

Imperial College of Science Technology and Medicine
Department of Mathematics

An Assessment of Some Methods for the Analysis of Binary Data
by
Silvia Ruiz Velasco Acosta

Thesis submitted to the University of London for the degree of Doctor of
Philosophy (Ph. D.)

March 1989

ABSTRACT

This thesis focuses on two aspects of the analysis of binary data. The first one is the presence of overdispersion in data that might at first sight be expected to have a binomial distribution. The primary object is to study the asymptotic relative efficiency of two different approaches to the problem. The first approach is by modelling the overdispersion using the beta binomial distribution and the second approach is the use of linear and quadratic statistics familiar from normal theory. The work focuses on estimating the parameters which define the overdispersion in different situations such as when the index of the binomial distribution varies from one individual to another or when there are explanatory variables present.

The second part of the thesis studies the relationship between two intimately related and yet conceptually different techniques: logistic regression and discriminant analysis. However there are situations in which the two techniques become very closely related, for instance, when the distributions of x are normal with the same covariance matrix. This work studies the asymptotic relative efficiency of the likelihood ratio test based on the conditional maximum likelihood estimators obtained from logistic regression to the common tests used in discriminant analysis.

Acknowledgments

I am deeply grateful to my supervisor, Professor Sir David Cox, for all his guidance, advice and encouragement throughout this work.

A very special Thanks to my Ph. D. colleagues Angela and Alan for all the helpful discussion, for their support during the bad moments and for sharing the good ones.

Financial support was provided by the Universidad Nacional Autónoma de México, to whom I extended my gratitude.

Table of Contents

Abstract	2
Acknowledgements	3
Table of Contents	4
Chapter 1: Introduction	6
Chapter 2: Overdispersion	9
2.1 Introduction	9
2.2 Efficiency of the estimators on the beta distribution	10
2.3 Efficiency of estimation for the beta binomial model	18
2.3.1 The model	18
2.3.2 Maximum likelihood estimators	20
2.3.3 Moment estimators approach	23
2.3.4 Quasilikelihood estimators	27
2.4 Comparison of the Methods	30
2.4.1 Equal sample size	30
2.4.2 Unequal sample size	36
2.5 Index of relative dispersion	43
2.6 Simulation study	46
2.7 Generalization to a binomial response	50
2.8 Dependence of the binary response on explanatory variables	57
2.9 Generalization to more than two categories	67
2.10 Generalization to more than two categories when the response depends on explanatory variables	79
2.11 Conclusions	80

Chapter 3: Asymptotic relative efficiency of the logistic regression	
estimators to the linear discriminant function	83
3.1 Introduction	83
3.1.1 General Formulation	83
3.1.2 General Comments	87
3.2 Properties of the estimators when the explanatory variables	
are normally distributed	89
3.2.1 Fisher's linear discriminant function	89
3.2.2 Logistic regression maximum likelihood estimators	94
3.3 Asymptotic relative efficiency in terms of coefficient estimation	98
3.3.1 The effect of a single variable	98
3.3.2 No effect from the explanatory variables ($\Delta^2 = 0$)	104
3.3.3 The effect of a subset of parameters	108
3.3.4 The ratio of linear combinations of the parameters	112
3.4 Relation of linear discriminant analysis and linear regression	114
3.5 Simulation study	116
3.6 Robustness	122
3.6.1 Introduction	122
3.6.2 Departures from the assumption of equal covariance	
matrix	123
3.6.3 Departures from the assumption of normality	132
3.6.3.1 Elliptical distributions	133
3.7 Conclusions	138
References	140

CHAPTER 1: INTRODUCTION

Suppose that for each individual or experimental unit the response, Y_i , can take only one of two possible values, denoted for convenience and without loss of generality, by 1 and 0 respectively. We may write

$$E(Y_i) = P(Y_i = 1) = \theta_i, \quad (1.1)$$

$$P(Y_i = 0) = 1 - \theta_i.$$

It is often convenient to call $Y_i = 1$ a "success" and $Y_i = 0$ a "failure". Examples of such variables are : an electronic component may be defective, or may be nondefective; a test animal may die from a specified dose of a poison, or may survive. It is reasonable to call such observations binary.

Suppose now, that for $i = 1, \dots, m$ observations are available on r_i individuals, or for each individual r_i independent observations are available. For example r_i might be the number of animals in a litter and the observations be whether they have survived after ten days from birth; in this case we could say that we have observations available on r_i individuals or for one individual, the mother, we have r_i independent observations.

The observations thus have the form y_i/r_i , where y_i is the number of positive responses, successes, on r_i trials $y_i = \sum_j y_{ij}$, for $j = 1, \dots, r_i$. Sometimes it is said that the observations are grouped, the group sizes being $r_1, \dots, r_m$.

Asymptotic approximations for grouped data can be based on one of two distinct asymptotes, either $m \rightarrow \infty$ or $r \rightarrow \infty$, with r being the minimum of r_i .

If the observations on all individuals are independent and if the probability of a positive response is constant for all individuals in the same group, or for each observation in the same individual, then the distribution of Y_i given r_i is binomial with index r_i and parameter θ_i . The likelihood function under these conditions is particularly simple. In many applications, however, observations in different groups may be independent or at least uncorrelated, but there is reason to expect positive correlation within groups. Under these circumstances the variance of Y_i will be larger than that under independence. This phenomenon, known as overdispersion, arises frequently in applications.

The same phenomenon arises if the probability of a success, θ_i , is different from group to group and is in fact a random variable.

In the first half of the present thesis we deal with this phenomenon from the second point of view. We are interested in the mean and variance of the distribution of θ_i . Assuming a beta distribution for the probability of success we obtain the maximum likelihood estimators and we obtain the asymptotic relative efficiency of moment based estimators relative to the maximum likelihood estimators. We consider the asymptotic properties of the estimators as $m \rightarrow \infty$, r_i small. Finally we generalize to different situations, such as more than two categories or binomial responses.

In most investigations, we have, associated with each individual or experimental unit, a vector of covariates or explanatory variables $(x_1, \dots, x_p)$. In a designed experiment this vector of covariates generally consists of indicator variables associated with blocking and treatment factors, together with quantitative covariates concerning various aspects of the experimental material. In non randomized studies the vector of covariates would consist of measured variables thought likely to influence the probability of success. The principal objective is to investigate the relationship between the response probability θ_i and the explanatory variables $(x_1, \dots, x_p)$.

Logistic regression analysis is the easiest way of relating a set of explanatory variables to a probability. Discriminant analysis is concerned with the optimal way of separating two populations based on a set of properties available for each individual. In logistic regression for each individual there is a vector, x , of explanatory variables and a well defined conditional probability that the binary response is one, i.e., the distribution of x is not directly relevant for the definition, whereas in discriminant analysis we assume two distinct populations and within each of these populations there is a distribution of a set of properties x ; the emphasis is on how the two distributions differ most sharply. Under certain circumstances the estimators obtained with one method might be used in the other one.

The second part of this thesis deals with the asymptotic efficiency of the logistic regression estimators relative to the linear discriminant function estimators when the explanatory variables, in the case of logistic regression, or the set of properties in the case of discriminant analysis, are normally distributed with the same covariance matrix and different means in each population. We focus on the efficiency of different tests of hypotheses about the coefficients. And finally we discuss the robustness of both methods to the violation of the assumptions, mainly equal covariance matrix and normal distribution.

CHAPTER 2: OVERDISPERSION

2.1 INTRODUCTION

The possibility of overdispersion in data that might at first sight be expected to have a binomial or Poisson distribution has long been recognized (Bartlett, 1936; Cochran, 1943). There are at least two consequences to be considered. One is the effect on the estimation of means and contrasts of means. The other is the direct study of the overdispersion and this is closely connected with the estimation of components of variance. The present chapter focuses on the second matter.

There are broadly two ways to proceed. One is by specific modelling, often involving the beta binomial or negative binomial distributions (Williams, 1975; Griffiths, 1973), the other is to use the linear and quadratic statistics familiar from normal theory. At least in simple problems the latter approach has the advantage of computational simplicity and interpretability under broad assumptions, whereas the former has the advantage of full statistical efficiency under the postulated assumptions.

The primary object of the present work is to study the asymptotic relative efficiency of the two approaches.

We start by considering the asymptotic relative efficiency, ARE, in the case of the beta distribution, which is the limit distribution in the case where the number of independently distributed binary (0,1) random variables, which jointly have a binomial distribution, tends to infinity.

We calculate the asymptotic efficiency of the pooled sample mean and the mean of group means relative to the maximum likelihood estimator as well as the asymptotic

efficiency of the quasilielihood estimator relative to the maximum likelihood estimator. For the estimation of the variance we suggest the components of variance estimator from the formal analysis of variance and calculate its asymptotic relative efficiency, as well as the asymptotic efficiency of the quasilielihood estimator of the variance, relative to the maximum likelihood estimator.

We extend the discussion to the case of binomial variables with parameters n_{ij} and θ_i and to more than two categories of the response . Also we discuss the case with explanatory variables.

2.2 EFFICIENCY OF THE ESTIMATORS OF THE BETA DISTRIBUTION

Let $\{Y_{ij}\}$, $i=1, \dots, m$; $j=1, \dots, r_i$ be independently distributed binary (0,1) random variables with $E(Y_{ij}) = \Theta_i$. Suppose then that $\Theta_1, \dots, \Theta_m$ are independently and identically distributed with density

$$\frac{\theta_i^{\alpha-1}(1 - \theta_i)^{\beta-1}}{B(\alpha, \beta)}, \quad (2.1)$$

with $B(\alpha, \beta)$ denoting the beta function.

For the beta distribution the mean and the variance are given by

$$\begin{aligned} \mu_{\theta} &= E(\Theta_i) = \frac{\alpha}{\alpha + \beta}, \\ \sigma_{\theta}^2 &= \text{var}(\Theta_i) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned} \quad (2.2)$$

If the r_i are all large then $X_i = Y_i / r_i$ has a beta distribution. It is helpful for interpretation, and of some intrinsic interest, to examine the estimates of the parameters μ_θ and σ_θ^2 for a random sample taken directly from the distribution (2.1).

We can express the maximum likelihood estimators of μ_θ and σ_θ^2 , $\hat{\mu}_\theta$ and $\hat{\sigma}_\theta^2$, and their variances in terms of the maximum likelihood estimators of α and β , $\hat{\alpha}$ and $\hat{\beta}$, and their variances.

The asymptotic covariance matrix of $\hat{\alpha}$ and $\hat{\beta}$ is given by m^{-1} times

$$\frac{1}{\Psi'(\alpha)\Psi'(\beta) + \Psi'(\alpha + \beta)\{\Psi'(\alpha) + \Psi'(\beta)\}} \begin{bmatrix} \Psi'(\beta) - \Psi'(\alpha + \beta) & \Psi'(\alpha + \beta) \\ \Psi'(\alpha + \beta) & \Psi'(\alpha) - \Psi'(\alpha + \beta) \end{bmatrix} \quad (2.3)$$

where $\Psi(\cdot)$ represents the digamma function.

Approximating $\Psi'(x)$ by $2/(2x - 1)$, we have that, for large values of α and β , the covariance matrix of $\hat{\alpha}$ and $\hat{\beta}$ is

$$\frac{1}{m} \begin{bmatrix} (2\alpha - 1)\alpha & 2\alpha\beta - (\alpha + \beta) + 1/2 \\ 2\alpha\beta - (\alpha + \beta) + 1/2 & (2\beta - 1)\beta \end{bmatrix}. \quad (2.4)$$

It is interesting that to this order the variance of, say, $\hat{\alpha}$ depends only on α itself.

We note that there are better approximations for $\Psi'(x)$ than this particular one, that is with a smaller relative error (where the relative error is defined as the ratio of the difference of the true value and the approximate value to the true value), especially for small values of x . In particular we investigate $\Psi'(x) = 1/(x - .5 + .1x^{-1})$ (Cox & Lewis, 1966, pp26), which gives a better approximation particularly for small values of x ($x < 1$).

For some values of x the values of the approximations are given in table 2.1.

Approximating $\Psi'(x)$ as $1/(x-.5 + .1x^{-1})$ gives better results if we are interested in the variance of $\hat{\alpha}$ and/or $\hat{\beta}$ and/or the covariance of $\hat{\alpha}$ and $\hat{\beta}$.

Table 2.1.
Approximations for the digamma function.

x	$\Psi'(x)$	$1/(x-.5+ .1x^{-1})$	$2/(2x-1)$
.5	4.935	5.000(.01)	not defined
1.0	1.645	1.667(.01)	2.000(.21)
1.5	0.935	0.937(*)	1.000(.07)
2.0	0.645	0.645(*)	0.667(.03)
2.5	0.490	0.490(*)	0.500(*)
3.0	0.395	0.395(*)	0.400(*)
3.5	0.330	0.330(*)	0.333(*)
4.0	0.284	0.284(*)	0.286(*)
4.5	0.249	0.249(*)	0.250(*)
5.0	0.221	0.221(*)	0.222(*)

The value in brackets represents the relative error of the approximation.

** represents a relative error smaller than .002*

Table 2.2.1.
Approximate value of the variance of $\hat{\alpha}$ or $\hat{\beta}$.

α	true value	using $\Psi'(x)=1/(x-.5+.1x^{-1})$	using $\Psi'(x)=2/(2x-1)$
1	1.712	1.294(.24)	1.000(.42)
2	7.245	6.338(.12)	6.000(.17)
5	47.754	45.732(.04)	45.000(.06)
20	790.252	780.393(.01)	780.000(.01)

The value in brackets represents the relative error of the approximation.

For instance, for some values of α (or β), the values of the two approximations to the variance of $\hat{\alpha}$ (or $\hat{\beta}$) are given in table 2.2.1; for estimates of the covariance of $\hat{\alpha}$ and $\hat{\beta}$ and to different values of the ratio α/β the values of the approximations are given in table 2.2.2.

Table 2.2.2.
Approximate value of the covariance of $\hat{\alpha}$ and $\hat{\beta}$.

α		true value	using $\Psi'(x)=1/(x-.5+.1x^{-1})$	using $\Psi'(x)=2/(2x-1)$
$\beta = \alpha$	1	1.104	1.059(0.04)	0.000(1.00)
	2	5.694	4.987(0.12)	4.000(0.30)
	5	43.236	40.986(0.05)	40.000(0.08)
	20	770.748	760.995(0.01)	760.000(0.01)
$\beta = \alpha/2$	1	0.475	0.871(0.83)	-0.500(2.05)
	2	2.523	2.169(0.14)	1.000(0.60)
	5	20.285	18.672(0.08)	17.500(0.14)
	20	378.641	371.199(0.02)	370.000(0.02)
$\beta = \alpha/3$	1	0.293	1.235(3.22)	-0.667(3.27)
	2	1.551	1.367(0.11)	0.000(1.00)
	5	12.818	11.362(0.11)	10.000(0.22)
	20	248.280	241.404(0.03)	240.000(0.03)
$\beta = \alpha/4$	1	0.209	2.2865(9.94)	-0.750(4.59)
	2	1.098	1.082(0.01)	-0.500(1.45)
	5	9.184	7.800(0.15)	6.250(0.36)
	20	183.273	176.607(0.03)	175.000(0.04)

The value in brackets represents the relative error of the approximation.

Even though approximating $\Psi'(x)$ by $1/(x-.5+.1x^{-1})$ seems to give better results for calculating the mean and variance of the parameter estimates, $\hat{\alpha}$ and $\hat{\beta}$, approximating $\Psi'(x)$ by $2/(2x-1)$ has been found to perform better in the specific context of calculating the variance of the maximum likelihood estimator of the mean of

the distribution, because we are approximating the variance of $\hat{\mu}_\theta$ by

$$\left(\frac{\partial \mu_\theta}{\partial \alpha}\right)^2 \text{var}(\hat{\alpha}) + \left(\frac{\partial \mu_\theta}{\partial \beta}\right)^2 \text{var}(\hat{\beta}) + 2\left(\frac{\partial \mu_\theta}{\partial \alpha}\right)\left(\frac{\partial \mu_\theta}{\partial \beta}\right) \text{cov}(\hat{\alpha}, \hat{\beta}), \quad (2.5)$$

where

$$\frac{\partial \mu_\theta}{\partial \alpha} = \frac{\beta}{(\alpha + \beta)^2},$$

and

$$\frac{\partial \mu_\theta}{\partial \beta} = -\frac{\alpha}{(\alpha + \beta)^2}.$$

Thus the contribution of the first two terms is positive whereas the contribution of the third one is negative, making the whole approximation perform better using $\Psi'(x) = 2/(2x - 1)$. For some values of α and β , approximated values of the variance of $\hat{\mu}_\theta$ are given in table 2.3.

Thus, approximating $\Psi'(x)$ by $2/(2x - 1)$, the variance of the maximum likelihood estimator of μ_θ can be approximated by m^{-1} times

$$\frac{\alpha\beta(\alpha + \beta - 1)}{(\alpha + \beta)^4}, \quad (2.7)$$

whereas the variance of the sample mean is m^{-1} times

$$\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \quad (2.8)$$

Table 2.3.
Approximated values of the variance of $\hat{\mu}_\theta$.

	α	true value	using $\Psi(x)=1/(x-.5+ x^{-1})$	using $\Psi(x)=2/(2x-1)$
$\beta = \alpha$	1	.0760	.0294(.61)	.0625(.17)
	2	.0484	.0422(.12)	.0469(.03)
	5	.0226	.0219(.03)	.0225(*)
$\beta = \alpha/2$	1	.0779	not defined	.0494(.36)
	2	.0529	.0348(.34)	.0494(.06)
	5	.0259	.0237(.08)	.0257(*)
$\beta = \alpha/3$	1	.0482	.0220(.54)	.0439(.08)
	2	.0241	.0209(.13)	.0239(.01)
	5	.0068	.0066(.02)	.0068(*)
$\beta = \alpha/4$	1	.0429	.0108(.58)	.0384(.10)
	2	.0218	.0178(.18)	.0215(.01)
	5	.0061	.0059(.03)	.0061(*)

The value in brackets represents the relative error.

** represents a relative error smaller than .01.*

Therefore except when α and β are small, the asymptotic efficiency of the sample mean relative to the maximum likelihood estimator is

$$1 - \frac{1}{(\alpha + \beta)^2}. \quad (2.9)$$

The asymptotic relative efficiency as $\alpha + \beta \rightarrow \infty$ is one. For some values of $\alpha + \beta$ the asymptotic relative efficiencies are given in table 2.4.

We note that if both α and β tend to infinity, keeping the ratio α/β constant, the beta distribution tends to normality, and when α is fixed and β tends to infinity the beta distribution tends to a gamma distribution. For both these limiting distributions the sample mean is an efficient estimator of the mean, and this to some extent explains the full efficiency as $\alpha + \beta \rightarrow \infty$.

Table 2.4.
Asymptotic relative efficiency of the sample mean to the maximum likelihood estimator in the case of Beta distributed random variables

$\alpha + \beta$	ARE
2	.912
5	.960
10	.990
15	.996
20	.977
50	1.000

In the case of the variance, it is convenient to obtain the asymptotic variance of the maximum likelihood estimator by

$$\left(\frac{\partial \sigma_{\theta}^2}{\partial \alpha}\right)^2 \text{var}(\hat{\alpha}) + \left(\frac{\partial \sigma_{\theta}^2}{\partial \beta}\right)^2 \text{var}(\hat{\beta}) + 2 \left(\frac{\partial \sigma_{\theta}^2}{\partial \beta}\right) \left(\frac{\partial \sigma_{\theta}^2}{\partial \alpha}\right) \text{cov}(\hat{\alpha}, \hat{\beta}),$$

(2.10)

where

$$\frac{\partial \sigma_{\theta}^2}{\partial \alpha} = \frac{\beta(\beta - 2\alpha)}{(\alpha + \beta)^4}$$

and

$$\frac{\partial \sigma_{\theta}^2}{\partial \beta} = \frac{\alpha(\alpha - 2\beta)}{(\alpha + \beta)^4}$$

(2.11)

In this case approximating $\Psi'(x)$ by $1/(x + .5 + .1x^{-1})$ gives better results; nevertheless since for α and β bigger than three the difference of the value obtained approximating $\Psi'(x)$ by $2/(2x - 1)$ is shown only in the fourth significant figure, then for consistency we use $\Psi'(x) = 2/(2x-1)$. It follows that, for large α and β , m times the asymptotic variance of the maximum likelihood estimator is

$$\frac{2\alpha^2\beta^2(\alpha + \beta)^2 - 3\alpha\beta(\alpha + \beta)(\alpha\beta - (\alpha - \beta)^2)}{(\alpha + \beta)^8}. \quad (2.12)$$

For large values of m , α and β , m times the variance of the sample variance is

$$\frac{6\alpha\beta(\alpha - \beta)^2 + 2\alpha^2\beta^2(\alpha + \beta)}{(\alpha + \beta)^7}. \quad (2.13)$$

Thus for large α and β the asymptotic relative efficiency for estimating σ_{θ}^2 is

$$\frac{3(\alpha - \beta)^2(\alpha + \beta) - 3\alpha\beta(\alpha + \beta) + 2\alpha\beta(\alpha + \beta)^2}{6(\alpha - \beta)^2(\alpha + \beta) + 2\alpha\beta(\alpha + \beta)^2}. \quad (2.14)$$

For the same values of α and β as in the case of the mean the values of the asymptotic relative efficiency of the sample variance to the maximum likelihood

estimator of the variance are given in table 2.5; because the approximation gives very bad results for α and β equal to or less than two the smallest value of α presented is three.

Table 2.5.
Asymptotic relative efficiency of the sample variance to the
maximum likelihood estimator.

α	$\beta=\alpha$	$\beta=\alpha/2$	$\beta=\alpha/3$	$\beta=\alpha/4$
3	.842	.871	.857	.842
5	.919	.920	.902	.887
10	.966	.959	.945	.934
20	.985	.979	.971	.964
50	.995	.992	.988	.985

The asymptotic relative efficiency as $\alpha + \beta \rightarrow \infty$ is one, but the convergence is slower than in the case of the mean.

Of course, in both cases, mean and variance, because they are a one to one transformation of α and β , the same result is obtained if first we reparameterize the beta distribution in terms of μ_θ and σ^2_θ and then apply maximum likelihood theory directly to the reparameterized distribution.

