

APPLICATION OF REGRESSION ANALYSIS AND THE ANALYSIS
OF VARIANCE TO FOREST DATA.

BY

GODWELL HENRY MWAMBI

This dissertation is submitted in partial fulfilment
for the degree of Master of Science in Mathematical
Statistics in the Department of Mathematics.

UNIVERSITY OF NAIROBI

JUNE, 1989.


University of NAIROBI Library




0420828 6

DECLARATION

This dissertation is my own work and has not been presented for a degree in any other University.

Signature: 
GODWELL HENRY MWAMBI

This dissertation has been submitted for examination with my approval as University Supervisor.

Signature: 
DR. J.W. ODHIAMBO

CONTENTS

	<u>PAGE</u>
Title	i
Declaration	ii
List of contents	iii
Summary of contents	vi
Acknowledgement	viii
 <u>CHAPTER I: INTRODUCTION</u>	
1.1 General introduction on forest management	1
1.2 Literature review	3
1.3 Statement of the problem	7
1.4 Significance of the study	8
 <u>CHAPTER II: REGRESSION ANALYSIS</u>	
2.1 Introduction	9
2.2 The linear regression model	10
 <u>CHAPTER III: ANALYSIS OF VARIANCE</u>	
3.1 Introduction	32
3.2 Partition of variance	32
 <u>CHAPTER IV: APPLICATION TO FOREST DATA</u>	
4.1 Introduction	41
4.2 Brief description on the experimental procedure	41
4.3 Regression analysis results	43
4.4 Analysis of variance	49
 <u>CHAPTER V: DISCUSSION AND CONCLUDING REMARKS</u>	
5.1 Regression analysis	64
5.2 Analysis of variance	66

	<u>LIST OF TABLES</u>	<u>PAGE</u>
Table I:	ANOVA table showing the entries in each column and the components of total variance.	38
Table 2 and 3:	Indicator variables for the pine and cypress species respectively. A given row shows the values of the indicator variables when an observation is from the district at the end of the row.	44
Table 4:	Analysis of variance table showing the partition of the total sum of squares into regression and residual sum of squares for the pine species.	46
Table 5:	The table gives the analysis of variance results for the regression model for the cypress species. The coefficient of determination together with the standard error are given.	48
Table 6:	ANOVA table for the Pine species giving the components of the total sum of squares and F-statistic.	50
Table 7:	Means arranged in descending order of magnitude for the pine species.	52
Table 8 and 9:	Table 8 shows pairs of districts with the corresponding calculated student t statistic and the attained significance levels while table 9 shows those pairs of districts with a significant difference in mean strength for the pine species.	53-55

	<u>PAGE</u>	
Table 10:	ANOVA table for the cypress species. It shows the components of the total sum of squares and finally the F- statistic is obtained.	56
Table 11:	Means arranged in descending order of magnitude for the cypress species.	57
Table 12 and 13:	Table 12 shows district pairs with the corresponding calculated student t statistic and the attained significance levels while table 13 shows the pairs of districts with a significance difference in mean strength for the cypress species.	59-63
LIST OF REFERENCES:		68-69

SUMMARY OF CONTENTS

Regression analysis and the analysis of variance are some of the most widely used statistical methods in biometrics. In this project, we use regression analysis to study the relationship between modulus of rupture of timber with density taking into account of site effects. This shall be done for two species of wood, namely the pine and cypress. The analysis of variance is used to study the variation of strength of these two species of wood with site.

Chapter I section 1.1, gives a general introduction on forest management. Section 1.2 gives previous studies done on forest management and techniques used. The statement and significance of the problem are contained in sections 1.3 and 1.4 respectively. We make use of sample regression methods because population parameters are not known. These techniques are under chapter II. Section 2.1 introduces the general linear regression, while section 2.2 describes the linear regression model and some of the major applicable results. In chapter III the Analysis of Variance is reviewed and the major results to be applied displayed.

Chapter IV gives the methodology on the applications of the above techniques to the data of the strength of the two species of wood, the Pine and Cypress grown in various districts in Kenya. Indicator variables to define the district levels are tabulated in tables 2 and 3 for the two species respectively. Results on the application of the two statistical techniques are tabulated in tables

4, 5, 6, 8, 9, 10, 12 and 13.

Finally Chapter V deals with the discussion and concluding remarks based on the results in chapter IV. In brief we discuss how the findings can be beneficial in efficient utilization of timber from the two species and in forest management in general. We also give recommendations for future studies in the area.

ACKNOWLEDGEMENTS

I wish to express my gratitude and appreciation to Dr. J.W. Odhiambo for his guidance and support during the course of this study and for all his advice during the preparation of this dissertation.

I would also like to acknowledge the Kenya Forest Research Institute through Mr. B. Chikamai, for providing me with data and other requirements for the study.

I convey my gratitude to the University of Nairobi and DAAD - the German Academic Exchange Service, for providing me with financial support for my postgraduate study at the Faculty of Science, University of Nairobi.

Finally I am very grateful to my parents Getrude and Mwaghania Mwambi and other members of our family for their wonderful show of love and understanding especially during the periods of study.

CHAPTER 1
INTRODUCTION

1.1 General Introduction on Forest Management

A forest Manager dealing with a renewable forest resource is usually concerned with the optimal harvesting strategy, when trees are classified by age structure or size structure. This would avoid unplanned depletion of the resource. Unfortunately data to carry out such a study is not available from the forest inventory of Kenya.

On the otherhand a forest utilization officer is interested in knowing characteristics of wood such as the modulus of rapture, modulus of elasticity, stress at limit of propotionality and many more. These properties of mature wood could help such an officer to calculate design stresses in order to set quality standards of the material concerned.

In this project we shall deal with two types of soft-woods grown in Kenyan forests for commercial use, these being the pine and cypress species. It is from these two species that some of the timber used in construction work is obtained. Each of these two species is grown in a given number of districts. The main property of interest under study for these two types of wood is their strengths. In particular the modulus of rapture which contributes greatly on deciding the strength of wood shall be studied. In the past a sample of size n_i was collected from district i , $i = 1, 2, \dots, v$ where v is the number of districts, then conclusions about design stresses made from

a pooled down sample. Modulus of rapture is the maximum load the experimental material can be able to support before it fails or breaks. Knowledge of material strength is a fundamental necessity in all structural designs. In Kenya strength values of the locally available wood are not reliably known.

Variation of strength of wood with site is one aspect which has not been given attention in the past. In this project we shall develop a one way analysis of variance model to study this. We shall take the various sites (districts) as the treatments. A general linear regression model of modulus of rapture on density shall be developed taking into account of site. This shall necessitate the use of indicator variables. The study shall be based on data obtained from the FOREST PRODUCTS RESEARCH PROGRAMME under KENYA FOREST RESEARCH INSTITUTE.

1.2 LITERATURE REVIEW

Forest management is an area in forestry where a lot has been done. Scientists have done studies on both management methods and properties of timber itself.

Biolley (1920, 1954) deals with a method of management of renewable resources which is called the check method. His system of management aims at producing as much timber as possible, consistent with the constraints of quality and conservation.

Selection forests as an example of renewable resource was first conceived by Gurnaud in the nineteenth century.

Colette (1934, 1960) considered methods of selection working and the exploitation of the stand was based on the results of periodic enumerations where records by species and circumference classes is a pre-requisite. Colette uses this information to calculate approximate probabilities of transition from one circumference class to the class above and figures used to calculate the exploitation. The stem-number curve forms a graphical check on the stand. The curve is compared with a theoretical smooth curve in which the number of trees in each successive class is represented by a decreasing geometric progression. Successive terms in this progression are related by the coefficient of "diminuation"

Another notable contributions towards the problem of forest management have been made by Usher. For example, Usher (1966) describes a method of calculating the

coefficient of diminuation. Taking into account that a manager of a selection forest knows the individual recruitment probabilities in each class to classes above, a theoretical structure can be determined which can be determined for any set of management objectives. The structure according to Usher is unique and it optimizes the yield from the resource over a long period of time. With the recruitment data available Usher came up with a matrix M that links the number in various size groups at time t to that at time $t+1$, and the relation is

$$Mn_t = n_{t+1}$$

From this a stable structure can be predicted. This structure is associated with a dominant latent root of the matrix, which is greater than unity. Associated with this latent root is a latent vector such that all its elements can be chosen positive. However the model did not tell whether the solution is unique or not.

In another paper Usher (1969) develops a model for the management of renewable resource. The mathematical development in this paper shows that there is only one solution of the model that is biologically meaningful. The solution is associated with the only latent root of matrix M which is greater than unity. This latent root has a latent vector such that all its components can be chosen as positive. He tested the data for a Scots pine forest.

Patterson (1971), in his study of the Kenyan wood, found that Kenya's timber strength may strongly be

related to

- (a) density and to some species to
- (b) moisture content and age.

He found that heavy timber are strong and light ones are weak. If S is to represent strength and W weight then

$$S = f(W)$$

which is an increasing function of W . This is because heavy timber contain more wood substance per unit volume, than light ones. He conducted an experiment on 46 species grown in Kenya and fitted graphs of strength versus weight. He pointed out that there was considerable variations about the mean curves he obtained when a species was considered on its own. Patterson found out that species like pencil cedar and Australian mountain Ash (*Eucalyptus regnaus*) are strong for their weight and others like Muirungi (*casearia*) and Muchichia (*Premna*) are weak for their weight. The variations are due to the anatomy of the timber. Moisture content causes variation only when it is below saturation point that is below 27%. When trees are still growing then age affects strength but after 25 years of age it is not very significant.

Burges (1962) suggested that the load-deflection relationship derived from certain tests on timber may be interpreted as a skewed normal integral

$$Y = \frac{k_1}{\sqrt{2\pi}} \int_0^x \frac{\exp(x^{\frac{1}{2}} - k_2)^2}{2k_3} dx$$

He investigated methods of fitting such a curve to experimental results and the effect of strain-rate was examined, by using data by Brokaw and Foster (1945). The approach is indicative of the physical representation of the mechanism of strain and failure, unlike the purely heuristic linear approach. A brief indication is given of arguments justifying the phenomenological study of apparently non-stochastic responses in stochastic terms.

Brister (1962), carried out an experiment on Kenyan pine timber where five thirty-year old trees were selected for test. Results showed that density and strength increases moving outward from the pith. Strength tests show a stronger correlation with distance from the pith than age. This paper shows that strength increases with density.

Sunley (1956) carried a research on modulus of rupture for the sitka spruce species obtained from various sources. He established the fact that modulus of rupture is described as the normal curve for a given population of a given species of trees. This idea supports the current work where it is assumed that modulus of rupture y is given by

$$Y = \mu + d + e \sim N(\mu + d, \sigma^2)$$
$$E(e) = 0 \quad \text{and} \quad E(e^2) = \sigma^2.$$

Where d is the effect of the district from which Y is observed.

1.3 STATEMENT OF THE PROBLEM

If we let y_{ij} to denote the j th observation (modulus of rapture) from district D_i ($i=1,2, \dots, v$) $j = 1, 2, \dots, n_i$. Then the problem is divided into two parts.

First using this information it is proposed to carry out a one way analysis of variance to see if there is a marked variation of modulus of rapture with site. From this we shall be able to say something about the variation of strength of wood in Kenya with site.

The second part of the problem is to fit a linear regression model with the response variable being the modulus of rapture (MOR) while the explanatory variable is the density. We shall make use of indicator (dummy) variables to take care of the site levels, which are qualitative variables. For a given district the model is of the form

$$Y(\text{MOR}) = \beta_0 + \beta_1 x + e$$

but with the inclusion of indicator variables it becomes a multiple regression model

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

The total number of observations is

$$N = \sum_{i=1}^v n_i$$

The results shall help to tell whether the relationship is appropriate or not. If not then we shall suggest reasons for that.

