518 -12509 -162967
8393 -8750 -211750
8928 -4500 -242500
9108 0 -253000
8928 4500 -242500
8393 8750 -211750
7518 12509 -162967
6328 15554 -99736
4858 17689 -26845
3153 18754 49946
1268 18634 124202
-732 17268 189176
-2772 14658 238238
-4767 10878 265370
-6622 6083 265727
-8232 518 236264
-9482 -5473 176429
-10247 -11438 88922
-10392 -16808 -19480
-9772 -20888 -137032
-8232 -22848 -246064
-5607 -21714 -321958
-1722 -16359 -332059
3608 -5494 -234520
10578 12341 22919
19393 38786 504218

```

```

;
```

```

/* Using IML procedure to handle matrix operations */

```

```

proc iml;
use cond;
read all var ccl into aal;
use ld;
read all var ldl into ldd;
use sq;
read all var sqi into sqq;
use cb;
read all var cbi into cbb;
use fr;

```

```

read all var fr1 into frr;
use m1;
read all var col1 col2 col3 col4 col5 col6 col7 col8 col9 col10
col11 col12 col13 col14 col15 col16 col17 col18 col19 col20
col21 col22 col23 col24 col25 col26 col27 col28 col29 col30
col31 col32 col33 col34 col35 col36 col37 col38 col39 col40
col41 col42 col43 col44 col45 into y;
use xd;
read all var x1 x2 x3 x4 into x;
use kd;
read all var k1 k2 k3 k4 into k1;
use cd;
read all var c1 c2 c3 into c2;
cns1 = aa1/55;
a1 = ldd/7590;
a2 = sqq/9203634;
a3 = cbb/92036340;
a4 = frr/2934936620;
c = cns1||a1||a2||a3||c2;

/* Standard error of the sample observations */

s=(y*y-y*x*inv(x*x)*x*y)/51;
print s;
b=inv(x*x)*x*y;
print b;

/* Calculation of the U test statistic */

h=(b*c2)*(x*x)*b*c2;
e=c2*(y*y-y*x*inv(x*x)*x*y)*c2;
v=det(e)/det(e+h);
print v;

/* Calculation of the correlation matrix of the transformed data */

r1=(y*c)*(1/55-x*inv(x*x)*x)*y*c/51;
d=diag(r1);
d2=sqrt(d);
v=inv(d2)*r1*inv(d2);
print v;

/* Inclusion of the covariates into the estimation process and */
/* the estimation of the parameters for the four groups */