2.3 EFFICIENCY OF ESTIMATION FOR THE BETA BINOMIAL MODEL

2.3.1 THE MODEL

Following the formulation of section 2.2 but assuming that all the sample sizes are small, we have that, conditional on θ_i , Y_i . ($= \sum_j Y_{ij}$) has a binomial distribution with density

$$\binom{r_i}{y_i} \theta_i^{y_i} (1 - \theta_i)^{r_i - y_i} . \quad (2.15)$$

The unconditional distribution of Y_i is the Beta binomial distribution, which in terms of the parameters of the beta distribution is written, writing Y_i for Y_i from now on

$$P(Y_i = y_i) = \binom{r_i}{y_i} \frac{\prod_{k=0}^{y_i-1} (\alpha + k) \prod_{k=0}^{r_i-y_i-1} (\beta + k)}{\prod_{k=0}^{r_i-1} (\alpha + \beta + k)} , \quad (2.16)$$

In terms of the expected value and the variance of $\{\Theta_i\}$, the mean and variance of the beta binomial distribution are written

$$\begin{aligned} E(Y_i) &= r_i \mu_\theta, \\ \text{var}(Y_i) &= r_i \mu_\theta (1 - \mu_\theta) + r_i (r_i - 1) \sigma_\theta^2, \end{aligned} \quad (2.17)$$

the extra term $r_i(r_i - 1) \sigma_\theta^2$ being the additional variance representing the overdispersion.

Prentice (1986) uses model (2.16) for correlated binary observations, where within an experimental unit the binary response probabilities have a common value θ_i and a pair of binary observations within the unit have a common correlation $\delta (= 1/(\alpha + \beta + 1))$. Then if there are r observations in the unit, the sum of the responses, i.e., the total number of successes in the unit, has mean $r\theta_i$ and variance $r\theta_i(1 - \theta_i)\{1 + (r - 1)\delta\}$, with δ being in the interval $(0, 1)$. Thus model (2.16) applies to positive values of correlation.

The Beta-binomial distribution might be used in the case when positive correlation among the variates exists, with $\delta = \sigma_\theta^2/(\mu_\theta(1 - \mu_\theta))$. The beta binomial distribution is still a distribution for a range of negative values of δ ; nevertheless that

case does not correspond to the formulation present here; positive values of δ give rise to overdispersion, whereas negative values of δ give underdispersion.. Underdispersion could not arise via a mixing distribution. Altham (1978) gives another way to generalize the binomial distribution in the case of correlation.

2.3.2 MAXIMUM LIKELIHOOD ESTIMATORS

It is convenient to parameterize the beta binomial distribution in terms of $\mu = \alpha/(\alpha + \beta)$ and $\pi = 1/(\alpha + \beta)$; while not exactly orthogonal, these parameters both have a clear interpretation and become orthogonal when $\alpha + \beta \rightarrow \infty$.

The likelihood in terms of μ and π is

$$P(Y_i = y_i) = \binom{r_i}{y_i} \frac{\prod_{k=0}^{y_i-1} (\mu + k\pi) \prod_{k=0}^{r_i-y_i-1} ((1-\mu) + k\pi)}{\prod_{k=0}^{r_i-1} (1 + k\pi)}. \quad (2.18)$$

We see that μ is identical to μ_θ , the mean of the distribution of $\{\Theta_i\}$ and therefore the maximum likelihood estimator of μ_θ is just the maximum likelihood estimator of μ and its variance is obtained directly from the information matrix, whereas

$$\frac{\text{var}(Y_i)}{r_i \mu_\theta (1 - \mu_\theta)} = 1 + \frac{(r_i - 1) \sigma_\theta^2}{\mu_\theta (1 - \mu_\theta)} = 1 + \frac{(r_i - 1) \pi}{1 + \pi} = 1 + (r_i - 1) \rho, \quad (2.19)$$

where ρ serves as an index of overdispersion.

We assume that the number of samples, m , is large while the numbers of

observations in each sample, r_i , are small and fixed as $m \rightarrow \infty$. In other words we are considering asymptotic variances of the estimators as the number of samples tends to infinity.

For small values of π (large values of $\alpha + \beta$), we write the likelihood of m samples as

$$\prod_{i=1}^m \binom{r_i}{y_i} \mu^{y_i} (1-\mu)^{r_i-y_i} \left\{ 1 - \pi \left(\frac{y_i(y_i-1)}{2\mu} - \frac{(r_i-y_i)(r_i-y_i-1)}{2(1-\mu)} + \frac{r_i(r_i-1)}{2} \right) \right\} \quad (2.20)$$

The covariance matrix of $\hat{\mu}$ and $\hat{\pi}$, using (2.20), is given by the inverse of

$$\begin{bmatrix} \sum_{i=1}^m \frac{r_i}{\mu(1-\mu)} - \frac{\pi r_i(r_i-1)}{\mu(1-\mu)} & \sum_{i=1}^m \frac{\pi r_i(r_i-1)(2\mu-1)}{2\mu(1-\mu)} \\ \sum_{i=1}^m \frac{\pi r_i(r_i-1)(2\mu-1)}{2\mu(1-\mu)} & \sum_{i=1}^m \frac{r_i(r_i-1)}{2} + C\pi \end{bmatrix}, \quad (2.21)$$

where C is the sum of C_i ; and C_i is given by

$$\begin{aligned} C_i = & -\frac{4r_i^2(r_i-1)\mu}{6(1-\mu)} + \frac{r_i^3(r_i-1)\mu}{3(1-\mu)} - \frac{r_i(r_i-1)(r_i-2)\mu^2}{(1-\mu)} + \frac{2r_i^2(r_i-1)(r_i-2)\mu^2}{3(1-\mu)} \\ & - \frac{9r_i(r_i-1)(r_i-2)\mu^2}{4(1-\mu)} + r_i(r_i-1)(r_i-2)(r_i-3)\mu + \frac{r_i^2(r_i-1)(r_i-2)^2\mu^2}{3} \\ & + \frac{r_i(r_i-1)(r_i-2)(r_i-3)}{4} + \frac{r_i^3(r_i-1)^2}{6} + \frac{r_i^2(r_i-1)^2}{4}. \end{aligned} \quad (2.22)$$

For small values of π , the variance of the maximum likelihood estimator of the mean is

$$\frac{\mu(1-\mu)}{\sum_{i=1}^m r_i} \left\{ 1 + \frac{\pi \sum_{i=1}^m r_i(r_i - 1)}{\sum_{i=1}^m r_i} \right\}. \quad (2.23)$$

As π tends to zero the variance of the maximum likelihood estimator tends to

$$\frac{\mu(1-\mu)}{\sum_{i=1}^m r_i}, \quad (2.24)$$

equal to the equivalent variance in the binomial model, whereas the term of order π corresponds to $\sum r_i(r_i-1)/\sum r_i$ times the variance of the maximum likelihood estimator of the mean of the beta distribution.

The maximum likelihood estimator of the variance in terms of μ and π is

$$\hat{\sigma}_{\theta}^2 = \hat{\mu}(1 - \hat{\mu}) \frac{\hat{\pi}}{(1 + \hat{\pi})}. \quad (2.25)$$

The variance of the maximum likelihood estimator of the variance, for small values of π , is

$$\frac{2\mu^2(1-\mu)^2}{\sum_{i=1}^m r_i(r_i-1)} \left\{ 1 - 2\pi \left(\frac{C}{\sum_{i=1}^m r_i(r_i-1)} + 4 \right) \right\}, \quad (2.26)$$

with C defined as in (2.21)

As $\pi \rightarrow 0$, the variance of the maximum likelihood estimator tends to

$$\text{var}(\hat{\sigma}_{\theta}^2) = \frac{2\mu^2(1-\mu)^2}{\sum_{i=1}^m r_i(r_i-1)}. \quad (2.27)$$

We will use this result to compare the moment estimator to the maximum likelihood estimator

2.3.3 MOMENT ESTIMATOR APPROACH

To estimate μ_θ we can either use the general sample mean, namely

$$\bar{y}_{..} = \frac{\sum_{i=1}^m \sum_{j=1}^{r_i} y_{ij}}{\sum_{i=1}^m r_i} = \frac{\sum_{i=1}^m y_{i.}}{\sum_{i=1}^m r_i}; \quad (2.28)$$

or the mean of the group means, which is written

$$\tilde{y}_{..} = \frac{1}{m} \sum_{i=1}^m \frac{y_{i.}}{r_i}. \quad (2.29)$$

Their variances in terms of μ and π are respectively

$$\frac{\mu(1-\mu)}{(\sum_{i=1}^m r_i)^2} \sum_{i=1}^m \frac{r_i(1+r_i\pi)}{(1+\pi)}, \quad (2.30)$$

$$\frac{\mu(1-\mu)}{m^2} \sum_{i=1}^m \frac{1+r_i\pi}{r_i(1+\pi)}. \quad (2.31)$$

When the sample sizes are equal, the two estimators are clearly the same.

It is clear that we could use other weighted estimators as the estimator of the mean. Kleinman (1973) suggests to use as a weight $w_i = r_i / (1 + (r_i - 1)\rho)$, where ρ is equal to $\sigma_\theta^2 / (\mu_\theta(1 - \mu_\theta))$ and assumed known; note that this estimator does not depend on the assumption of the beta binomial distribution. This estimator corresponds to the quasiliikelihood estimator that we will discuss later. Cochran (1943) suggests, instead of a pooled sample mean, using a weighted sum of the group means, the weights being the sample sizes until an upper limit is reached, where the upper limit is defined either as the average of the sample sizes or a fixed percent of the sample, and the rest of weights being constant, for instance, the nearest integer to the average of the sample sizes. He calculates the efficiency of these estimators relative to the estimator calculated with the true weights, where the true weights are defined as $r_i / (\theta_i(1 - \theta_i) + r_i\sigma_\theta^2)$ in the particular case of equal values of θ_i and ten sample of sizes 1, 2, 6, 8, 10, 20, 30, 40, 50 and 80, and for a series of different values of $u = (\theta_i(1 - \theta_i)/\bar{r}) / ((\theta_i(1 - \theta_i)/\bar{r}) + \sigma_\theta^2)$, from zero to one, with $\bar{r}$ the mean of the sample sizes. A difficulty in using this estimator is the choice of the upper limit and the constant weight to use in order to have more efficiency than the pooled sample mean.

For the variance we concentrate on estimation of σ_θ^2 by means of the formal mean square between groups in an analysis of variance.

The analysis of variance table is

source	d.f.	Sum of squares	E(mean square)
Between	$m-1$	$\sum_{i=1}^m r_i (\bar{y}_i - \bar{y}_{..})^2$	$\mu_\theta(1 - \mu_\theta) - \sigma_\theta^2 + \frac{(\sum_{i=1}^m r_i)^2 - \sum_{i=1}^m r_i^2}{(\sum_{i=1}^m r_i)(m-1)} \sigma_\theta^2$
Within	$\sum_{i=1}^m (r_i - 1)$	$\sum_{i=1}^m \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_i)^2$	$\mu_\theta(1 - \mu_\theta) - \sigma_\theta^2$
Total	$\sum_{i=1}^m r_i - 1$	$\sum_{i=1}^m \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{..})^2$	

Then for large m , we estimate σ_θ^2 as $m \sum r_i$ times the difference of the mean of squares between and the mean of squares within divided by $((\sum r_i)^2 - \sum r_i^2)$. Writing the mean of squares between as the difference of the total sum of squares and the sum squares within divided by $m - 1$ we have that

$$\tilde{\sigma}_\theta^2 = \frac{n_1}{n_1^2 - n_2} \left(\frac{n_1}{n_1 - m} \sum_{i=1}^m \frac{y_{i.}^2}{r_i} - \frac{m}{n_1 - m} y_{..}^2 - \frac{1}{n_1} y_{..}^2 \right), \quad (2.32)$$

where $n_j = \sum r_i^j$, $y_{i.}$ is the total number of ones in a particular group and $y_{..}$ is the total number of ones in the whole sample.

When $r_1, \dots, r_m$ are all equal to r , this reduces to

$$\tilde{\sigma}_\theta^2 = \frac{r}{m(r-1)} \sum_{i=1}^m \bar{y}_i^2 - \frac{1}{(r-1)} \bar{y}_{..}^2 - \bar{y}_{..}^2, \quad (2.33)$$

where $\bar{y}_i$ is the mean of the observations in the i th sample and $\bar{y}_{..}$ is the mean of the whole sample and questions of different systems of weighting depending on r_i do not arise.

The variance of this estimator is a linear combination of the first four moments, $\mu_1, \dots, \mu_4$, of the distribution of $\{\Theta_i\}$ where the coefficients are functions of the sample sizes; in general for large numbers of samples the variance of the estimator is given by

$$\begin{aligned}
& \left(\frac{n_1}{n_1^2 - n_2} \right)^2 \left\{ \sum_{i=1}^m \left(\frac{m^2 r_i}{(n_1 - m)^2} \mu_\theta (1 - \mu_\theta) + \left(\frac{n_1}{(n_1 - m) r_i} - \frac{1}{n_1} \right)^2 \left((-2r_i^2(r_i - 1)(\mu_\theta^3 + \mu_\theta \mu_2) \right. \right. \right. \\
& + 7r_i(r_i - 1)(\mu_\theta^2 + \mu_2) + r_i \mu_\theta (1 - \mu_\theta) + 6r_i(r_i - 1)(r_i - 2)(\mu_\theta^3 + 3\mu_\theta \mu_2 + \mu_3) \\
& - r_i^2(r_i - 1)^2(\mu_\theta^4 + \mu_2^2 + 2\mu_\theta^2 \mu_2) + r_i(r_i - 1)(r_i - 2)(r_i - 3)(\mu_\theta^4 + 4\mu_\theta \mu_3 + 6\mu_\theta^2 \mu_3 + \mu_4) \left. \right. \\
& \left. \left(\frac{2m}{n_1(n_1 - m)} - \frac{2n_1 m}{(n_1 - m)^2 r_i} \right) \left(r_i \mu_\theta - r_i^2 \mu_\theta^2 + 3r_i(r_i - 1)(\mu_\theta^2 + \mu_2) - r_i^2(r_i - 1)(\mu_\theta^3 + \mu_\theta \mu_2) \right. \right. \\
& + r_i(r_i - 1)(r_i - 2)(\mu_\theta^3 + 3\mu_\theta \mu_2 + \mu_3) \left. \right) \left. + \sum_{i=1}^m \sum_{j \neq i}^m \left(\left(\frac{4}{(n_1 - m)} - \frac{4}{n_1^2} \right) (2r_i r_j (r_i - 1) \mu_\theta^3 (1 - \mu_\theta) \right. \right. \right. \\
& + r_i r_j \mu_\theta^2 (1 - \mu_\theta) \left. \right) - \frac{4m}{n_1(n_1 - m)} r_i r_j \mu_\theta^2 (1 - \mu_\theta) + \sum_{i=1}^m \sum_{j>i}^m \frac{4}{n_1^2} (r_i r_j \mu_\theta^2 (1 - \mu_\theta) \\
& + r_i r_j (r_i + r_j) \mu_\theta^3 (1 - \mu_\theta) \left. \right) + \sum_{i=1}^m \sum_{j>i}^m \sum_{k>j}^m \frac{24}{n_1^2} r_i r_j r_k \mu_\theta^3 (1 - \mu_\theta) \left. \right\} , \quad (2.34)
\end{aligned}$$

and in the limit as $\alpha + \beta \rightarrow \infty$, using the fact that all the moments of the beta distribution but the first one tend to zero, the variance is

$$\begin{aligned}
& \left(\frac{n_1}{n_1^2 - n_2} \right)^2 \left\{ \sum_{i=1}^m \frac{m^2 r_i}{(n_1 - m)^2} \mu(1 - \mu) + \left(\frac{n_1}{(n_1 - m) r_i} \right)^2 \left(7r_i(r_i - 1)\mu^2 - r_i \mu^2 \right. \right. \\
& + 6r_i(r_i - 1)(r_i - 2)\mu^3 - 2r_i^2(r_i - 1)\mu^3 - r_i^2(r_i - 1)^2 \mu^4 + r_i(r_i - 1)(r_i - 2)(r_i - 3)\mu^4 \left. \right) - \frac{2n_1 m}{(n_1 - m)^2 r_i} \left(r_i \mu \right. \\
& - r_i^2 \mu^2 + 3r_i(r_i - 1)\mu^2 + 2r_i(r_i - 1)\mu^3 \left. \right) - \sum_{i=1}^m \sum_{j \neq i}^m \frac{4}{(n_1 - m)} \left(2r_i r_j (r_i - 1)\mu^3 (1 - \mu) + r_i r_j \mu^2 (1 - \mu) \right) \\
& + \frac{4m}{n_1(n_1 - m)} r_i r_j \mu^2 (1 - \mu) + \sum_{i=1}^m \sum_{j>i}^m \sum_{k>j}^m \frac{24}{n_1^2} r_i r_j r_k \mu^3 (1 - \mu) \left. \right\} . \quad (2.35)
\end{aligned}$$

In the case of equal sample sizes the expression for the variance reduces to

$$\begin{aligned} \text{var}(\tilde{\sigma}_{\theta}^2) = \frac{1}{mr(r-1)} \{ & 2\mu_1^2(1-\mu_1)^2 + 2\mu_2 + 4(r-2)\mu_3 + (r-2)(r-3)\mu_4 \\ & - 8(r-2)\mu_3\mu_1 + 4(r-4)\mu_2\mu_1(1-\mu_1) - r(r-1)\mu_2^2 \}, \end{aligned} \quad (2.36)$$

For small values of π , that is large values of $\alpha + \beta$, the variance is given by

$$\text{var}(\tilde{\sigma}_{\theta}^2) = \frac{1}{mr(r-1)} \{ 2\mu_1^2(1-\mu_1)^2 + \pi(2\mu_1(1-\mu_1) + 4(r-4)\mu_1^2(1-\mu_1)^2) \}, \quad (2.37)$$

and as $\alpha + \beta \rightarrow \infty$, assuming θ_i beta distributed, the variance tends to

$$\text{var}(\tilde{\sigma}_{\theta}^2) = \frac{2\mu_{\theta}^2(1-\mu_{\theta})^2}{mr(r-1)}. \quad (2.38)$$

We shall use this formula to calculate the asymptotic efficiency of the components of variance estimator relative to the maximum likelihood estimator in section 2.4.

2.3.4 QUASILIKELIHOOD ESTIMATORS

Hill & Tsai (1988) investigate the efficiency of quasilikelihood estimation for a class of a mixture models. Specifically, they assume that conditional on the sampling means, θ_i , the data, y_i , have independent distributions belonging to a natural exponential family with quadratic variance function, and that the means θ_i are independent with conjugate mixture distribution.

If we assume that the natural exponential family is the binomial and the mixture distribution is the beta distribution, we have the beta binomial model.

Assuming that the parameter, ρ , which represents the index of overdispersion is known, the quasiloglikelihood is written

$$\sum_{i=1}^m \frac{r_i}{(1 + (r_i - 1)\rho)} \left\{ y_i \log \left(\frac{\mu}{1 - \mu} \right) - \log(1 - \mu) \right\} \quad (2.39)$$

and the quasilielihood estimator of the mean is

$$\frac{\sum_{i=1}^m \frac{r_i y_i}{(1 + (r_i - 1)\rho)}}{\sum_{i=1}^m \frac{r_i}{(1 + (r_i - 1)\rho)}}, \quad (2.40)$$

which corresponds to the sample mean when the sample sizes are all equal and to the moment estimator proposed by Kleinman(1973) when they are different.

Following the results of Wedderburn(1974) and McCullagh(1983), the variance of the quasilielihood estimator is given by the expected value of the second derivative of the log quasilielihood with respect to μ .

The variance of the quasilielihood estimator of μ is then

$$\frac{\mu(1 - \mu)}{\sum_{i=1}^m \frac{r_i}{(1 + (r_i - 1)\rho)}}. \quad (2.41)$$

When ρ is unknown, which is the most realistic case, the extended

quasilikelihood estimation suggested by Nelder & Pregibon (1987) might be used. The extended quasilikelihood is defined by

$$Q'(y, \mu) = -\frac{1}{2} \log\{2\pi\rho V(y)\} - \frac{1}{2}D(y, \mu)\rho^{-1}, \quad (2.42)$$

where

$$D(y, \mu) = -2 \int_y^\mu \frac{y-u}{V(u)} du, \quad (2.43)$$

as proposed by Wedderburn (1974), and ρ is the dispersion parameter.

As for quasilikelihood theory the use of extended quasilikelihood does not assume a full distributional assumption but only the form of the two first moments.

To estimate μ and ρ we use an iterative technique, that holds one fixed and estimates the second one, and then holds the second one fixed and estimates the first one, until Q' is maximized.

In this particular case the maximum likelihood estimator of μ is the same one as in the case ρ known, and the extended quasilikelihood estimator of ρ is

$$\tilde{\rho} = \frac{\sum_{i=1}^m \frac{(r_i - 1)}{(1 + (r_i - 1)\rho)^2} \{r_i(y_i \log(\frac{\tilde{\mu}}{1-\tilde{\mu}}) - \log(1-\tilde{\mu})) - 1\}}{\sum_{i=1}^m \frac{(r_i - 1)^2}{(1 + (r_i - 1)\rho)^2}}. \quad (2.44)$$

The quasi-expected Fisher Information matrix, defined as the expected value of the second derivative of the quasilikelihood, is given by

$$\text{diag} \left(\sum_{i=1}^m \frac{r_i}{\mu(1-\mu)(1 + (r_i - 1)\rho)}, \sum_{i=1}^m \frac{(r_i - 1)^2}{2(1 + (r_i - 1)\rho)^2} \right). \quad (2.45)$$

On the basis of the properties of quasi-likelihood estimation, we calculate the quasi-likelihood estimator of σ_θ^2 as

$$\tilde{\sigma}_\theta^2 = \tilde{\mu}(1 - \tilde{\mu})\tilde{\rho}, \quad (2.46)$$

with variance

$$\text{var}(\tilde{\sigma}_\theta^2) = \frac{\partial \sigma_\theta^2}{\partial \mu} \text{var}(\tilde{\mu}) + \frac{\partial \sigma_\theta^2}{\partial \rho} \text{var}(\tilde{\rho}), \quad (2.47)$$

where

$$\begin{aligned} \frac{\partial \sigma_\theta^2}{\partial \mu} &= \rho(1 - 2\mu), \\ \frac{\partial \sigma_\theta^2}{\partial \rho} &= \mu(1 - \mu); \end{aligned} \quad (2.48)$$

therefore

$$\text{var}(\tilde{\sigma}_\theta^2) = \frac{\mu\rho^2(1 - \mu)(1 - 2\mu)^2}{\sum_{i=1}^m \frac{r_i}{(1 + r_i\rho)}} + \frac{\mu^2(1 - \mu)^2}{\sum_{i=1}^m \frac{(r_i - 1)^2}{2(1 + (r_i - 1)\rho)}}. \quad (2.49)$$

In the next section we will calculate the asymptotic relative efficiency of this estimator to the maximum likelihood estimator.