1.4 SIGNIFICANCE OF THE STUDY

The analysis of variance results shall help to tell whether in future design stresses for constructional purposes should be done district by district or not. This will help in the efficient utilization of the material. The Kenya Bureau of Standards may use the results in setting standards of the Kenyan timber.

From the regression model we shall be able to decide whether much of the variability in the modulus of rupture is explained by density or not. This is by making use of the sample coefficient of determination, r^2 .

If not much variability is explained then we shall propose that other independent variables be included in the model in a future study. Then a better predictive regression model can be developed. The study is important in the sense that it can serve as a starting point for future studies. The results of the study shall indicate whether a better method of collecting the data should be adopted or not.

Finally the work of the study may prove valuable in that, it will provide reading and reference material for research scientist both in Physical and Biological Sciences and Social Sciences. Biostatisticians may find the work very useful in their research activities.

CHAPTER II
REGRESSION ANALYSIS

2.1 INTRODUCTION

Regression analysis is a statistical method which deals with the study of the relationship between measurable variables. In regression analysis we usually deal with two types of variables namely the response variable(s) also called the dependent variable(s); and the explanatory variables also known as independent variables.

Let Y denote the response variable and let x_1, x_2, \dots, x_p denote the explanatory variables. Then a usual assumption in regression analysis is that observations on response variables are subject to error but observations on the explanatory variables are made without error. Assuming a functional relationship between the response variable Y and the explanatory variables x_1, x_2, \dots, x_p .

We can write

$$Y = f(x_1, x_2, \dots, x_p) \quad (2.1)$$

where f is some function. This is called the regression function of Y on x_1, x_2, \dots, x_p . If f is a linear function then we say the regression is linear. If f is a non-linear function then the regression is non linear.

One of the main reasons for fitting regression models to observed data is to describe the relationship between the response and explanatory variables and to predict the values of the response variable. Regression analysis has been applied in varied fields. These include social and economic sciences, physical and biological sciences, technological applications and many others.

2.2

THE LINEAR REGRESSION MODEL

Suppose we have a population of individuals each of which has $p+1$ characteristics, say

$$Y, x_1, \dots, x_p$$

For example with human beings we might have height (Y), weight (x_1) and girth (x_2). The whole population may be thought of as forming a cluster in a $p+1$ dimensional space \mathbb{R}^{p+1} . Frequently we are interested in questions of the type (i) how much of the variation in Y can be attributed to the variation in x_1, x_2, \dots, x_p

OR (ii) what can we say about Y for an individual given that it has specified values for x_1, x_2, \dots, x_p .

Looked at in this way we consider what function of x_1, x_2, \dots, x_p should be used to predict Y and what the error of such prediction will be. We shall confine ourselves to linear functions of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (2.2)$$

This is called a linear predictor for the characteristic Y . Polynomial predictors are included as a special case since we could have x_2 equals x_1^2 and so on.

The constants $\beta_1, \beta_2, \dots, \beta_p$ are called regression coefficients. In particular β_j ($j=1, 2, \dots, p$) is the coefficient of partial regression of Y on X_j . It measures the rate of change of Y w.r.t. x_j when the other variables are fixed.

It is reasonable to define the best predictor of Y from x_1, x_2, \dots, x_p as that linear function

$$f(x_1, x_2, \dots, x_p) = \beta_1 x_1 + \dots + \beta_p x_p$$

for which the average value of

$$\left[Y - f(x_1, x_2, \dots, x_p) \right]^2 \tag{2.3}$$

is minimum. This is the predictor which will result in minimum mean square error of prediction. It is called the least squares predictor.

To obtain the values of $\beta_0, \beta_1, \dots, \beta_p$ we have to minimize

$$Q = E(Y - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p)^2 \tag{2.4}$$

Setting the derivative of Q w.r.t β_0 equal to zero and writing $E(x_i) = \mu_i$ and $E(Y) = \mu_y$ we have

$$\mu_y - \beta_0 - \beta_1 \mu_1 - \beta_2 \mu_2 - \dots - \beta_p \mu_p = 0$$

that is

$$\beta_0 = \mu_y - \beta_1 \mu_1 - \beta_2 \mu_2 - \dots - \beta_p \mu_p$$

if we substitute this value of β_0 into Q we obtain

$$E \left[(Y - \mu_y) - \beta_1 (x_1 - \mu_1) - \beta_2 (x_2 - \mu_2) - \dots - \beta_p (x_p - \mu_p) \right]^2 \tag{2.5}$$

Differentiating equation (2.5) w.r.t. $\beta_1, \beta_2, \dots, \beta_p$ and equating each derivative to zero we obtain P equations.

These are

$$\left. \begin{aligned} \sigma_{1y} &= \beta_1 \sigma_1^2 + \beta_2 \sigma_{12} + \dots + \beta_p \sigma_{1p} \\ \sigma_{2y} &= \beta_1 \sigma_{12} + \beta_2 \sigma_2^2 + \dots + \beta_p \sigma_{2p} \\ &\vdots \\ \sigma_{py} &= \beta_1 \sigma_{1p} + \beta_2 \sigma_{2p} + \dots + \beta_p \sigma_p^2 \end{aligned} \right\} \quad (2.6)$$

where

$$\begin{aligned} \sigma_{iy} &= E \left[(x_i - \mu_i) (Y - \mu_y) \right] & i = 1, 2, \dots, p \\ \sigma_{ij} &= E \left[(x_i - \mu_i) (x_j - \mu_j) \right] & i \neq j = 1, 2, \dots, p \\ \sigma_i^2 &= \text{var}(x_i) & \text{and} \quad \sigma_y^2 = \text{var}(Y) \end{aligned}$$

To obtain the values of the regression coefficients $\beta_1, \beta_2, \dots, \beta_p$ We solve the system of linear equations (2.6) simultaneously. For known values of σ_{ij} 's, σ_{iy} 's and σ_i^2 's these are the population regression coefficients.

Multiple Correlation:

Suppose that Y, x_1, \dots, x_p are jointly normally distributed then the conditional mean of Y given x_1, \dots, x_p will be a linear function of the form

$$E(Y/x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

which can be written as

$$E(Y/x_1, \dots, x_p) = \mu_y + \beta_1 (x_1 - \mu_1) + \dots + \beta_p (x_p - \mu_p) \quad (2.7)$$

Then a measure of the linear relationship between Y and x_1, x_2, \dots, x_p is given by the multiple correlation coefficient. This can be calculated as the ordinary correlation

between

$E(Y/x_1, \dots, x_p)$ and Y . Let

$$\Sigma = \begin{bmatrix} \sigma_{yy} & \sigma_{1y} & \dots & \sigma_{py} \\ \sigma_{1y} & \sigma_{11} & \dots & \sigma_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{py} & \sigma_{1p} & \dots & \sigma_{pp} \end{bmatrix} = \left[\begin{array}{c|c} \sigma_{yy} & \sigma_{y2}' \\ \hline \sigma_{y2} & \Sigma_{22} \end{array} \right]$$

where

$$\sigma_{y2} = \begin{bmatrix} \sigma_{1y} \\ \sigma_{2y} \\ \vdots \\ \sigma_{py} \end{bmatrix} \quad \Sigma_{22} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma_{pp} \end{bmatrix}$$

Then it is easily shown that

$$R_{y.123\dots p}^2 = \frac{\sigma_{y2}' \Sigma_{22}^{-1} \sigma_{y2}}{\sigma_{yy}} \quad (2.8)$$

and

$$\text{Var}(Y/x_1, \dots, x_p) = \sigma_{yy} (1 - R_{y.123\dots p}^2) \quad (2.9)$$

The quantity $\sigma_{yy} (1 - R_{y.12\dots p}^2)$ is called the residual variance of Y when the effects of x_1, x_2, \dots, x_p are eliminated.

It is easily seen that the residual variance will be zero if and only if

$$R_{y.12\dots p}^2 = 1$$

This means that the linear relationship between Y and x_1, \dots, x_p will be perfect if

$$R_{y.12\dots p}^2 = 1$$

If $R_{Y.12\dots p}^2 = 0$ then there is no linear relationship between Y and x_1, \dots, x_p . Thus $R_{Y.12\dots p}^2$ can be used as a measure of the degree of the linear relationship between Y and x_1, x_2, \dots, x_p .

SAMPLE LINEAR REGRESSION:-

In a practical situation we do not deal with the whole population but instead a sample is collected and the corresponding response and explanatory variables determined for each sample point. Suppose $n(>p)$ observations are available, and let y_i denote the i th observation on the response variable and x_{ij} denote the i th level of the j th explanatory variable x_j . Assuming that the observed responses are subject to experimental errors and the explanatory variables have fixed levels, we can write the linear model as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + e_i \\ &= \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + e_i \quad (i=1, 2, \dots, n) \quad (2.10) \end{aligned}$$

We assume $E(\underline{\xi}) = \underline{0}$ and $\text{Var}(\underline{\xi}) = \sigma^2 I$, where $\underline{\xi}$ is shown below:-

In matrix notation we may write

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

or in an obvious compact notation

$$\underline{Y} = X \underline{\beta} + \underline{\xi} \quad (2.11)$$

where

\underline{Y} is $n \times 1$, X is $n \times (p+1)$, $\underline{\beta}$ is $(p+1) \times 1$ and $\underline{\xi}$ is $n \times 1$

To obtain the sample estimate of the regression vector $\underline{\beta}$, we use the least squares method. That is we obtain $\underline{\beta}$ which minimizes

$$\begin{aligned} S(\beta) &= \sum_{i=1}^n e_i^2 = \underline{\xi}' \underline{\xi} = (\underline{Y} - X \underline{\beta})' (\underline{Y} - X \underline{\beta}) \\ &= \underline{Y}' \underline{Y} - \underline{\beta}' X' \underline{Y} - \underline{Y}' X \underline{\beta} + \underline{\beta}' X' X \underline{\beta} \\ &= \underline{Y}' \underline{Y} - 2 \underline{\beta}' X' \underline{Y} + \underline{\beta}' X' X \underline{\beta} \end{aligned}$$

Differentiating $S(\underline{\beta})$ w.r.t $\underline{\beta}$ and equating to zero we obtain $p+1$ normal equations given by

$$X' X \underline{\beta} = X' \underline{Y}$$

which gives the least squares estimate of $\underline{\beta}$ as

$$\hat{\underline{\beta}} = (X' X)^{-1} X' \underline{Y} \quad (2.12)$$

provided that $(X' X)$ is non singular.

Note that $\underline{\beta}$ is unbiased estimator of $\underline{\beta}$ since

$$\begin{aligned} E(\underline{\hat{\beta}}) &= E \left[(X'X)^{-1} X' \underline{Y} \right] \\ &= E \left[(X'X)^{-1} X' (X\underline{\beta} + \underline{\xi}) \right] \\ &= E \left[(X'X)^{-1} X' X \underline{\beta} + (X'X)^{-1} X' \underline{\xi} \right] \\ &= \underline{\beta} \quad \dots\dots(2.13) \end{aligned}$$

and $E(\underline{\xi}) = \underline{0}$.

$$\text{Also } \text{Cov}(\underline{\hat{\beta}}) = \text{Var} \left[(X'X)^{-1} X' \underline{Y} \right] = (X'X)^{-1} \sigma^2 \quad (2.14)$$

The matrix $C = (X'X)^{-1}$ is sometimes called the unscaled covariance matrix. When $(X'X)^{-1}$ does not exist then we use the g-inverse of $X'X$, normally denoted by $(X'X)^*$

Then a least squares estimator of $\underline{\beta}$ is

$$\underline{\hat{\beta}}^* = (X'X)^* X' \underline{Y} \quad (2.15)$$

This estimator is not unique because a g inverse of a matrix is not unique.

