

UNIVERSITY OF NAIROBI

COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES

SCHOOL OF MATHEMATICS

PROJECT WORK

MULTIPLE REGRESSION: BAYESIAN INFERENCE

**A PROJECT SUBMITTED TO THE SCHOOL OF MATHEMATICS IN PARTIAL
FULFILLMENT FOR THE AWARD OF A MASTER OF SCIENCE DEGREE IN STATISTICS**

SEPTEMBER 2011

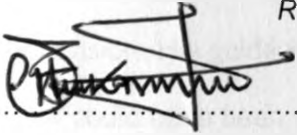
OMUKAMI HOWARD

Declaration

I the undersigned declare that the project is my original work and to the best of my knowledge has not been submitted for award of a degree in any other university.

OMUKAMI HOWARD

Reg. No: 156/72144/2008



4th August 2011

Signature

Date

Declaration by Supervisor

This project has been submitted with my approval as supervisor.

Dr. F.NJUI

School of Mathematics

University of Nairobi

P.O Box 30197 Nairobi

KENYA



August 4th. 2011

Signature

Date

Acknowledgement

The content and flow of this research work has been greatly influenced by a number of people and their contribution cannot be overlooked. I am greatly indebted to those who contributed to this success.

My deepest gratitude goes to Dr. Njui for allowing me to do the work under his tireless supervision. His guidance and commitment to see me succeed has been overwhelming. Without him it could have been very difficult to finish this work within the stipulated time frame. Allow me to mention his availability at all times and his ingenuity in various topics that has been so inspiring.

Deep appreciation goes to the School of Mathematics teaching staff for the M.S.C program who in their able hands I have acquired knowledge, skills and values through constructive, brief and critical approach at all levels.

To my classmates; Rose, Mecha, Tonui and Otieno for their willingness to see all of us succeed. You have been such a challenge.

Finally, to my friends Adagala and Amaheno for their wits that have seen me through this work.

Dedicated to:

My beloved dear wife Esther Tsindoli for her constant encouragements, inspirations and support throughout the entirety of my academic traversals.

Abstract

Bayesian approach is based on subjective interpretation of probability. It views probability as a degree of belief concerning an uncertainty. It also regards an unknown parameter as an uncertainty on which a degree of belief can be expressed and then revised based on sample information. A parameter is viewed as a random variable which prior to sample evidence is assigned a prior distribution. When sample evidence is obtained, the prior distribution is revised and posterior distribution obtained.

The Bayesian approach also uses data to update the uncertainty distribution for unknown parameters then draw conclusions using the updated distributions. Bayesian inferences help in decision making such that the gains made out of a study are maximized and the risks minimized. Data is used to generate posterior distribution depending on assumed prior distribution. This is done in terms of a likelihood function corresponding to the observed data. There are many situations especially in business, industry and technology where a careful analysis of a decision making is beneficial.

In this project, the data analyzed was collected from a pig breeding research centre. Various factors affecting the weights of pigs were determined by a regression equation after computing various Bayesian point estimates.

Keywords

Parameters, Bayesian approach, minimum and maximum risk, posterior and priori distribution, likelihood function, regression, Bayesian point estimate.

Table of Contents

Declaration.....	i
Acknowledgement.....	ii
Dedication.....	iii
Abstract.....	iv
INTRODUCTION	1
1.1 Statistical Decision Theory	1
1.2 Review of Bayes Theorem.....	2
1.3 Prior distribution	4
1.4 Posterior distribution.....	5
1.5 The Loss function	6
Risk function	7
1.6 Bayes solutions of a statistical decision problem	7
against a specified priori distribution	7
1.7 Literature Review.....	10
1.9 Aims and Objectives.....	14
CHAPTER 2	15
PROBLEM IDENTIFICATION AND DECISION RULES	15
2.1 Introduction.....	15
2.2 Randomized and Non-Randomized decision Rules.....	15
2.3 Construction of Bayes Decision functions	18
2.4 Comparison of Bayes Estimates with other Estimates.....	18
2.5 Convergence of Bayes Estimates	19
2.6 Bayes Point estimates.....	21
2.7 Bayesian confidence Regions	21
2.8 Empirical Bayes methods	23
2.9 Bayes strategy	25
CHAPTER 3	27
Multiple Regression; A Bayesian inference.....	27
3.1 Introduction.....	27
3.2 A Bayesian Multiple Regression Model.....	29
3.2.1 A Bayesian multiple Regression model with a conjugate prior.....	30
3.3 Marginal Posterior Density of β	32
3.4 Bayesian point and Interval Estimates of Regression Coefficients	34
CHAPTER 4	37
Data Analysis and Conclusion	37
Conclusion.....	39
Areas of Further Research.....	41
REFERENCES.....	42

CHAPTER 1

INTRODUCTION

1.1 Statistical Decision Theory

In most practical problems, a decision maker should be able to make a choice from among several different acts or subjects to a wide variety of states or conditions over which he or she has no control. Classical decision procedures do not recognize the validity of any information pertaining to a decision that does not exist in the form of empirical data and results from a process of sampling.

A procedure for using data as an aid to decision making will involve a set of instructions that assign one action to a possible value of dataset. In any study there are consequences of making wrong decisions. Bayesian decision making is based on data from which optimal decisions can be made. It provides a model for decision making in situations that involve multiple states of a parameter or nature. It also incorporates economic consequences of taking certain decisions especially where a study is done on an industrial process.

Bayesian decision theory allows use of information for prior or sampling experimentation, whether the decision is in the form of empirical data or is subject to assessment by the decision maker. In any organization the main function of the executive is to make decisions. The decisions can be made on new products to be introduced in the market, the number of units to be produced and how the products will be marketed. The decision theory is also applied in financial management in order to make decisions that will minimize costs and maximize profits.

A decision may be needed where two or more alternative courses of action are available and where only one must be taken. If there's only one course available, then the line of action is clear. The line of action is one with minimal time wastage and minimal cost as long as the action taken is the best among several others which may be available.

Uncertainty may occur when the outcome of an action is not known in advance, for example when there are many possible outcomes of an event one cannot predict with certainty what would happen if a particular line of action is taken. Probability is used to identify the best line of action.

Sometimes decisions have to be made under *risk* and *uncertainty*. When the state of nature is known and objective or empirical data is available so that the decision maker can use this data to assign probabilities to various states of nature; then the decision is said to have been *made under Risk*. When the state is unknown and there is no objective information on which probabilities can be based, then the *decision is said to have been made under uncertainty*.

1.2 Review of Bayes Theorem

To make use of Bayes Principle in statistical decision problems, the decision maker must be able to assign probabilities to the states of nature. Suppose a researcher is conducting an experiment in which he is aware that the result of interest will be affected by any of the existing alternatives say $\beta_1, \beta_2, \dots, \beta_n$. He may not be certain about which one of these alternatives will ultimately prevail; he may have some information from which he can make subjective judgments concerning probabilities in n alternatives. Thus, he assigns probabilities $P(\beta_1), P(\beta_2), \dots, P(\beta_n)$ to the alternatives before obtaining experimental evidence. Since the probabilities primarily reflect the researcher's subjective judgment or degree of belief they constitute *prior probabilities*.

The researcher can obtain experimental evidence by collecting data from a given study. Let us denote this data by A , also let the data be observed under a specific alternative β_j . The conditional probability $P(A|B)$ may be computed. This will allow the determination of the probability of an alternative β_j given the experimental evidence A .

The likelihood function $P(\beta_j | A)$ represents the likelihood of a sample result β_j given A. The posterior distribution is obtained when prior information is combined with sample information and the result revised using a set of data. This combination is done according to Bayes Theorem.

Since information concerning random variables can be periodically revised as additional sample evidences are obtained, the current posterior distributions become the future prior distribution.