2.4 COMPARISON OF THE METHODS

2.4.1 EQUAL SAMPLE SIZE

To compare the estimators we use the asymptotic relative efficiency of the moment estimators and the quasiliquelihood estimators to the maximum likelihood

estimators.

In the particular case when all the samples are of size two the likelihood of the beta binomial model can be written as

$$(1-\mu)(1-\mu+\pi) \left(\frac{4\mu^2(1-\mu)^2}{\mu(\mu+\pi)(1-\mu)^3(1-\mu+\pi)^3} \right)^{\sum_{i=1}^m y_i} \left(\frac{((1-\mu)(1-\mu+\pi))^{1/2}}{2\mu(1-\mu)(\mu(\mu+\pi))^{1/2}} \right)^{\sum_{i=1}^m y_i^2} . \quad (2.50)$$

It can be seen easily from the likelihood that the sum of the observations and the sum of the observations squared are the joint minimal sufficient complete statistics for μ_θ and σ_θ^2 . We have that the moment estimators are the solutions of

$$\begin{aligned} \sum_{i=1}^m y_i &= mr\mu_\theta , \\ \sum_{i=1}^m y_i^2 &= mr^2\mu_\theta^2 + mr\mu_\theta(1 - \mu_\theta) + mr(r-1)\sigma_\theta^2 , \end{aligned} \quad (2.51)$$

i.e., they are functions of the minimal sufficient statistics and are unbiased estimators of the mean and variance; therefore by the Lehmann-Scheffe theorem we have that they are the uniformly minimum variance unbiased estimators and hence the maximum likelihood estimators.

When the sample sizes are bigger than two there is no reduction to two sufficient statistics and the result no longer applies.

The asymptotic efficiency of the sample mean relative to the maximum likelihood estimator of the mean for some values of r and $\alpha + \beta$, which is the function of α and β that influences it, are given in table 2.6.

Table 2.6.
Asymptotic relative efficiency of the sample mean to the maximum likelihood estimator when the sample sizes are equal.

$\alpha + \beta$	$r = \text{sample size}$				
	$r = 3$	$r = 5$	$r = 10$	$r = 20$	$r = 50$
2	.996	.989	.973	.957	.938
5	.999	.998	.995	.991	.987
10	1.000	1.000	.999	.998	.996
15	1.000	1.000	1.000	.999	.998
20	1.000	1.000	1.000	1.000	.999

The asymptotic relative efficiency increases when $\alpha + \beta$ increases and decreases as r increases but nevertheless the limiting value of the asymptotic relative efficiency as $\alpha + \beta$ tends to infinity does not depend on r and it is always one. This is explained, to some extent, because as $\alpha + \beta$ tends to infinity the beta binomial distribution tends to the binomial distribution and the sample mean is fully efficient in this case.

With equal sample sizes the quasilikelihood estimator of the mean is the sample mean and therefore its asymptotic efficiency relative to the maximum likelihood estimator of the mean is the same.

In the case of the variance the relative efficiency depends on the ratio α/β as well as on $\alpha + \beta$. Some results for different values of r and α/β are given in table 2.7.

The ratio α/β influences the relative efficiency, the components of variance based estimator becoming relatively more efficient as the ratio α/β is closer to one, that is, in terms of the distribution of θ_1 , when the distribution is closer to being symmetrical.

Table 2.7.
Asymptotic relative efficiency of the components of variance estimator to
the maximum likelihood when the sample sizes are equal.

$r = 3$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	1.000	.997	.993	.990
3	1.000	.998	.996	.995
5	1.000	.999	.998	.997
10	1.000	.999	.999	.999
20	1.000	1.000	1.000	1.000
$r = 5$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	.989	.989	.982	.976
3	.998	.993	.989	.986
5	.999	.996	.994	.993
10	1.000	.999	.998	.997
$r = 10$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	.957	.970	.964	.955
3	.992	.985	.975	.970
5	.998	.991	.986	.983
10	1.000	.997	.994	.994
$r = 20$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	.915	.947	.945	.935
3	.982	.974	.960	.951
5	.994	.984	.975	.970
10	.999	.994	.990	.988
$r = 50$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	.864	.918	.921	.911
3	.968	.959	.940	.927
5	.987	.975	.960	.951
10	.997	.989	.981	.977
20	.999	.996	.993	.991

When the sample size per group increases the asymptotic relative efficiency of the components of variance estimator to the maximum likelihood estimator of the variance decreases, but the decrease is not major. As $r \rightarrow \infty$ the limiting values of the asymptotic relative efficiencies are those corresponding to the beta distribution; we found in section 2, that for values of α and/or β less than five the efficiency was under 90%, but nevertheless for values of r as large as 50 the efficiency is over 90% when α and β are bigger than one. The limiting value of the relative efficiency as $\alpha + \beta \rightarrow \infty$ is one, but the convergence is slower than in the case of the mean.

The quasiliikelihood estimator is less efficient than the components of variance estimator, except for very small values of μ and large values of r . For small values of π , large values of $\alpha + \beta$, the asymptotic relative efficiency of the quasiliikelihood estimator to the components of variance estimator is

$$\frac{r-1}{r} \left\{ 1 + \pi \left(\frac{1}{\mu(1-\mu)} + (r-7) \right) \right\}. \quad (2.52)$$

Table 2.8 gives the values of the asymptotic relative efficiency of the quasiliikelihood estimator to the maximum likelihood estimator for the same values of sample size, α and β , as table 2.7.

As $\pi \rightarrow 0$ the variance of the quasiliikelihood estimator of the variance tends to

$$\frac{2\mu^2(1-\mu)^2}{m(r-1)^2}. \quad (2.53)$$

Table 2.8.
Asymptotic relative efficiency of the quasilielihood estimator to the
maximum likelihood when the sample sizes are equal.

$r = 3$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	.320	.355	.568	.662
3	.494	.570	.629	.662
5	.552	.605	.644	.670
10	.605	.635	.655	.669
20	.635	.650	.661	.668
50	.654	.660	.664	.667
$r = 5$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	.433	.532	.688	.808
3	.581	.680	.655	.807
5	.656	.725	.772	.805
10	.723	.761	.787	.804
20	.761	.780	.793	.802
50	.784	.792	.797	.801
$r = 10$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	.373	.585	.788	.945
3	.632	.759	.855	.920
5	.725	.810	.871	.911
10	.808	.854	.885	.905
20	.854	.877	.893	.903
50	.882	.891	.897	.901
$r = 20$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	.363	.599	.824	.995
3	.646	.789	.897	.972
5	.750	.848	.917	.964
10	.846	.898	.933	.955
20	.899	.924	.943	.955
50	.930	.939	.948	.951
$r = 50$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	.345	.596	.833	1.010
3	.643	.797	.914	.992
5	.757	.863	.939	.989
10	.862	.920	.958	.976
20	.920	.948	.960	.979
50	.956	.972	.978	.984

The asymptotic relative efficiency with respect to the maximum likelihood estimator (or the components of variance estimator) is

$$\frac{r - 1}{r}. \quad (2.54)$$

Thus the smaller the sample sizes are the worse the estimator is. For relatively big sample sizes, the loss of efficiency, for large values of α and β , is not important, especially when β is equal to 4α ; but apart from being more efficient the components of variance estimator is easier to calculate because it is based on the sum and the sum of the squares of the observations, i.e., sums of zeros and ones while the quasiliikelihood requires an iterative solution.

2.4.2 UNEQUAL SAMPLE SIZES

If we use the pooled sample mean to estimate μ_θ the asymptotic relative efficiency, as $\alpha + \beta \rightarrow \infty$, is one, whereas if we use the unweighted mean of the group means the asymptotic relative efficiency for small values of π is

$$\frac{m^2}{\sum_{i=1}^m r_i \left(\sum_{i=1}^m \frac{1}{r_i} \right)} \left\{ 1 + \pi \left(\frac{\sum_{i=1}^m r_i^2}{\sum_{i=1}^m r_i} - \frac{m}{\sum_{i=1}^m \frac{1}{r_i}} \right) \right\}, \quad (2.55)$$

i.e., the relative efficiency depends on how different are the values of the $\{r_i\}$; the more homogeneous they are the greater is the relative efficiency. Writing $r_i = r(1 + d_i)$ with $\sum d_i = 0$, then provided the d_i are small, the asymptotic relative efficiency as $\alpha + \beta \rightarrow \infty$ is approximately

$$1 - \frac{\sum_{i=1}^m d_i^2}{m}, \quad (2.56)$$

and hence the smaller the d_i are the bigger the relative efficiency is, the maximum being obtained when $d_i = 0$ for all values of i , corresponding to the case of equal sample sizes.

As $\alpha + \beta \rightarrow \infty$, σ_{θ}^2 tends to zero, i.e., the overdispersion is negligible and the pooled sample mean is fully efficient, whereas the mean of the group means corresponds to a weighted estimator, where the weights are the inverses of the sample sizes. Again the full efficiency of the pooled mean in the limit case is explained because the limit case corresponds to the binomial distribution and the pooled mean is fully efficient in this case.

Table 2.9 shows the asymptotic efficiency of the mean of group means, the pooled mean and the quasiliikelihood estimator of the mean relative to the maximum likelihood estimator in the case when $m/2$ sample sizes are equal to r and $m/2$ sample sizes are equal to $2r$, i.e., the ratio of the smallest sample size to the biggest sample size is equal to two; a difference from the case of equal sample sizes is that the ratio α/β influences the asymptotic relative efficiency, therefore we consider the asymptotic relative efficiency for four different values of this ratio. The asymptotic relative efficiency of the mean of group means as $\alpha + \beta$ tends to infinity is 88.9%. The asymptotic relative efficiency of the quasiliikelihood estimator to the maximum likelihood estimator as $\alpha + \beta$ tends to infinity is one.

Now if we have $m/3$ samples of size r , $m/3$ samples of size $2r$ and $m/3$ samples of size $3r/2$, i.e., the ratio of the smallest sample size to the biggest sample size is equal to two, the efficiency for the three estimators increases compared with the two different sample sizes case. For equivalent values of α and β bigger then one the efficiency of the pooled mean and the quasiliikelihood estimator is over 97%. The asymptotic relative

efficiency of the mean of group means as $\alpha + \beta$ tends to infinity is 92.3%. This is explained because the r_i are more homogeneous, i.e., the mean of the sample sizes is the same but the variance is smaller, in other words $\sum d_i^2$ is smaller and the efficiency is bigger.

Table 2.9.
*Asymptotic efficiency of the mean of group means, pooled mean and
quasilikelihood estimator relative to the maximum likelihood
estimator of the mean.*

α	mean of group means	pooled mean	quasilikelihood estimator
$\beta = \alpha$			
1	.976	.960	.996
3	.946	.988	1.000
5	.931	.995	1.000
10	.914	.998	1.000
20	.902	1.000	1.000
$\beta = 2\alpha$			
1	.967	.973	.998
3	.934	.994	1.000
5	.920	.997	1.000
10	.907	.999	1.000
20	.898	1.000	1.000
$\beta = 3\alpha$			
1	.959	.980	.981
3	.926	.996	.996
5	.914	.998	.998
10	.903	.999	.999
20	.896	1.000	1.000
$\beta = 4\alpha$			
1	.952	.985	.985
3	.920	.997	.997
5	.909	.999	.999
10	.900	1.000	.999
20	.895	1.000	1.000

The sample sizes are $m/2$ of them equal to r and $m/2$ of them equal to $2r$.

If we increase the ratio of the smallest sample size to the biggest sample size the asymptotic relative efficiency decreases, whereas if we increase the number of different sample sizes in between the smallest and the biggest the relative efficiency of any of the estimators increases.

In general we find that the pooled mean and the quasilielihood estimators perform better than the mean of group means (even though for small values of α and β the mean of group means could be better, the difference is negligible) and of these two the quasilielihood estimator involves the index of overdispersion that in general needs to be estimated. Therefore the use of the pooled mean is recommended because it is the simplest to calculate, its efficiency when α and β are bigger than one is over 95% when the values of the sample sizes are not very different and even when its efficiency is affected by the ratio α/β it improves (a difference from the quasilielihood estimator) as the ratio is further than one. Cochran's estimator is obviously more efficient than the pooled mean if we know where is the best place to fix the upper limit and what is the best constant to use. In general if there is great variation among the sample sizes it would be worth to try to use it but otherwise the pooled mean is as efficient as it is.

In the case of the variance, using the fact that the moments of the beta distribution can be expressed as functions of μ and π and that the limit, as π tends to zero, of all but the first one is zero, the asymptotic relative efficiency of the components of variance estimator as $\alpha + \beta \rightarrow \infty$ is given by the ratio of the limit of the variance of the maximum likelihood estimator given by expression (2.26) and the limit of the variance of the components of variance estimator given by expression (2.35).

When the sample sizes are small the relative efficiency depends on the ratio α/β , but when all the sample sizes are bigger than or equal to twenty it does not depend on this ratio. In all cases the relative efficiency depends on how different are the sample sizes. Again writing (as in the mean case) $r_i = r(1 + d_i)$ with $\sum d_i = 0$, if in addition $\sum d_i^2/m < (r - 1)/r$, the asymptotic relative efficiency as $\alpha + \beta \rightarrow \infty$, is

$$1 - \frac{\sum_{i=1}^m d_i^2}{m^2} \left\{ \frac{6(r-1)\mu}{(1-\mu)r^2} + \frac{1}{2(r-1)(1-\mu)^2} + \frac{3}{(r-1)} - \frac{8r^2\mu + 2(r-1)}{(1-\mu)} \right\}. \quad (2.57)$$

The smaller the d_i are, the bigger the asymptotic relative efficiency is and, as in the mean case, the maximum is obtained when all the d_i are equal to zero.

In the case of the quasilikelihood estimator the asymptotic relative efficiency as $\alpha + \beta$ tends to infinity is

$$\frac{\sum_{i=1}^m (r_i - 1)^2}{\sum_{i=1}^m r_i (r_i - 1)}, \quad (2.58)$$

and writing again $r_i = r(1 + d_i)$ with $\sum d_i = 0$, we have (assuming $r \sum d_i^2 / (m(r-1)) < 1$) that the asymptotic efficiency of the quasilikelihood estimator relative to the maximum likelihood estimator is

$$\frac{r-1}{r} \left\{ 1 + \frac{\sum_{i=1}^m d_i^2}{m} \left(\frac{r}{(r-1)^2} \right) \right\}. \quad (2.59)$$

Thus the asymptotic relative efficiency is always less than one and because of the restrictions imposed on the d_i the asymptotic relative efficiency is always bigger than $(r-1)/r$.

We calculate the relative efficiency of the components of variance estimator and the quasilikelihood estimator for different ratios of the smallest to the biggest sample sizes, as in the mean case. Table 2.10 gives the values of the asymptotic relative efficiency of both estimators when we just have two different sample sizes each one

m/2 times and the ratio is equal to two.

Table 2.10.
Asymptotic efficiency of the components of variance estimator and the quasilielihood estimator relative to the maximum likelihood estimator.

α	CVE $\beta = \alpha$	QLE $\beta = \alpha$	CVE $\beta = 2\alpha$	QLE $\beta = 2\alpha$	CVE $\beta = 3\alpha$	QLE $\beta = 3\alpha$	CVE $\beta = 4\alpha$	QLE $\beta = 4\alpha$
1	.943	.321	.953	.453	.948	.579	.936	.678
3	.944	.512	.925	.596	.901	.659	.876	.706
5	.931	.577	.908	.637	.881	.679	.856	.709
10	.911	.639	.889	.672	.862	.694	.836	.710
20	.896	.674	.876	.692	.850	.704	.825	.709
50	.885	.696	.867	.702	.843	.710	.817	.711
100	.881	.704	.864	.706	.840	.709	.815	.711

The sample sizes are two different values and the ratio of the bigger to the smaller is 2.

When we have two different sample sizes where the ratio of the bigger to the smaller is two, the asymptotic relative efficiency of the quasilielihood estimator as $\alpha + \beta$ tends to infinity is 71.4%, whereas the asymptotic relative efficiency of the components of variance estimator depends on the ratio α/β ; for the ratio equal to one, .5, .33, .25 the asymptotic relative efficiencies are 88%, 86.2%, 84% and 81.2% respectively.

Even though the components of variance estimator loses efficiency as $\alpha + \beta$ increases and the quasilielihood estimator is more efficient, in general, for values of α as big as one hundred and β four times α the components of variance estimator is still more efficient.

If we increase the number of different sample sizes, the asymptotic relative efficiency of the components of variance estimator increases, whereas the quasilielihood one decreases. This can be explained because the sample sizes are more homogeneous and therefore the sum of the d_i squared is closer to zero and the asymptotic efficiency in the case of the components of variance estimator is closer to

one; in the case of the quasilikelihood estimator the asymptotic efficiency is closer to $(r-1)/r$.

For a ratio equal to three with just two different sample sizes the asymptotic relative efficiency of the components of variance estimator decreases compared to the case of the ratio equal to two, whereas the asymptotic relative efficiency of the quasilikelihood one increases; for the same values of α and β as before the result are given in table 2.11.

Table 2.11.
Asymptotic efficiency of the components of variance estimator and the quasilikelihood estimator relative to the maximum likelihood estimator.

α	CVE $\beta = \alpha$	QLÉ	CVE $\beta = 2\alpha$	QLÉ	CVE $\beta = 3\alpha$	QLÉ	CVE $\beta = 4\alpha$	QLÉ
1	.869	.342	.903	.497	.906	.643	.894	.759
3	.894	.566	.867	.667	.825	.744	.785	.797
5	.872	.648	.832	.719	.786	.769	.742	.803
10	.833	.723	.792	.762	.745	.788	.702	.808
20	.801	.764	.764	.784	.720	.799	.679	.807
50	.767	.790	.745	.798	.704	.804	.664	.809
100	.742	.800	.738	.803	.698	.807	.659	.810

The sample sizes are two different values and the ratio of the bigger to the smaller is 3.

The asymptotic efficiencies as $\alpha + \beta$ tends to infinity for the ratio α/β equal to 1, .5, .33, .25 of the components of variance estimator relative to the maximum likelihood estimator are 73.8%, 73%, 69.1%, 65.2% respectively. The asymptotic efficiency of the quasilikelihood estimator is 81.25%.

If we increase the number of different sample sizes the relative efficiency of the quasilikelihood estimator decreases and that of the components of variance increases, therefore it is not easy to decide which one to use; when the ratio of the largest to the smallest is two the components of variance is a better choice, but when this ratio is three or more the decision depends on how many different values of sample size there

are; for about five different values of sample size the components of variance estimator would be a better choice but otherwise the quasilielihood estimator would give better results.

2.5 INDEX OF RELATIVE DISPERSION

The estimation of the ratio $\rho = \sigma_\theta^2 / \{\mu_\theta(1 - \mu_\theta)\}$ is of interest because it represents the ratio of the variance due to the overdispersion to the variance of the binomial model. Under the present assumptions, it takes values between 0 and 1, zero as $\alpha + \beta \rightarrow \infty$ and one as $\alpha + \beta \rightarrow 0$.

The maximum likelihood estimator of this ratio is

$$\frac{\hat{\pi}}{1 + \hat{\pi}}, \quad (2.60)$$

and its asymptotic variance is

$$\frac{1}{(1 + \pi)^4} \text{var}(\hat{\pi}), \quad (2.61)$$

which for small values of π , i. e., big values of $\alpha + \beta$, is

$$\frac{2}{\sum_{i=1}^m r_i(r_i - 1)} \left\{ 1 + \pi \frac{2 \sum_{i=1}^m C_i}{\sum_{i=1}^m r_i(r_i - 1)} \right\}, \quad (2.62)$$

where C_i is given by expression (2.21); then the limit as $\pi \rightarrow 0$ is

$$\frac{2}{\sum_{i=1}^m r_i(r_i - 1)} . \quad (2.63)$$

Alternatively, we can estimate the ratio by

$$\frac{\tilde{\sigma}_{\theta}^2}{\bar{y}_{..}(1 - \bar{y}_{..})}, \quad (2.64)$$

which is asymptotically unbiased, and its variance, in the case of equal sample sizes is

$$\begin{aligned} & \frac{1}{\mu_1^2(1-\mu_1)^2mr(r-1)} \left\{ 2\mu_1^2(1-\mu_1)^2 + 2\mu_2 + 4(r-2)\mu_3 + (r-2)(r-3)\mu_4 - 8(r-2)\mu_3\mu_1 \right. \\ & \quad \left. + 4(r-4)\mu_2\mu_1(1-\mu_1) - r(r-1)\mu_2^2 \right\} + \frac{(1-4\mu_1(1-\mu_1))\mu_2^2}{rm\mu_1^4(1-\mu_1)^4} (\mu_1(1-\mu_1) + (r-1)\mu_2) \\ & + \frac{(1-4\mu_1(1-\mu_1))\mu_2}{mr(r-1)\mu_1^3(1-\mu_1)^3} (\mu_1(1-\mu_1) + (r-1)\mu_2) - \frac{2(1-2\mu_1)\mu_2}{m(r-1)\mu_1^3(1-\mu_1)^3} \left(\frac{3(r-1)}{r}(\mu_1^2 + \mu_2) \right. \\ & \left. + \frac{(r-2)(r-1)}{r}(\mu_3 + \mu_1^3 + \mu_1\mu_2) + \frac{\mu_1}{r} - (r-1)(\mu_1^3 + \mu_1\mu_2) - \mu_1^2 \right\} . \quad (2.65) \end{aligned}$$

For small values of π , the variance in terms of μ and π is

$$\frac{2}{mr(r-1)} \left\{ 1 + \pi \left(\frac{1}{(1-\mu)^2} - \frac{\mu^2(2r-1)}{(1-\mu)^2} - \frac{1-\mu}{2\mu} - 3 \right) \right\} , \quad (2.66)$$

and the limit as $\pi \rightarrow 0$ is

$$\frac{2}{mr(r-1)}, \quad (2.67)$$

therefore the asymptotic relative efficiency as $\alpha + \beta$ tends to infinity is one.

When the sample sizes are all two both estimators are identical, since, as we noted in section 2.4.1, both the numerator and the denominator of the latter estimator are maximum likelihood estimators.