Let us denote the residual sum of squares by SS_E ;

$$\begin{aligned} SS_E &= \underline{e}' \underline{e} = (Y - X \underline{\hat{\beta}})' (Y - X \underline{\hat{\beta}}) \\ &= \underline{Y}' \underline{Y} - \underline{\hat{\beta}}' X' \underline{Y} \quad (2.16) \end{aligned}$$

since the residual sum of squares has $n - (p+1)$ degrees of freedom, the residual mean sum of squares is given by

$$MS_E = \frac{SS_E}{n - p - 1} \quad (2.17)$$

NOW

$$\begin{aligned} \underline{y} &= X \hat{\underline{\beta}} \\ &= X (X'X)^{-1} X' \underline{y} \\ &= H \underline{y} \end{aligned}$$

Where $H = X (X'X)^{-1} X'$

Then
$$\begin{aligned} \underline{e} &= \underline{y} - X \hat{\underline{\beta}} \\ &= \underline{y} - H \underline{y} \\ &= (I - H) \underline{y} \end{aligned}$$

Clearly $I-H$ is both symmetric and idempotent, therefore

$$SS_E = \underline{e}'\underline{e} = \underline{y}'(I-H)\underline{y} \quad (2.18)$$

Since

$$\begin{aligned} \text{Var}(\underline{y}) &= \sigma^2 I \\ E(SS_E) &= \text{tr}(I-H) \sigma^2 + [\underline{E}(\underline{y})]' (I-H) [\underline{E}(\underline{y})] \\ &= (n - p - 1) \sigma^2 + \underline{\beta}' X' (I - X(X'X)^{-1} X') X \underline{\beta} \\ &= (n - p - 1) \sigma^2 + 0 \\ &= (n - p - 1) \sigma^2 \end{aligned}$$

Thus

$$E\left(\frac{SS_E}{n - (p+1)}\right) = \sigma^2$$

So an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{SS_E}{n - (p+1)} = MS_E \quad (2.19)$$

Sample Multiple Correlation:-

We shall let $r_{y.12....p}^2$ denote the sample multiple correlation between Y and x's.

Let

$$S = \begin{bmatrix} s_{yy} & s_{1y} & \dots & s_{py} \\ s_{1y} & s_{11} & \dots & s_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{py} & s_{1p} & \dots & s_{pp} \end{bmatrix}$$

where

$$s_{yy} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

$$s_{jy} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_{ij} - \bar{x}_j)}{n}$$

$$s_{jj} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n}$$

$$s_{jj'} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})}{n}$$

Further let

$$s_{y2} = \begin{bmatrix} s_{1y} \\ s_{2y} \\ \vdots \\ s_{py} \end{bmatrix}, \quad S_{22} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{12} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \dots & s_{pp} \end{bmatrix}$$

Then taking S as the maximum likelihood estimator for the population variance-covariance matrix Σ we have

$$r_{y.12\dots p}^2 = \frac{S'_{y2} S_{22}^{-1} S_{-y2}}{S_{yy}} \quad (2.20)$$

The quantity $r_{y.12\dots p}^2$ is called the coefficient of determination and is used to measure the degree of the linear relationship between Y and the x_i 's obtained from the sample.

Testing for Regression

Under the assumption that the error vector $\underline{\xi}$ is normally distributed with mean vector $\underline{0}$ and variance $\sigma^2 I$ that is $\underline{\xi} \sim N(\underline{0}, \sigma^2 I)$ it follows that the observations y_i are normally and independently distributed with mean $\beta_0 + \sum_{j=1}^p \beta_j x_{ij}$ and variance σ^2 . Since the least squares estimator $\underline{\beta}$ is a linear combination of the observations it follows that

$$\underline{\beta} \sim N \left[\underline{\beta}, (X'X)^{-1} \sigma^2 \right] \quad (2.21)$$

Now let us denote $(X'X)^{-1} \sigma^2$ by $\sigma^2 C$.

We next consider the transformation $\underline{\alpha} = (\sigma^2 C)^{-\frac{1}{2}} (\underline{\hat{\beta}} - \underline{\beta})$

Then it follows that

$$\underline{\alpha} \sim N \left[\underline{0}, I \right]$$

and since

$$(\hat{\underline{\beta}} - \underline{\beta}) = (\sigma^2 C)^{\frac{1}{2}} \underline{\alpha}$$

Then

$$\begin{aligned}
Q &= \underline{\alpha}' (\sigma^2 C)^{\frac{1}{2}} (\sigma^2 C)^{-1} (\sigma^2 C)^{\frac{1}{2}} \underline{\alpha} \\
&= \underline{\alpha}' \underline{\alpha} \\
&= \sum_{i=1}^{p+1} \alpha_i^2
\end{aligned}$$

But each $\alpha_i \sim N(0, 1)$ which means that

$$Q \sim \chi^2(p+1)$$

That is

$$Q = \frac{(\hat{\underline{\beta}} - \underline{\beta})' (X'X) (\hat{\underline{\beta}} - \underline{\beta})}{\sigma^2} \sim \chi^2(p+1) \quad (2.22)$$

assuming σ^2 is known.

Now, we wish to test the hypothesis

$$H_0: \underline{\beta} = \underline{0} \text{ against } H_a: \underline{\beta} \neq \underline{0}$$

under H_0 the expression in (2.22) reduces to:

$$Q_0 = \frac{\hat{\underline{\beta}}' (X'X) \hat{\underline{\beta}}}{\sigma^2}$$

Now $E(Q_0) = 0$ when H_0 is true and

> 0 when H_0 is not true

Hence to test the above hypothesis for known σ^2

we compute

$$Q_0 = \frac{\hat{\underline{\beta}}' (X'X) \hat{\underline{\beta}}}{\sigma^2} \quad (2.24)$$

from the sample data and reject H_0 whenever

$$P = \text{prob} (Q > Q_0/H_0) \quad (2.25)$$

is small. Now from (2.18) we have that the sum of squares due to error denoted by

SS_E is given by

$$SS_E = \underline{y}' (I-H) \underline{y}$$

where

$$H = X(X'X)^{-1} X'$$

and

$$E(SS_E) = (n-p-1) \sigma^2$$

$$E\left(\frac{SS_E}{n-p-1}\right) = \sigma^2$$

Now from the model $\underline{y} = X\underline{\beta} + \underline{\xi}$ we deduce that

$$E(\underline{y}) = X \underline{\beta} \text{ since } \underline{\xi} \sim N[\underline{0}, \sigma^2 I]$$

Then it follows that

$$\underline{y} \sim N[X \underline{\beta}, \sigma^2 I]$$

and

$$(\underline{y} - X \underline{\beta}) \sim N[\underline{0}, \sigma^2 I]$$

Now we shall consider the quadratic form

$$(\underline{y} - X\underline{\beta})' (I-H) (\underline{y} - X\underline{\beta}) \tag{2.26}$$

Then we see that

$$\begin{aligned} E\left\{(\underline{y} - X\underline{\beta})' (I-H) (\underline{y} - X\underline{\beta})\right\} &= \text{trace} (I-H) \sigma^2 \\ &= (n-p-1) \sigma^2 \end{aligned}$$

Therefore

$$\frac{(\underline{y} - X\underline{\beta})' (I-H) (\underline{y} - X\underline{\beta})}{n-p-1}$$

is an unbiased estimator of σ^2 (2.27)

But $(\underline{y} - X\underline{\beta})' (I-H) (\underline{y} - X\underline{\beta})$

$$\begin{aligned} &= \underline{y}' (I-H) \underline{y} - \underline{y}' (I-H) X \underline{\beta} - \underline{\beta}' X' (I-H) \underline{y} + \underline{\beta}' X' (I-H) X \underline{\beta} \\ &= \underline{y}' (I-H) \underline{y} \end{aligned}$$

because the last three terms in the expansion above are equal to zero.

Therefore the quadratic form in (2.26) gives the sum of squares due to error, SS_E . In other words

$$\frac{SS_E}{(n-p-1)} \quad (2.28)$$

is an unbiased estimator for σ^2 .

Now

$$(i) \quad H' = \left[X(X'X)^{-1}X' \right]' = X(X'X)^{-1}X'$$

meaning H is symmetric

$$(ii) \quad H^2 = \left(X(X'X)^{-1}X' \right) \left(X(X'X)^{-1}X' \right) = X(X'X)^{-1}X'$$

hence H is idempotent.

From these two properties of H it follows that $I-H$ is also symmetric and idempotent. Then from Cochran's theorem (1934) the quadratic form

$$(\underline{y} - X\underline{\beta})' (I-H) (\underline{y} - X\underline{\beta})$$

has the Wishart distribution with a scale parameter σ^2 and $n-p-1$ degrees of freedom denoted by:

$$w_1(\sigma^2, n-p-1) = \sigma^2 \chi^2_{(n-p-1)}$$

that is

$$\underline{y}' (I-H) \underline{y} \sim \sigma^2 \chi^2_{(n-p-1)}$$

and so

$$\frac{SS_E}{\sigma^2} \sim \chi^2_{(n-p-1)} \quad (2.29)$$

where σ^2 is known.

From (2.22)

$$\frac{(\underline{\beta} - \underline{\beta})' (X'X) (\underline{\beta} - \underline{\beta})}{\sigma^2} \sim \chi^2_{(p+1)}$$

when σ^2 is known.

It then follows that

$$F = \frac{(\hat{\underline{\beta}} - \underline{\beta})' (X' X) (\hat{\underline{\beta}} - \underline{\beta})}{\sigma^2 (p+1)} \sim \chi^2(p+1)$$

$$\frac{SS_E}{(n-p-1) \sigma^2} \quad (2.30)$$

has the Fishers F distribution with $p+1$ degrees of freedom in the numerator and $n-p-1$ degrees of freedom in the denominator.

That is $F \sim F(p+1, n-p-1)$

The expression (2.30) reduces to

$$F = \frac{(\hat{\underline{\beta}} - \underline{\beta})' (X' X) (\hat{\underline{\beta}} - \underline{\beta})}{(p+1) \frac{SS_E}{n-p-1}} \quad (2.31)$$

which is independent of the unknown σ^2 .

The random variable F can be used to test the hypotheses

$$H_0 : \underline{\beta} = \underline{0} \text{ against } H_a : \underline{\beta} \neq \underline{0}$$

Under H_0 that is when $\underline{\beta} = 0$ the expression in (2.31)

reduces to

$$F_0 = \frac{\hat{\underline{\beta}}' (X' X) \hat{\underline{\beta}}}{(p+1) \frac{SS_E}{n-p-1}} = \frac{\hat{\underline{\beta}}' (X' X) \hat{\underline{\beta}}}{(p+1) MSE} \sim F(p+1, n-p-1)$$

In regression analysis the quantity $\hat{\underline{\beta}}' (X' X) \hat{\underline{\beta}}$ is called the sum of squares due to regression denoted by SS_{reg}

Therefore

$$F_0 = \frac{SS_{reg}}{(p+1)MS_E} \sim F(p+1, n-p-1) \quad (2.32)$$

Now $E(F_0/H_0) = 0$ and $E(F_0/H_a) > 0$

Hence to test the above hypothesis we compute

$$F_c = \frac{SS_{reg}}{(p+1)MS_E}$$

from the sample data and reject H_0 if the attained significance level

$$P = \text{prob} (F > F_c / H_0) \quad (2.33)$$

is small.