Let $P_Y(y)$ and $f_Y(y)$ be the prior probability or probability density functions of Y and $f(x | y)$ be the likelihood function. The posterior probability of Y is given by the sample evidence x is given as follows:

$$P(y | x) = \frac{f(x | y) P_Y(y)}{\sum_Y f(x | y) P_Y(y)} \quad (1.2.1)$$

for a discrete case. Similarly, for a continuous case we obtain

$$f(x | y) = \frac{f(x | y) f_Y(y)}{\int_Y f(x | y) f_Y(y) dy} \quad (1.2.2)$$

The following example illustrates the Bayes theorem in a study to establish the relationship between smoking and lung cancer. Scientists in a large medical center suspected that of all the smokers who were suspected to have lung cancer, 90% did, while only 5% of the non-smokers who were suspected to have lung cancer did. The proportion of smokers was 0.45. We can use

Bayes theorem to calculate the probability that a lung cancer patient who was selected by chance was a smoker as follows.

Let k_1 and k_2 be events that a patient is a smoker or a non-smoker respectively.

Let c be the event that a patient has lung cancer, also let k_1 and k_2 are the alternatives that may prevail.

Their prior probabilities are 0.45 and 0.55 respectively. Whether a patient has lung cancer or not he or she may be affected by whichever of the two alternatives. The conditional probability of c given k_1 is

$$P(c|k_1) = 0.9$$

And that of c given k_2 is

$$P(c|k_2) = 0.05$$

We wish to determine the posterior probability of selecting a smoker.

For a discrete case we have the following computation

$$\begin{aligned} P(k_1|c) &= \frac{(0.45)(0.9)}{(0.45)(0.9) + (0.55)(0.05)} \\ &= 0.9364 \end{aligned}$$

Therefore the probability that a randomly selected lung cancer patient is a smoker is 0.9364

1.3 Prior distribution

The prior distribution has two consequences on the unknown parameter, first the probability density function will often contain information about the unknown parameter to the extent that the information is correct thus sharpening the inference about the parameter. Second, since the

unknown parameter is a random variable, the method to be used for analysis can be greatly clarified.

Essentially, prior distributions can be interpreted as frequency distributions. They also have normative and objective representation of what is rational to believe about a parameter usually in a situation of ignorance and as a subjective measure of what a particular individual actually believes.

Sometimes the parameter value may be generated by a stable physical random mechanism whose properties are either known or can be inferred by analysis of suitable data, for example if the parameter is a measure of the properties of a batch of material in an individual inspection problem. Observations on previous batches allow the estimation of prior distribution.

If the relative frequency of occurrence of the states of nature is available prior to obtaining any observations, prior probabilities can be used to weigh the average loss for each state to obtain an expected loss for each strategy

1.4 Posterior distribution

The posterior distributions are influenced by the correctness of the prior and data collected.

They are usually convenient to use because specifications have to be made. For posterior distributions that are nearly normal, the mean and standard deviation can be used for analysis. A transformation of the parameter may be required if it is sufficiently simple. When parameters are multidimensional, the mean and covariance matrix can be used if the posterior distributions are approximately normal.

If the parameter is usually uni-dimensional and the posterior distribution unimodal, the upper and lower confidence limits are constructed to be used in analysis.

1.5 The Loss function

The success of a given function is the accomplishment of its objective. The loss function measures the penalty that arises from taking a particular action depending on a given decision function. The Bayesian approach considers the loss function because the posterior density cannot be used to determine a point estimate of a given parameter. The loss function is a non negative function and is a function of a random parameter.

Let $f_{\theta}(\theta)$ be the prior density of a random parameter θ . Also let $L(x_1, \dots, x_n | \theta)$ be the likelihood function of a random sample of n variables.

Let $f(\theta | \underline{x})$ be the posterior density and let $l(d, \theta)$ be the loss function. The Bayes estimate for θ say

$T = U(x_1, x_2, \dots, x_n)$ is one in which the expectation of the loss function given by

$$E(l(d, \theta)) = \int_{\theta} l(d, \theta) f(\theta | \underline{x}) d\theta \quad (1.5.1)$$

is a minimum.

To determine the Bayes estimator, one must specify a loss function. Specification is difficult because economic consequences are not easily measurable. For many applied problems a reasonable argument can be made using a loss function in the form

$l(d, \theta) = (d - \theta)^2$ which gives the **squared error or quadratic loss function**. For a quadratic loss function, the Bayes estimate of θ is the posterior expectation $E(\theta | \underline{x})$ of θ .

Risk function

In a decision theory and estimation theory, the expected value of the loss function gives the **Risk function** which is denoted by $R(\theta, \delta)$. The main objective of a decision maker is to minimize the risk function. For a decision function to exist, the risk must be greater than zero. If prior distribution for θ is unknown then the risk function automatically becomes a random variable. The expected value of the risk function measures the overall effectiveness of the decision function.

1.6 Bayes solutions of a statistical decision problem against a specified priori distribution

Suppose a value θ occurs in accordance with some apriori probability function $q(\theta)$, $\theta = \theta_1, \dots, \theta_n$. Then for a given decision function d in the available class D , the risk function $R(\theta, \delta)$ would be averaged over θ with respect to a prior distribution $q(\theta)$ to yield the average risk $\bar{r}(q, d)$ as follows:

$$\bar{r}(q, d) = \sum_{i=1}^n r(\theta_i, d) q(\theta_i) \quad (1.6.1)$$

$$= \sum_{t=1}^n \sum_{i=1}^n l(\theta_i, d)(x_t) p(x_t | \theta_i) q(\theta_i) \quad (1.6.2)$$

Where

$$\bar{r}(q, d^*) \leq \bar{r}(q, d) \quad (1.6.3)$$

A decision function which minimizes $\bar{r}(q, d)$ is the Bayes solution relative to a particular priori distribution $q(\theta)$

Since $q(\theta_i) > 0$ $i = 1, \dots, n$ and

$$q(\theta_1) + \dots + q(\theta_n) = 1 \text{ then}$$

$q(\theta_1) \dots q(\theta_n)$ can be expressed as a point on the $(n-1)$ dimensional simplex G in the Euclidean space R_n which is spanned by n points $(1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (0, \dots, 0, 1)$

We can then minimize the risk function as follows

$$\bar{r}(q, d) = \sum_{t=1}^n \sum_{i=1}^n l(\theta_i, d)(x_t) Q(\theta_i | x_t) p_q(x_t) \quad (1.6.4)$$

Where

$$Q(\theta_i | x_t) = \frac{p_q(x_t | \theta_i) q(\theta_i)}{p_q(x_t)} \quad (1.6.5)$$

And

$$p_q(x_t) = \sum_{i=1}^n p_q(x_t | \theta_i) q(\theta_i) \quad (1.6.6)$$

$Q(\theta_i | x_t)$ is a posterior probability that $\theta = \theta_i$ and $\mathcal{X} = x_t$ given a priori distribution $q(\theta_i)$.

For a given $q(\theta_i)$ and t , let

$$\sum_{i=1}^n l(\theta_i, d^*)(x_t) Q(\theta_i | x_t)$$

for all points $(l(\theta_1, d)(x_t), \dots, l(\theta_n, d)(x_t))$ in F for a fixed t .

For a fixed t , $(Q(\theta_1 | x_t), \dots, Q(\theta_n | x_t))$ is a point in G . Thus the sequence of points

x_1, x_2, \dots, x_n in the sample space R determines a sequence of vectors in F and a corresponding sequence in G .