As α and β increase the relative efficiency is very similar to the relative efficiency in the case of the estimation of the variance, since the major contribution comes from the first term of the expression for the variance.

This is also true in the case of different sample sizes, because any weighted mean is asymptotically unbiased (as $m \rightarrow \infty$) for $\mu_\theta(1 - \mu_\theta)$. The asymptotic relative efficiency as $\alpha + \beta \rightarrow \infty$ is identical to that in the case of the estimation of σ_θ^2 .

The quasilielihood estimator is given by expression (2.42) and its variance is

$$\frac{2}{\sum_{i=1}^m \frac{(r_i - 1)^2}{(1 + (r_i - 1)\rho)^2}}. \quad (2.68)$$

In general the asymptotic relative efficiency is very similar to that of estimating the variance. The limit as $\alpha + \beta$ tends to infinity is given by expression (2.54), which is the limit of the asymptotic relative efficiency for estimating the variance. The comments made for asymptotic relative efficiencies of variances apply to the asymptotic relative efficiencies of this index.

2.6 SIMULATION STUDY

In the previous section we have considered only the asymptotic variances of the estimators. If we are interested to know the small sample efficiencies of the moment estimators relative to the maximum likelihood estimators or, in other words, how big has to be the sample for the asymptotic efficiencies to be valid, we can proceed in two different ways. The first one is to calculate the terms of order m^{-2} for the variances of the estimators; while in the moment estimators this is not difficult, in the case of the maximum likelihood estimators it is quite complicated. The second way is simulation of small samples with the characteristics we have been studying.

For a series of values of α , β and r , in the case where the r_i are equal, we simulate three hundred samples of size $m = 20$ and thirty samples of size $m = 30$. We calculate the variance of these thirty values and compare the variance of the sample mean to the variance maximum likelihood estimator of the mean. The results for ten observations for each individual ($r = 10$) and sample size twenty ($m = 20$) are given in table 2.12.

Table 2.12.
Relative efficiency of the sample mean to the maximum likelihood estimator based on 300 simulated samples of size 20.

$r = 10$				
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$
1	1.411	.951	1.026	.998
3	.997	1.055	.982	1.065
5	.957	.983	1.000	.995
10	.998	1.001	.999	1.001
20	.996	1.002	1.003	1.000

The standard error of the simulation is less than 7×10^{-3} .

For all the values of r considered the efficiency of the sample mean relative to the maximum likelihood estimator is close to one; therefore the use of the moment estimator is recommended even when m is small. The theoretical values in terms of $\alpha + \beta$ are given in table 2.6, we can see that for $r = 10$ the asymptotic relative efficiency is bigger than .995 for $\alpha + \beta$ bigger than 5, which is true for the simulated values with the exception of $\alpha + \beta$ equal to six.

In the case of the variance, we calculate the components of variance estimator, the quasilielihood estimator and the maximum likelihood estimator of the variance, and the relative efficiency for different values of r and m ; some of the results for some values are shown in tables 2.13 and 2.14.

*Table 2.13.
Relative efficiency of the components of variance estimator and
quasilielihood estimator to the maximum likelihood estimator of the
variance, based on 300 simulated samples of size 20.*

$r = 3$								
α	$\beta = \alpha$		$\beta = 2\alpha$		$\beta = 3\alpha$		$\beta = 4\alpha$	
	CVE	QLE	CVE	QLE	CVE	QLE	CVE	QLE
1	.975	.547	.923	.533	.813	.520	.727	.436
3	.761	.383	.720	.383	.504	.343	.618	.489
5	.622	.352	.591	.348	.623	.402	.537	.346
10	.753	.398	.450	.264	.624	.417	.344	.205
20	.318	.083	.449	.292	.405	.320	.315	.179
100	.452	.246	.509	.336	.379	.269	.241	.290
$r = 10$								
1	.915	.480	.930	.567	.902	.933	.870	.968
3	.869	.579	.901	.682	.799	.751	.780	.769
5	.944	.667	.945	.587	.484	.402	.600	.574
10	.883	.651	.652	.543	.572	.594	.646	.468
20	.519	.443	.541	.334	.555	.457	.686	.509
100	.348	.291	.340	.245	.407	.260	.335	.207

The standard error of the simulation for both estimators is always less than 3×10^{-3} .

In general, in the case of the variance for $m = 20$ the relative efficiency of the

components of variance estimator is smaller than the corresponding asymptotic relative efficiency (for the same values of α and β). Nevertheless for small values of α and β , less than or equal to three, and for r greater than four it is over 85%; for r equal to two or three this is true for α and β less than or equal to one and α/β bigger than .3. In general if we increase m to 30 the relative efficiency for values of r bigger than or equal to four and α and β less than or equal to 10 is greater than 80%. For bigger values of α and β a bigger sample size is required to improve the efficiency (when the overdispersion is smaller).

*Table 2.14.
Relative efficiency of the components of variance estimator and
quasilikelihood estimator to the maximum likelihood estimator of the
variance, based on 300 simulated samples of size $m = 30$.*

$r = 3$								
α	$\beta = \alpha$		$\beta = 2\alpha$		$\beta = 3\alpha$		$\beta = 4\alpha$	
	CVE	QLE	CVE	QLE	CVE	QLE	CVE	QLE
1	.959	.502	.955	.536	.903	.543	.768	.767
3	.654	.341	.637	.349	.247	.195	.602	.475
5	.658	.350	.673	.403	.655	.412	.436	.450
10	.656	.339	.513	.296	.428	.307	.436	.489
20	.225	.116	.500	.344	.490	.397	.508	.391
100	.175	.092	.342	.222	.322	.234	.415	.341
$r = 10$								
1	.882	.486	.802	.548	.973	1.054	.961	.880
3	.961	.581	.891	.696	.943	.824	.809	.799
5	.967	.608	.867	.760	.604	.464	.765	.565
10	.948	.669	.668	.496	.696	.621	.609	.479
20	.608	.484	.547	.388	.458	.365	.689	.464
100	.477	.413	.192	.144	.235	.187	.338	.287

The standard error of the simulation for both estimators is always less than 3×10^{-6} .

Another use of the simulation is to see how robust are the maximum likelihood estimators to departures from the assumption of a beta distribution for θ_i . We simulate samples where the distribution of θ_i is a two point distribution. We focus on two different situations, the first one where θ_i takes each of the values with probability of a

half and the second one where θ_1 takes one of the values with probability $p_1 = .1$ and the other with probability $p_2 = .9$. The results for the relative efficiency of the sample mean to the maximum likelihood estimator of the mean for different sets of values of (θ_1, θ_2) are presented in table 2.15.

Table 2.15.
Relative efficiency of the sample mean to the maximum
likelihood estimator for a two point distribution of θ_i
300 simulated samples.

θ_1	θ_2	$r = 3$		$r = 10$	
		$p_1=.5$	$p_1=.1$	$p_1=.5$	$p_1=.1$
$m = 20$					
.4	.6	.990	1.005	.997	.998
.3	.7	.985	1.015	.993	1.079
.1	.9	1.261	1.028	3.8075	.948
$m = 30$					
.4	.6	.998	.995	.995	.998
.3	.7	.987	1.003	1.011	1.002
.1	.9	1.226	.941	4.267	1.812

The standard error is always less than 2×10^{-3} .

In general, the efficiency of the sample mean relative to the maximum likelihood estimator is close to one, that is the maximum likelihood estimator is robust to departures from the beta distribution for the θ_1 but when the two possible values of θ_1 are .1 and .9 the sample mean is more efficient than the maximum likelihood estimator.

For the variance we calculate the components of variance estimator, the quasiliikelihood estimator and the maximum likelihood estimator. The results for some values of r , m and different sets of values of θ_1 and θ_2 are shown in table 2.16.

In general when the values of θ_1 and θ_2 are .1 and .9 the relative efficiency of the components of variance estimator to the maximum likelihood estimator is close to one or bigger than one, and for values of θ_1 and θ_2 equal to .3 and .7 the efficiency of the components of variance estimator is over 50%. Thus we can say that the maximum

likelihood estimator is robust to departures from the assumption of beta distribution, but the components of variance estimator is a good option when the values of θ_i are not close to .5.

*Table 2.16.
Relative efficiency of the components of variance estimator and the
quasilikelihood estimator to the maximum likelihood estimator of the variance,
for a two points distribution of θ_i based on 300 simulated samples.*

θ_1	θ_2	$p_1=.5$		$p_1=.1$		$p_1=.5$		$p_1=.1$	
		$r = 3$				$r = 10$			
		$m = 20$							
		CVE	QLE	CVE	QLE	CVE	QLE	CVE	QLE
.4	.6	.445	.239	.465	.250	.789	.624	.346	.287
.3	.7	.790	.421	.529	.337	.765	.571	.830	.615
.1	.9	.925	.478	.971	.974	1.285	.635	1.677	2.274
		$m = 30$							
.4	.6	.438	.226	.393	.212	.800	.673	.351	.301
.3	.7	.765	.390	.654	.352	.783	.548	.637	.527
.1	.9	.990	.526	1.030	.768	2.095	.9715	2.783	3.306

The standard error is always less than 2×10^{-3} .

2.7 GENERALIZATION TO A BINOMIAL RESPONSE

Let $\{y_{ij}\}$ $i = 1, \dots, m$, $j = 1, \dots, r_i$ be independently distributed binomial random variables with $E(y_{ij}) = n\theta_i$ (i.e. we are observing the response of n independent binary trials done at the same time), and assume that n is the same for each observation and each individual. Conditional on θ_i , $y_i = \sum y_{ij}$ is a binomially distributed random variable with parameters nr_i , θ_i .

Suppose that $\theta_1, \dots, \theta_m$ are independent identically distributed random variables

we assume that θ_i follows a beta distribution with parameters $\beta(= (1 - \mu)\pi)$ and $\alpha(=\mu\pi)$; therefore the unconditional distribution of y_i , writing, for simplicity, y_i from here on is beta binomial given by

$$\binom{nr_i}{y_i} \frac{\prod_{k=0}^{y_i-1} (\mu + k\pi) \prod_{k=0}^{nr_i-y_i-1} (1-\mu + k\pi)}{\prod_{k=0}^{nr_i-1} (1 + k\pi)}. \quad (2.69)$$

It is easy to see that the likelihood corresponds to the likelihood in the binary case changing r_i for nr_i and therefore we can obtain the maximum likelihood estimators of the parameters and their variances replacing r_i by nr_i in the expressions for the binary case. This comment also applies to the quasilielihood estimators.

In the case of moment based estimators, the sample mean is the obvious estimator in the case of equal sample sizes; if the sample sizes are different we could use any weighted estimator. In particular we focus, as in the binary case, on the pooled sample mean and the mean of the group means, defined by expressions (2.28) and (2.29) in section 2.3.3.

Their variances in terms of μ and π are respectively

$$\frac{\mu(1 - \mu)}{n(1 + \pi)(\sum_{i=1}^m r_i)^2} \sum_{i=1}^m (r_i + nr_i^2 \pi), \quad (2.70)$$

$$\frac{\mu(1 - \mu)}{nm^2(1 + \pi)} \sum_{i=1}^m \frac{1 + nr_i \pi}{r_i}. \quad (2.71)$$

It is clear that both estimators reduce to the same when the sample sizes are equal.

The variances of the maximum likelihood estimator and the quasilielihood

estimator are respectively

$$\frac{\mu(1-\mu)}{n \sum_{i=1}^m r_i} \left\{ 1 + \pi \frac{\sum_{i=1}^m r_i (nr_i - 1)}{\sum_{i=1}^m r_i} \right\}, \quad (2.72)$$

$$\frac{\mu(1-\mu)}{\sum_{i=1}^m \frac{nr_i}{1 + (nr_i - 1)\rho}}. \quad (2.73)$$

For any value of α , β , n and r_i , we have that the asymptotic efficiency of any moment based estimator of the mean, including the quasiliikelihood estimator, relative to the maximum likelihood estimator is identical to that obtained in the binary case for the same values of α , β and r_i .

In the case of the estimation of σ_θ^2 , we can obtain an estimator from the formal mean square between groups in an analysis of variance.

The analysis of variance table is

source	d.f.	sum of squares	$E(\text{mean squared})$
Between	$m - 1$	$\sum_{i=1}^m r_i (\bar{y}_i - \bar{y}_{..})^2$	$n\mu_\theta(1 - \mu_\theta) - n\sigma_\theta^2 + n^2 \frac{(\sum_{i=1}^m r_i)^2 - \sum_{i=1}^m r_i^2}{(m-1)\sum_{i=1}^m r_i} \sigma_\theta^2$
Within	$\sum_{i=1}^m r_i - m$	$\sum_{i=1}^m \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_i)^2$	$n\mu_\theta(1 - \mu_\theta) - n\sigma_\theta^2$
Total	$\sum_{i=1}^m r_i - 1$	$\sum_{i=1}^m \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{..})^2$	

Then, for large values of m , we estimate σ_θ^2 , as in the binary case, as $m \sum r_i$

times the ratio of difference of the mean square between and the mean square within to $\{ (\sum r_i)^2 - \sum r_i \}$. Again substituting the mean square between by $1/(m - 1)$ times the difference of the sum of squares total and the sum of squares within the components of variance estimator of $\sigma_\theta^2, \tilde{\sigma}_\theta^2$ is

$$\frac{\sum_{i=1}^m r_i}{n^2 \left(\left(\sum_{i=1}^m r_i \right)^2 - \sum_{i=1}^m r_i^2 \right)} \left\{ \frac{\sum_{i=1}^m r_i}{\sum_{i=1}^m (r_i - 1)} \sum_{i=1}^m \frac{y_{i.}^2}{r_i} - \frac{m}{\sum_{i=1}^m (r_i - 1)} \sum_{i=1}^m \sum_{j=1}^{r_i} y_{ij}^2 - \frac{y_{..}^2}{\sum_{i=1}^m r_i} \right\}, \quad (2.74)$$

which reduces to

$$\frac{1}{n^2} \left\{ \frac{\sum_{i=1}^m \sum_{s=1}^r \sum_{t=1}^r y_{is} y_{it}}{mr(r-1)} - y_{..}^2 \right\}, \quad (2.75)$$

when all the sample sizes are equal to r .

The variance in terms of the moments of the beta distribution is

$$\begin{aligned} & \frac{2\mu_1^2(1-\mu_1)^2}{n^2 mr(r-1)} + \frac{2\mu_2}{mr} \left(\frac{1}{n^2(r-1)} + 2(nr-1) \right) + \mu_3 \left(\frac{4}{n mr} - \frac{4}{n mr(r-1)} \right) \\ & + \frac{\mu_4}{n^2 mr(r-1)} (n^2 r(r-1) - 4n(r-1) + 2) + \frac{12\mu_1\mu_2}{n^2 mr(r-1)} (n(r-1) - 1) \\ & + \mu_1^2\mu_2 \left(\frac{12}{n^2 mr(r-1)} - \frac{8}{mr} - \frac{2}{m} \right) + \frac{8\mu_1\mu_3}{n^2 mr(r-1)} (1 - n(r-1)). \end{aligned} \quad (2.76)$$

which for small values of π (large values of $\alpha + \beta$) is

$$\begin{aligned} & \frac{2\mu^2(1-\mu)^2}{n^2 mr(r-1)} + \frac{\pi\mu}{n^2 mr(r-1)} (12\mu(1-\mu)(n(r-1)-1) + 2\mu^2(1-\mu)(n^2(r-1)(r+4)+6) \\ & + 2(1-\mu)(2n^2(nr-1)(r-1)+1)). \end{aligned} \quad (2.77)$$

The limit as $\alpha + \beta$ tends to infinity is

$$\frac{2\mu^2(1-\mu)^2}{n^2 mr(r-1)} \quad (2.78)$$

and hence the asymptotic relative efficiency as $\alpha + \beta$ tends to infinity is

$$\frac{nr - \eta}{nr - 1}. \quad (2.79)$$

For some values of n , r , α and β , the asymptotic relative efficiency is given in table 2.17.

The efficiency increases with r having n constant and decreases with n having r constant; therefore if we have a large number of responses at each observation we need to have a large number of observations in each sample. If n is bigger than or equal to the number of observations the use of the components of variance estimator is not recommended.

The quasilielihood estimator of the variance is given by expression (2.46) and its variance is given by

$$\rho^2(1-2\mu)^2\mu(1-\mu)\left(\sum_{i=1}^m \frac{nr_i}{1+(nr_i-1)\rho}\right)^{-1} + \mu^2(1-\mu)^2\left(\sum_{i=1}^m \frac{(nr_i-1)^2}{2(1+(nr_i-1)\rho)^2}\right)^{-1} \quad (2.80)$$

Table 2.17.
Asymptotic relative efficiency of the components of variance
estimator to the maximum likelihood estimator of the variance when
the observations are binomially distributed.

		n = 2		r = 2	
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$	
1	.871	.836	.822	.811	
3	.797	.765	.747	.735	
5	.761	.735	.721	.711	
10	.722	.705	.696	.691	
20	.697	.687	.682	.679	
50	.679	.675	.673	.672	
100	.673	.671	.667	.669	
		n = 2		r = 5	
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$	
1	.948	.891	.883	.887	
3	.967	.931	.922	.915	
5	.960	.937	.924	.917	
10	.941	.925	.916	.911	
20	.922	.919	.906	.903	
50	.904	.900	.897	.895	
100	.897	.894	.893	.892	
		n = 5		r = 2	
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$	
1	.902	.821	.800	.797	
3	.852	.791	.753	.729	
5	.797	.740	.706	.683	
10	.715	.672	.647	.631	
20	.649	.621	.606	.596	
50	.597	.584	.577	.573	
100	.577	.570	.566	.564	
		n = 2		r = 10	
α	$\beta = \alpha$	$\beta = 2\alpha$	$\beta = 3\alpha$	$\beta = 4\alpha$	
1	.914	.828	.812	.823	
3	.998	.940	.921	.916	
5	.986	.961	.948	.942	
10	.983	.969	.961	.957	
20	.974	.966	.961	.958	
50	.962	.957	.955	.953	
100	.956	.953	.952	.951	

which tends to

$$\frac{2\mu^2(1 - \mu)^2}{\sum_{i=1}^m (nr_i - 1)^2} \quad (2.81)$$

as $\pi \rightarrow 0$.

In the case of equal sample sizes the asymptotic relative efficiency of the quasiliikelihood estimator with respect to the maximum likelihood estimator is given by

$$\frac{(nr - 1)^2}{nr(nr - 1)} \quad (2.82)$$

The efficiency of the quasiliikelihood estimator depends on the product of the number of responses and the number of observations in each sample, and the fact that the number of responses might be bigger than the number of observations does not affect the efficiency.

If the number of responses is smaller than the number of observations in each sample the components of variance estimator is more efficient than the quasiliikelihood estimator, whereas if the number of responses is bigger than or equal to the number of observations in each sample the quasiliikelihood estimator is more efficient. This can be seen, for example in the case $n = 2$ and $r = 5$; the asymptotic efficiency of the quasiliikelihood estimator relative to the maximum likelihood estimator is given by table 2.8 taking the sample size equal to ten, and if we compare it with the value in the table 2.17, for the same values of α and β , we see that the components of variance estimator performs better. Conversely for the case $n = 5$ and $r = 2$, we have that the asymptotic relative efficiency of the quasiliikelihood estimator to the maximum likelihood estimator is given by the same value in table 2.8 and the performance is better than the performance of the components of variance estimator.

When the sample sizes are different, the asymptotic relative efficiency follows the same pattern as in the case of binary variables, i.e., in the case of the components of variance estimator as $\alpha + \beta$ increases the asymptotic relative efficiency decreases, whereas for the quasiliikelihood estimators the opposite happens.

2.8 DEPENDENCE OF THE BINARY RESPONSE ON EXPLANATORY VARIABLES

In most investigations there exists a vector of covariates or explanatory variables associated with each individual, i.e., treatment factors or quantitative variables concerning aspects of the individual.

Usually it is assumed that the dependence of θ_i on $(x_1, \dots, x_p)$ is through the linear combination

$$\eta = \sum_{i=1}^m \beta_i x_i, \quad (2.83)$$

for $\beta_1, \dots, \beta_p$ unknown coefficients. We also assume that

$$g(\theta_i) = \eta_i = \sum_{j=1}^p \beta_j x_{ij}, \quad (2.84)$$

where g maps the interval $(0,1)$ into the real line; equivalently

$$h(\eta_i) = \theta_i. \quad (2.85)$$

We introduce the dependence of θ_i on $(x_1, \dots, x_p)$ through the likelihood

$$\prod_{i=1}^m \binom{r_i}{y_i} \theta_i^{y_i} (1 - \theta_i)^{r_i - y_i}, \quad (2.86)$$

assuming, as before, that we have r_i trials for each individual and $y_i (= \sum y_{ij})$ is the number of positive responses.

Even when all the available explanatory variables have been fitted, the residual variation may be greater than that which can be attributed to the binomial sampling variation assumed by the model, i.e., overdispersion is present. We can have different approaches to introduce it.

One possibility is to assume that there are factors affecting θ_i whose effects are additive on the same scale as the explanatory variables, that is to assume that

$$\theta_i = h\left(\sum_{j=1}^p \beta_j x_{ij} + \sigma u_i\right), \quad (2.87)$$

where the u_i are independently and identically distributed random variables which take values in the interval $(-\infty, \infty)$ and $\sigma \geq 0$ is unknown.

Pierce and Sands(1975) assume that u_i are independently Normal(0,1) and obtain the maximum likelihood estimator of β and σ , and propose a test for σ equal to zero. They show that assuming that h is the logit function a simpler estimator may be obtained.

Im & Gianda(1988) assume a two way nested layout with allowance for binomial response, dependent on covariates; they assume that an underlying variable, z_{ijk} , associated with the response follows the mixed linear model

$$z_{ijk} = \mu_{ij} + s_i + \delta_{ij} + \epsilon_{ijk}, \quad (2.88)$$

where μ_{ij} is a location parameter, that can be written as $X'\beta$, where X is a fixed and known matrix and β is an unknown fixed vector; s_i is a random effect and δ_{ij} is nested on it and random itself. Assuming that s_i and δ_{ij} are normally distributed with mean zero and variances σ_1^2 and σ_2^2 respectively and ε_{ijk} are symmetrical and identically distributed they suggest to estimate the parameters using maximum likelihood theory.