MARKING PREDICTIONS

Consider the linear model

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

Then a particular variable y is given by

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + e$$

From the earlier assumptions on e it follows

$$E(y) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (2.34)$$

Let

$$\mu = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \dots + \beta_p x_p^*$$

and

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \dots + \hat{\beta}_p x_p^* = \underline{x}^* \hat{\underline{\beta}}$$

where

$$\underline{x}^* = (1, x_1^*, \dots, x_p^*)'$$

is a point in the space of explanatory variables.

We wish to obtain confidence bounds on μ .

Now

$$E(\hat{\mu}) = E(\underline{x}'\hat{\beta}) = \underline{x}'\beta$$

$$\text{var}(\hat{\mu}) = \text{var}(\underline{x}'\hat{\beta}) = \underline{x}' \text{var}(\hat{\beta}) \underline{x} = \sigma^2 \underline{x}'(X'X)^{-1} \underline{x}$$

if we let

$$v^* = \underline{x}'(X'X)^{-1} \underline{x}$$

Then

$$\text{var}(\hat{\mu}) = \sigma^2 v^* \quad (2.35)$$

Since $\hat{\beta}$ is normally distributed it follows that

$$\mu \sim N(\underline{x}'\beta, \sigma^2 v^*)$$

and so

$$Z = \frac{\hat{\mu} - \mu}{\sqrt{\sigma^2 v^*}} \sim N(0,1)$$

since

$$\frac{SS_E}{\sigma^2} \sim \chi^2_{(n-p-1)}$$

Then the random variable

$$T = \frac{\hat{\mu} - \mu}{\sqrt{MS_E v^*}}$$

where $v^* = \underline{x}'(X'X)^{-1} \underline{x}$, has the student t distribution with $n-p-1$ degrees of freedom.

Therefore confidence bounds on μ are given by the probability statement

$$P\left(-\frac{t_\alpha}{2} \leq t \leq \frac{t_\alpha}{2}\right) = 1 - \alpha$$

which implies

$$P\left(-\frac{t_\alpha}{2} \leq \frac{\hat{\mu} - \mu}{\sqrt{MS_E v^*}} \leq \frac{t_\alpha}{2}\right) = 1 - \alpha$$

which gives

$$P\left(\hat{\mu} - \frac{t_\alpha}{2} \hat{\sigma} \sqrt{v^*} \leq \mu \leq \hat{\mu} + \frac{t_\alpha}{2} \hat{\sigma} \sqrt{v^*}\right) = 1 - \alpha \quad (2.36)$$

since $\hat{\sigma}^2 = MS_E$

Thus the $(1-\alpha)$ 100% confidence interval for

$$\underline{x}^* \underline{\beta} \text{ is } \left(\underline{x}^* \underline{\beta} - t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{v^*}, \underline{x}^* \underline{\beta} + t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{v^*} \right) \quad (2.37)$$

Note that the standard error of the fitted y which is \hat{y} is

$$s.e(\hat{y}) = \hat{\sigma} \sqrt{\underline{x}^{*'} (X' X)^{-1} \underline{x}^*} \quad (2.38)$$

Hence the model is useful for prediction only for vectors \underline{x}^* near the centre of the region defined by the initial set of explanatory points used to fit the linear regression model ,

$$\underline{y} = \underline{X} \underline{\beta} + \underline{\xi}$$

As an example we shall consider the simple linear model

$$y = \alpha + \beta x + e$$

In this case

$$E(\hat{\mu}) = E(\hat{\alpha} + \hat{\beta} x^*) = \alpha + \beta x^*$$

and

$$\text{var}(\mu) = \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \sigma^2$$

and so

$$t = \frac{\hat{\mu} - \mu}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}} \sim t(n-2)$$

where

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{n-2}$$

Hence

$$\sum_{i=1}^n \frac{(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{\sigma^2} \sim \chi^2 (n-2)$$

Therefore confidence bounds on μ are obtained from

$$P\left(-t_{\frac{\alpha}{2}} \leq t \leq t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

i.e.

$$P\left(-t_{\frac{\alpha}{2}} \leq \frac{\hat{\mu} - \mu}{\hat{\sigma} \sqrt{\frac{1 + (x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}} \leq t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

which gives

$$P\left(\hat{\mu} - t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} < \mu < \hat{\mu} + t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}\right) = 1-\alpha$$

Thus the $(1-\alpha)$ 100% confidence interval for $\mu = \alpha + \beta x^*$ is

$$\left(\hat{\alpha} + \hat{\beta}x^* - t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \hat{\alpha} + \hat{\beta}x^* + t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}\right)$$

The standard error of the estimated y which is \hat{y} is

$$s.e(\hat{y}) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

at any point x , where $\hat{y} = \hat{\alpha} + \hat{\beta}x$ the behaviour of $s.e(\hat{y})$ as x deviates from the centre point is illustrated in the graph sketched in figure 1.

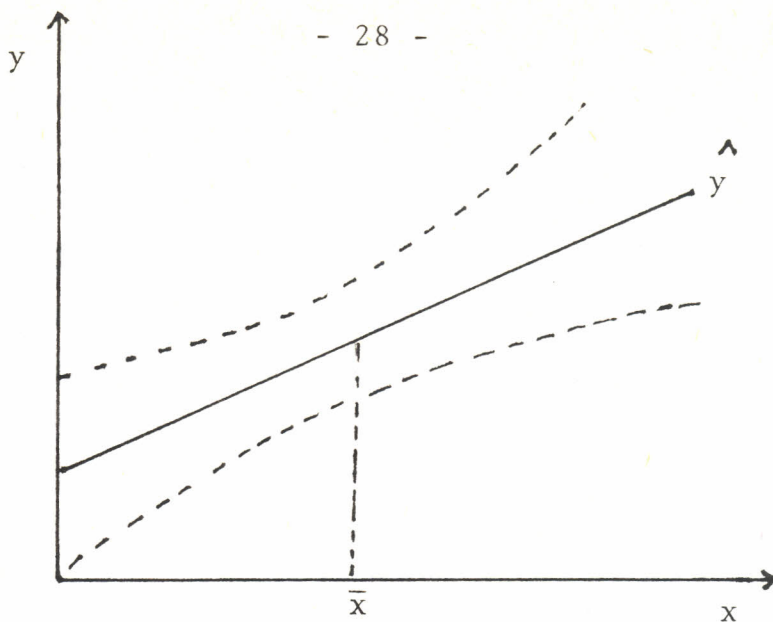


Figure 1.

We observe that the simple linear model is useful for prediction only near the centre of the x- values that are near \bar{x} .

Prediction Intervals

Let y_k denote the predicted value of y at the point

$$\underline{x}_k = (1, x_1^{(k)}, x_2^{(k)}, \dots, x_p^{(k)})'$$

that is y_k is a future observation given by

$$y_k = \beta_0 + \beta_1 x_1^{(k)} + \beta_2 x_2^{(k)} + \dots + \beta_p x_p^{(k)} + \xi_k \tag{2.39}$$

for $k > n$

Then y_k is estimated by

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(k)} + \dots + \hat{\beta}_p x_p^{(k)} = \underline{x}_k' \hat{\underline{\beta}}$$

We wish to determine the probability bounds on y_k .

Now consider $y_k - \hat{y}_k$

Then

$$E(y_k - \hat{y}_k) = 0 \tag{2.40}$$

and since \hat{y}_k is based on observations excluding y_k ,
the two are independent;
therefore

$$\begin{aligned} \text{var}(y_k - \hat{y}_k) &= \text{Var}(y_k) + \text{Var}(\hat{y}_k) \\ &= \sigma^2 + \text{var}(\underline{x}_k' \hat{\beta}) \\ &= \sigma^2 + \sigma^2 \underline{x}_k' (X'X)^{-1} \underline{x}_k \\ &= \sigma^2 (1 + \underline{x}_k' (X'X)^{-1} \underline{x}_k) \end{aligned}$$

Hence

$$y_k - \hat{y}_k \sim N \left[0, \sigma^2 (1 + \underline{x}_k' (X'X)^{-1} \underline{x}_k) \right] \quad (2.41)$$

This means that

$$Z = \frac{y_k - \hat{y}_k}{\sigma \sqrt{1 + \underline{x}_k' (X'X)^{-1} \underline{x}_k}} \sim N(0, 1) \quad (2.42)$$

but we also know that

$$\frac{(n-p-1) MS_E}{\sigma^2} \sim \chi^2_{(n-p-1)}$$

$$t = \frac{y_k - \hat{y}_k}{\sqrt{MS_E (1 + \underline{x}_k' (X'X)^{-1} \underline{x}_k)}} \sim t_{(n-p-1)}$$

or

$$t = \frac{y_k - \hat{y}_k}{\hat{\sigma} \sqrt{1 + \underline{x}_k' (X'X)^{-1} \underline{x}_k}} \sim t_{(n-p-1)} \quad (2.43)$$

since

$$\hat{\sigma}^2 = MS_E$$

and so

$$P(|t| \leq \frac{t_\alpha}{2}) = 1 - \alpha$$

which implies that

$$P \left(\frac{|y_k - \hat{y}_k|}{\hat{\sigma} \sqrt{1 + \underline{x}'_k (X'X)^{-1} \underline{x}_k}} \leq \frac{t_{\alpha}}{2} \right) = 1 - \alpha$$

that is:

$$P \left(\hat{y}_k - \frac{t_{\alpha}}{2} \hat{\sigma} \sqrt{1 + \underline{x}'_k (X'X)^{-1} \underline{x}_k} \leq y_k \leq \hat{y}_k + \frac{t_{\alpha}}{2} \hat{\sigma} \sqrt{1 + \underline{x}'_k (X'X)^{-1} \underline{x}_k} \right) = 1 - \alpha$$

Thus the $(1-\alpha)$ 100% prediction interval for y_k is

$$\left(\frac{\underline{x}'_k \hat{\beta} - t_{\alpha}}{2} \hat{\sigma} \sqrt{1 + \underline{x}'_k (X'X)^{-1} \underline{x}_k}, \quad \frac{\underline{x}'_k \hat{\beta} + t_{\alpha}}{2} \hat{\sigma} \sqrt{1 + \underline{x}'_k (X'X)^{-1} \underline{x}_k} \right) \quad (2.44)$$

The standard error of the predicted y_k which is \hat{y}_k is

$$s.e(\hat{y}_k) = \hat{\sigma} \sqrt{1 + \underline{x}'_k (X'X)^{-1} \underline{x}_k}$$

Hence as in the case of estimating the mean response the model is useful for predicting only for vectors \underline{x}_k near the centre of the region jointly defined by the original levels of the regressors;

$$(x_{i1}, x_{i2}, \dots, x_{ip}) \quad i=1,2, \dots, n$$

INDICATOR VARIABLES

We consider the case of simple linear regression where N observations can be formed into v groups with the v^{th} group having n_v observations. The most general model consists of v separate equations, such as

$$y = \beta_{0v} + \beta_{1v}x + \xi, \quad v=1,2, \dots, v \quad (2.45)$$

It is often of interest to compare the general model to a more restrictive one. Indicator variables are helpful in this regard. To fit the reduced model, define $v-1$ indicator variables D_1, D_2, \dots, D_{v-1} corresponding to v groups and fit

$$y = \beta_0 + \beta_1 x + \beta_2 D_1 + \beta_3 D_2 + \dots + \beta_v D_{v-1} \quad (2.46)$$

each D_i $i=1,2, \dots, v-1$ can only take value 0 or 1. In particular they all take value 0 if an observation is from group 1, if D_1 takes value 1 and the rest zero then that observation is from group 2. In general if D_{i-1} takes value 1 and the rest zero then that observation is from group i . Then by the use of the F- test we can make conclusions on whether or not

$$\beta_2 = \beta_3 = \dots = \beta_v = 0$$

and hence determine if it is needless to fit the reduced model or if the reduced model is valid. It is proposed to use this idea on the forest data with the groups being the districts.