If F is extended to include vectors $l(\theta_1, d(x_t) Q(\theta_1 | x_t), \dots, l(\theta_n, d(x_t) Q(\theta_n | x_t)$

for each t and for all d in D , and also the minimizing vector

$$l(\theta_1, d^*(x_t) Q(\theta_1 | x_t), \dots, l(\theta_n, d^*(x_t) Q(\theta_n | x_t)) \quad (1.6.7)$$

for each t , then E and hence D are closed and provide a minimum such that D will contain at least one d^* so that the equation below is obtained.

$$\sum_{t=1}^n \sum_{i=1}^n l(\theta_i, d^*(x_t) Q(\theta_i | x_t)) p_q(x_t) \leq \bar{r}(q, d) \quad (1.6.8)$$

For any d in D Each d^* provides a Bayes solution for each of the decision problem against a priori distribution $q(\theta)$

1.7 Literature Review

Inferences need to be made by combining information provided by prior probabilities with that given sample data. The combination is achieved by repeated use of Bayes theorem and the final inferences expressed by the posterior probabilities. This concept was outlined by Lindley (1972) on page xi of his research work. He further described a two sided test of a simple hypothesis testing. He discovered that opinions differed in appropriate form and difficulties in interpretation arose. He also studied that if the prior distribution is sensibly constant over a given range of θ for which the likelihood function is obtained and not too large over that range for which the likelihood function is small, then the posterior distribution is approximately equal to the normalized likelihood function.

Jeffrey (1961) in Chapter 5 of his book explains that for effective hypothesis testing, prior information of the test statistic should be in two parts; a prior discrete probability and posterior. He proposed that in the absence of any prior knowledge, the posterior priori should be divided equally. He also discussed the Bayes-Laplace's principle of insufficient reason which stated that "...if there is no reason to believe one hypothesis rather than another, the probabilities obtained are equal, to say that the probabilities are equal is equal to have no good ground for choosing between alternatives."

Brown L (1966) discusses the admissibility of invariant estimators of one or more location parameters.

James W and Stein C (1961) analyzed estimation procedures using the quadratic loss. The theory of statistical decision functions was initiated and developed by Wald (1950). He also showed that under general conditions the set of admissible decision functions forms a complete class and that the set of all Bayes solutions considered later is complete. This theory followed the development of the theory of likelihood functions developed by Von Neumann and Morganstern (1947).

Hard (1986c, 1987a) on studying Bayesian analysis of disease prevalence discovered that the aims of screening and estimation of prevalence are different, so a threshold must be chosen with a particular aim, for prevalence estimation the aim was to minimize variance whereas with screening the aim was to maximize accuracy. He investigated the choice of threshold separately for each aim using a different sample scheme.

Geisser (1982) reviewed the Bayesian approach to discriminant analysis. His approach was based on the concept of predictive density of the likelihood function.

According to Aitchison and Dunsmore (1975) in Chapter 11 of his work, the predictive approach of prior probabilities was first explicitly presented by Geisser (1964) for multivariate normal group conditional distributions. Dunsmore (1966) gave similar results on studying relatively large samples and estimative approaches to posterior probabilities. However, for small samples there can be dramatic differences. Aitchison and Kay (1975) infer that large samples and small samples are compared under normal models, the estimates produced by the estimative approach are corrected for bias, and then the differences between the approaches considerably reduced.

Rao C.R (1968) discussed the prepositions characterizing complete and minimal classes of decision rules. He suggested that every admissible rule is a Bayes rule with some prior distribution. He also found out that if the loss vector corresponds to a rule, then the totality of vectors varies until all its components are negative.

The 2008 Bayesian applications modeling workshop identified that in constructing a Bayesian Model, probabilistic information is required for establishing the numerical-parameters of the model. The probabilistic information judgments are known to be biased as a result of heuristics human use of assessing probabilities.

Bayesian and Non-Bayesian approaches were studied by Patrick Lam (1981). He found that in Non-Bayesian approach parameters are fixed and their true values unknown, objective notion of

probability is based on sampling and large sample properties have asymptotic approximations. In Bayesian approach, parameters are random variables with distributions attached to them. Subjective notion of probability (prior) combined with data does not require large sample approximations.

Lindsay and Smith (1972) studied the technical property that allows one to treat parameters as random variables called exchangeability of the observation units. International Association of Bayesian Analysis (ISBA) 1992 was founded to promote the development and application of Bayesian Analysis in the solution of theoretical and applied problems in science and industry.

DeGroot (1970) gave a careful account of the axiomatic basis of subjective probability as well as the general development linked to decision making. He discussed of parameters in multivariate models using discriminant analysis.

Alba and Van Ryzin (1979a) introduced an approach to outlier detection based on non standard empirical Bayes framework. The model used considered that out of a given sample of size n , $(n-k)$ of the random variables had equal mean and variance while the remaining k random variables had the same mean but larger variance.

Robins considered the compound decision problem where a statistician was interested in minimizing the overall frequency of errors in n identical but unrelated decision problems, n values were considered as a sample of n independent observations on a single random variable with a fixed priori distribution.

1.8 Problem statement

One of the most important tasks a researcher faces is proper analysis of data. Sometimes subjective judgment prevails leading to bias while making inferences in a given study. There are situations where unknown parameters cannot be realistically determined, for example a manufacturing process is likely to fluctuate depending on a number of factors like the efficiency of machines, type of raw materials that are used in the production process among others. Different tests can be performed on a given set of numerical data in order to obtain more reliable results that can be used in decision making and quality control especially in an industrial process. Companies employ different strategies to reduce the occurrence of errors in the daily production.

Most firms and companies are affected by poor decision making sometimes leading to loss of revenues. The Bayesian approach has been greatly favored in cases where there are many factors affecting an outcome and a decision has to be made on a fixed quantity.

The present research work has concentrated on how the weight of a given species of a pig is affected by the levels of calcium, carbohydrates and proteins in its blood. The data was collected from a breeding research centre. Bayes multiple regression has been used to model the relationship between the three variables and formulate a regression equation for predicting future weights based on the three predictor variables.

1.9 Aims and Objectives

- i. To use Bayesian inference and multiple regression to model the relationship between the predictor and response variables that determine the weight of a certain species of a pig.
- ii. To discuss how prior and posterior distributions affect parameter estimation in Bayesian inferencing

CHAPTER 2

PROBLEM IDENTIFICATION AND DECISION RULES

2.1 Introduction

Bayes rule requires perfect knowledge of a priori distribution which can be assumed in practice. Therefore even if extensive past experience is available one would in most cases use Bayes rule for a slightly slipped distribution from a true priori distribution. We shall investigate the problem of delivering allocation procedures if the prior distribution is assumed to be known. If the prior distribution depends on some unknown parameter an adaptive two stage allocation procedure will be proposed, that is an observation is made after each observation or group of observations.

A decision function is a function defined on a sample space whose range consists of actions to be taken during a research or an experiment. A decision function with the most desired properties shall be selected.

The chapter also deals with Bayesian estimation, convergence of Bayes estimation and comparison of Bayes estimates with other estimates.

2.2 Randomized and Non-Randomized decision Rules

A **non-randomized decision rule** assigns to each possible value of $x \in X$ one of these two decisions and therefore divides the sample space into two corresponding regions ω_0 and ω_1 . If x falls into ω_0 a given hypothesis is accepted, otherwise it is rejected. ω_0 is therefore the region of acceptance and ω_1 the region of rejection.

When performing a test one may arrive at the correct decision. This may not always be the case because sometimes **type I** and **type II** errors occur.

A **randomized test** selects among the decisions, rejection or acceptance certain probabilities that depend on x . The probabilities are denoted by $\phi(x)$ and $1 - \phi(x)$ respectively. If this value of x is taken, a random experiment is performed with two possible outcomes R or \bar{R} . If in the experiment, R occurs the hypothesis is rejected with probability $\phi(x)$ otherwise it is accepted.

A randomized test is therefore completely characterized by the critical function ϕ which varies from 0 to 1 for all values of x .

Statistical inference is concerned with using probability concepts to deal with uncertainty in decision making. It involves selecting and using a sample statistic to draw inference about a parameter. It also involves hypothesis testing and estimation. A hypothesis is a supposition made as a basis for reasoning. It also aims at arriving at conclusions or decisions concerning parameters of a population on the basis of information contained in a sample. We thus make a statistical inference when we estimate.

Problem of statistical inferences have been classified into problems of estimation and problems testing hypothesis.