Another possibility could be to assume that there are factors which affect θ_i whose effects are additive to θ_i , i.e., to assume

$$\theta_i = \lambda h \left(\sum_{j=1}^p \beta_j x_{ij} \right) + (1 - \lambda) \varepsilon_i, \quad (2.89)$$

where ε_i is a random variable taking values in the interval (0,1) and λ could be a known or unknown parameter. We could assume for instance that ε_i is a beta distributed random variable. The main difficulty is likely to be finding a simple expression for the likelihood.

A third possibility is to assume that θ_i follows a distribution and that μ_i , the mean of the distribution, is a function of the explanatory variables. One possibility is to assume that θ_i follows a beta distribution with parameters α_i and γ_i ; then

$$\begin{aligned} E(\theta_i) &= \frac{\alpha_i}{\alpha_i + \gamma_i} = h \left(\sum_{j=1}^p \beta_j x_{ij} \right) = \mu_i, \\ \text{var}(\theta_i) &= \frac{\alpha_i \gamma_i}{(\alpha_i + \gamma_i)^2 (\alpha_i + \gamma_i + 1)} = \mu_i (1 - \mu_i) \rho_i = \sigma_i^2. \end{aligned} \quad (2.90)$$

We could assume different constraints in the parameters, e.g. Crowder(1978)

1) μ_i and σ_i are different from sample to sample, in this case we have $m(p + 1)$ parameters to be estimated from the likelihood.

2) μ_i are different from sample to sample and $\rho_i = \rho \forall i$ or in terms of the notation in section 2.3 $\pi_i = \pi$; we have $mp + 1$ parameters to be estimated from the likelihood which is written as expression (2.18) replacing μ by μ_i .

3) μ_i and ρ_i are different from sample to sample but $\sigma_i = \mu_i(1 - \mu_i)\rho_i = \sigma \forall i$; in this case we have as well $mp + 1$ parameters to be estimated from the likelihood which is written

$$\frac{\left(\frac{r_i}{y_i} \right) \prod_{k=0}^{n-1} \left(\mu_i + \frac{k\sigma}{\sigma - \mu_i(1 - \mu_i)} \right) \prod_{k=0}^{n-1} \left(1 - \mu_i + \frac{k\sigma}{\sigma - \mu_i(1 - \mu_i)} \right)}{\prod_{i=1}^m \frac{\sigma(1 + k) - \mu_i(1 - \mu_i)}{\sigma - \mu_i(1 - \mu_i)}}. \quad (2.91)$$

The main difference between 2 and 3 is that in 2 we are assuming that ρ , the parameter which represents the index of relative overdispersion (variance of the Beta distribution to the variance of the binomial distribution), is constant from sample to sample, whereas in 3 we are assuming that the total variance is constant from sample to sample. The choice of one or the other depends on the specific problem.

Another approach is quasilielihood estimation. In this case we just need to assume the form of the unconditional expected value and variance of y_i . For instance, if we assume that conditional on θ_i , y_i has mean $r_i\theta_i$ and variance $r_i\theta_i(1 - \theta_i)$ and unconditionally $E(y_i) = r_i\mu_i$ and $\text{var}(y_i) = r_i(1 + (r_i - 1)\rho)\mu_i(1 - \mu_i)$, i.e., we assume that $E(\theta_i) = \mu_i$ and $\text{var}(\theta_i) = \rho\mu_i(1 - \mu_i)$, we reproduce the second model obtained assuming θ_i is beta distributed.

Firth(1983) studies the efficiency, when overdispersion is present, of the quasilielihood estimators of the parameters of the linear predictor relative to the maximum likelihood estimators for the members of the natural exponential family with quadratic variance, which includes the binomial distribution.

We focus on the estimation of the parameter which in some way represents the overdispersion, either σ or ρ , especially on the asymptotic efficiency of moment based estimators relative to the maximum likelihood estimator.

Under the assumption of π_i equal to π for all i , the ^{asymptotic} covariance matrix of the maximum likelihood estimators of μ_i and π is given by the inverse matrix of

$$\begin{bmatrix} A & B \\ B & \frac{\sum_{i=1}^m r_i(r_i - 1)}{2} + C\pi \end{bmatrix}, \quad (2.92)$$

where C is defined as in section 2.3.1, A is a $p \times p$ matrix whose elements are of the form

$$a_{kj} = \sum_{i=1}^m \frac{r_i \mu'_{ik} \mu'_{ij}}{\mu_i(1 - \mu_i)} - \pi \sum_{i=1}^m \frac{r_i(r_i - 1) \mu'_{ik} \mu'_{ij}}{\mu_i(1 - \mu_i)}, \quad (2.93)$$

and B is the vector whose elements are given by

$$b_j = \sum_{i=1}^m \frac{\pi r_i(r_i - 1)(2\mu_i - 1) \mu'_{ij}}{2\mu_i(1 - \mu_i)}, \quad (2.94)$$

where

$$\mu'_{ij} = \frac{\partial \mu_i}{\partial \beta_j}. \quad (2.95)$$

For small values of π (large value of $\alpha + \beta$), the asymptotic variance of $\hat{\pi}$ is given by

$$\text{var}(\hat{\pi}) = \frac{2}{\sum_{i=1}^m r_i(r_i - 1)} \left\{ 1 - \frac{2\pi C}{\sum_{i=1}^m r_i(r_i - 1)} \right\} + \frac{4}{\left(\sum_{i=1}^m r_i(r_i - 1) \right)^2} B' A^{-1} B, \quad (2.96)$$

and as π tends to zero the variance tends to

$$\frac{2}{\sum_{i=1}^m r_i(r_i - 1)}, \quad (2.97)$$

independently of the link function and of the values of the explanatory variables.

The asymptotic variance of $\hat{\rho}$ is given by

$$\frac{(1 - 2\rho) \text{var}(\hat{\pi})}{(1 - \rho)^4}, \quad (2.98)$$

its limit as $\alpha + \beta$ tends to infinity (ρ tends to zero), is

$$\frac{2}{\sum_{i=1}^m r_i(r_i - 1)}. \quad (2.99)$$

In order to define an alternative estimator for ρ , we notice first that if all the sample sizes are equal to r , the unconditional variance of y_i reduces to $\mu_i(1 - \mu_i)$ times $\tau\phi$, where

$$\phi = 1 + \rho(r - 1), \quad (2.100)$$

and can be estimated by

$$\frac{r}{mr - p} \sum_{i=1}^m \frac{(y_i - \hat{\mu})^2}{\hat{\mu}_i(1 - \hat{\mu}_i)}, \quad (2.101)$$

(McCullagh & Nelder (1983pp173), Wedderburn(1974)). This is essentially the components of variance estimator with a correction in the denominator because we are

estimating more parameters and its asymptotic variance is

$$\frac{2\phi^2(1 + \frac{\bar{\gamma}_2}{2})}{m - p}, \quad (2.102)$$

where $\bar{\gamma}_2$ is the average of the standardized fourth cumulant of the data. Therefore the variance depends, as with the maximum likelihood estimator, on the particular set of data.

The variance of $\tilde{\rho}$ is

$$\frac{2(1 + \rho(r - 1))^2(1 + \frac{\bar{\gamma}_2}{2})}{(r - 1)^2(m - p)}. \quad (2.103)$$

The asymptotic relative efficiency depends up to the fourth standardized moment of the data. If we assume that as m tends to infinity the standardized fourth sample moment tends to zero, which is not a very strong assumption, the limit as ρ tends to zero ($\alpha + \beta \rightarrow \infty$) is

$$\frac{2}{(r - 1)^2}, \quad (2.104)$$

that corresponds to the variance of the quasiliikelihood estimator in the case of no explanatory variables and does not involve the particular link function.

When the sample sizes are different this estimator does not work. Williams (1982) suggests a procedure to obtain an equivalent estimator. It is based on equating an approximation to the chi square goodness of fit statistic to its expectation, in this case the estimator is given initially by

$$\tilde{\rho} = \frac{X^2 - (m - p)}{\sum_{i=1}^m (r_i - 1)(1 - v_i^* q_i)}, \quad (2.105)$$

where q_i is the i th element of $X'(X'WV^*X)^{-1}X'$, $W = \text{diag}(w_i)$, $V^* = \text{diag}(v_i^*)$, $w_i = \{1 + \hat{\rho}(r_i - 1)\}^{-1}$, $v_i^* = r_i \mu_i (1 - \mu_i)$ and X^2 is the goodness of fit statistic

$$X^2 = \sum_{i=1}^m \frac{w_i (y_i - r_i \hat{\mu}_i)^2}{r_i \hat{\mu}_i (1 - \hat{\mu}_i)}. \quad (2.106)$$

We compare X^2 with a chi squared with $m - q$ degrees of freedom ; if X^2 is larger than the value of the chi squared we use $\tilde{\rho}$ as the estimator of ρ , if X^2 is smaller than or equal to it then using new weights we recalculate the estimators of the parameters, the goodness of fit statistic and we estimate ρ as

$$\tilde{\rho} = \frac{X^2 - \sum_{i=1}^m w_i' (1 - w_i' v_i^* q_i)}{\sum_{i=1}^m w_i' (r_i - 1)(1 - w_i' v_i^* q_i)}, \quad (2.107)$$

until the condition is reached.

Williams(1982)' gives a straightforward procedure to obtain the estimator using GLIM; the problem is that it is not easy to investigate the properties of the estimator obtained.

In a more general way, we might define the extended quasilielihood estimator of ρ , $\tilde{\rho}$, as in the case $\mu_i = \mu$. The variance of $\tilde{\rho}$ is given by expression (2.68) and is independent of any link function or value of the explanatory variables, but as Nelder & Pregibon(1988) point out the values of the variances of the estimators are conditional on the other set (μ_i or β_{ij}) set equal to their estimate.

It is clear that in this case we can not construct a components of variance estimator of the parameter σ_θ^2 because the assumption of heterogeneity of variance does not hold.

Therefore we concentrate on the asymptotic efficiency of the quasilielihood estimator of ρ relative to the maximum likelihood estimator.

The asymptotic variance of the estimator of ρ as ρ tends to zero ($\alpha + \beta$ tends to infinity) is the same as in the case of no explanatory variables given by expression (2.58). For small values of ρ the relative efficiency depends on the values of the explanatory variables and on the particular link function.

If the assumption made is that the variance is constant from sample to sample, we suggest to estimate σ_θ^2 based on the formal analysis of variance. In this case and assuming the same number of observations for each individual the sum of squares between and the sum of squares within from the table of analysis of variance are given by

source	d.f.	sum of squares	$E(\text{mean squared})$
Between	$m-1$	$\sum_{i=1}^m r_i (\bar{y}_i - \bar{y}_{..})^2$	$(r-1)\sigma_\theta^2 + \sum_{i=1}^m \left\{ \frac{\mu_i(1-\mu_i)}{m} + \frac{r\mu_i^2}{m-1} \right\} + \frac{r}{m(m-1)} \left(\sum_{i=1}^m \mu_i \right)^2$
Within	$m(r-1)$	$\sum_{i=1}^m \sum_{j=1}^r (y_{ij} - \bar{y}_i)^2$	$\sum_{i=1}^m \frac{\mu_i(1-\mu_i)}{m} - \sigma_\theta^2$

We suggest to estimate σ_θ^2 as $1/r$ times the difference of the mean squared between and the mean squared within, expressions (2.32) for different sample sizes and (2.33) for equal sample sizes. A difference from the case with no explanatory variables is that this estimator is biased, the expected value in the case of equal sample sizes being

$$\sigma_\theta^2 + \frac{1}{m} \left\{ \sum_{i=1}^m \mu_i^2 - \frac{\left(\sum_{i=1}^m \mu_i \right)^2}{m} \right\} ; \quad (2.108)$$

obviously the bias depends on how different are the mean values. If we assume that $\mu_i = \mu(1 + d_i)$, with $\sum d_i = 0$, we have that the expected value of the estimator can be written as

$$\sigma_{\theta}^2 + \frac{\mu^2}{m} \sum_{i=1}^m d_i^2 . \quad (2.109)$$

To calculate the variance of the estimator we need to make assumptions about the third and fourth moments of the distribution (such as that they are constant from individual to individual); nevertheless for small values of π_i the variance is given by

$$\begin{aligned} & \frac{1}{m^2 r(r-1)} \sum_{i=1}^m \{ 2(r-3)\mu_i^4 + 2\mu_i^2 \} - \frac{4}{m^3 r(r-1)} \sum_{i=1}^m \sum_{j \neq i}^m \{ (r-2)\mu_i^2 \mu_j + (r-1)(r-3)\mu_i^3 \mu_j \} \\ & + \frac{4}{m^4 r} \sum_{i=1}^m \sum_{j \neq i}^m \sum_{k \neq j, i}^m \mu_i (1 - \mu_i) \mu_j \mu_k + \frac{8\sigma_{\theta}^2}{m^3 r} \sum_{i=1}^m \sum_{j \neq i}^m \mu_i \mu_j \\ & \frac{2\sigma_{\theta}^2}{mr(r-1)} + \frac{\sigma_{\theta}^2}{m^2 r(r-1)} \left\{ \sum_{i=1}^m 4\mu_i(2r-5) + 6\mu_i^2(r-2)(r-3) - \mu_i^3 r(r-1) \right\} . \quad (2.110) \end{aligned}$$

As $\alpha_i + \beta_i$ tends to infinity keeping α_i/β_i constant all the moments but the first ones tend to zero and therefore the limiting asymptotic variance is given by expression (2.110) without the last three terms.

In the case of the maximum likelihood estimator we obtain an estimator maximizing expression (2.91); we can always solve the problem numerically but finding an analytically expression for the variance is quite complicated. The asymptotic relative efficiency depends on the particular link function.

2.9 GENERALIZATION TO MORE THAN TWO CATEGORIES

In order to generalize to more than two categories we have to take account of whether the categories have an order or not. It is obvious that, for the models considered here, in the case of binary variables the order does not matter and the categories are interchangeable without altering the fit of the model or effect of different variables.

When the categories are not ordered they are fully interchangeable without altering results or inferences from a fitted model, whereas when order is present the only possible permutation of the categories is the one which reverses the order; not all the models for ordinal data have this property but a large number of them do. The lack of this property could be or could not be of importance depending on the specific context.

We discuss first the case of categories without order. For simplicity we present the case of three categories and we assume m samples of size r ; it is clear that the generalization to more than three categories or to different sample sizes is straightforward.

Without implying any order we might label the categories as 0, 1 and 2 and define

$$\begin{aligned} P(y_{ij} = 1 \mid p_i, q_i) &= p_i, \\ P(y_{ij} = 2 \mid p_i, q_i) &= q_i, \\ P(y_{ij} = 0 \mid p_i, q_i) &= 1 - p_i - q_i. \end{aligned} \tag{2.111}$$

If we define y_{ip} to be the variable whose value is the total number of ones in sample i and similarly y_{iq} to be the variable whose value is the total number of twos in

sample i , the joint distribution of y_{ip} and y_{iq} given p_i and q_i is multinomial, given by

$$f(y_{iq}, y_{ip} | p_i, q_i) = \frac{r! p_i^{y_{ip}} q_i^{y_{iq}} (1 - p_i - q_i)^{r - y_{ip} - y_{iq}}}{y_{ip}! y_{iq}! (r - y_{ip} - y_{iq})!}, \quad (2.112)$$

with $0 \leq y_{ip} \leq r$, $0 \leq y_{iq} \leq r$, $y_{ip} + y_{iq} \leq r$.

Assuming that p_i and q_i are random variables with joint distribution, dropping the subscript i for simplicity, given by

$$f_{p,q}(p, q) = \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} p^{(\alpha-1)} q^{(\beta-1)} (1 - p - q)^{(\gamma-1)}, \quad (2.113)$$

$0 \leq p \leq 1$, $0 \leq q \leq 1$, $0 \leq p + q \leq 1$, then the compound distribution for one observation is given by

$$\int_0^1 \int_0^q \frac{r! p^{y_p} q^{y_q} (1 - p - q)^{r - y_p - y_q}}{y_p! y_q! (r - y_p - y_q)!} \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} p^{\alpha-1} q^{\beta-1} (1 - p - q)^{\gamma-1} dp dq, \quad (2.114)$$

and the unconditional distribution of y_p and y_q is

$$\frac{B(\alpha + y_p, \beta + y_q, \gamma + r - y_p - y_q)}{(r + 2)(r + 1)B(\alpha, \beta, \gamma)B(y_p + 1, y_q + 1, r - y_p - y_q + 1)}, \quad (2.115)$$

$$\text{where } B(r, s, t) = \frac{\Gamma(r)\Gamma(s)\Gamma(t)}{\Gamma(r + s + t)},$$

which can be written as

$$\frac{r! \prod_{k=0}^{y_p-1} (\alpha + k) \prod_{k=0}^{y_q-1} (\beta + k) \prod_{k=0}^{r-y_p-y_q-1} (\gamma + k)}{y_p! y_q! (r - y_p - y_q)! \prod_{k=0}^{r-1} (\alpha + \beta + \gamma + k)} \quad (2.116)$$

The distribution (2.113) is known as the Dirichlet distribution and can be seen as a multivariate generalization of the beta distribution. The marginal distributions of p_i and q_i are beta with parameters $\alpha, \beta + \gamma$ and $\beta, \alpha + \gamma$ respectively.

As in the binary case, we are interested in the parameters of the distribution of p and q ; we call them $\mu_p, \mu_q, \sigma_p^2, \sigma_q^2, \sigma_{pq}^2$, where

$$\mu_p = \frac{\alpha}{\alpha + \beta + \gamma}, \quad \mu_q = \frac{\beta}{\alpha + \beta + \gamma},$$

$$\sigma_p^2 = \frac{\alpha(\beta + \gamma)}{(\alpha + \beta + \gamma)^2 (\alpha + \beta + \gamma + 1)} = \mu_p (1 - \mu_p) \frac{\pi}{1 + \pi},$$

$$\sigma_q^2 = \frac{\beta(\alpha + \gamma)}{(\alpha + \beta + \gamma)^2 (\alpha + \beta + \gamma + 1)} = \mu_q (1 - \mu_q) \frac{\pi}{1 + \pi},$$

$$\sigma_{pq}^2 = - \frac{\alpha\beta}{(\alpha + \beta + \gamma)^2 (\alpha + \beta + \gamma + 1)} = - \mu_p \mu_q \frac{\pi}{1 + \pi}, \quad (2.117)$$

and $\pi = 1/(\alpha + \beta + \gamma)$.

Note that a disadvantage of the use of the Dirichlet distribution as a mixing distribution is that does not allow for positive correlation of p and q .

Reparameterising the distribution in terms of μ_p, μ_q and π , the likelihood is

written

$$\frac{r! \prod_{k=0}^{y_p-1} (\mu_p + k\pi) \prod_{k=0}^{y_q-1} (\mu_q + k\pi) \prod_{k=0}^{r-y_p-y_q-1} (1 - \mu_p - \mu_q + k\pi)}{y_p! y_q! (r - y_p - y_q)! \prod_{k=0}^{r-1} (1 + k\pi)} . \quad (2.118)$$

For small values of π , large values of $\alpha + \beta + \gamma$, the likelihood can be written as

$$\mu_p^{y_p} \mu_q^{y_q} (1 - \mu_p - \mu_q)^{r-y_p-y_q} \left\{ 1 + \pi \left(\frac{y_p(y_p-1)}{2\mu_p} + \frac{y_q(y_q-1)}{2\mu_q} + \frac{(r-y_p-y_q)(r-y_p-y_q-1)}{2(1-\mu_p-\mu_q)} - \frac{r(r-1)}{2} \right) \right\} . \quad (2.119)$$

Using (2.119) the covariance matrix of the maximum likelihood estimates is given by m^{-1} times the inverse matrix of

$$\begin{bmatrix} \frac{r(1-\mu_q)}{\mu_p(1-\mu_p-\mu_q)} + \pi A & \frac{r}{(1-\mu_p-\mu_q)} + \pi F & \pi D \\ \frac{r}{(1-\mu_p-\mu_q)} + \pi F & \frac{r(1-\mu_p)}{\mu_q(1-\mu_p-\mu_q)} + \pi B & \pi E \\ \pi D & \pi E & r(r-1) + \pi C \end{bmatrix} , \quad (2.120)$$

where

$$H = \sum_{i=1}^m h_i , \quad (2.121)$$

for $H = A, B, C, D, E, F$ and $h = a, b, c, d, e, f$ and, dropping the subscript i for simplicity

$$\begin{aligned}
a = & \frac{2r(r-1)}{\mu_p} + \frac{3r(r-1)}{2(1-\mu_p-\mu_q)} - \frac{r(r-1)\mu_q}{\mu_p} - \frac{r(r-1)\mu_p}{(1-\mu_p-\mu_q)(1-\mu_q)} \\
& - \frac{r(r-1)\mu_q}{(1-\mu_p-\mu_q)(1-\mu_p)} - \frac{r(r-1)(1-\mu_p-\mu_q)}{\mu_p(1-\mu_q)} - r(r-1), \quad (2.122)
\end{aligned}$$

$$\begin{aligned}
c = & r(r-1) \left\{ \frac{3}{2\mu_p} + \frac{3}{2\mu_q} + (\mu_p + \mu_q) \left(\frac{(r-2)(r-3)}{4} - \frac{2r(r-2)}{3} \right) + \mu_p \mu_q \frac{(r-2)(r-3)(r-4)}{6} \right. \\
& + \frac{1}{12(1-\mu_p-\mu_q)} (5r^2 - 3r) + \frac{\mu_p^2 + \mu_q^2}{(1-\mu_p-\mu_q)} + \frac{\mu_p + \mu_q}{3(1-\mu_p-\mu_q)} (r-1)^2 (r-3) + \frac{\mu_p}{6\mu_q} (r-2) + \frac{\mu_q}{6\mu_p} (r-2) \\
& + \frac{\mu_p \mu_q}{(1-\mu_p-\mu_q)} \left(\frac{(r-1)(r-2)(r-3)}{2} + (r-2)(r-3) \right) - \frac{\mu_p^2(1-\mu_q) + \mu_q^2(1-\mu_p)}{12(1-\mu_p-\mu_q)} (r-3)(r-4) + 2 \\
& - \frac{\mu_p \mu_q (\mu_p + \mu_q)}{2(1-\mu_p-\mu_q)} (r-2)(r-3) + \frac{3(r-3)(r-4)}{2} - \frac{2r^2(r-1) + r(r-1) + 6(r-2)}{12} \\
& \left. + \frac{\mu_p + \mu_q}{4(1-\mu_p-\mu_q)} (2r-1)(r-2) + \frac{\mu_p^2 + \mu_q^2}{12(1-\mu_p-\mu_q)} (2r-1)(r-2)(r-3) \right\}, \quad (2.123)
\end{aligned}$$

$$d = r(r-1)^2 - \frac{r(r-1)}{\mu_p} + \frac{r(r-1)(\mu_p + \mu_q)}{(1-\mu_p-\mu_q)}, \quad (2.124)$$

$$f = \frac{r(r-1)}{\mu_p} + \frac{r(r-1)}{\mu_q} + \frac{r(r-1)(2r-1)}{3} - \frac{r(r-1)(r-2)}{3}, \quad (2.125)$$

b is equivalent to a and e to d interchanging μ_p and μ_q .