From model given by (2.46) we note that

$$E(y|x, \text{ the data is from group } i) = \beta_0 + \beta_1 x + \beta_i D_{i-1} \quad (2.47)$$

so that when $D_{i-1} = 1$; meaning the data point is from group i we get the fitted simple model for group i as

$$\hat{y}_i = (\hat{\beta}_0 + \hat{\beta}_i) + \hat{\beta}_1 x \quad (2.48a)$$

$$i=2,3, \dots, v$$

and that for group 1 is

$$\hat{y}_1 = \beta_0 + \beta_1 x \quad (2.48b)$$

CHAPTER III

ANALYSIS OF VARIANCE

3.1 Introduction

The analysis of variance is perhaps the most widely used computational procedure in biometrics and analysis of quantitative inheritance. The procedure has widely different functions. The most important of these functions include the following:-

- (1) The study of the variance in a population, by decomposing the total variation into distinct components such as as the division of variance into genotypic and environmental components. In this respect the analysis of variance is essentially a procedure for estimating statistical parameters.
- (ii) Testing hypothesis or constructing tests of significance for any of the populations mentioned above.
- (iii) Testing the statistical significance of formulae which give the dependence of one variate on other variates e.g. in regression analysis.

3.2 PARTITION OF VARIANCE

If an observation Y is determined additively by two effects G and E such that

$$Y = G + E \quad (3.1)$$

it is preferable for a statistician to redefine the equation as

$$Y = \mu + G + E \quad (3.2)$$

where μ is a constant while G and E are now redefined as deviations which sum to zero over the whole population of G and E . consider a one-way fixed effects model

$$y_{ij} = \mu + t_i + e_{ij} \quad \begin{matrix} i=1,2, \dots, v \\ j=1,2, \dots, n_i \end{matrix} \quad (3.3)$$

and $\sum_{i=1}^v n_i = n$

where y_{ij} is the j th response on the i -th treatment

t_i is the effect due to treatment i .

e_{ij} is the experimental error associated with the ij -th response.

We shall assume

$$e_{ij} \sim N(0, \sigma^2) \quad \text{and} \quad y_{ij} \sim (\mu + t_i, \sigma^2) \quad (3.4)$$

where σ is the error variance, assumed to be unknown.

Then to test the significance of the treatment effects

t_1, t_2, \dots, t_v we test the null hypothesis

$$H_0: t_1 = t_2 = \dots = t_v \quad (3.5)$$

against the alternative hypothesis

$$H_1: t_i \neq t_j, \text{ for some } i \text{ and } j$$

This is an analysis of variance problem and is tested by partitioning the total variation into two components, namely variation due to treatment effects and variation due to error.

Let

$$\bar{y}_{i..} = \frac{1}{n} \sum_{j=1}^{n_i} y_{ij}$$

be the mean response on the i -th treatment. Then by the use of least squares method with the condition

$$\sum_{i=1}^v t_i = 0 \quad (3.6)$$

we get the estimates of μ and t_i as

$$\begin{aligned}\hat{\mu} &= \bar{y}_{..} \\ \hat{t}_i &= \bar{y}_{i.} - \bar{y}_{..}\end{aligned}$$

and

$$\hat{\xi}_{ij} = \bar{y}_{ij} - \bar{y}_{i.}$$

substituting these estimates in (3.3) we get

$$(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}) \quad (3.7)$$

squaring and summing both sides in (3.7) we get

$$\begin{aligned}\sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 &= \sum_{i=1}^v \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \\ &= \sum_{i=1}^v n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2\end{aligned}$$

Thus the total sum of squares

$$SS_T = \sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

has been partitioned into two components

$$SS_t = \sum_{i=1}^v n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

called the treatment sum of squares and

$$SS_E = \sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

called the error sum of squares. For computational

purposes we shall use the formulae,

$$SS_T = \sum_{i=1}^v \sum_{j=1}^{n_i} y_{ij}^2 - \frac{G^2}{n}$$

$$SS_t = \sum_{i=1}^v \frac{T_i^2}{n_i} - \frac{G^2}{n}$$

and

$$SS_E = SS_T - SS_t$$

where

$$T_i = \sum_{j=1}^{n_i} y_{ij} \quad i=1,2, \dots, v$$

$$G = \sum_{i=1}^v \sum_{j=1}^{n_i} y_{ij}$$

The Null distribution of SS_t

Here we wish to find the distribution of

$$SS_t = \sum_{i=1}^v n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

under the null hypothesis given by equation (3.5).

From the model

$$y_{ij} = \mu + t_i + e_{ij}$$

and

$$y_{ij} \sim N(\mu + t_i, \sigma^2), \quad i = 1, 2, \dots, v$$

$$j = 1, 2, \dots, n_i$$

Now

$$\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

This implies that

$$\bar{y}_{i.} \sim N\left(\mu + t_i, \frac{\sigma^2}{n_i}\right) \quad i=1, 2, \dots, v$$

if we let $\mu_i = \mu + t_i$, then we have that

$$\bar{y}_{i.} \sim N\left(\mu_i, \sigma^2/n_i\right) \quad i=1, 2, \dots, v \quad (3.9)$$

Since one of the assumptions in the model is that the v subpopulations are independent it follows that

$\bar{y}_{1.}, \bar{y}_{2.}, \dots, y_{v.}$ are independently distributed as

$$\bar{y}_{i.} \sim N(\mu_i, \sigma^2/n_i) \quad i=1,2, \dots, v$$

and since

$$\bar{y}_{..} = \frac{1}{v} \sum_{i=1}^v \bar{y}_{i.}$$

it follows that under H_0

$$\sum_{i=1}^v \frac{n_i (\bar{y}_{i.} - \bar{y}_{..})^2}{\sigma^2} \sim \chi^2(v-1)$$

OR

$$\sum_{i=1}^v n_i (\bar{y}_{i.} - \bar{y}_{..})^2 \sim \sigma^2 \chi^2(v-1) \quad (3.10)$$

The distribution of SS_e

$$SS_e = \sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

Again if we let $\mu_i = \mu + t_i$ then

$$Y_{ij} \sim N(\mu_i, \sigma^2) \quad \begin{matrix} i=1,2, \dots, v \\ j=1,2, \dots, n_i \end{matrix}$$

For a given i we have that

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_{i.}$$

and since $y_{ij}, j=1,2, \dots, n_i$ is a random sample of size n_i it follows that

$$\text{Var}(\bar{y}_{i.}) = \frac{\sigma^2}{n_i}$$

but we also know that

$$\sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_{i.})^2}{n_i - 1}$$

is an unbiased estimator of σ^2

Therefore

$$\sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_{i.})^2}{\sigma^2} \sim \chi^2_{(n_i - 1)} \quad (3.11)$$

And by the additive property of ch-square variables which can be stated as

$$\sum_{\alpha=1}^k \chi^2_{(K\alpha)} \sim \chi^2_{\left(\sum_{\alpha=1}^k K\alpha\right)}$$

it follows that

$$\sum_{i=1}^v \sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_{i.})^2}{\sigma^2} \sim \chi^2_{(n-v)}$$

OR

$$SS_e \sim \sigma^2 \chi^2_{(n-v)} \quad (3.12)$$

Since $SS_T = SS_t + SS_e$

we conclude that

$$SS_T / \sigma^2 \sim \chi^2_{(n-1)}$$

Therefore when H_0 is true then the statistic

$$F = \frac{SS_t / v - 1}{SS_e / n - v} \quad (3.13)$$

has the Fishers distribution with degrees of freedom $v-1$ and $n-v$ in the numerator and in the denominator respectively.

The computation needed for the above analysis can be arranged in an Analysis of variance table as follows:-

TABLE 1: ANOVA TABLE SHOWING THE ENTRIES IN EACH COLUMN AND THE COMPONENTS OF TOTAL VARIANCE

SOURCE OF VARIATION	DEGREES OF FREEDOM (D.F)	SUM OF SQUARES (SS)	MEAN SUM OF SQUARES (MSS)	THE F STATISTIC (F)
SOURCE OF	v-1	$\sum_{i=1}^v n_i (\bar{y}_{i.} - \bar{y}_{..})^2$ = SS_t	$\frac{SS_t}{v-1} = MS_t$	$\frac{MS_t}{MS_e}$
WITHIN TREATMENTS	n-1	$\sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$ = SS_e	$\frac{SS_e}{n-v} = MS_e$	
Total	n-1	$\sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$		

To test the hypothesis H_0 given in (3.5), the procedure is to compute the sample value of F which we call F_c and reject H_0 whenever

$$P = \text{Prob} (F > F_c / H_0) \text{ is small} \quad (3.14)$$

Paired Comparisons

More often it is of interest to know whether a given pair of treatments have same effects on the observation or not. That is we wish to test the hypothesis.

$$H_0 : t_i - t_k = 0.$$

against

$$H_a : t_i - t_k = 0 \quad i \neq k = 1, 2, \dots, v \quad (3.15)$$

Now we know that the estimates of t_i and t_k are given by

$$\hat{t}_i = \bar{y}_{i.} - \bar{y}_{..}$$

and

$$\hat{t}_k = \bar{y}_{k.} - \bar{y}_{..}$$

Therefore

$$\hat{t}_i - \hat{t}_k = \bar{y}_{i.} - \bar{y}_{k.} \quad (3.16)$$

is the estimate of $t_i - t_k$ from the samples

Now

$$\begin{aligned} \text{Var}(\bar{y}_{i.} - \bar{y}_{k.}) &= \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_k} \\ &= \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_k} \right) \end{aligned}$$

and under H_0

$$E(\bar{y}_{i.} - \bar{y}_{k.}) = 0 \quad (3.17)$$

Therefore

$$Z = \frac{\bar{y}_{i.} - \bar{y}_{k.}}{\sqrt{\sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}} \sim N(0,1) \quad (3.18)$$

When H is true and σ^2 is known. However in most practical problems σ^2 is not known. In this case we replace σ^2 in (3.18) by its unbiased estimator, the pooled sample error variance, which is the mean sum of squares due to error denoted by MS_e to obtain

$$t = \frac{\bar{y}_{i.} - \bar{y}_{k.}}{\sqrt{MS_e \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}} \sim t(n-v) \quad (3.19)$$

Therefore to test the hypothesis given by (3.15) we compute the value of the statistic t from the sample. We shall call this value t_c , that is we shall compute

$$t_c = \frac{\bar{y}_{i.} - \bar{y}_{k.}}{\sqrt{MS_e \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}} \quad (3.20)$$

and reject H_0 whenever

$$|t_c| = \frac{|\bar{y}_{i.} - \bar{y}_{k.}|}{\sqrt{MS_e \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}} > t_{\frac{\alpha}{2}} \quad (3.21a)$$

In the case of

$$n_1 = n_2 = \dots = n_v = n$$

the quantity

$$t_{\frac{\alpha}{2}} \sqrt{\frac{2MS_e}{n}} \quad (3.21b)$$

is called the least significance difference (Lsd).

However the rejection criterion can be made more flexible by computing the quantity

$$P = \text{pr} (t(n-v) < -t_c \text{ or } t(n-v) > t_c / H_0) \quad (3.22)$$

then reject H_0 whenever P is small. P is called the attained significance level or the P -value.