k=k1||c2;
delt = inv(x*x)*x*y*inv(c)*k*inv(k*inv(s)*k);
print delt;

```

APPENDIX D

/Data file of subject with identity number = 0001506 /

OBS	AGE	WEIGHT	NUMBER	TIME	VE	VO2_KG	HR	SMOKE	LIFESTL
1	52	85	0001506	0.1	31	9.3	88	1	0
2	52	85	0001506	0.2	29	7.3	102	1	0
3	52	85	0001506	0.3	36	10.6	124	1	0
4	52	85	0001506	0.4	36	11.2	129	1	0
5	52	85	0001506	0.5	45	15.0	133	1	0
6	52	85	0001506	0.6	48	14.5	130	1	0
7	52	85	0001506	0.7	51	15.7	133	1	0
8	52	85	0001506	0.8	54	16.6	151	1	0
9	52	85	0001506	0.9	59	18.3	156	1	0
10	52	85	0001506	1.0	62	17.9	143	1	0
11	52	85	0001506	1.1	65	18.4	134	1	0
12	52	85	0001506	1.2	69	18.2	133	1	0
13	52	85	0001506	1.3	70	18.7	133	1	0
14	52	85	0001506	1.4	72	18.4	136	1	0
15	52	85	0001506	1.5	72	19.0	139	1	0
16	52	85	0001506	1.6	63	17.2	133	1	0
17	52	85	0001506	1.7	68	19.2	146	1	0
18	52	85	0001506	1.8	69	18.3	142	1	0
19	52	85	0001506	1.9	74	19.9	131	1	0
20	52	85	0001506	2.0	73	19.0	134	1	0
21	52	85	0001506	2.1	76	18.5	132	1	0
22	52	85	0001506	2.2	77	19.1	138	1	0
23	52	85	0001506	2.3	78	18.7	139	1	0
24	52	85	0001506	2.4	77	19.1	140	1	0
25	52	85	0001506	2.5	74	18.1	138	1	0
26	52	85	0001506	2.6	78	18.8	140	1	0
27	52	85	0001506	2.7	80	19.0	145	1	0
28	52	85	0001506	2.8	77	19.2	146	1	0
29	52	85	0001506	2.9	83	19.4	146	1	0
30	52	85	0001506	3.0	82	19.0	149	1	0
31	52	85	0001506	3.1	88	20.1	144	1	0
32	52	85	0001506	3.2	84	19.5	147	1	0
33	52	85	0001506	3.3	84	19.8	145	1	0
34	52	85	0001506	3.4	82	18.6	145	1	0
35	52	85	0001506	3.5	86	19.3	153	1	0
36	52	85	0001506	3.6	87	19.5	156	1	0

REFERENCES

- Amer, F.A. and Williams, W.T. (1957). Leaf-area growth in *Pelargonium zonale*. *Ann. Botany* (London) 21, 339-342.
- Baksalary, J.K., Corsten, L.C.A. and Kala, R. (1975). Reconciliation of two different views on estimation of growth curve parameters. *Biometrika*, 65, 662-665.
- Bock, R.D., and Thissen, D. (1978) Statistical problems in fitting individual growth curves. Technical Report, University of Chicago.
- Bux, C.E.P., and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and control* (2nd Ed). Holden-Day, San Francisco.
- Brillinger, D.K. (1975). *Time Series : Data Analysis and Theory* Holt, Rinehart and Winston, New York.
- Dobson, A.J. (1983). *Introduction to Statistical modelling* London : Chapman and Hall.
- Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*. J. Wiley and sons, New York (2nd edition):
- Du Toit, S.H.C. (1979). *Analysis of growth curves* Unpublished Phd thesis, University of Pretoria.
- Everitt, B. (1987). *Introduction to optimization methods and their application in statistics*. Chapman and Hall.
- Fearn, T. (1975). A Bayesian approach to growth curves *Biometrika* 62 : 89-100.
- Goldfeldt, S.M. and Quandt, R.E. (1972): *Nonlinear Methods in Econometrics* Amsterdam: North-Holland Publishing Company.
- Gompertz, B. (1925). On the nature of the function expressive of the law of human mortality. *Phil. Trans. Royal. Soc. London* 36, 513-585.
- Gregory, F.G. (1928). Studies in energy relations of plants *Annals of Botany*, 42, 469-507.
- Grizzle, J.E. and Allen, D.M. (1969). Analysis of growth and dose response curves. *Biometrics*, 25, 357-381.
- Jenkins, G.M. and Watts, D.G. (1968). *Spectral Analysis and its applications*. San Francisco: Holden-Day.
- Khatri, C.G. (1966). A note on a MANOVA model applied to problems in growth curves. *Ann. Inst. Stat. Math.* 18, 75-86.
- Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model. *J. Royal Stat. Soc. Series B* 34 1-41.

- Malinvaud, E. (1970): *Statistical Methods of Econometrics*. Amsterdam: North-Holland Publishing Company.
- Medawar, P.B. (1940). Growth, growth energy, and ageing of the chicken's heart *Proc. Roy. Soc. B129*, 332-355.
- Pearl, R. and Reed, L.J. (1920). On the rate of growth of the population of the U.S. since 1790 and its mathematical representation. *Proceedings of the National Academy of Science U.S.* 6, 275-288.
- Pearson, E.S. and Hartley, H.O. (1966). *Biometrika Tables for Statisticians*. Cambridge University Press.
- Potthoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 51, 313-326.
- Rao, C.R. (1965). The theory of least squares when the parameters are stochastic and its application in the analysis of growth curves. *Biometrika*, 52, 447-458.
- Rao, C.R. (1966). Covariance adjustment and related problems in multivariate analysis. In: P. Krishnaiah, ed., *Multivariate Analysis*, Academic Press, New York, 87-103.
- Rao, C.R. (1967). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. *Proc. Fifth. Berkely Symp. Math. Stat. Prob.* 1, 355-372.
- Reed and Holland (1919). The growth rate of an annual plant. *Proc. Natl. Acad. Sci. U.S.* 5, 135-144.
- Richards, F.J. (1969). The quantitative analysis of plant growth. In F.C. Steward (Ed) *Plant Physiology Vol. V.A.* Academic Press, New York.
- Rijven, A. H.G. (1952). In-vitro studies on the embryo of *Capsella bursa-pastoris*. *Acta Botan. Neerl.* 1, 157-200.
- Seber, G.A.F. (1977). *Linear Regression analysis*. Wiley: New York.
- Seber, G.A.F. (1984). *Multivariate observations*. Wiley: New York.
- Steyn, A.G.W., Smit, C.F. and Du Toit, S.H.C. (1981) *Moderne Statistiek vir die Praktyk*. J.L. & Schaik: Pretoria.
- Timm, N.H. (1975). *Multivariate Analysis with Applications in Education and Psychology*. Brooks-Cole: Monterey California.

Author: Paiker Mark Joel.

Name of thesis: Growth Curve Analysis For Cardio-pulmonary Responses To Physical Exercise.

PUBLISHER:

University of the Witwatersrand, Johannesburg

©2015

LEGALNOTICES:

Copyright Notice: All materials on the University of the Witwatersrand, Johannesburg Library website are protected by South African copyright law and may not be distributed, transmitted, displayed or otherwise published in any format, without the prior written permission of the copyright owner.

Disclaimer and Terms of Use: Provided that you maintain all copyright and other notices contained therein, you may download material (one machine readable copy and one print copy per page) for your personal and/or educational non-commercial use only.

The University of the Witwatersrand, Johannesburg, is not responsible for any errors or omissions and excludes any and all liability for any errors in or omissions from the information on the Library website.