The variances of $\hat{\mu}_p$ and $\hat{\mu}_q$ are obtained directly from the matrix of covariance, whereas the variance of $\hat{\sigma}_p^2$, $\hat{\sigma}_q^2$, $\hat{\sigma}_{pq}^2$ are

$$\text{var}(\hat{\sigma}_l^2) = \frac{(1-2\mu_l)^2\pi^2}{(1+\pi)^2}\text{var}(\hat{\mu}_l) + \frac{\mu_l^2(1-\mu_l)^2}{(1+\pi)^4}\text{var}(\hat{\pi}) + \frac{2(1-2\mu_l)(1-\mu_l)\mu_l\pi}{(1+\pi)^3}\text{cov}(\hat{\mu}_l, \hat{\pi}), \quad l=p,q$$

(2.126)

$$\begin{aligned} \text{var}(\hat{\sigma}_{pq}^2) = & \frac{\mu_p^2\mu_q^2}{(1+\pi)^4}\text{var}(\hat{\pi}) + \frac{\mu_q^2\pi^2}{(1+\pi)^2}\text{var}(\hat{\mu}_p) + \frac{\mu_p^2\pi^2}{(1+\pi)^2}\text{var}(\hat{\mu}_q) + \frac{2\mu_p\mu_q\pi^2}{(1+\pi)^2}\text{cov}(\hat{\mu}_p, \hat{\mu}_q) \\ & + \frac{2\mu_p\mu_q^2\pi}{(1+\pi)^3}\text{cov}(\hat{\mu}_p, \hat{\pi}) + \frac{2\mu_p^2\mu_q\pi}{(1+\pi)^3}\text{cov}(\hat{\mu}_q, \hat{\pi}). \end{aligned}$$

(2.127)

As π tends to zero the covariance matrix tends to m^{-1} times

$$\begin{bmatrix} \frac{\mu_p(1-\mu_p)}{r} & -\frac{\mu_p\mu_q}{r} & 0 \\ -\frac{\mu_p\mu_q}{r} & \frac{\mu_q(1-\mu_q)}{r} & 0 \\ 0 & 0 & \frac{1}{r(r-1)} \end{bmatrix}. \quad (2.128)$$

In order to define alternative estimators, we suggest a components of variance estimator from a multivariate analysis of variance. We define the bivariate binary vector y_{ij} , with components y_{ijp} and y_{ijq} , in the following way: y_{ijp} is equal to one if observation j of individual i belongs to category one and zero otherwise and y_{ijq} is equal to one if observation j of individual i belongs to category two and zero otherwise, giving

$$\begin{aligned}
E(y_{ijp} | p_i, q_i) &= p_i, & E(y_{ijq} | p_i, q_i) &= q_i, \\
\text{var}(y_{ijp} | p_i, q_i) &= p_i(1 - p_i), & \text{var}(y_{ijq} | p_i, q_i) &= q_i(1 - q_i).
\end{aligned}
\tag{2.129}$$

We also define

$$\begin{aligned}
E(p_i) &= \mu_p, & E(q_i) &= \mu_q, \\
\text{var}(p_i) &= \sigma_p^2, & \text{var}(q_i) &= \sigma_q^2, \\
\text{cov}(p_i, q_i) &= \sigma_{pq}^2.
\end{aligned}
\tag{2.130}$$

Then unconditionally

$$\begin{aligned}
E(y_{ip}) &= r\mu_p, & E(y_{iq}) &= r\mu_q, \\
\text{var}(y_{ip}) &= r\mu_p(1 - \mu_p) + r(r - 1)\sigma_p^2, & \text{var}(y_{iq}) &= r\mu_q(1 - \mu_q) + r(r - 1)\sigma_q^2, \\
\text{cov}(y_{ip}, y_{iq}) &= -r\mu_p\mu_q + r(r - 1)\sigma_{pq}^2,
\end{aligned}
\tag{2.131}$$

where $y_{il} = \sum y_{ijl}$, $l = p, q$.

The Multivariate Analysis of Variance table is given by

source	d.f.	sum of squares
Between	$m - 1$	$B = \sum_{i=1}^m r_i (\bar{y}_i - \bar{y}_{..}) (\bar{y}_i - \bar{y}_{..})'$
Within	$\sum_{i=1}^m r_i - m$	$W = \sum_{i=1}^m \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_i) (y_{ij} - \bar{y}_i)'$
Total	$\sum_{i=1}^m r_i - 1$	$\sum_{i=1}^m \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{..}) (y_{ij} - \bar{y}_{..})'$

where $\bar{y}_i = (\sum y_{ijp}/r_i, \sum y_{ijq}/r_i)' = (\bar{y}_{ip}, \bar{y}_{iq})$ and $\bar{y}_{..} = (\sum y_{ip}/\sum r_i, \sum y_{iq}/\sum r_i)' = (\bar{y}_{..p}, \bar{y}_{..q})$.

Assuming that the sample sizes are equal, we have

$$E\left(\frac{1}{m-1}B - \frac{1}{m(r-1)}W\right) = r \begin{bmatrix} \sigma_p^2 & \sigma_{pq}^2 \\ \sigma_{pq}^2 & \sigma_q^2 \end{bmatrix}, \quad (2.132)$$

and therefore, for large m , is

$$\tilde{\sigma}_l^2 = \frac{r}{m(r-1)} \sum_{i=1}^m (\bar{y}_{il}^2 - \bar{y}_{..l}^2) - \frac{1}{r-1} \bar{y}_{..l} (1 - \bar{y}_{..l}), \quad l = p, q, \quad (2.133)$$

is an unbiased estimator for σ_l^2 .

An unbiased estimator for σ_{pq}^2 is

$$\tilde{\sigma}_{pq}^2 = \frac{r}{m(r-1)} \sum_{i=1}^m \bar{y}_{ip} \bar{y}_{iq} - \bar{y}_{..p} \bar{y}_{..q}, \quad (2.134)$$

which, for large m , is equivalent to

$$\tilde{\sigma}_{pq}^2 = \frac{1}{mr(r-1)} \sum_{i=1}^m y_{ip} y_{iq} - \frac{1}{m^2 r} \sum_{i=1}^m \sum_{j \neq i}^m y_{ip} y_{jq}. \quad (2.135)$$

The variances of $\tilde{\sigma}_p^2$ and $\tilde{\sigma}_q^2$ correspond to the variance of the estimator of $\tilde{\sigma}_\theta^2$ in the binary case, replacing the moments of the beta distribution by the moments of the marginal beta distributions of p and q respectively. The variance of $\tilde{\sigma}_{pq}^2$ is given by

$$\begin{aligned} & \frac{1}{mr(r-1)} \{ \mu_p \mu_q + \sigma_{pq}^2 + \mu_p^2 \mu_q + \mu_p \mu_q^2 + (r-2)(r-3) \mu_{22} + r(r-1) \mu_p^2 \mu_q^2 - r(r-1) \sigma_{pq}^2 \\ & + 2(r-1)^2 \sigma_{pq}^2 \mu_p \mu_q + (r-1)^2 (\mu_p^2 \sigma_q^2 + \mu_q^2 \sigma_p^2) - 2(r-1)(r-2) (\mu_p \mu_{12} + \mu_q \mu_{21}) \}. \end{aligned} \quad (2.136)$$

Asymptotically, as $\alpha + \beta + \gamma \rightarrow \infty$, all the moments but the first ones, μ_p and μ_q , of the marginal beta distributions of p and q tend to zero, whereas for the moments of the Dirichlet distribution μ_{12} tends to $\mu_p \mu_q^2$, μ_{21} to $\mu_p^2 \mu_q$, μ_{22} to $\mu_p^2 \mu_q^2$ where μ_{st} means $E(p^s q^t)$. Therefore, as $\alpha + \beta + \gamma \rightarrow \infty$, the variance of $\tilde{\sigma}_{pq}^2$ tends to

$$\frac{1}{mr(r-1)} \{ \mu_p \mu_q - \mu_p^2 \mu_q - \mu_p \mu_q^2 + 2\mu_p^2 \mu_q^2 \}. \quad (2.137)$$

For some values of α , β and γ the relative efficiency of the components of variance estimator of σ_p^2 (or σ_q^2) are given in table 2.18.

Table 2.18.
Asymptotic relative efficiency of the components of variance estimator to the maximum likelihood estimator.

α	$\beta = \gamma = \alpha$ $\mu_1 = 1/3$	$\beta = \gamma = 2\alpha$ $\mu_1 = 1/4$	$\beta = 2\gamma = 2\alpha$ $\mu_1 = 1/5$	$\beta = \gamma = \alpha/2$ $\mu_1 = 1/2$
$r = 2$				
1	.547	.618	.576	.552
3	.522	.552	.533	.535
5	.514	.534	.521	.527
10	.508	.518	.511	.516
20	.504	.509	.506	.509
100	.501	.502	.501	.502
$r = 10$				
1	.481	.573	.506	.531
3	.495	.536	.498	.580
5	.500	.525	.498	.588
10	.503	.515	.498	.576
20	.502	.508	.499	.553
100	.500	.502	.500	.514

The asymptotic relative efficiency of the moment based estimator of σ_p^2 (or σ_q^2) with respect to the maximum likelihood estimator as $\alpha + \beta + \gamma$ tends to infinity tends to .5 for any value of the sample size when all are equal. It is possible to show that in general, if k is the number of possible categories of response, the limit of the

asymptotic efficiency, as $\alpha + \beta + \gamma$ tends to infinity, of the components of variance estimator relative to the maximum likelihood estimator tends to $1/(k-1)$. This can be explained by the fact that we are not using all the information available; for instance, in the three categories case, when y_{ip} is equal to zero the observation might belong to category zero or to category two and we are not incorporating that information into the estimator. This does not happen in the binary case because all the information is contained in the estimator.

In general as the sample size, r , increases the efficiency increases and its convergence to the limit value is slower.

It is clear that estimating μ_p by $\sum \sum y_{ijp}/(mr)$, and analogously μ_q by $\sum \sum y_{ijq}/(mr)$ is as efficient as in the binary case.

In the case of the covariance we find that the asymptotic efficiency of the moment estimator is always less than or equal to .25, but whereas the maximum likelihood estimator can only take negative values, the components of variance estimator is not restricted. Thus if for some reason we think that the correlation can only be negative it is better to use the maximum likelihood estimator; otherwise the components of variance estimator might be preferable.

We notice that as in the binary case, we can express the covariance matrix of the observations in terms of the mean of the population and a parameter which can be seen as an index of overdispersion. Writing the matrix of covariance as m^{-1} times

$$r(1 + (r-1)\rho) \begin{bmatrix} \mu_p(1 - \mu_p) & -\mu_p\mu_q \\ -\mu_p\mu_q & \mu_q(1 - \mu_q) \end{bmatrix}, \quad (2.138)$$

which corresponds to $(1 + (r-1)\rho)$ times the variance expected under a multinomial model, therefore ρ can be seen as an index of relative dispersion. It is clear that we can have different moment based estimators of ρ , such as the ones based on the ratio $\tilde{\sigma}_l^2 / \{ \bar{y}_{..l}(1 - \bar{y}_{..l}) \}$, for $l = p, q$, or alternatively the ratio $\tilde{\sigma}_{pq}^2 / \{ \bar{y}_{..p}\bar{y}_{..q} \}$. The first

ones are preferable because they are more efficient than the second one relative to the maximum likelihood estimator; the choice of $l = p$ or $l = q$ depends on the particular values of μ_p and μ_q , the more efficient estimator being the one based on the bigger value, i.e., if $\bar{y}_{..p}$ is bigger than $\bar{y}_{..q}$ the estimator based on $\tilde{\sigma}_p^2$ and $\bar{y}_{..p}$. Again, as in the binary case, the comments referring to the asymptotic relative efficiency of the components of variance estimator to the maximum likelihood estimator of the variance apply for the estimation of ρ .

One alternative estimator of ρ is the quasiliikelihood estimator; it is not difficult to see that we could use extended quasiliikelihood and could estimate ρ by

$$\tilde{\rho} = \frac{1}{\sum_{i=1}^m \frac{(r_i-1)^2}{(1+(r_i-1)\rho)^2}} \left\{ \sum_{i=1}^m \frac{(r_i-1)}{(1+(r_i-1)\rho)^2} \left\{ r_i \left(\frac{\bar{y}_{ip}}{\bar{y}_{iq}} (\log(1-\bar{y}_{ip}-\bar{y}_{iq}) - \log(\bar{y}_{ip}) + (1-\bar{y}_{iq})\log(1-\bar{y}_{ip}-\bar{y}_{iq})) \right. \right. \right. \\ \left. \left. \left. - \frac{\bar{y}_{ip}}{\tilde{\mu}_q} (\log(1-\tilde{\mu}_p-\tilde{\mu}_q) - \log(\tilde{\mu}_q)) + (1-\bar{y}_{iq})\log(1-\tilde{\mu}_p-\tilde{\mu}_q) - 1 \right) \right\} \right\} , \quad (2.139)$$

(or the equivalent interchanging $\bar{y}_{ip}$ with $\bar{y}_{iq}$ and $\tilde{\mu}_p$ with $\tilde{\mu}_q$) where $\tilde{\mu}_p$ and $\tilde{\mu}_q$ are the quasiliikelihood estimators of the mean given by expression (2.40). The variance of both estimators is given by expression (2.68) and the limit of the asymptotic efficiency, as $\alpha + \beta + \gamma \rightarrow \infty$ is

$$\frac{\sum_{i=1}^m (r_i - 1)^2}{2 \sum_{i=1}^m r_i (r_i - 1)} , \quad (2.140)$$

which is always less than .5. When the number of observations in each sample is small, the estimator based in the components of variance estimator and the sample mean is preferable to the quasiliikelihood.

In the case when the categories of response are ordered, in the simple situation, with no explanatory variables, we could use the same models. In this case to assign to the categories numbers, like 0, 1, 2 is not just a label, we have to assign them in natural order according to the order of the categories and the only permutation that might be possible is that one which reverses the order of the categories. But even in the ordinal scale, the use of the labels 1,..., k does not imply that the spacing or distance between categories is in any sense regular.

The same moment based estimators are obtained by equating the first two sample moments and the first mixed sample moment to their expectation. In the case of three categories and using the same notation as before we have

$$\begin{aligned}
m_{10} &= \sum_{i=1}^m \frac{y_{ip}}{m}, & E(m_{10}) &= r\mu_p, \\
m_{01} &= \sum_{i=1}^m \frac{y_{iq}}{m}, & E(m_{01}) &= r\mu_q, \\
m_{11} &= \sum_{i=1}^m \frac{y_{ip}y_{iq}}{m}, & E(m_{11}) &= r(r-1)\mu_p\mu_q + r(r-1)\sigma_{pq}^2, \quad (2.141) \\
m_{20} &= \sum_{i=1}^m \frac{y_{ip}^2}{m}, & E(m_{20}) &= r\mu_p + r(r-1)\mu_p^2 + r(r-1)\sigma_p^2, \\
m_{02} &= \sum_{i=1}^m \frac{y_{iq}^2}{m}, & E(m_{02}) &= r\mu_q + r(r-1)\mu_q^2 + r(r-1)\sigma_q^2.
\end{aligned}$$

Equating the moments to their expected values and solving for $\mu_p, \mu_q, \sigma_p^2, \sigma_q^2$ and σ_{pq}^2 we have the same estimators as before.

2.10 GENERALIZATION TO MORE THAN TWO CATEGORIES WHEN THE RESPONSE PROBABILITY DEPENDS ON A LINEAR PREDICTOR

To develop a suitable statistical model it is essential to distinguish between the different types of scale measurement. We consider models for ordinal and nominal categories following McCullagh & Nelder (1978, pp102)

The models for ordinal scale of measurement can be classified into two types, the first one being the models which are invariant under the group of adjacent response categories. These models are based on the cumulative distribution probabilities $\gamma_j (=p_1 + \dots + p_j)$, $j=1, \dots, q$. The simplest models in this case involve parallel regressions on the appropriate chosen scale.

$$g(\gamma_{ij}) = \alpha_j - \beta'x_i \quad j = 1, \dots, q; i = 1, \dots, m \quad (2.142)$$

where g is any suitable function mapping the interval $(0,1)$ onto the whole real line.

The parameters α_j and β are generally unknown and the main interest is in β and not in the α_j .

The second class of models are those where the categories are of interest themselves, but because examples are not very common we do not present them explicitly.

The better known and more useful model for nominal scale of measurements is obtained by taking a linear logistic relation for the comparison of the probabilities of any two levels of response.

In general it is written

$$p_{ij} = \frac{\exp(\eta_{ij})}{\sum_{k=0}^q \exp(\eta_{ik})}, \quad (2.143)$$

where $\eta_{ij} = \beta'_j x_i$ and in order to avoid overparameterisation we set $\beta_q = 0$.

The likelihood for both models is based on the multinomial likelihood and it can be expressed as a function of the unknown parameters by substituting the appropriate systematic model into the likelihood.

When overdispersion is present, we can assume a Dirichlet distribution and the same kind of assumptions as in the binary case, of equal index of overdispersion or, in this case, equal covariance matrix of the variables. In this case it is more sensible to assume that the index of overdispersion is constant from sample to sample, because with the assumption of equal covariance matrix the likelihood becomes very complicated. We could use moment based estimators, but the relative efficiency of the moment estimators of the parameters which define the overdispersion to the maximum likelihood estimators is not very good.

2.11 CONCLUSIONS

When overdispersion is present and the sample sizes in all the individuals or groups are the same the sample mean is a good estimator, in terms of asymptotic efficiency relative to the maximum likelihood estimator, of the mean value of the probabilities of success; the degree of overdispersion is unimportant in this situation. In this case the easiest way to estimate the variance is using the components of variance estimator given by expression (2.33) and its asymptotic efficiency relative to the maximum likelihood estimator is over 95% for values of the sample size in each

individual or group less than or equal to ten. As the sample size in each group increases the asymptotic efficiency decreases, but nevertheless the limiting values as r (the sample size in each group or individual), tends to infinity is still over 80%, for any amount of overdispersion. This estimator is more efficient than the quasiliikelihood estimator of the variance, particularly when the sample sizes and the overdispersion are small.

When the sample sizes are different for each group, the pooled mean is a good estimator, in terms of efficiency, of the mean value of the probabilities of success. The quasiliikelihood estimator is more efficient than the pooled mean but, in general, the difference is not large, unless the sample sizes are very different; in this case Cochran's estimator is the best option.

In the case of the variance the components of variance estimator and the quasiliikelihood estimator are not as efficient as in the equal sample size case. The efficiency of the components of variance estimator decreases as the heterogeneity of the sample sizes increases and also as the overdispersion decreases, but nevertheless the efficiency is still quite high in a wide range of situations.

When instead of a binary response we have a categorical one, neither the quasiliikelihood estimator nor the components of variance estimator are good to estimate the covariance matrix of the vector of probabilities. The efficiency decreases when the number of categories increases. Therefore especially when the overdispersion is small the use of the maximum likelihood estimators is recommended with the exception of the estimation of the covariance between the probabilities of two different categories, because the maximum likelihood estimator is restricted to be negative.

When explanatory variables are present more consideration must be made and more assumptions about the distribution of the random probability of success have to be made. Nevertheless when the overdispersion is small moment based estimators perform well, and they are easier to calculate.

Potthoff & Whittinghill (1966) suggest a test for the presence of overdispersion based on the assumption of a beta distribution for the probability of success, which in the case of the mean of the beta distribution being known is the locally most powerful test; the test statistic is based on sums of the observations, and is therefore straightforward to calculate. In the case that this mean is unknown, the most common one, they suggest a test which is based as well on sums of the observations. This test is conservative, i.e., the actual probability of type I error is always less than or equal to the stated one. As a result in some cases the test loses power. They investigate the power of this test and the traditional chi square test for homogeneity. Finally they extend this test to the case of more than two categories of response, using the assumption that the multinomial probabilities are Dirichlet distributed.

CHAPTER 3 : ASYMPTOTIC EFFICIENCY OF THE LOGISTIC REGRESSION ESTIMATORS RELATIVE TO THE LINEAR DISCRIMINANT FUNCTION

3.1 INTRODUCTION

3.1.1 GENERAL FORMULATION

Logistic regression and discriminant analysis are conceptually two different techniques and yet they are intimately related. We are talking of two different problems, and the solutions proposed for one are different from those proposed for the other. Nevertheless in some situations estimators obtained in one of them might be used in the other one.

In discriminant analysis there are two distinct populations, which we will denote throughout by 1 and 0 (usually two intrinsically different groups). Within each of these populations there is a set of properties x . That is there are two probability densities $f_1(x)$ and $f_0(x)$ and the focus in discriminant analysis is on how the two distributions differ most sharply. In a different form we may wish to assign new individuals, in some optimal way, to one of the two groups given new vectors of variables from new individuals .

Note that the emphasis is strongly on the distributions of x within the two populations, which are supposed to be stable well-defined characteristics of the two populations.

On the other hand logistic regression presupposes a stable statistical relation such that once a vector of explanatory variables, x , is given a probability (that a binary response, y , is equal to one) is determined. The distribution of x is not directly relevant to the definition. Usually y is in some sense determined by x and ideally the vector x would be enough to determine y .

The linear logistic regression is a particular specification of this relation and is conceptually just a special case of a regression or dependency relation.

If a binary variable y , is related to a vector of explanatory variables, x , by means of this model, we have

$$P(y = 1|x) = \frac{e^{\alpha + \beta x}}{(1 + e^{\alpha + \beta x})}, \quad (3.1)$$

$$P(y = 0|x) = \frac{1}{(1 + e^{\alpha + \beta x})}.$$

Within the framework of discriminant analysis, there are two different situations for consideration. In the first the relative frequencies with which the two populations generate data are not defined. We can not represent y by a random variable, i.e., we cannot consider its probability distribution either marginally or conditionally on x . Thus logistic regression is not applicable. The statistic for assessing x in this case is the log likelihood ratio

$$\log(f_1(x)) - \log(f_0(x)), \quad (3.2)$$

with large positive values pointing to the first population and large negative values to the second one.