CHAPTER IV

APPLICATION TO FOREST DATA

4.1 INTRODUCTION

In this chapter we will apply the statistical methods described in chapters two and three. As earlier on stated, we have two species of trees which are grown for commercial use. These are the Pine and Cypress. The data for the Pine specie was obtained from nine districts which we represent by D_i ($i=1,2, \dots,9$). These are Nakuru, Kiambu, Nyeri, Kericho, Baringo, Meru, Laikipia, Nyandarua and Murang'a respectively.

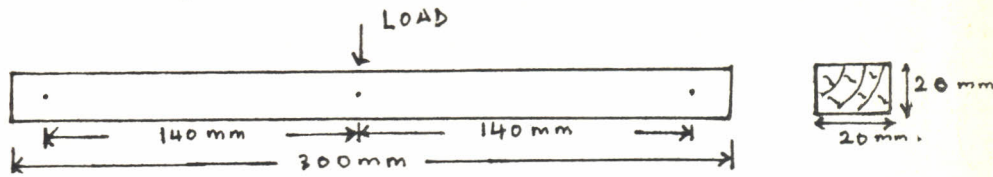
The Cypress data was obtained from eleven districts again denoted by D_i ($i=1,2, \dots,11$). These are Nakuru, Kiambu, Nyeri, Kericho, Baringo, Elgeyo, Uasin Gishu, Meru, Laikipia, Nyandarua and Muranga respectively. For each of the species, a number of logs n_i was obtained from district D_i . These samples were then transported to Karura forest laboratory for testing.

4.2 BRIEF DESCRIPTION ON THE EXPERIMENTAL PROCEDURE

For each of the n_i logs from district D_i a maximum of two pieces of size 20 mm by 20 mm by 300 mm were obtained, to give two specimens, one and two. Each specimen is supported on a span of 280 mm and the force applied at mid-span using loading heads. A deflectometer, located on the neutral axis of the specimen, is used to record the central deflection of the specimen relative to a span of 280 mm.

The diagram below explains the above.

Diagram 1:



From the experiment the following quantities were determined:-

- b = width of specimen
- d = depth of specimen
- L = loading span and span of deflectometer
- P' = load at limit of proportionality
- P = maximum load
- Δ = deflection at limit of proportionality

From the above quantities the following measures of strength for each specimen were computed:-

- (i) Stress at limit of proportionality = $3P'L/2bd^2$
- (ii) Modulus of rupture = $3PL/2bd^2$
- (iii) Modulus of elasticity = $P'L^3/4\Delta bd^3$

In this project the property of wood that is going to be studied is Modulus of rupture in (ii) above. This was measured in Mega-pascals. Another quantity of interest which was also determined for each specimen is the density for green timber, called the basic density in grams/cm^3 .

The density was computed from the formulae:-

$$\text{Basic density (green timber)} = \frac{W 100}{bdL(100+M)}$$

where

- W = mass of specimen
- M = moisture content of specimen.

We shall let y_{ij} denote the j -th modulus of rapture value from district i $j=1,2, \dots, n_i$ and $i=1,2, \dots, v$. In other words y_{ij} is the ij -th observation. v shall be nine or eleven depending on whether we are considering the Pine or the Cypress species.

4.3 REGRESSION ANALYSIS RESULTS

The dependent or response variable is Modulus of rapture and the independent variable is the Basic density. We will fit a linear models of the form.

$$y = \beta_{0v} + \beta_{1v}x + \xi \quad (4.1)$$

where

$$\begin{aligned} v &= 1,2, \dots, 9 \quad \text{in the case of Pine species} \\ v &= 1,2, \dots, 11 \quad \text{in the case of Cypress species} \end{aligned}$$

Instead of fitting the linear models separately we make use of indicator variables, and fit a model of the form

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 \dots + \beta_vx_v + \beta_{v+1}x_{v+1} + \xi \quad \dots (4.2)$$

The variables x_2, x_3, \dots, x_v are the indicator variables, take values 0 or 1. In particular they all take the value zero if the data point y_{ij} is from district 1. The full set of indicator variables is shown below.

TABLE 2: INDICATOR (DUMMY) VARIABLES FOR THE PINE SPECIES. A
GIVEN ROW SHOWS THE VALUES OF THE INDICATOR VARIABLES
WHEN AN OBSERVATION IS FROM THE DISTRICT AT THE END
OF THE ROW.

x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	DISTRICT
0	0	0	0	0	0	0	0	NAKURU (1)
1	0	0	0	0	0	0	0	KIAMBU (2)
0	1	0	0	0	0	0	0	NYERI (3)
0	0	1	0	0	0	0	0	KERICHO (4)
0	0	0	1	0	0	0	0	BARINGO (5)
0	0	0	0	1	0	0	0	MERU (6)
0	0	0	0	0	1	0	0	LAIKIPIA (7)
0	0	0	0	0	0	1	0	NYANDARUA (8)
0	0	0	0	0	0	0	1	MURANG'A (9)

TABLE 3: INDICATOR (DUMMY) VARIABLES FOR THE CYPRESS SPECIES

x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	DISTRICT
0	0	0	0	0	0	0	0	0	0	NAKURU (1)
1	0	0	0	0	0	0	0	0	0	KIAMBU (2)
0	1	0	0	0	0	0	0	0	0	NYERI (3)
0	0	1	0	0	0	0	0	0	0	KERICHO (4)
0	0	0	1	0	0	0	0	0	0	BARINGO (5)
0	0	0	0	1	0	0	0	0	0	ELEGEYO (6)
0	0	0	0	0	1	0	0	0	0	U.GISHU (7)
0	0	0	0	0	0	1	0	0	0	MERU (8)
0	0	0	0	0	0	0	1	0	0	LAIKIPIA (9)
0	0	0	0	0	0	0	0	1	0	NYANDARUA (10)
0	0	0	0	0	0	0	0	0	1	MURANG'A (11)

The variable x_{v+1} was introduced to cater for the two specimens from a given log.

That is

$$x_{v+1} = \begin{cases} 0 & \text{if the observation is specimen 1} \\ 1 & \text{if the observation is specimen 2} \end{cases} \quad (4.3)$$

The model given by (4.1) becomes a multiple linear regression models after the introduction of the indicator variables shown above

From the solution given in equation (2.12) of chapter 2 that is

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

We get the models for each species as below;

REGRESSION ANALYSIS FOR THE PINE SPECIES

We shall let MOR denote the Modulus of Rapture For this species the indicator variables are x_2, \dots, x_9 and x_{10} to serve the purpose given by (4.3) above. The fitted linear regression model of MOR on density was found to be

$$\begin{aligned} \text{MOR} = & 37.0675 + 9.3770 \text{ DENSITY} - 3.1824x_2 + 3.7590x_3 \\ & + 2.9981x_4 - 5.0891x_5 - 5.5916x_6 - 11.4224x_7 \\ & - 5.4467x_8 + 8.2375x_9 - 0.62151x_{10}. \end{aligned} \quad (4.4)$$

To get the simple linear regression model for district i , say $i=7$ all we need to do is to set all the x_j 's = 0 except x_7 ($j=2,3, \dots, 9$). That is the linear regression model for Laikipia district is

$$\text{MOR} = (37.0675 - 11.4224) + 9.3670 \text{ DENSITY} - 0.62151x_{10}$$

$$= 25.6451 + 9.3670 \text{ DENSITY} - 0.62151x_{10} \quad (4.5)$$

The same can be done for other districts.

TABLE 4: ANALYSIS OF VARIANCE TABLE SHOWING THE PARTITION OF THE TOTAL SUM OF SQUARES INTO REGRESSION AND RESIDUAL SUM OF SQUARES

SOURCE VARIATION	D.F	SUM OF SQUARES	MEAN SQUARE	F
REGRESSION	10	7734.8533	773.4853	5.1846
RESIDUAL	190	28345.7824	149.1883	
TOTAL				

Coefficient of determination $r_{y.12\dots p}^2 = 0.21438$ (4.6)

which implies $r_{y.12\dots p}^2 = 0.46301$

standard error which is given by

$$\text{s.e.} = \sqrt{149.1883} = 12.21427. \quad (4.7)$$

COMMENTS:-

Since $P(F(10,190) > 3.137) = 0.001$,

It follows that

$$P(F(10,190) > 5.1846) < 0.001, \quad (4.8)$$

hence we reject the hypothesis that

$$\beta_1 = \beta_2 = \beta_3 = \dots = \beta_{10} = 0$$

at level of significance 0.001.

From the coefficients of the sample linear regression model we find that Murang'a, Nyeri, Kericho and Nakuru produce timber of higher strength for the Pine species than the others.

REGRESSION ANALYSIS RESULTS FOR THE CYPRESS SPECIES

For this species we had x_2, x_3, \dots, x_{11} to represent the indicator (Dummy) variables to take care of the district levels as explained in table 3 of this chapter. It was found that the difference between the two specimens for a particular log was not significant hence the model does not have an extra indicator variable x_{12} .

There were 572 observations in all, that were used to fit the sample linear regression model. The calculated model for this species is given by:-

$$\begin{aligned} \text{MOR} = & 34.7294 + 14.5099 \text{ DENSITY} - 1.6614x_2 + 1.7186x_3 \\ & + 5.9319x_4 + 3.3781x_5 + 5.6254x_6 + 4.5034x_7 \\ & - 2.4627x_8 + 8.8309x_9 + 4.5593x_{10} - 3.7619x_{11} \end{aligned} \quad (4.9)$$

To obtain the linear regression model for district i , say $i=3$ we set all the $x_{j,s} = 0$

except x_3 ($j=2,3, \dots, 11$)

That is the linear regression model for NYERI district is given by:-

$$\begin{aligned} \text{MOR} &= 34.7294 + 14.5099 \text{ DENSITY} + 1.7186 \\ &= 36.4480 + 14.5099 \text{ DENSITY} \end{aligned} \quad (4.10)$$

TABLE 5: The table gives the analysis of variance results for this regression model. Beneath, we give the coefficient of determination together with the standard error.

SOURCE OF VARIATION	D.F	SUM OF SQUARES (SS)	MEAN SQUARE (MSS)	F
REGRESSION	11	10107.2865	918.8442	11.7780
RESIDUAL	560	43687.6181	78.0136	
TOTAL				

$$\text{Coefficient of determination } r_{y.12\dots p}^2 = 0.18789$$

$$\text{Standard error} = 8.83253$$

COMMENTS:

$$\text{Since } \text{Prob} (F(11,560) > 2.9061) = 0.001 ,$$

It follows

$$\text{Prob} (F(11,560) > 11.7780) < 0.001, \quad (4.11)$$

hence we reject the hypothesis that

$$\beta_1 = \beta_2 = \beta_3 = \dots = \beta_{11} = 0$$

at significance level 0.001.

From the coefficients of the sample linear regression model for the Cypress species, we may say that Laikipia, Elgeyo, Kericho, Uasin Gishu, Nyandarua, Nyeri, Baringo and Nakuru districts produce timber of quite a high strength. While Murang'a, Kiambu and Meru districts are not doing well as far as the strength of the product is concerned.