Problem of estimation require that we provide values for unknown parameters of distributions while hypothesis testing deals with reaching conclusion or decisions concern assumed values of parameters.

Statistical estimation is concerned with the methods by which the population characteristics are estimated from sample information. There are two types of estimates, point estimates and interval estimates.

A **point estimate** is a single number which is used as an estimate of the unknown population parameter. For this case, a random sample of n observations x_1, x_2, \dots, x_n from a population $f(x, \theta)$ is selected.

Suppose (x_1, \dots, x_n) is an n-dimensional random variable from a cumulative distribution function $F_n(x_1, \dots, x_n; \theta)$ where θ is a one-dimensional real parameter with a parameter space Ω .

Let $\tilde{\theta}(x_1, \dots, x_n)$ or more briefly $\tilde{\theta}$ be a function of (x_1, \dots, x_n) where $\tilde{\theta}$ itself is a random variable. If the observed value of $\tilde{\theta}$ corresponding to the observed value of (x_1, \dots, x_n) is used for θ_0 , the true value of θ . The random variable $\tilde{\theta}$ is called a point estimate or estimator for θ_0 .

If an estimator $\tilde{\theta}$ converges in a probability to θ_0 and $n \rightarrow \infty$ it constitutes a **consistent estimator** for θ_0 . If $\tilde{\theta}$ is an unbiased estimator for θ_0 having finite variance and no other unbiased has a smaller variance, then $\tilde{\theta}$ is an **efficient estimator** of θ_0 .

If $\tilde{\theta}$ is a statistic such that for any other statistic $\tilde{\theta}'$, the distribution of the conditional random variable $\tilde{\theta}' | \tilde{\theta}$ does not depend on θ_0 then it is a **sufficient value** of θ_0 .

An **interval estimate** of a population parameter is a statement of two variables between which it is estimated that the parameter lies. It makes use of two end points which are functions of observed random variables.

Suppose (x_1, \dots, x_n) is a random variable having cumulative distribution function $F_n(x_1, \dots, x_n; \theta)$. Let $\underline{\theta}(x_1, \dots, x_n)$, $\bar{\theta}(x_1, \dots, x_n)$ be two functions (random variables) from a sample of elements such that $\underline{\theta} < \bar{\theta}$. If the sample $\underline{\theta}$ and $\bar{\theta}$ can be chosen so that for a given value Y

$$p(\underline{\theta} < \theta < \bar{\theta} | \theta) = Y$$

Where

$$p(\underline{\theta} < \theta < \bar{\theta} | \theta) = Y \text{ denotes the indicated probability } F_n(x_1, \dots, x_n; \theta) \text{ then}$$

$(\underline{\theta}, \bar{\theta})$ is called a 100Y % confidence interval for θ and $\underline{\theta}$ and $\bar{\theta}$ are the lower and upper

confidence limits for θ . Y is called the confidence coefficient. $(\theta - \bar{\theta})$ is a two dimensional random variable such that the probability is Y and the interval $(\theta - \bar{\theta})$ contains the true value of θ in $F_n(x_1, \dots, x_n; \theta)$.

2.3 Construction of Bayes Decision functions

Suppose that for a given probability density function it is desired that a decision function δ which minimizes the risk function $f(x, \delta)$ is to be computed

There are three factors supporting the use of Bayes Decision Rule as a reasonable prescription for action. These factors are:

- (i) It has formal optimality in utility theorem terms.
- (ii) It has inevitable inadmissibility – a decision rule $\delta(x)$ is said to be admissible if there's no other decision rule which dominates it.
- (iii) It has intuitive appeal when tangible prior information about θ is available.

Minimax Decision Rules

In the absence of a specified prior distribution for θ , a principle is sometimes advanced for singling out a decision rule, which consists of choosing that decision rule for which the maximum risk over Ω is as small as possible. This is referred to as **minimax decision rule** or **minimax principle**. The minimax principle considers the expected loss when population distribution is fixed, a function of a sample random variable or a function of the estimator. The expected loss should be as small as possible.

2.4 Comparison of Bayes Estimates with other Estimates

Under very weak restrictions, Bayes estimates and their limits form a class. The weak function holds, thus it should be possible to exhibit prior measures or sequences leading to any and all risk functions obtained by other methods like the maximum likelihood estimation and maximum probability estimation. If the Bayes estimate of a population mean with respect to a Dirichlet prior with parameter x has given rise to the interpretation that $\alpha(X)$ is the prior sample size and if $\alpha(X)$ is made to tend to zero, then the Bayes estimate mathematically converges to the classical estimator called the sample mean.

2.5 Convergence of Bayes Estimates

We shall be mainly interested in the limits of the Bayes estimates of various functions say

$\phi(p)$ as $\alpha(X) \rightarrow 0$.

where $\alpha(X)$ is the prior sample size

We will therefore make the following assumptions

$$\alpha_r(X) \rightarrow 0 \text{ and } \sup_A |\bar{\alpha}_r(A) - \bar{\alpha}_0(A)| \rightarrow 0 \quad (2.5.1)$$

\sup_A denotes the supremum of A.

Where $\bar{\alpha}_0$ is the probability measure in p.

We will also be interested in β a special class of functions $\phi(p)$ as defined below.

Let g be a permutation invariant measurable function from X^k into R^1 such that

$$\int |g(x_1, \dots, x_1, x_2, \dots, x_1, \dots, x_m, \dots, x_m)| d_{\bar{\alpha}}(x_1), \dots, d_{\alpha}(x_m) < \infty \quad (2.5.2)$$

For all possible combinations of arguments $(x_1, \dots, x_1, x_2, \dots, x_1, \dots, x_m, \dots, x_m)$ from all distinct ($m=k$) to all identical ($m=1$). g vanishes whenever any two co-ordinates are equal and the condition (2.5.2) reduces to the simple condition

$$\int |g(x_1, \dots, x_k)| d_{\bar{\alpha}}(x_1), \dots, d_{\alpha}(x_k) < \infty \quad (2.5.3)$$

We shall define a parametric function

$$\phi_g(p) = g(x_1, \dots, x_m) dp(x_1), \dots, dp(x_m) \quad (2.5.4)$$

for all those p's for which it exists. Let p have D_{α} as the prior distribution and let (x_1, \dots, x_n) be a sample from p. Under further assumptions concerning the second moment of g under $\bar{\alpha}^k$, the Bayes estimate of (with respect to the squared error loss) of $\phi_g(p)$ based on this sample is

$$\hat{\phi}_{g, \alpha}^n = \mathbb{E}_{D_\alpha} (\phi_g(p) \mid x_1, \dots, x_n) \quad (2.5.5)$$

and based on no samples is

$$\hat{\phi}_{g, \alpha}^0 = \mathbb{E}_{D_\alpha} (\phi_g(p)) \quad (2.5.6)$$

Since the conditional distribution of p given (x_1, \dots, x_n) is $D_{\alpha + nF_n}$,

where F_n is the empirical distribution function of (x_1, \dots, x_n) we have

$$\hat{\phi}_{g, \alpha}^n = \hat{\phi}_{g, \alpha + nF_n}^0 \quad (2.5.7)$$

Suppose that we substitute $\alpha = \alpha_r$ where $\{\alpha_r\}$ satisfies (2.6.1)

From the results we obtain

$$D_{\alpha_r} \longrightarrow \delta_0 \text{ weakly} \quad (2.5.8)$$

and

$$D_{\alpha_r + nF_n} \longrightarrow D_{nF_n} \text{ as } r \longrightarrow \infty \quad (2.5.9)$$

The main result shows the convergence of Bayes estimates

$$\hat{\phi}_{g, \alpha_r}^0 \quad \text{and} \quad \hat{\phi}_{g, \alpha_r + nF_n}^0$$

2.6 Bayes Point estimates

There may be situations where it may be convenient to choose a single value as an estimate of θ . The value with the highest posterior probability is thus chosen. Consequently a Bayesian Point estimate is defined as the quantity $\bar{\theta}(x)$ which maximizes $\pi(\theta|x)$. The highest of the posterior distribution is mostly preferred in choosing the value of θ . The estimate is not invariant with respect to the transformations in the parameter space.