Now if the densities come from the same exponential family with canonical statistic x and with two different parameter values, then it is easy to see that (3.2) is a linear combination of the components of x . We call the resulting function a linear discriminant function. The most important case is when $f_1(x)$ and $f_0(x)$ are multivariate normal densities with the same covariance matrix Σ and means μ_1 and μ_0 . Then (3.2) becomes

$$-\frac{1}{2}(\mu_1 - \mu_0)' \Sigma^{-1}(\mu_1 + \mu_0) + (\mu_1 - \mu_0)' \Sigma^{-1} x, \quad (3.3)$$

which is the population linear discriminant function.

The second situation within the framework of discriminant analysis is that there are physically defined probabilities π_1 and π_0 , with $\pi_1 + \pi_0 = 1$, so that we can represent the population for an arbitrary individual by a random variable Y , which will take the value one when the individual is from population one and zero when the individual is from population zero. The full properties of such an individual are represented by a random variable (Y, X) , the functions $f_1(x)$ and $f_0(x)$ now specifying conditional densities of X given y equal to one or zero. For a new individual of known x but unknown Y , we have by Bayes' theorem that

$$P(Y = 1 | X = x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_0 f_0(x)}, \quad (3.4)$$

so that

$$\log \left(\frac{P(y = 1 | x)}{P(y = 0 | x)} \right) = \log \left(\frac{\pi_1}{\pi_0} \right) + \log \left(\frac{f_1(x)}{f_0(x)} \right), \quad (3.5)$$

defining a logistic regression equation in which the prior probabilities are isolated into a

single term.

In this formulation the logistic regression emerges as a natural way of describing the uncertainty in attaching a value of y to a new individual with a known value of x ; alternatively, if that particular problem is somewhat notional we can regard the right hand side of (3.5) as a function of x describing the criterion in terms of which the two distributions of X most differ.

One approach to linear discriminant analysis which in a sense is purely descriptive is Fisher's Linear Discriminant Function (LDF) which is constructed as the linear combination of the components of x , $\lambda'x$, which maximizes the ratio of the difference between the means to the standard deviation within populations, assuming a common but arbitrary covariance matrix for the components of x , i.e., the linear combination of the components of x which most strongly separates the two groups (Fisher, 1936); in this case $\lambda \propto \Sigma^{-1} (\mu_1 - \mu_0)$. This approach was developed distribution free and λ might be estimated by replacing μ_i by the sample mean of population i , $\bar{x}_i$, where again μ_i and $\bar{x}_i$ represent column vectors, and Σ by the pooled estimator $S = \{ (n_1 - 1)S_1 + (n_0 - 1)S_0 \} / (n - 2)$, where n_i is the number of individuals in population i , $n (= n_1 + n_0)$ is the total number of individuals in the two populations, and S_i is the sample covariance matrix of the individuals coming from population i .

In the particular case of x being normally distributed, with mean μ_i and variance Σ , $\bar{x}_i$ is the maximum likelihood estimator, $\hat{\mu}_i$, of μ_i and S is the maximum likelihood estimator, $\hat{\Sigma}$, of Σ , and therefore when we replace μ_i by $\bar{x}_i$ and Σ by S we obtain a multiple of the maximum likelihood estimator of the linear discriminant function defined by equation (3.3) and the linear combination function defined by equation (3.4), replacing π_i by n_i/n .

When we are interested in discrimination it is not difficult to see that, for any positive c , $c\lambda (= c\Sigma^{-1}(\mu_1 - \mu_0))$, gives the same rule of classification and therefore the same probability of misclassification; we will see that in terms of coefficient estimation the value of c does not affect inference about the parameters.

In terms of maximum likelihood estimation, the contribution of the i th individual, dropping the subscript for simplicity, can be written as

$$P(Y = y)f_y(x), \quad y = 0,1 \quad (3.6)$$

or

$$P(Y = y|X = x)f(x), \quad (3.7)$$

where in expression (3.6) $f_y(x)$ represents the conditional distribution of x given y , and in expression (3.7) $f(x)$, represents the unconditional distribution of x . Maximization of either expression leads to the same result, in the case of a normal population to Fisher's Linear Discriminant Function, but if in (3.7) we maximize only the first factor we would be using techniques of logistic regression analysis. It follows that logistic regression is inefficient, under the full assumptions, in that the second factor of (3.7) does contain information about relevant parameters and this information has been discarded.

3.1.2 GENERAL COMMENTS

An advantage of logistic regression is that the estimators are conditional on the observed values of the x and do not require an assumption of a distributional form for the x , or if the x are in fact random variables they can have any distribution.

Another point is that the logistic regression model is satisfied for a wide range of conditional distributions of x (among them multivariate binary observations and members of the family of densities $f_y(x)=g(\theta_y,\eta)h(x,\eta)\exp(x'\theta_y)$, with η an arbitrary nuisance parameter). In the case of the linear discriminant function, we have

that when the observations are normally distributed the linear discriminant function estimators, formed by replacing the parameters by their moment estimators, are the maximum likelihood estimators (or a multiple of them), but this is not necessarily true for any distribution; however it is true in the case of observations from any elliptical distribution, which covers a fairly wide range of distributions including the normal.

In terms of discrimination Efron (1975) proved that the linear discriminant function is more efficient than logistic regression when the observations have a multivariate normal distribution, and that this efficiency depends on the Mahalanobis' distance between two normal populations (Mahalanobis (1936)), defined by Δ , where $\Delta^2 = (\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0)$, as well as on the number of individuals in each population (or the probability of an individual being drawn from a specific population). His work is based on comparing the error rates of the misclassification obtained with both methods.

Several authors have compared the two estimators, among them O'Neill (1978), who compared the maximum likelihood estimators of the linear discriminant function with the logistic regression estimators for several distributions of the explanatory variables, in terms of probability of misclassification. Not surprisingly he found that the maximum likelihood estimators are more efficient than the logistic regression. This is not necessarily true (even in terms of misclassification) when we use the moment estimator of the linear discriminant function (Fisher (1936)).

Some other work on this subject has been done either with simulation studies or empirical studies under several distributions of the explanatory variables (see, for instance Press & Wilson(1978) , Amemiya & Powell(1983)), but most of them have been in terms of probabilities of misclassification.

In terms of coefficient estimation Byth & McLachlan (1980) compare the relative efficiency, in terms of mean square error, of the logistic regression estimators to the linear discriminant function for particular distributions of the explanatory variables which satisfy the logistic regression model, in the sense that the logarithm of

the ratio of conditional probabilities is linear in x . They find, that in most of the cases the logistic regression estimators have less bias (order $1/n$) than the maximum likelihood estimators, but nevertheless in terms of mean squared error the maximum likelihood estimators are in general superior.

We start by reviewing both estimators and their properties in the particular case when the p explanatory variables follow a multivariate normal distribution. We calculate the asymptotic relative efficiency for different tests associated with the parameters and finally we discuss the robustness of the logistic regression estimators and the linear discriminant function estimators to different violations of the assumptions, first when the covariance matrix is not the same in both populations and later when the explanatory variables do not follow a multivariate normal but an elliptical or spherical distribution.

3.2 PROPERTIES OF THE ESTIMATORS WHEN THE EXPLANATORY VARIABLES ARE NORMALLY DISTRIBUTED

3.2.1 FISHER'S LINEAR DISCRIMINANT FUNCTION

An estimator of Fisher's linear discriminant function is given by $cS^{-1}d$, where S is a $p \times p$ matrix which estimates Σ , the covariance matrix (assumed equal in both populations) defined as

$$S = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)(x_i - \bar{x}_1)' + \sum_{i=n_1+1}^{n_1+n_0} (x_i - \bar{x}_0)(x_i - \bar{x}_0)'}{n - 2}, \quad (3.8)$$

where

$$\bar{x}_1 = \frac{\sum_{j=1}^{n_1} x_j}{n_1}, \quad \bar{x}_0 = \frac{\sum_{j=n_1+1}^{n_1+n_0} x_j}{n_0}, \quad (3.9)$$

$$d = \bar{x}_1 - \bar{x}_0, \quad (3.10)$$

$$n = n_1 + n_0, \quad (3.11)$$

with n_i the number of individuals in the sample from population i ($i = 0, 1$), c any positive constant and, without loss of generality, we assume that the first n_1 individuals are from population one and the remaining n_0 from population zero.

When the observations are normally distributed $(n-2)S$ is a Wishart distributed random matrix $W(\Sigma; n-2, p)$, independent of $\bar{x}_1$, $\bar{x}_0$ and $\bar{x}_1 - \bar{x}_0$; therefore using properties of the Wishart distribution we have that

$$E(cS^{-1}d) = cE(S^{-1})E(d) = \frac{c(n-2)}{n-p-3} \Sigma^{-1}(\mu_1 - \mu_0) = \frac{c(n-2)}{n-p-3} \beta'. \quad (3.12)$$

To obtain the covariance matrix of the linear discriminant function estimators we follow closely the work of Das Gupta (1968) who obtained it for the case of one population and $\Sigma = I$.

Writing $(n-2)S$ as $\bar{S}$, the linear discriminant function is defined as

$$b = c(n-2)\bar{S}^{-1}(\bar{x}_1 - \bar{x}_0). \quad (3.13)$$

Assuming that the observations from population i have a normal distribution with mean μ_i and variance Σ , it is clear that $\bar{x}_i$ has a normal distribution with mean μ_i and variance $n_i^{-1}\Sigma$, and therefore $\bar{x}_1 - \bar{x}_0$ is normally distributed with mean $\mu_1 - \mu_0$

and variance $(n_1^{-1} + n_0^{-1}) \Sigma$. The covariance matrix of b is given by

$$\text{cov}(b) = E(bb') - E(b)E(b'). \quad (3.14)$$

To obtain these expectations we note first that if we write $\Sigma = AA'$, and define $S^* = A^{-1}\bar{S}(A')^{-1}$; then S^* is Wishart distributed $W(I_p; n-2, p)$. If we define

$$b^* = S^{*-1}y, \quad (3.15)$$

where $y = A^{-1}(\bar{x}_1 - \bar{x}_0) = A^{-1}d$, we have that

$$b = (n-2)(A')^{-1}b^* \quad (3.16)$$

and the covariance of b is then given by

$$\text{cov}(b) = (n-2)^2(A')^{-1}\text{cov}(b^*)A^{-1}. \quad (3.17)$$

Das Gupta shows that if S^* is Wishart distributed $W(I_p; m, p)$ and v is a $p \times 1$ vector

$$E(S^{*-1}vv'S^{*-1}) = \frac{v'v I_p + (m-p-1)vv'}{(m-p)(m-p-1)(m-p-3)}; \quad (3.18)$$

in particular, if v is a random vector we replace $v'v$ and vv' by their expected values.

Therefore noting that

$$E(b^*) = (n-2) \frac{A^{-1}(\mu_1 - \mu_0)}{(n-p-3)}, \quad (3.19)$$

we have that

$$\text{cov}(b^*) = (n-2)^2 \left\{ \frac{E(y'y) I_p + (m-p-1)E(yy')}{(m-p)(m-p-1)(m-p-3)} - \frac{A^{-1}\mu\mu'(A^{-1})'}{(m-p-1)^2} \right\}, \quad (3.20)$$

where $\mu = \mu_1 - \mu_0$ and $m = n-2$. Because y is a normally distributed random variable with mean $A^{-1}\mu$ and variance $(n_1^{-1} + n_0^{-1})I$, we have that

$$E(yy') = A^{-1}\mu\mu'(A^{-1})' + \left(\frac{1}{n_1} + \frac{1}{n_0} \right) I_p, \quad (3.21)$$

and

$$E(y'y) = \left(\frac{1}{n_1} + \frac{1}{n_0} \right) p + \mu'(A^{-1})'A^{-1}\mu = \left(\frac{1}{n_1} + \frac{1}{n_0} \right) p + \mu'\Sigma^{-1}\mu. \quad (3.22)$$

Therefore

$$\text{cov}(b^*) = \frac{m^2 \left((m-1) \left(\frac{1}{n_1} + \frac{1}{n_0} \right) I_p + (\mu'\Sigma^{-1}\mu) I_p + A^{-1}\mu\mu'(A^{-1})' \left(\frac{m-p+1}{m-p-1} \right) \right)}{(m-p)(m-p-1)(m-p-3)}, \quad (3.23)$$

and

$$\text{cov}(cS^{-1}d) = \frac{c^2(n-2)^2 \left((\mu'\Sigma^{-1}\mu) \Sigma^{-1} + (n-3) \frac{n}{n_1 n_0} \Sigma^{-1} + \Sigma^{-1} \mu \mu' \Sigma^{-1} \frac{n-p-1}{n-p-3} \right)}{(n-p-2)(n-p-3)(n-p-5)}. \quad (3.24)$$

We notice that the limit distribution, as $n \rightarrow \infty$, of $\sqrt{n}(b_n - \Sigma^{-1}\mu)$, (where we write b_n to emphasize the fact that the linear discriminant function has been calculated

from n observations) is equal to the limit distribution of $(A')^{-1} \sqrt{n} (b_n^* - v)$, where $v = A^{-1}\mu$.

To obtain the asymptotic distribution of $\sqrt{n} (b_n^* - v)$, we note that

$$\begin{aligned} \sqrt{n} (b_n^* - v) &= (n-2)S^{*-1} \left(\frac{n}{n_1}\right)^{1/2} \sqrt{n_1} (\bar{x}_1^* - v_1) + (n-2)S^{*-1} \left(\frac{n}{n_0}\right)^{1/2} \sqrt{n_0} (\bar{x}_0^* - v_0) \\ &\quad - \left(\frac{n}{n-2}\right)^{1/2} \sqrt{(n-2)} \left(\frac{S^*}{n-2} - I\right) ((n-2)S^{*-1})v. \end{aligned} \quad (3.25)$$

As $n \rightarrow \infty$,

$$\begin{aligned} (n-2)S^{*-1} &\rightarrow I_p \text{ in probability,} \\ \sqrt{n_i}(\bar{x}_i^* - v_i) &\rightarrow N(0, I_p) \text{ in distribution,} \end{aligned} \quad (3.26)$$

and

$$\sqrt{(n-2)}((n-2)^{-1}S^* - I_p)v \rightarrow N(0, vv' + v'v I_p) \text{ in distribution,}$$

and we have that $\sqrt{n} (b_n^* - v)$ tends to $N(0, vv' + (n^2/(n_1 n_0) + v'v)I_p)$. Hence the limiting distribution of $\sqrt{n} (b_n - \Sigma^{-1}\mu)$ is normal with mean zero and variance given by

$$\frac{1}{n} \left(c^2 \left(\Delta^2 + \frac{n^2}{n_1 n_0} \right) \Sigma^{-1} + \Sigma^{-1} \mu \mu' \Sigma^{-1} \right). \quad (3.27)$$

This result will be used to compare the linear discriminant function to the logistic regression estimators.

3.2.2 LOGISTIC REGRESSION MAXIMUM LIKELIHOOD ESTIMATORS

In order to obtain estimators of the parameters of the logistic regression model we maximize the conditional likelihood given by

$$l_{Y|X}(\alpha, \beta; y | x) = \prod_{i=1}^n \frac{e^{(\alpha + \beta x_i)y_i}}{(1 + e^{\alpha + \beta x_i})} , \quad (3.28)$$

where in order to have a unique solution the explanatory variables must satisfy that the rank of X , the matrix whose rows are the individuals and whose columns are the explanatory variables has rank equal to the number of explanatory variables, provided that there are more individuals than explanatory variables (O'Neill, 1980).

Asymptotically the estimators are normally distributed with mean α and β' and variance given by the inverse of

$$n \int_{-\infty}^{\infty} \frac{e^{\alpha + \beta'x}}{(1 + e^{\alpha + \beta'x})^2} (1, x)'(1, x) dF(x) , \quad (3.29)$$

where $dF(x) = (\pi_1 f_1(x) + \pi_0 f_0(x))dx$. In the particular case of x being normally distributed $f_i(x) = (2\pi)^{-1/2} |\Sigma|^{-1/2} \exp(-(x - \mu_i)' \Sigma^{-1} (x - \mu_i) / 2)$.

In order to calculate the asymptotic variance of the estimators we use the well known fact that if x is a normally distributed random variable with mean μ and variance Σ , $x^* = Ax + a$ is a normally distributed random variable with mean $A\mu + a$

and variance $A\Sigma A'$. Based on this, and for simplicity, we define a standard situation. By a standard situation we understand that x (dropping the superscript for simplicity of notation) is normally distributed with mean μ_s and variance I . In this particular work we define $\mu_{s_1} = (\Delta/2, 0, \dots, 0)$ and $\mu_{s_0} = (-\Delta/2, 0, \dots, 0)$. The reason for this particular choice is that in this case $E(\alpha) = \log(\pi_1 / \pi_2)$ and $E(\beta') = (\Delta, 0, \dots, 0)$, simplifying the integrals which define the variance as much as possible.

In order to have this standard situation, A must be such that $A'A = \Sigma^{-1}$, and for some a , $A\mu_1 + a = (\Delta/2, 0, \dots, 0)$ and $A\mu_0 + a = (-\Delta/2, 0, \dots, 0)$. In other words, if we call v_{ij} the element (i, j) of Σ^{-1} , the elements of A are the solution of the following equations

$$\sum_{k=1}^p a_{ki}^2 = v_{ii}, \quad i = 1, \dots, p, \quad (3.30)$$

$$\sum_{k=1}^p a_{kj} a_{ki} = v_{ij}, \quad i, j = 1, \dots, p, \quad j \neq i, \quad (3.31)$$

$$\sum_{k=1}^p a_{1k} \mu_{(1)} = \Delta, \quad (3.32)$$

$$\sum_{k=1}^p a_{jk} \mu_{(k)} = 0, \quad j = 2, \dots, p \quad (3.33)$$

with $\mu_{(j)}$ equal to the j th component of μ ($= \mu_1 - \mu_0$).

There are more equations than variables; therefore a solution might not exist.

We obtain a solution by setting

$$a_{1j} = \frac{\beta_j}{\Delta}, \quad j = 1, \dots, p-1, \quad (3.34)$$

$$a_{1p} = \frac{\Delta}{\mu_{(p)}} - \sum_{k=1}^{p-1} a_{1k} \frac{\mu_{(k)}}{\mu_{(p)}} \quad (3.35)$$

and letting

$$\sum_{k=2}^p a_{kj}^2 = v_{jj} - a_{1j}^2, \quad j = 1, \dots, p, \quad (3.36)$$

and

$$\sum_{k=2}^p a_{kj} a_{ki} = v_{ij} - a_{1j} a_{1i}, \quad i, j = 1, \dots, p, j \neq i. \quad (3.37)$$

The conditions define by equations (3.32) and (3.33) are satisfied when we solve equations (3.36) and (3.37), having imposed values for a_{1j} , $j=1, \dots, p$. Because Σ^{-1} is a positive definite matrix there exists A such that $A'A = \Sigma^{-1}$; therefore a solution for equations (3.36) and (3.37) must exist.

Hence if we transform the variables in this way, we have that $\mu_i = A^{-1}(\mu_{s_i} - a)$, and therefore

$$\beta = A'AA^{-1}(\mu_{s1} - \mu_{s2}) = A'(\mu_{s1} - \mu_{s2}) = A'\beta_s. \quad (3.38)$$

The covariance matrix of the asymptotic distribution of $\hat{\beta}$ is expressed in terms of the covariance matrix of the asymptotic distribution of $\hat{\beta}_s$ as

$$A' \text{cov}(\hat{\beta}_s) A. \quad (3.39)$$

For the standard case the covariance of the asymptotic distribution of α_s and β_s is given by the inverse of the information matrix based on the transform variables. The information matrix is given by

$$n \int_{-\infty}^{\infty} \frac{\frac{\pi_1}{\pi_0} e^{\Delta x_1}}{\left(1 + \frac{\pi_1}{\pi_0} e^{\Delta x_1}\right)^2} \begin{pmatrix} 1 \\ x \end{pmatrix} dF(x), \quad (3.40)$$

where $dF(x) = (\pi_1(2\pi)^{-p/2} \exp(-(x^2 - 2\Delta x + \Delta^2)/2) + \pi_0(2\pi)^{-p/2} \exp(-(x^2 + 2\Delta x + \Delta^2)/2)) dx$.

If we define

$$v_j = \int_{-\infty}^{\infty} \frac{\frac{\pi_1}{\pi_0} e^{\Delta x_1}}{\left(1 + \frac{\pi_1}{\pi_0} e^{\Delta x_1}\right)^2} x_1^{j-1} dF(x), \quad (3.41)$$

the covariance of the asymptotic distribution of $\hat{\beta}_s$ is given by the inverse $(p \times p)$ matrix of n times

$$\text{diag}[v_3, v_1, \dots, v_1] - (v_2, 0, \dots, 0)' v_1^{-1} (v_2, 0, \dots, 0), \quad (3.42)$$

which is equal to

$$\frac{1}{n} \text{diag} \left[\frac{v_1}{v_1 v_3 - v_2^2}, \frac{1}{v_1}, \dots, \frac{1}{v_1} \right]. \quad (3.43)$$

The element (i, j) of the covariance matrix of the asymptotic distribution of $\hat{\beta}$ is then

$$a_{1i} a_{1j} \frac{v_1}{v_1 v_3 - v_2^2} + \frac{1}{v_1} \sum_{k=2}^p a_{ki} a_{kj}. \quad (3.44)$$

We will use this result to compare the logistic regression estimators with the linear discriminant function estimators.

3.3 ASYMPTOTIC RELATIVE EFFICIENCY IN TERMS OF COEFFICIENT ESTIMATION

3.3.1 THE EFFECT OF A SINGLE VARIABLE

If we are interested in the effect on the response of a particular explanatory variable, say, without loss of generality, x_1 , in the presence of other explanatory variables, one way of comparing the efficiency of the above two approaches is by comparing the efficiency of the test of the hypothesis $H: \beta_1 = 0$, under the two procedures.

Because both estimators are asymptotically normally distributed in both cases to use as a test statistic the ratio of the difference of the estimator and the null value to the standard deviation of the estimator is appropriate, and both test statistics have asymptotically a standard normal distribution.