4.4. ANALYSIS OF VARIANCE

In this section we wish to study the variation of Modulus of Rapture with location (District). We need to tell whether or not location affects the modulus of Rapute and hence the strength of timber. We shall adopt the model

$$y_{ij} = \mu + d_i + \xi_{ij}$$

$i=1,2, \dots, 9$ for the Pine species

$i=1,2, \dots, 11$ for the Cypress species

$j=1,2, \dots, n_i$

Here μ = location parameter common to all observations

d_i = effect peculiar to i th district

ξ_{ij} = normally distributed random variable with mean zero and variance σ^2

We shall need the following quantities in order to set the ANOVA table appropriately:-

$$(i) \quad G = \sum_{i=1}^v \sum_{j=1}^{n_i} y_{ij} \quad \text{and} \quad n = \sum_{i=1}^v n_i$$

(ii) For each district D_i we computed the following

$$T_i = \sum_{j=1}^{n_i} y_{ij}; \quad \text{total for district } D_i$$

$$\bar{y}_i = T_i/n_i; \quad \text{sample mean for district } D_i$$

$$\sum_{j=1}^{n_i} y_{ij}^2 = \text{Sum of squares for district } D_i$$

$i=1,2, \dots, 9 \text{ or } 11$

Then after this the following were computed as required;

The total sum of squares TSS given by

$$TSS = \sum_{i=1} \sum_{j=1} y_{ij}^2 - \frac{G^2}{N}$$

Sum of squares due to site (treatments) given by

$$SS_t = \sum_{i=1} T_i^2/n_i - \frac{G^2}{N}$$

Then from the two sum of squares above the within sum of squares or error sum of squares was obtained by subtraction, that is

$$ESS = TSS - SS_t$$

The following are the results for the Pine species.

TABLE 6: ANOVA table for the Pine species. It gives the components of the total sum of squares and finally the F-statistic is obtained

SOURCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE
TREATMENTS (SITE)	3890.085	8	486.2606
WITHIN DISTRICTS (ERROR)	14224.095	92	154.6097
TOTAL	18114.180	100	

From the table, the calculated $F(8,92)$ is given by:-

$$F_c = \frac{486.2606}{154.6097} = 3.1451$$

Then using the Fishers tables with 8 and 92 degrees of freedom in the numerator and demoninator respectively we get

$$\text{Prob} (F(8,92) > 2.9929) = 0.005$$

which implies

$$P = \text{Prob} (F(8,92) > 3.1451) < 0.005 \quad (4.12)$$

Hence we reject the hypothesis

$$H_0: =d_1=d_2\dots\dots\dots = d_9$$

at significance level 0.005. Meaning that the effect due to districts on Modulus of rapture is different for at least two districts or $d_i=d_j$ for $i=j$.

Next we carry out a paired comparison test for the Pine species.

PAIRED COMPARISONS FOR THE PINE SPECIES

Below we arrange the sample means from each district which Pine is grown in descending order of magnitude. This will facilitate the computations for the calculated t statistics.

TABLE 7: MEANS ARRANGED IN DESCENDING ORDER OF MAGNITUDE

SAMPLE MEAN	SAMPLE NO	DISTRICT
$\bar{y}_{1.} = 47.8956$	$n_1 = 18$	NAKURU
$\bar{y}_{3.} = 46.7098$	$n_3 = 13$	NYERI
$\bar{y}_{9.} = 46.0466$	$n_9 = 6$	MURANG'A
$\bar{y}_{4.} = 40.4684$	$n_4 = 6$	KERICHO
$\bar{y}_{5.} = 38.3521$	$n_5 = 8$	BARINGO
$\bar{y}_{8.} = 37.9624$	$n_8 = 6$	NYANDARUA
$\bar{y}_{6.} = 35.3212$	$n_6 = 7$	MERU
$\bar{y}_{2.} = 34.8838$	$n_2 = 30$	KIAMBU
$\bar{y}_{7.} = 28.1457$	$n_7 = 7$	LAIKIPIA

To test the hypothesis numbered (3.15) in Chapter III we compute the sample t statistic given by

$$t_c = \frac{|\bar{y}_{i.} - \bar{y}_{k.}|}{MS_e \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}$$

and reject H when the value

$P = \text{Prob} (t(n-v) > t_c \text{ or } t(n-v) < - t_c/H_0)$ is small.

This value is called the attained significance level.

Below we give the table of results.

TABLE 8: SHOWS PAIRS OF DISTRICTS WITH THE
CORRESPONDING CALCULATED STUDENT t STATISTIC
AND THE ATTAINED SIGNIFICANCE LEVELS

DISTRICT PAIRS	CALCULATED STUDENT $t = t_c$	$P = P [t(92) > t_c \text{ or } t(92) < -t_c]$
NAKURU - NYERI	0.2560	0.70 < P < 0.80
NAKURU - MURANG'A	0.3154	0.60 < P < 0.70
NAKURU - KERICHO	1.2671	0.20 < P < 0.30
NAKURU - BARINGO	1.8063	0.05 < P < 0.10
NAKURU - NYANDARUA	1.6946	0.05 < P < 0.10
NAKURU - MERU	2.2703	0.02 < P < 0.05
NAKURU - KIAMBU	3.5099	P < 0.001
NAKURU - LAIKIPIA	3.5658	P < 0.001
NYERI - MURANG'A	0.1081	0.90 < P < 0.95
NYERI - KERICHO	1.0858	0.20 < P < 0.30
NYERI - BARINGO	1.3771	0.10 < P < 0.20
NYERI - NYANDARUA	1.4254	0.10 < P < 0.20
NYERI - MERU	1.9537	0.05 < P < 0.10
NYERI - KIAMBU	2.8643	0.002 < P < 0.01
NYERI - LAIKIPIA	3.1846	0.001 < P < 0.002
MURANG'A - KERICHO	0.777P	0.40 < P < 0.50
MURANG'A - BARINGO	1.1458	0.20 < P < 0.30
MURANG'A - NYANDARUA	1.1261	0.20 < P < 0.30
MURANG'A - MERU	1.5504	0.10 < P < 0.20

DISTRICT PAIRS	CALCULATED STUDENT $t = t_c$	$P = P [t(92) > t_c \text{ or } t(92) < -t_c]$
MURANG'A - KIAMBU	2.0074	0.40 < P < 0.05
MURANG'A - LAIKIPIA	2.5876	0.01 < P < 0.02
KERICHO - BARINGO	0.3151	0.70 < P < 0.80
KERICHO - NYANDARUA	0.3491	0.70 < P < 0.80
KERICHO - MERU	0.7441	0.40 < P < 0.50
KERICHO - KIAMBU	1.0043	0.30 < P < 0.40
KERICHO - LAIKIPIA	1.7813	0.05 < P < 0.10
BARINGO - NYANDARUA	0.0580	0.97 < P < 0.98
BARINGO - MERU	0.4709	0.60 < P < 0.70
BARINGO - KIAMBU	0.7010	0.40 < P < 0.50
BARINGO - LAIKIPIA	1.5657	0.10 < P < 0.20
NYANDARUA - MERU	0.3818	0.60 < P < 0.70
NYANDARUA - KIAMBU	0.5536	0.50 < P < 0.70
NYANDARUA - LAIKIPIA	1.4191	0.10 < P < 0.20
MERU - KIAMBU	0.0838	0.90 < P < 0.95
MERU - LAIKIPIA	1.0796	0.20 < P < 0.30
KIAMBU - LAIKIPIA	2.29110	0.10 < P < 0.20

As earlier stated we reject the hypothesis

$$H_0 : d_i - d_k = 0$$

against

$$H_0 : d_i - d_k \neq 0$$

whenever the p-value in column III of table 8 is small. Taking $P < 0.10$ to be small, we conclude that the mean strengths of timber from the district pairs listed in table 9 below differ significantly.

TABLE 9: PAIRS OF DISTRICTS WHICH SHOW A SIGNIFICANT DIFFERENCE IN MEAN STRENGTH FOR THE PINE SPECIES

DISTRICT PAIRS		P- VALUE
NAKURU	- BARINGO	0.05 < P < 0.10
NAKURU	- NYANDARUA	0.05 < P < 0.10
NAKURU	- MERU	0.02 < P < 0.05
NAKURU	- KIAMBU	P < 0.001
NAKURU	- LAIKIPIA	P < 0.001
NYERI	- MERU	0.05 < P < 0.10
NYERI	- KIAMBU	0.02 < P < 0.01
NYERI	- LAIKIPIA	0.001 < P < 0.002
MURANG'A	- KIAMBU	0.02 < P < 0.05
MURANG'A	- LAIKIPIA	0.01 < P < 0.02
KERICHI	- LAIKIPIA	0.05 < P < 0.1

ANALYSIS OF VARIANCE RESULTS FOR THE CYPRESS SPECIES

In this case there were eleven districts and from computation similar to those with the PINE species lead to the following ANOVA table.

Table 10: ANOVA table for the Cypress species. It shows the components of the total sum of squares and finally the F-statistics is obtained.

SOURCE OF VARIATION	SUM OF SQUARES (SS)	DEGRESS OF FREEDOM	MEAN SQUARE (MSS)
TREATMENTS (SITE)	4242.2314	10	424.2231
WITIN SITES (ERROR)	23573.6230	281	83.8914
TOTAL	27815.854	291	F(10,281)

From the table the calculated $F(10,281)$ statistic is given by:-

$$F_C = \frac{424.2231}{83.8919}$$
$$= 5.0568$$

Since $\text{Prob} (F(10,250) > 3.0893) = 0.001$

It follows that

$$P = \text{Prob} (F(10,281) > 5.0568) < 0.001 \quad (4.13)$$

The meaning of the statement in (4.13) is that the hypothesis,

$$H_0 : d_1 = d_2 = \dots = d_{11}$$

is rejected, at level of significance of 0.001. This implies that the quantity under study namely the Modulus of Rapture, a measure of the strength of wood, is different for at least two districts that is

$$d_i \neq d_k \text{ for } i \neq k = 1, 2, \dots, 11$$

The next step is to carry out pairwise comparisons to discover which pair of districts bring about this difference.

PAIRED COMPARISONS FOR THE CYPRESS SPECIES

Below we arrange the sample means from the districts which CYPRESS is grown in descending order of magnitude. This will facilitate the computations for the calculated student t statistics.

TABLE 11: MEANS ARRANGED IN DESCENDING ORDER OF MAGNITUDE

SAMPLE MEAN	SAMPLE NO	DISTRICT
$\bar{y}_9 = 48.6767$	$n_9 = 15$	LAIKIPIA
$\bar{y}_6 = 47.3294$	$n_6 = 25$	ELGEYO
$\bar{y}_4 = 46.6421$	$n_4 = 35$	KERICHO
$\bar{y}_7 = 45.8388$	$n_7 = 25$	UASIN GICHU
$\bar{y}_{10} = 44.5428$	$n_{10} = 16$	NYANDARUA
$\bar{y}_3 = 43.2006$	$n_3 = 34$	NYERI
$\bar{y}_5 = 42.2048$	$n_5 = 30$	BARINGO
$\bar{y}_1 = 41.3058$	$n_1 = 39$	NAKURU

TABLE 11: CONTINUED:

SAMPLE MEAN	SAMPLE NO	DISTRICT
$\bar{y}_{11.}$ = 38.0558	n_{11} = 15	MURANG'A
$\bar{y}_{2.}$ = 37.8601	n_2 = 38	KIAMBU
$\bar{y}_{8.}$ = 35.5560	n_8 = 20	MERU

TABLE 12: SHOWS DISTRICT PAIRS WITH THE CORRESPONDING CALCULATED STUDENT t STATISTIC AND THE ATTAINED SIGNIFICANCE LEVELS, FOR THE CYPRESS SPECIES.