2.7 Bayesian confidence Regions

A more informative $\pi(\theta|x)$ for practical purposes is obtained by stating that θ lies in some region Ω with a prescribed probability. The region say

$S_\alpha(x)$ is a $100(1 - \alpha)\%$ Bayesian confidence region for θ if

$$\int_{S_\alpha(x)} \pi(\theta|x) = 1 - \alpha \quad (2.7.1)$$

$S_\alpha(x)$ can be chosen to satisfy 2.5.1

$1 - \alpha$ is called the confidence interval. We must always ensure that

$$P[\theta \in S_\alpha(x)] \geq 1 - \alpha; \quad (2.7.2)$$

$1 - \alpha$ is called the confidence level which occurs in cases where θ has a discrete component.

We shall choose a small value of α . For example 90%, 95%, 99% among confidence although the chance is arbitrary.

If $\pi(\theta|x)$ is unimodal, a finite interval for $S_\alpha(x)$ is obtained. Here $S_\alpha(x)$ takes the form

$$(\underline{\theta}_\alpha(x), \bar{\theta}_\alpha(x)) \text{ where}$$

$$\int_{\underline{\theta}_\alpha(x)}^{\bar{\theta}_\alpha(x)} \pi(\theta|x) = 1 - \alpha \quad (2.7.3)$$

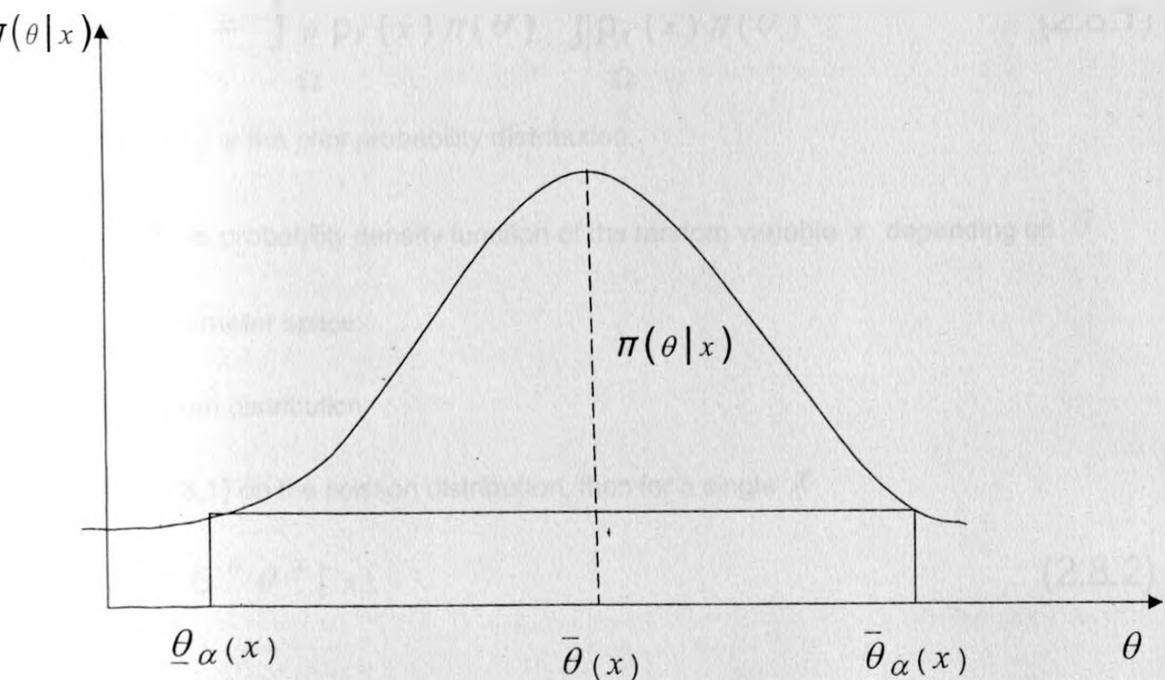
and

$$\pi(\theta|x) \geq \pi(\theta'|x) \quad (2.7.4)$$

for any

$$\theta \in S_\alpha(x), \theta' \in S_\alpha(x) \quad (2.7.5)$$

The Bayesian confidence region may be graphically represented as follows:



The interval $[\underline{\theta}_\alpha(x), \bar{\theta}_\alpha(x)]$ on the graph represents the relative likelihood that a true value θ of lies in the region prior to the observations of any $\pi(\theta|x)$.

2.8 Empirical Bayes methods

This was studied by Robbins and Maritz (1970). Maritz broadly defines the empirical Bayes approach as follows

“It may be regarded as part of the development towards more effective utilization of all relevant data in statistical analysis. Its field of application is expected to lie where there’s no conceptual difficulty in postulating the existence of a prior distribution that is capable of a frequency interpretation and where data suitable for estimation of the prior distribution may be accumulated.”

Empirical Bayes approach often employs classical methods of estimation for finding estimates, for example the prior distribution based on prior sample data. To illustrate Bayes empirical procedure we shall consider estimating the mean of a possible distribution. With a quadratic loss structure, the optimal point estimator of θ is the mean of the posterior distribution. That is we would estimate θ by

$$\bar{\theta}_{\pi}(x) = \frac{\int_{\Omega} \theta p_{\theta}(x) \pi(\theta) d\theta}{\int_{\Omega} p_{\theta}(x) \pi(\theta) d\theta} \quad (2.8.1)$$

Where $\pi(\theta)$ is the prior probability distribution.

$p_{\theta}(x)$ is the probability density function of the random variable x depending on θ

Ω is the parameter space.

θ is the mean distribution

Applying (2.8.1) on the poisson distribution, then for a single x

$$p_{\theta}(x) = e^{-\theta} \theta^x / x! \quad (2.8.2)$$

and

$$\bar{\theta}_{\pi}(x) = (x + 1) \phi_{\pi}(x + 1) / \phi_{\pi}(x) \quad (2.8.3)$$

Where

$$\bar{\theta}_{\pi}(x) = \int_{\Omega} p_{\theta}(x) \pi(\theta) = \frac{1}{x!} \int_{\Omega} \theta^x e^{-\theta} \pi(\theta) \quad (2.8.4)$$

Which is the likelihood function smoothed by the prior distribution θ .

If the prior distribution was known we would have in (2.8.2) a reasonable estimate of θ .

In typical empirical Bayes situation we assume that we have, in addition to the current observations when the parameter value θ is a set of previous observations x_1, x_2, \dots, x_n obtained when the parameter values were $\theta_1, \theta_2, \dots, \theta_n$.

θ_i arises as a random sample from the prior distribution $\pi(\theta)$ and x_i ($x_i = 1, \dots, n$) are independent sample observations.

Suppose in a given set of data an observation i occurs $f_n(i)$ times ($i = 0, 1, \dots$) the

x_i may be regarded as a random sample from the smoothed likelihood function 2.8.3 since θ_i are assumed to arise at random from $\pi(\theta)$. Thus a simple classical estimate $\phi_n(i)$ is given by

$$f_n(i) | (n+1) \quad \text{for } i \neq x \quad \text{or } [1 + f_n(x) | (n+1)] \quad \text{for } i = x \quad (2.8.5)$$

The Bayes point estimate is then estimated by

$$\bar{\theta}_n(n, x) = (x + 1) f_n(x) | [1 + f_n(x)] \quad (2.8.6)$$

Efficiency or optimality of such an empirical Bayes procedure are complicated since they must take into account possible variations in the parameter values as well as sampling fluctuations in $\phi_n(n, x)$ arising from different sets of previous data which might be encountered.

2.9 Bayes strategy

For multivariate estimation of parameters many characteristics of phenomenon shall be observed by say p . Let the characteristics collected correspond to a $p \times 1$ vector z . Furthermore U , different populations shall be given from which the z vectors z originate. Let there be U classes $\omega_1, \dots, \omega_u$. Thus each vector z has to be allocated to one of these classes.

Let the occurrence of the class ω_i be random which implies that ω has to be considered as a discrete random variable whose values ω_i have probabilities $p(\omega_i)$, hence $p(\omega_i)$ is the probability density function of ω . Furthermore, let z be a random vector. It's conditional density function given that class ω_i occurs is denoted by $p(z | \omega_i)$.

Illustration

If the vector z contains the coordinates of grid points of capital letters into which the letters have been divided for an automated reading, then z has to be assigned to one of the 26 Classes ω_i . The letter E appears in English more than the letter F. The probability $p(\omega_i)$ of the occurrence of different classes therefore varies. The conditional density function $p(z | \omega_i)$ of certain arrangements of grid points also varies as a function of the letters. For example, suppose the density function of the configuration which expresses the letter D, is greater for the letters C, E and G than for the letters F, D and J.

If $p(\omega_i)$ and $p(z | \omega_i)$ are known and if a vector z of characteristics has been observed, then with reference to Bayes formula for a discrete case we obtain

$$p(\omega_i | z) = \frac{p(z | \omega_i) p(\omega_i)}{\sum_{j=1}^u p(z | \omega_j) p(\omega_j)} \quad \text{for } i \in (1, \dots, u) \quad (2.9.1)$$

Intuitively one will assign the vector \mathbf{z} to the class ω_i for which $p(\omega_i | \mathbf{z})$ will be a minimum. With a decision, a loss may be connected if e_i denotes the decision for the class ω_i based on the vector \mathbf{z} characteristics, then the loss $k(e_i | \omega_i)$ will be assumed to occur. The loss will be small for a correct decision and large for a wrong decision. Often one chooses the simple loss function.

$$k(e_i | \omega_i) = \begin{cases} 0 & \text{for } i = j \quad i, j \in (1, \dots, u) \\ c & \text{for } i \neq j \end{cases} \quad (2.9.2)$$

which imposes no loss for the correct classification and equal loss c for a wrong classification. The expected value $R(e_i | \mathbf{z})$ of the loss decision for decision e_i will be computed.

$$R(e_i | \mathbf{z}) = \sum_{j=1}^u k(e_i | \omega_j) p(\omega_j | \mathbf{z}) \quad (2.9.3)$$

$R(e_i | \mathbf{z})$ is called the conditional risk.

The decision based e_i on the vector \mathbf{z} of the characteristics establish the decision rule $\mathbf{e}(\mathbf{z})$, which for every vector \mathbf{z} taken on one of the u values e_1, \dots, e_u , the overall risk R to be expected for applying the decision rule $R(\mathbf{z})$ follows with the density function given by the equation below

$$p(\mathbf{z}) = \sum_{j=1}^u k(e_i | \omega_j) p(\omega_j | \mathbf{z}) \quad (2.9.4)$$

R is expressed as follows

$$R = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R(e(z)|z) p(z) dz \quad (2.9.5)$$

The Bayes strategy now consists of choosing a decision rule such that the risk R attains a minimum. Since $R(e(z)|z)$ and $p(z)$ are positive, the loss function is also positive. A minimum of R follows from $R(e(z)|z)$ which is obtained with (2.9.2) and (2.9.3) by

$$R(e_i|z) = \sum_{j \neq 1} c p(\omega_j|z) = c(1 - p(\omega_1|z)) \quad (2.9.6)$$

The expression attains a minimum if the posterior probability $p(\omega_j|z)$ of the class ω_j has a maximum. Hence the Bayes classification follows with:

Decide for ω_i with $i \in \{1, \dots, u\}$, if $p(\omega_i|z) > p(\omega_j|z)$

For all $j \in \{1, \dots, u\}$ with $i \neq j$

CHAPTER 3

Multiple Regression; A Bayesian inference

3.1 Introduction

In this chapter we shall fit data to a probability model and summarize the results by a probability distribution on the parameters of the model and on the unobserved quantities such as predictions for new observations. We shall also investigate and model the relationship between variables using multiple regressions. Bayesian data analysis requires knowledge about an underlying scientific problem and the data collection process. Bayesian inference is based on two equations from the Bayes theorem. In these equations as presented below θ is a vector of m continuous parameters, y is a vector of n continuous observations and f, g, h, k, p, r and t are probability density functions. The first equation is the conditional density of θ given y .

$$g(\theta|y) = \frac{k(y, \theta)}{h(y)} \quad (3.1.1)$$

where $k(y, \theta)$ is the joint density of y_1, y_2, \dots, y_n and $\theta_1, \theta_2, \dots, \theta_m$. Using equation (1.2.1) we can write $k(y, \theta) = f(y|\theta)p(\theta)$, therefore (3.1.1) becomes

$$g(\theta|y) = \frac{f(y|\theta)p(\theta)}{h(y)} \quad (3.1.2)$$

The marginal density of $h(\theta)$ can be obtained by integrating θ out of $k(y, \theta) = f(y|\theta)p(\theta)$ so that (3.1.2) becomes

$$g(\theta|y) = \frac{f(y|\theta)p(\theta)}{\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(y|\theta)p(\theta)d(\theta)} \quad (3.1.3)$$

$$= cf(y|\theta)p(\theta)$$

$$d(\theta) = d\theta_1, \dots, d\theta_m$$

The definite integral in (3.1.3) can be replaced by a constant c after integration because it no longer involves the random vector θ . C is chosen such that the posterior density integrates to 1 thus it's called a normalizing constant.

Rearranging (3.1.2) and the joint density function $f(y|\theta)$ of the data as the likelihood function $L(\theta|y)$, we obtain

$$g(\theta|y) = cp(\theta)L(\theta|y) \quad (3.1.4)$$

The second general equation for Bayesian inference shall consider a future observation y_0 . In Bayesian approach, y_0 is not independent of y because its density depends on θ . Since y_0 and θ are jointly distributed, the posterior predictive density of y_0 given y is obtained by integrating θ out of the joint conditional density of y_0 and θ given y as shown below.

$$\begin{aligned} r(y_0|y) &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} t(y_0, \theta|y) d(\theta) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} q(y_0|\theta, y) g(\theta|y) d(\theta) \quad (3.1.5) \end{aligned}$$

where $q(y_0|\theta, y)$ is the conditional density function of the sampling distribution for a future observation y_0 . Since y_0 is dependent on y only through θ , $q(y_0|\theta, y)$ simplifies and we

Obtain the following equation.

$$r(y_0|y) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} q(y_0, \theta|y) g(\theta|y) d(\theta) \quad (3.1.6)$$

3.2 A Bayesian Multiple Regression Model

Bayesian multiple regression models includes specifications of the prior distributions for parameters. Prior specification may be conjugate or diffuse. If prior distributions are formulated with very small variances so that the prior knowledge strongly influences posterior distribution of the parameters in the model, then they are called informative priors. On the other hand if the priors are formulated using large variances they may have very little effect on the posterior distributions thus they are called diffuse priors.