DISTRICT PAIRS	CALCULATED STUDENT $t = t_c$	$P = \text{Prob}(t(281) > t_c \text{ or } t(281) < -t_c)$
LAIKIPIA - ELGEYO	0.4503	$0.65 < P < 0.66$
LAIKIPIA - KERICHO	0.7198	$0.47 < P < 0.48$
LAIKIPIA - U.GISHU	0.9487	$0.34 < P < 0.35$
LAIKIPIA - NYANDARUA	1.2558	$0.21 < P < 0.22$
LAIKIPIA - NYERI	1.9288	$0.05 < P < 0.06$
LAIKIPIA - BARINGO	2.2345	$0.02 < P < 0.03$
LAIKIPIA - NAKURU	2.6488	$P < 0.01$
LAIKIPIA - MURANG'A	3.1757	$P < 0.01$
LAIKIPIA - KIAMBU	3.8728	$P < 0.01$
LAIKIPIA - MERU	4.1940	$P < 0.01$
ELGEYO - KERICHO	0.2866	$0.77 < P < 0.78$
ELGEYO - U.GISHU	0.5754	$0.56 < P < 0.57$
ELGEYO - NYANDARUA	0.9503	$0.34 < P < 0.35$
ELGEYO - NYERI	1.7110	$0.08 < P < 0.09$
ELGEYO - BARINGO	2.0661	$0.03 < P < 0.04$
ELGEYO - NAKURU	2.5668	$0.01 < P < 0.02$
ELGEYO - MURANG'A	3.1001	$P < 0.01$
ELGEYO - KIAMBU	4.0147	$P < 0.01$
ELGEYO - MERU	4.2847	$P < 0.01$

TABLE 12: CONTINUED

DISTRICT PAIRS	CALCULATED STUDENT $t = t_c$	P= Prob($t(281) >$ or $t(281) < -t$)
KERICHO - U.GISHU	0.3349	0.73 < P < 0.74
KERICHO - NYANDARUA	0.7595	0.44 < P < 0.45
KERICHO - NYERI	1.5604	0.11 < P < 0.12
KERICHO - BARINGO	1.9471	0.05 < P < 0.06
KERICHO - NAKURU	2.5022	0.01 < P < 0.02
KERICHO - MURANG'A	3.0377	P < 0.01
KERICHO - KIAMBU	4.0926	P < 0.01
KERICHO - MERU	4.3180	P < 0.01
U. GISHU - NYANDARUA	0.4420	0.65 < P < 0.67
U. GISHU - NYERI	1.0933	0.26 < P < 0.27
U. GISHU - BARINGO	1.4670	0.14 < P < 0.15
U. GISHU - NAKURU	1.9317	0.05 < P < 0.06
U. GISHU - MURANG'A	2.6018	P < 0.01
U. GISHU - KIAMBU	3.3827	P < 0.01
U. GISHU - MERU	3.7422	P < 0.01
NYANDARUA - NYERI	0.4834	0.62 < P < 0.63
NYANDARUA - BARINGO	0.8246	0.40 < P < 0.41
NYANDARUA - NAKURU	1.1904	0.23 < P < 0.24
NYANDARUA - MURANG'A	1.9706	0.04 < P < 0.05
NYANDARUA - KIAMBU	2.4482	0.01 < P < 0.02
NYANDARUA - MERU	2.9253	P < 0.01

TABLE 12: CONTINUED

DISTRICT PAIRS	CALCULATED STUDENT $t = t_c$	P= Prob($t(281)$ or $t(281) <$
NYERI - BARINGO	0.4340	0.66 < P < 0.6
NYERI - NAKURU	0.8817	0.37 < P < 0.3
NYERI - MURANG'A	1.8122	0.06 < P < 0.0
NYERI - KIAMBU	2.4700	0.01 < P < 0.0
NYERI - MERU	2.9618	P < 0.0
BARINGO - NAKURU	0.4042	0.68 < P < 0.69
BARINGO - MURANG'A	1.4325	0.66 < P < 0.67
BARINGO - KIAMBU	1.9422	0.05 < P < 0.06
BARINGO - MERU	2.5146	0.01 < P < 0.02
NAKURU - MURANG'A	1.1679	0.24 < P < 0.25
NAKURU - KIAMBU	1.6504	0.09 < P < 0.10
NAKURU - MERU	2.2825	0.02 < P < 0.03
MURANG'A - KIAMBU	0.0701	0.94 < P < 0.43
MURANG'A - MERU	0.7990	0.36 < P < 0.37
KIAMBU - MERU	0.9106	0.36 < P < 0.37

Note:

To compute the probabilities in the last column the student t with 281 degrees of freedom was approximated to a standard normal distribution.

We recall that the hypothesis we want to test is

$$H_0 : d_i - d_k = 0$$

against

$$H_1 : d_i - d_k \neq 0$$

that is to test whether the mean strength of timber for the Cypress species due to districts D_i and D_k is the same or not. We reject the hypothesis H_0 when the attained significance level is small. If we regard $P < 0.10$ as small, we obtain the list in table 13 below for which the mean strength of wood differ significantly.

TABLE 13: PAIRS OF DISTRICTS WHICH SHOW A SIGNIFICANT DIFFERENCE IN MEAN STRENGTH FOR THE CYPRESS SPECIES

DISTRICT PAIRS	P - VALUE
LAIKIPIA - NYERI	0.05 < P < 0.06
LAIKIPIA - BARINGO	0.02 < P < 0.03
LAIKIPIA - NAKURU	P < 0.01
LAIKIPIA - MURANG'A	P < 0.01
LAIKIPIA - KIAMBU	P < 0.01
LAIKIPIA - MERU	P < 0.01
ELGEYO - NYERI	0.08 < P < 0.09
ELGEYO - BARINGO	0.03 < P < 0.04
ELGEYO - NAKURU	0.01 < P < 0.02
ELGEYO - MURANG'A	P < 0.01
ELGEYO - KIAMBU	P < 0.01
ELGEYO - MERU	P < 0.01
KERICHO - NAKURU	0.01 < P < 0.02
KERICHO - MURANG'A	P < 0.01
KERICHO - KIAMBU	P < 0.01

TABLE 13 CONTINUED

DISTRICT PAIRS		P - VALUE	
U. GISHU	- MURANG'A		P < 0.01
U. GISHU	- KIAMBU		P < 0.01
U. GISHU	- MERU		P < 0.01
NYANDARUA	- MURANG'A	0.04	< P < 0.05
NYANDARUA	- KIAMBU	0.01	< P < 0.02
NYANDARUA	- MERU		P < 0.01
NYERI	- MURANG'A	0.06	< P < 0.07
NYERI	- KIAMBU	0.01	< P < 0.02
NYERI	- MERU		P < 0.01
BARINGO	- KIAMBU	0.05	< P < 0.06
BARINGO	- MERU	0.01	< P < 0.02
NAKURU	- KIAMBU	0.09	< P < 0.10
NAKURU	- MERU	0.02	< P < 0.03
U. GISHU	- NAKURU	0.05	< P < 0.06

CHAPTER 5

DISCUSSION AND CONCLUDING REMARKS

5.1 REGRESSION

For both wood species, the Pine and the Cypress we rejected the hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_v = 0$$

where v is 10 and 11 for the Pine and Cypress species respectively. Given that in both the regression models most independent variables were indicator variables to take care of site levels, we conclude here that there is high dependence of strength of timber with site apart from density of the material

The small values of the sample coefficients of determinations imply that little variability of the Modulus of rupture is explained by the main independent variable the density.

This was 21.43% for the case of the Pine species and only 18.8% for the case of the Cypress Species.

A course for these low values could be due to errors made in the determination of the measurable variables, the response variable which was the modulus of rupture and the explanatory variable which was the basic density. This is so, because the other independent variables were qualitative, the site levels, in this case the districts.

The model can be improved in a future study by including more explanatory variables.

Now, looking at the regression models closely we find very

useful information. We can easily rank the Districts according to quality (strength) of timber. From the regression models we can retrieve the simple linear regression models for each district. The simple linear regression models have same gradient (slopes) but different intercepts. Let us consider the Pine species case; whose calculated multiple regression model is given by equation (4.4). In this model, the indicator variables all take the value zero when the data point is from Nakuru district, hence the simple linear model for the District. Taking Nakuru district as the reference point we find that Murang'a, Nyéri, Kericho produce timber of a higher strength than Kiambu, Baringo, Meru, Laikipia and Nyandarua districts.

Similarly with the case of Cypress we get the simple linear model for Nakuru when we put all the indicator variables to take value zero in the model given by (4.9). Hence taking Nakuru district as the reference point we find that Laikipia, Elgeyo, Kericho, Uasin Gishu, Nyandarua, Nyeri, Baringo and Nakuru districts produce Cypress wood of quite a higher strength than Murang'a, Kiambu and Meru districts.

The findings here are important to a forest manager because if the interest is the strength of wood then emphasis can be put in the Districts which show a tendency of producing timber of higher quality (strength). The manager can even decide to substitute a species of lower quality with one of higher quality for a given District. For a particular species, the manager can decide to follow a sequence of harvesting

to avoid depletion of that species. Since if several districts produce timber of higher quality for a given species then harvesting can be planned in such a way that more harvesting is done for those districts which have a bigger number than those with a lower number.

5.2 ANALYSIS OF VARIANCE:

From the analysis of variance model

$$y_{ij} = \mu + d_i + e_{ij} \quad \begin{matrix} i=1, 2, \dots, v \\ j=1, 2, \dots, n_i \end{matrix}$$

we rejected the hypotheses

$$H_0 : d_1 = d_2 = \dots = d_9$$

for the Pine species, and

$$H_0 : d_1 = d_2 = \dots = d_{11}$$

for the Cypress species, where the d_i 's are the effects of districts D_i , $i=1, 2, \dots, v = 9$ or 11 on Modulus of Rapture.

The conclusion here is that the mean strengths of wood for a given species are not the same for at least a pair of Districts. This was verified by the paired comparison. The recommendation therefore is that in future constructional standards for these species of wood should be calculated per population in this case the districts, as opposed to the past where the districts were taken as one population. This may mean that even the prices of wood from district D_i may not be the same as that of district D_i .

The results of the paired comparison can further help in management because if wood from district D_i and that of district D_k show same effects on the strength of wood, then if wood of a given strength is required and is not available in District D_i then it can be obtained from district D_k .

The study in this project shall go a long way towards helping the Kenya Bureau of Standards in setting standards of the Kenyan wood. This is because in the past standards were based on a pooled sample from the various districts, but from the findings it appears that there is need to treat the districts individually.

On improving the carrying out of the study it is recommended that in future in case of similar study an equal number of observations per district be obtained since the analysis of variance carried out here was aimed at comparing means of v populations, where v is the number of districts. This will make the computations for paired comparison easier by making use of the least significance difference which is common for all pairs.

It is believed that the study has answered some of the questions intended to be answered as per the aim of study.

REFERENCES

1. Burges H.J. (1962): A new approach to stress-strain time relations of wood; paper presented to the British Commonwealth Forestry Conference, 1962.
2. Brister G.H. and Fry G. (1962): Kenya Pine timber-structure and strength; paper presented to the British Commonwealth Forestry Conference, 1962.
3. Charles L.L. and Richard J.H. (1974): Solving least squares problems; Prentice - Hall, INC.
4. Mack J.J. (1979): Australian methods of mechanically testing small clear specimens of timber, division of building research technical paper No.31, Commonwealth Scientific and Industrial Organization, Australia.
5. Oscar K. (1957): An introduction to genetic statistics, John Wiley and Sons, Inc., New York
6. Paterson D.N. (1963): The strengths of Kenya timbers their derivation and application; Kenya Forest Department Research Bulletin No. 23.

7. Sunley J.G. (1956): Working stress for structural woods; Forest products research Bulletin No. 37, Department of Scientific and Industrial Research, London.

8. Usher M.B (1966): A matrix approach to management of renewable resources, with special references to selection forests, J. Appl. Ecol. 3 355-67.

9. Usher M.B. (1969): A matrix model for Forest Management; Biometrics, 25, 309-315.

10. Woodward I.O. (1982): Modelling Population growth in stage-grouped organisms: a simple extension to the Leslie model; Australian Journal of Ecology, 7, 389-94.