3.2.1 A Bayesian multiple Regression model with a conjugate prior

Bayesian models can be parameterized using the precision (τ) rather than variance δ^2

Where

$$\tau = \frac{1}{\delta^2}$$

Using the parameterization, let

$$y | \beta, \tau \text{ be } N_n(X\beta, \frac{1}{\tau} I),$$

$$\beta | \tau, \text{ be } N_{k+1}(\phi, \frac{1}{\tau} I),$$

$$\tau \text{ be gamma } (\alpha, \delta)$$

We shall assume that ϕ, V, α, δ are known parameters of prior distributions. The prior density for $\beta | \tau$ is given as shown below. K denotes the number of predictor variables so that the rank of X is $K+1$. N denotes the number of observations.

$$p(\beta | \tau) = \frac{1}{(2\pi)^{(k+1)/2} |\tau^{-1} V|^{1/2}} e^{-\tau(\beta - \phi)'V^{-1}(\beta - \phi) / 2} \tag{3.2.1}$$

According to Gelman et.al 2004 the prior density of τ is the gamma density given as

$$P_2(\tau) = \frac{\delta^\alpha}{\Gamma \delta} \tau^{\alpha-1} e^{-\delta\tau} \tag{3.2.2}$$

$\alpha > 0, \delta > 0$, and by definition

$$\Gamma(\delta) = \int_0^{\infty} x^{\delta-1} e^{-x} dx \quad (3.3.3)$$

$$E(T) = \alpha / \delta \text{ and } \text{var}(T) = \alpha / \delta^2$$

If V in (3.2.1) were a diagonal matrix with the very large diagonal elements and if δ in (3.2.2) were very close to zero then the priors would diffuse. The prior specifications in (3.2.1) and (3.2.2) have the mathematical properties illustrated in the theorem below. The joint prior for β and T are conjugate priors because they result in a posterior distribution of the same form as the prior. The following theorem will enable us to understand the properties of the prior and posterior distributions of T and β .

Theorem

Consider the Bayesian regression model in which $y|\beta, T$ is $N_n(X\beta, T^{-1})$, $\beta|T$ is $N_{k+1}(T^{-1}, V)$ and T is gamma (α, δ) . The joint prior distribution is conjugate, that is $g(\beta, T|y)$ is of the form $p(\beta, T)$

Proof

I have proved the theorem by combining (3.2.1) and (3.2.2). The joint density is given by

$$\begin{aligned} p(\beta, T) &= p_1(\beta|T) p(T) \\ &= c_1 T^{(k+1)/2} e^{-T(\beta - \phi)' V^{-1}(\beta - \phi) / 2} T^{\alpha-1} e^{-\delta T} \\ &= c_1 T^{(\alpha_* + k + 1) / 2} e^{-T(\beta - \phi)' V^{-1}(\beta - \phi) + \delta_* / 2} T^{\alpha-1} e^{-\delta T} \end{aligned} \quad (3.3.4)$$

Where $\alpha_* = 2\alpha - 2$, $\delta_* = 2\delta$; all the other factors not involving random variables are collected in the normalizing constant which we shall denote by c_1 . Using equation (3.1.4) the Joint posterior density is given by

$$= c_2 T^{(\alpha_* + k + 1) / 2} e^{-T(\beta - \phi)' V^{-1}[(\beta - \phi) + \delta_*] / 2} T^{n/2} e^{-T[y - X\beta]}$$

$$(y - X\beta) / 2$$

$$= c_2 T^{(\alpha_* + k + 1) / 2} e^{-T(\beta - \phi)' V^{-1}(\beta - \phi) + (y - X\beta)' (y - X\beta) + \delta_*} / 2$$

where $\alpha_{**} = 2\alpha - 2 + n$

and all the factors not involving random variables are also collected in the normalizing constants c_2 . By expanding and completing the square in the exponent we obtain

$$g(\beta, \tau | y) = c_2 \tau^{(\alpha_{**} + k + 1) / 2} e^{-\tau(\beta - \phi_*)' V^{-1}(\beta - \phi_*) \delta_{**}} / 2$$

where

$$V_* = [(v_*^{-1} + X'X)^{-1} \phi_*] = V_*(V_*^{-1} \phi_* + X'y) \quad (3.3.5)$$

$$\delta_{**} = \phi_*' V_*^{-1} \phi_* + \phi_*' V_*^{-1} \phi_* + y'y + \delta_*$$

3.3 Marginal Posterior Density of β

In order to carry out inferences for β , the marginal posterior density of β must be obtained by integrating τ out of the posterior density in (3.3.5). The following lemma illustrates the form of this marginal distribution. The following lemma is a byproduct of the theorem above and will help us to understand further how to compute the diagonal matrices of the posterior densities.

Lemma

Consider the Bayesian multiple regression model in which $y | \beta, \tau$ is $N_n(x\beta, \tau^{-1}I)$, $\beta | \tau$ is $N_{k+1}(\phi, \tau^{-1}V)$ and τ is gamma (α, δ) . The marginal posterior distribution $u(\beta | y)$ is a multivariate t distribution with parameters $(n+2\delta, \phi_*, W_*)$ where

$$\phi_* = (v_*^{-1} + X'X)^{-1}(v_*^{-1} \phi_* + X'y) \quad (3.3.6)$$

and

$$W_* = \left[\frac{(y - X\phi)^1 [I + X'v_*X]^{-1} (y - X\phi) + 2\delta}{n + 2\delta} \right] (v_*^{-1} + X'X)^{-1}$$

roof

the marginal distribution of $\beta|y$ is obtained by integration as

$$U(\beta|y) = \int_0^{\infty} g(\beta, \tau|y) d\tau$$

By (3.3.5) this becomes

$$c_2 \int_0^{\infty} \tau^{(\alpha_{..} + k + 1)/2} e^{-[(\beta - \phi_*)' V^{-1} (\beta - \phi_*) + \delta_{..}] \tau / 2} d\tau \quad (3.3.7)$$

$$U(\beta|y) = \int_0^{\infty}$$

Using (3.3.3) together with integration by substitution, the integral in this expression gives the posterior distribution of $\beta|y$ as

$$U(\beta|y) = c_2 \frac{\Gamma(\frac{\alpha_{..} + 2 + k + 1}{2})}{2} \frac{[(\beta - \phi_*)' V^{-1} (\beta - \phi_*) + \delta_{..}]^{-(\alpha_{..} + 2 + k + 1)/2}}{2}$$

$$= c_3 [(\beta - \phi_*)' V^{-1} (\beta - \phi_*) - \phi' V^{-1} \phi + y'y + \delta_{..}]^{-(\alpha_{..} + 2 + k + 1)/2} \quad (3.3.8)$$

$U(\beta|y)$ is a multivariate t density. Thus it becomes

$$U(\beta|y) = c_3 [(\beta - \phi_*)' V^{-1} (\beta - \phi_*) + (y - X\phi)' (I + X'V^{-1}X)^{-1} (y - X\phi) + 2\delta]^{-(\alpha_{..} + 2 + k + 1)/2}$$

dividing the expression in the square brackets by $(y - X\phi)'(I + XVX')^{-1}(y - X\phi) + 2\delta$, and multiplying the normalizing constant accordingly and replacing $2\delta - 2 + n$, we obtain the following equation

$$J(\beta|y) = C_4 \left[\frac{(\beta - \phi_*)' V_*^{-1} (\beta - \phi_*) / (n + 2\delta)}{[(y - X\phi)'(I + XVX')^{-1}(y - X\phi) + 2\delta] / (n + 2\delta)} \right]^{-(n + 2\delta + k + 1)/2}$$

$$C_4 \left(\frac{1 + (\beta - \phi_*)' W_*^{-1} (\beta - \phi_*)}{n + 2\delta} \right)^{-(n + 2\delta + k + 1)/2}$$

where w_* is as given in (3.3.7).

The equation can be recognized as the density of the multivariate t distribution with ϕ_* as the mean vector and

$$[(n + 2\delta) / (n + 2\delta - 2)] w_*$$

as the covariance matrix of $\beta|y$. This lemma will be used in computing the Bayesian point estimates of the data collected using equation (3.3.6).

3.4 Bayesian point and Interval Estimates of Regression Coefficients

A Bayesian point estimator of β is the mean of the marginal posterior density.

$$\phi_* = (V^{-1} + X'X)^{-1} (V^{-1}\phi + X'y) \quad (3.3.8)$$

a $100(1 - \omega)\%$ Bayesian confidence region for β is the highest density region in Ω such that

$$C_4 \int_{\Omega} \dots \int \left[\frac{1 + (\beta - \phi_*)' W_*^{-1} (\beta - \phi_*)}{n + 2\delta} \right]^{-(n + 2\delta + k + 1)/2} d\beta = 1 - \omega$$

(3.3.9)

Linear functions of the random vector follow the univariate t distribution, thus an important special case is given by the following expression

$$\frac{\beta_i - \phi_{*i}}{W_{*ii}}$$

where ϕ_{*i} is the i^{th} element of ϕ .

W_{*ii} is the i^{th} diagonal element of W .

A Bayesian point estimate of β_i is ϕ_{*i} .

A $100(1 - \omega)\%$ Bayesian confidence interval for β_i is given by

$$\phi_{*i} \pm t_{\omega/2, n+2} \delta W_{*ii}$$

The Bayesian estimation of β can be obtained using the generalized least squares method.

After studying the Bayesian linear models, we have considered to use a diffuse prior with $\phi = 0$, $V =$ diagonal matrix with all diagonal elements equal to a large constant say (10^6) . We have also taken ϕ and δ to be equal to a small constant say (10^{-6}) .

In this case V^{-1} is close to 0 and so is ϕ_* .

The Bayesian point estimate of (3.3.8) is approximately equal to

$$(X^T X)^{-1} X^T y$$

which is the least squares estimate.

We have illustrated how the Bayesian point estimates of β can be calculated from the data we collected from a pig breeding research center for a certain species of pigs showing the weights of pigs denoted by y and the composition of proteins(x_1), carbohydrates(x_2) and calcium(x_3) in their feeds.

Table 1

y	X ₁	X ₂	X ₃
80	356	124	55
97	289	117	76
105	319	143	105
90	356	199	108
90	323	240	143
86	381	157	165
100	350	221	119
85	301	186	105
97	379	142	98
97	296	131	94
91	353	221	53
87	306	178	66
78	290	136	142
90	371	200	93
86	312	208	68
80	393	202	102
90	364	152	76
99	359	185	37
85	296	116	60
90	345	123	50
90	378	136	47
88	304	134	50
95	347	184	47
94	327	192	50
92	386	279	91
74	365	228	235
98	365	145	158
100	352	172	140
86	325	179	145
98	321	222	99
70	360	134	90
99	336	143	105
75	352	169	32
90	353	263	165
85	373	174	78
99	376	134	80
100	361	182	54
78	335	241	175
106	396	128	80

98	227	222	186
102	378	165	117
90	360	182	160
94	291	94	71
80	269	121	29
93	318	73	42
86	328	106	56

to which are the predicted
The weight y which



\bar{Y} have been computed using the 3 plus software. The values of
 \bar{Y} are the best estimates of μ which are the Bayes estimates
 227, 378, 360
 222, 165, 182
 186, 117, 160
 71, 29, 42
 56

CHAPTER 4

Data Analysis and Conclusion

From table 1, we can form the design matrix X using x_1, x_2 and x_3 which are the predictor variables of proteins, carbohydrates and calcium respectively. The weight y shall constitute the response variable

y =	80	x =	356	124	55
	97		289	117	76
			319	143	105
	.		.	.	
	.		.	.	
	.		.	.	
86					
328	328	106	56		

$(X^T X)^{-1}$ and $X^T y$ have been computed using the S plus software. The values of

$B = (X^T X)^{-1} X^T y$ gives the point estimates of β which are the Bayes estimates

$$B_0 = 80.22477158$$

$$B_1 = 0.02924758$$

$$B_2 = -0.0194718$$

$$B_3 = 0.03703557$$

These Bayes estimates are a weighted average of the prior means and the least squares estimates.

The coefficients of β 's are estimated as random effects.

In multiple linear regression, the observation consist of a response variable in a vector y and predictor variables in matrix form. The matrix y has n elements corresponding to n observations.

Regressions are often fitted in order to make predictions. The predictive distribution is , expressed in the equation.

$$Y_{ij} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + e_i \quad (4.1.1)$$

Or

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

β_0 = Regression constant $e_i \sim N(0, \delta^2)$

β_1 = Regression coefficient for variable x_i

β_k = Regression coefficient for variable x_k

Conclusion

Bayesian posterior combines information from the prior and the data of the prior is non-informative (diffuse). The sample data dominates the prior and the Bayesian results would be close to the Maximum Likelihood Estimate results.

For any fixed prior when the sample size gets bigger, the data model increasingly dominates the prior, so Bayesian results move towards Maximum Likelihood Estimate results.

Bayesian inferences for parameters can be carried out using sample statistics of empirical joint posterior distribution for example a Bayesian point estimate of τ could be calculated as the sample mean or media of the draws of τ from the joint posterior distribution.

Using the values of β_i 's obtained the following regression equation is obtained.

$$Y = 80.225 + 0.02925X_1 - 0.01942X_2 + 0.03704X_3 + \sum_i$$

Where this equation can be used to predict Y (the weight) given the response variables x_1 , x_2 , and, x_3 of a pig

A week later the model was confirmed and tested to be true because x_1 , x_2 , and x_3 were varied. The weights of different pigs were computed using the regression equation (4.1.1) above. The results obtained were affected by some margin of error.

Areas of Further Research

Inferences regarding τ and δ^2 require knowledge of the marginal posterior distribution of $\tau|y$. Future research can concentrate on how to derive the posterior density of $\tau|y$. This research focused on how to obtain Bayesian point estimates from a given set of data. Further research can focus on interval estimation and also incorporate hypothesis tests for regression coefficients in Bayesian inferences.

This research work employed the use of the diffuse priors in Bayesian inferencing. Further work can be done using conjugate priors.

REFERENCES

1. Abraham Wald (1950) **Statistical Decision Functions** pp 10 – 14, 85 – 99
2. C. Radhackerishna Rao (1965) **Linear Statistical inference and its application** pp 491 – 497, 574 – 587
3. Wilfrid J. Dixon (1951) **Introduction to Statistical Analysis** pp 75 -88
4. Vic Barnett (1973) **Comparative statistical inference** pp 164 – 282
5. Jaya N. Srivastava (1988) **Probability and Statistics** pp 201 – 205
6. K. Matusita (1980) **Recent developments in Statistical Inference and data Analysis** pp 33-39
7. D.R. Cox (1974) **Theoretical Statistics** pp 364-406
8. Cochran W.G (1943) **On the History of Probability and Statistics** pp 297 -310
9. Yadolah Dodge (1989) **Statistical Data Analysis and Inferences** pp 61 – 70, 125
10. Shanti S. Gupta and James Berger (1982) **Statistical Decision Theory and Related topic** pp 456 – 465
11. Ian Hacking (1965) **Logics of Statistical Inferences** pp190 – 207
12. D.A.S. Fraser (1973) **Inference and Decision** pp 1 -16
13. Demetrius G Lainiotis (1974) **Estimation theory** pp 128 – 151
14. Haid A. (1981) **Statistical Theory of Sampling Inspection by Attributes** pp 125 – 177
15. Mike Goddard (1965) **Introduction to Bayesian Statistics** pp 10 – 14, 136 – 141
16. Ferguson T.S. (1973) **A Bayesian Analysis of some non parametric problems** pp209 – 230
17. George C. Cavawos (1984) **Applied probability and Statistics 2nd Edition** pp 70 – 79
18. F.A. GrayBill (1963) **Introduction to the theory of statistics 2nd Edition** pp 70 – 79
19. Bolstad William M. (1943) **Introduction to Bayesian Statistics** pp 6 – 7
20. Patrick Lam (1970) **Introduction to Bayesian Statistics** pp 2 – 3
21. Meyer P. (1970) **Introductory Probability and Statistical applications** pp 33 – 41
22. Gammerman A. (1995) **Probabilistic Reasoning and Bayesian belief Networks** pp 1 -24
23. Mary Rouncefield and Peter Holmes (1989) **Practical Statistics** pp 164 – 177
24. Lindley D.V. (1972) **Bayesian Statistics, A Review** pp 64 – 74
25. *Bayesian network(2008) modern trends in Bayesian statistics*