In testing a given hypothesis against a given alternative, the simplest way of comparing two tests is by comparing power functions. In other words, for the same statistical hypothesis, the relative efficiency of the second test with respect to the first is given by the ratio n_1/n_2 , where n_2 is the sample size of the second test required to achieve the same power for a given alternative as is achieved by the first test with respect to the same alternative when using a sample size n_1 .

In general the ratio n_1/n_2 will depend on the particular alternative chosen, that in some way is undesirable; however if we restrict attention to asymptotic results this fact can be avoided, and we can obtain a single measure of efficiency.

Both tests are α -consistent tests against the class of alternative hypotheses in the problem, i.e., for any fixed α , level of significance of the tests, the tests have power

approaching one, for any fixed alternative; thus the simple comparison of the ratio of the power is no longer useful.

We use the definition of efficiency, which seems to be due to Pitman (Noether,1955). The general idea is to define sequences of alternatives changing with n (approaching the value tested), in such way that as $n \rightarrow \infty$ the power of the corresponding sequences of tests converge to some number less than one, and define the asymptotic relative efficiency as the limit of the corresponding ratios n_1/n_2 .

Consider test of the null hypothesis $H: \theta = \theta_0$ against alternatives $H: \theta \neq \theta_0$, based on the statistic $T_n (= T(x_1, \dots, x_n))$, with $E(T_n) = \psi_n(\theta)$ and $\text{var}(T_n) = \sigma_n^2(\theta)$ (we write the argument to emphasize that either the mean or variance or both might depend on the value of the parameter), satisfying the following conditions:

$$1.- \quad \lim_{n \rightarrow \infty} n^{-m\rho} \frac{\psi_n^{(m)}(\theta_0)}{\sigma_n(\theta_0)} = c, \quad (3.45)$$

for some positive c and ρ ; where m indicates the first non zero derivative of ψ evaluated at θ_0 .

2.- For an alternative $H_1: \theta_n = \theta_0 + k/n^\rho$, where k is an arbitrary positive constant

$$\lim_{n \rightarrow \infty} \frac{\psi_n^{(m)}(\theta_n)}{\psi_n^{(m)}(\theta_0)} = 1, \quad \lim_{n \rightarrow \infty} \frac{\sigma_n(\theta_n)}{\sigma_n(\theta_0)} = 1. \quad (3.46)$$

3.- The original condition on Pitman's theorem is that as n tends to infinity the distribution of $(T_n - \psi_n(\theta))/\sigma_n(\theta)$ tends to the standard normal distribution both under the null and alternative hypothesis, but from the proof of the theorem it is possible to see that this condition can be relaxed to the distribution

of $(T_n - \psi_n(\theta))/\sigma_n(\theta)$ tends to a regularly behaved distribution function (continuous, strictly increasing function) both under the null and alternative hypothesis.

Then Pitman's theorem establishes that the asymptotic relative efficiency of two such tests, with the same value of m and ρ , is given by the limit as $n \rightarrow \infty$ of the ratio of the efficacy of the two tests, where the efficacy of a test is defined by

$$\left(\frac{\psi_{in}^{(m)}(\theta_0)}{\sigma_{in}(\theta_0)} \right)^{1/mp} . \quad (3.47)$$

A proof is given by Noether(1955). For a discussion of the consequences, in terms of the asymptotic relative efficiency, for tests satisfying (3.45) for different values of m and ρ see Kendall & Stuart(vol. 2 pp 288).

If, in addition,

$$\lim_{n \rightarrow \infty} \frac{\psi_{2n}^{(m)}(\theta_n)}{\psi_{1n}^{(m)}(\theta_0)} = 1 \quad (3.48)$$

is satisfied and $\psi_{1n}(\theta) = \psi_{2n}(\theta)$ then the asymptotic relative efficiency of the tests reduces to asymptotic relative efficiency of the two test statistics (the ratio of their variances). In particular if T_1 and T_2 are unbiased estimators of the parameter θ , the asymptotic relative efficiency of the two tests is the same as the asymptotic relative efficiency of T_2 and T_1 used as estimators of θ .

Blomqvist(1950), define the asymptotic relative efficiency of tests as the ratio of the derivatives of the power functions; this criterion is equivalent to Pitman's when the additional condition

$$\lim_{n \rightarrow \infty} \frac{\sigma_n^2(\theta_0)}{\psi_n'(\theta_0)} = 0 \quad (3.49)$$

is satisfied.

We will see that both estimators, the linear discriminant function and logistic regression, satisfy the conditions of Pitman's theorem and we will find the asymptotic relative efficiency of the tests.

First, in the case of linear discriminant analysis, we have that the expected value of the linear discriminant function, as suggested by Fisher, i.e., using the moment estimators of the parameters, as the number of observations tends to infinity is $c\beta$ and the variance of the estimator of β_1 is given by

$$\frac{c^2}{n} \left(\frac{1}{\pi_1 \pi_2} + \Delta^2 \right) v_{11} + \frac{c^2}{n} \beta_1' \beta_1, \quad (3.50)$$

where v_{11} represents the first element of Σ^{-1} .

Obviously the first derivative of the expected value of the linear discriminant function evaluated at the null value is different from zero, being equal to c for any value of the parameter, and the variance evaluated at the null value $\beta_1 = 0$ reduces to the first term in expression (3.50).

It is clear that the conditions one and two in Pitman's theorem are satisfied with $\rho = 1/2$ and $m = 1$. The third condition is obviously satisfied because the linear discriminant function is asymptotically normal. Therefore the efficacy of the test is given by

$$\frac{n \pi_1 \pi_0}{(1 + \pi_1 \pi_0 \Delta^2) v_{11}}, \quad (3.51)$$

for any positive value c . From here on when we refer to the linear discriminant function estimators we are considering c equal to one, without loss of generality. Because we are assuming that the observations are normally distributed the estimators are the maximum likelihood estimators.

In the case of the logistic regression estimators, we have seen that they are asymptotically unbiased and with asymptotic covariance matrix given by expression (3.44). Obviously the first derivative of the expected value evaluated at the null value is different from zero; in fact it is equal to one for any value of the parameter.

From expression (3.44) in section 3.3.2, we have that the variance of β_1 is given by

$$a_{11}^2 \frac{v_1}{v_1 v_3 - v_2^2} + \frac{1}{v_1} \sum_{k=2}^p a_{k1}^2 . \quad (3.52)$$

When β_1 is equal to zero, we have that, from the construction of the matrix A in section 3.3.2, $a_{11} = 0$ and $\sum a_{k1}^2 = v_{11}$ and therefore the variance of the conditional maximum likelihood estimator evaluated at the null value is given by v_{11}/v_1 , where v_1 corresponds to the integral

$$n \int_{-\infty}^{\infty} \frac{\frac{\pi_1}{\pi_0} e^{\Delta x}}{\left(1 + \frac{\pi_1}{\pi_0} e^{\Delta x}\right)^2} dF(x) , \quad (3.53)$$

with $dF(x) = \pi_1 (2\pi)^{-p/2} \exp(-(x^2 - 2\Delta x + \Delta^2)/2) + \pi_0 (2\pi)^{-p/2} \exp(-(x^2 + 2\Delta x + \Delta^2)/2) dx$; it is not difficult to see that (3.53) is equal to

$$\frac{n \pi_1 \pi_0 e^{-\Delta^2/8}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\pi_1 e^{\Delta x/2} + \pi_0 e^{-\Delta x/2}} dx . \quad (3.54)$$

It is clear that the two first conditions of Pitman's theorem are satisfied for $m=1$ and $\rho = 1/2$. The third one is satisfied from the fact that the asymptotic distribution of the estimator is normal. Therefore the asymptotic relative efficiency of the two tests corresponds to the asymptotic relative efficiency of the estimators of the parameter and is given by

$$\frac{1 + \pi_1 \pi_0 \Delta^2}{\sqrt{2\pi}} e^{-\Delta^2/8} \int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\pi_1 e^{\Delta x/2} + \pi_0 e^{-\Delta x/2}} dx, \quad (3.55)$$

which corresponds to the value obtain by Efron(1975), as the asymptotic relative efficiency in terms of misclassification errors.

The asymptotic relative efficiency depends on the Mahalanobis' distance as well as on the values of π_1 and π_0 ; for some combinations of different values of these quantities we have

Table 3.1.
Asymptotic relative efficiency of the test $\beta_1 = 0$ based on the logistic regression estimator to the test based on the linear discriminant function estimators.

$\pi_1 \setminus \Delta^2$	0	.5	1	1.5	2	2.5	3	3.5	4
.1	1.000	.998	.977	.911	.801	.660	.511	.370	.252
.2	1.000	.998	.980	.928	.836	.710	.566	.421	.293
.3	1.000	.999	.988	.949	.870	.752	.608	.458	.321
.4	1.000	.1.000	.993	.963	.892	.778	.633	.479	.338
.5	1.000	1.000	.999	.968	.899	.786	.641	.486	.343

As the Mahalanobis' distance increases the efficiency of the logistic regression estimators decreases; also when the difference of the probabilities of being in one or the other population increases the efficiency decreases, and this is independent of p , the dimension of the explanatory variables.

3.3.2 NO EFFECT FROM THE EXPLANATORY VARIABLES ($\Delta^2 = 0$)

We might be interested in another kind of hypothesis, for instance, in testing that the explanatory variables have no effect on the response or in terms of discrimination that the variables do not discriminate among the two populations, i.e., in testing $H: \beta' = 0$; because $\beta' = \Sigma^{-1}(\mu_1 - \mu_0)$, and Σ represents the covariance matrix, i.e., a positive definite matrix, the test is equivalent to test $H: \mu_1 - \mu_0 = 0$ (or $\mu_1 = \mu_0$).

Hotelling(1931), proposed the T^2 statistic for testing hypotheses about the mean vector of a multivariate normal distribution; later this test was extended to numerous applications. One of its uses is to test whether the mean vectors of two normal distributions with the same covariance matrix are equal against the alternative that they are different.

In this case the T^2 statistic is defined as

$$T^2 = \frac{n_1 n_0}{n} (\bar{x}_1 - \bar{x}_0)' S^{-1} (\bar{x}_1 - \bar{x}_0) . \quad (3.56)$$

It is possible to show that (see, for example, Muirhead, 1982, pp98)

$$\frac{T^2(n-p-1)}{(n-2)p} \sim F_{p, n-p-1}(\delta_1) , \quad (3.57)$$

where $F_{p, n-p-1}(\delta_1)$ represents the noncentral F distribution with p and $n-p-1$

degrees of freedom and noncentrality parameter δ_1 given by

$$\frac{n_1 n_0}{n} (\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0) . \quad (3.58)$$

When H is true the noncentrality parameter is equal to zero. Therefore testing $H: \beta_1 = 0$ is equivalent to testing $H: \delta_1 = 0$, or equivalently we can write $\delta_1 = n_1 n_0 \delta / n$, with $\delta = (\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0)$, and then the test is equivalent to testing $H: \delta = 0$.

This test is the uniformly most powerful invariant test and is also equivalent to the likelihood ratio test (see for example Muirhead (1982, pp211)). Kshirsagar (1972, pp124) gives the distribution of T^2 , both under the null and alternative hypotheses.

In the case of the logistic regression estimators, the test we use is the conditional maximum likelihood ratio statistic, defined as

$$e^{(1/2)W} = \frac{l(\hat{\beta}; y|x)}{l(\beta_0; y|x)} , \quad (3.59)$$

where β_0 represents the value under the null hypothesis and in this case is equal to the vector whose components are all zero, $l(\beta_0; y|x)$ represents the loglikelihood evaluated at β_0 and $l(\hat{\beta}; y|x)$ represents the loglikelihood evaluated at the conditional maximum likelihood estimator of β , $\hat{\beta}$. Strictly we should write $l(\hat{\alpha}, \beta_0; y|x)$ instead of $l(\beta_0; y|x)$ and $l(\hat{\alpha}, \hat{\beta}; y|x)$ instead of $l(\hat{\beta}; y|x)$ but because the value of α is not of interest, and for simplicity of notation we omit it.

Under $H: \beta = \beta_0$ ($\beta_0 = 0$), W has a central chi- square distribution with degrees of freedom equal to the dimension of β (Cox & Hinkley (1974, pp34)).

This test is asymptotically equivalent to

$$W_e = (\hat{\beta} - \beta_0)' i.(\beta_0) (\hat{\beta} - \beta_0), \quad (3.60)$$

and

$$W_u = U'(\beta_0) i^{-1}(\beta_0) U(\beta_0), \quad (3.61)$$

where $i(\beta_0)$ and $U(\beta_0)$ represent respectively the expected value of the second and the first derivative of the log likelihood evaluated at the null value. Again we are writing $i(\beta_0)$ instead of $i(\beta_0; \hat{\alpha})$ but the value is exactly the same as can be seen from expression (3.42) and (3.43).

When H is not true the distribution of W , W_e and W_u is a noncentral chi square with noncentrality parameter δ_2 equal to $(\beta - \beta_0)' i(\beta_0) (\beta - \beta_0)$. Therefore the test is equivalent to $H: \delta_2 = 0$.

Because of the definition of the covariance of the conditional maximum likelihood estimators in section 3.1, we have that when β' is equal to zero, the elements of the covariance matrix are of the form

$$\frac{1}{v_1} \sum_{k=2}^p a_{ki} a_{kj}, \quad (3.62)$$

where v_1 is defined as in section 3.3.1. Then

$$\text{cov}(\beta_0) = \frac{1}{v_1} \Sigma, \quad (3.63)$$

and therefore the noncentrality parameter, δ_2 , is

$$v_1 (\beta - \beta_0)' \Sigma (\beta - \beta_0), \quad (3.64)$$

or equivalently, when $\beta_0 = 0$,

$$v_1 (\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0); \quad (3.65)$$

thus the non centrality parameter is equal to $v_1 \delta$, and the test is equivalent to $H: \delta = 0$

We are interested on comparing the performance of both tests, which again we do by means of the power.

Firstly, it is a well known fact that the limit as n tends to infinity, keeping d constant, of a noncentral F distribution, $F_{d,n}(\delta)$, with d and n degrees of freedom and noncentrality parameter δ is d times a noncentral F distribution with d and infinity degrees of freedom and the same noncentrality parameter, $dF_{d,\infty}(\delta)$ or equivalently a noncentral chi-square distribution with non centrality parameter δ , $X_d^2(\delta)$.

Therefore asymptotically both tests are X_p^2 , the only difference being in the noncentrality parameters. To compare them we might use the power function, which is a function of the noncentrality parameter, given by $P_\delta(X_p^2(\delta) \geq X_p^2(\alpha))$, where α represents the significance level; but because both tests are equivalent to testing the hypothesis that the non centrality parameter is zero and both tests are α -consistent we can apply Pitman's theorem to the tests $H: \delta = 0$.

The expected value in terms of the noncentrality parameter in the case of the discriminant function is $p + \delta n_1 n_0 / n$ compared to $p + v_1 \delta$ in the case of the logistic regression estimators. The derivatives of the expected value evaluated at the null value (in fact evaluated at any value) are $n_1 n_0 / n$ and v_1 respectively.

The variance, evaluated at the null value is equal to $2p$ for both tests. Therefore conditions one and two of Pitman's theorem are satisfied. The third one is satisfied because the distributions of the tests are $X_d^2(\delta)$ and as the alternative value tends to the null value the distributions of the tests tends to a central chi square, which is a regularly behaved distribution. Hence the asymptotic efficiency of the likelihood ratio test based on the logistic regression estimators relative to Hotelling's test is $nv_1/n_1 n_0$, where v_1 , when β' is equal zero is equal to

$$\frac{n\pi_1\pi_0}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx = n\pi_1\pi_0, \quad (3.66)$$

and therefore the asymptotic relative efficiency of the tests is one. Obviously the efficiency corresponds to the efficiency of testing that the effect of a particular variable is zero when the Mahalanobis' distance is zero, independent of the dimension of the vector of explanatory variables x .

3.3.3 THE EFFECT OF A SUBSET OF THE PARAMETERS

Perhaps the most realistic situation is to consider hypotheses of the kind $H: \beta_{(2)} = 0$, where $\beta_{(2)}$ is a $q \times 1$ vector representing a subset of the vector of parameters, with $1 \leq q \leq p$.

Obviously we want to use tests that in the case of $q = 1$ reduce to the normal tests and in the case of $q = p$ to the Hotelling's or likelihood ratio test.

Without loss of generality we assume that $\beta_{(2)}$ is the vector whose components are the last q components of β' . To test that $\beta_{(2)}$ is equal to zero is equivalent to testing that Δ_p^2 , the Mahalanobis distance when all the p variables are used, is equal to Δ_k^2 , the Mahalanobis distance when the first k ($= p - q$) variables are used. To see this equivalence we partition the elements of $\mu (= \mu_1 - \mu_0)$ and $d (= \bar{x}_1 - \bar{x}_0)$, into two subsets $(\mu = (\mu_{(1)}, \mu_{(2)})'$ and $d = (d_{(1)}, d_{(2)})'$) the first one of dimension k and the second one of dimension q , and we partition S and Σ , as follows:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}, \quad (3.67)$$

where Σ_{11} and Σ_{22} are square matrices of dimension $k \times k$ and $q \times q$ respectively,

$\Sigma_{12}=(\Sigma_{21})'$, is a $k \times q$ matrix and equivalently for the elements of the partition of S .

Now if we define

$$\gamma = \Sigma_{21}(\Sigma_{11})^{-1}, \quad \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}, \quad (3.68)$$

and similarly for the sample covariance matrix

$$g = S_{21}(S_{11})^{-1}, \quad S_{22.1} = S_{22} - S_{21}(S_{11})^{-1}S_{12}, \quad (3.69)$$

we find that

$$\beta_{(2)} = (\Sigma_{22.1})^{-1}(\mu_{(2)} - \gamma\mu_{(1)}), \quad (3.70)$$

and

$$\Delta_p^2 = \Delta_k^2 + (\mu_{(2)} - \gamma\mu_{(1)})' (\Sigma_{22.1})^{-1}(\mu_{(2)} - \gamma\mu_{(1)}), \quad (3.71)$$

with the appropriate equivalent definitions for the sample data. Hence $H: \beta_{(2)} = 0$ is equivalent to $H': \mu_{(2)} = \gamma\mu_{(1)}$ and to $H'': \Delta_p^2 = \Delta_k^2$.

A test for H'' , that from here on we call H , was proposed by Rao(1946,1948), and consists of testing that the difference of the Mahalanobis' distance calculated from the p explanatory variables to the one calculated with the first k explanatory variables is zero when the Mahalanobis' distance for the first k explanatory variables is fixed.

This test is given, with D_j^2 representing the sample Mahalanobis' distance, by

$$F = \frac{n-p-1}{q} \cdot \frac{\frac{n_1 n_0}{n(n-2)}(D_p^2 - D_k^2)}{(1 + \frac{n_1 n_0}{n(n-2)}D_k^2)}, \quad (3.72)$$

which has a noncentral F distribution with q and $n - p - 1$ degrees of freedom and

noncentrality parameter equal to δ/ν , where

$$\delta = \mu_{(2)}' (\Sigma_{22.1})^{-1} \mu_{(2)} \quad (3.73)$$

and

$$\nu = 1 + \frac{n_1 n_0}{n(n-2)} \Delta_k^2 \quad (3.74)$$

When H is true the noncentrality parameter is zero and therefore the hypothesis is equivalent to the hypothesis that the noncentrality parameter is zero (as in the case of Hotelling's test statistic), or equivalently to the hypothesis that δ is equal to zero. As n tends to infinity the distribution of F tends to a non central chi square $X_q^2(\delta/\nu)$.

It is clear that when $q = p$, the expression (3.72) reduces to Hotelling's statistic given by (3.57), and when $q = 1$, it reduces to the normal test.

In the case of logistic regression we use the likelihood ratio test W defined as

$$e^{W/2} = \frac{l(\hat{\beta}; y|x)}{l(\hat{\beta}_{(1)}, \beta_{(2)0}; y|x)} \quad (3.75)$$

where $l(\hat{\beta}_{(1)}, \beta_{(2)0}; y|x)$ represents the conditional likelihood evaluated at the conditional maximum likelihood estimator of $\beta_{(1)}$, $\hat{\beta}_{(1)}$, having $\beta_{(2)} = \beta_{(2)0}$ fixed. Note that α is included in the vector $\beta_{(1)}$.

This test is asymptotically equivalent to

$$W_e = (\hat{\beta}_{(2)} - \beta_{(2)0})' i. (\hat{\beta}_{(2)}; \hat{\beta}_{(1)}) (\hat{\beta}_{(2)} - \beta_{(2)0}), \quad (3.76)$$

and to

$$W_u = U' (\beta_{(2)0}, \hat{\beta}_{(1)}) i. \beta_{(2)} \beta_{(2)} (\beta_{(2)0}, \hat{\beta}_{(1)}) U (\beta_{(2)0}, \hat{\beta}_{(1)}), \quad (3.77)$$

where $i.(\beta_{(2)}; \beta_{(1)})$ represents the inverse of the covariance matrix $i. \beta_{(2)} \beta_{(2)} (\beta_{(1)}, \beta_{(2)})$ of the limiting normal distribution of $\beta_{(2)}$, given by

$$i.(\beta_{(2)}; \beta_{(1)}) = i. \beta_{(2)} \beta_{(2)} (\beta_{(2)}, \beta_{(1)}) - i. \beta_{(2)} \beta_{(1)} (\beta_{(2)}, \beta_{(1)}) i. \beta_{(1)} \beta_{(1)}^{-1} (\beta_{(2)}, \beta_{(1)}) i. \beta_{(1)} \beta_{(2)} (\beta_{(2)}, \beta_{(1)}), \quad (3.78)$$

where $i. \beta_{(i)} \beta_{(j)} (\beta_{(2)}, \beta_{(1)})$ represents the element (i, j) $i, j = 1, 2$ of the partition of the matrix $i. \beta \beta (\beta)$.

All the tests are asymptotically equivalent and their asymptotic distribution is $X_q^2(\delta_2)$, where δ_2 , the non centrality parameter, is equal to

$$(\beta_{(2)} - \beta_{(2)0})' i. \beta_{(2)} \beta_{(1)} (\beta_{(2)0}; \beta_{(1)}) (\beta_{(2)} - \beta_{(2)0}), \quad (3.79)$$

which under the null hypothesis is equal to zero; therefore the test is equivalent to testing that the noncentrality parameter is equal to zero.

When $\beta_{(2)}$ is equal to zero, we can see from the definition of the covariance of β in section 3.3.2 that if we partition the covariance matrix as follows

$$\frac{1}{v_1} \begin{bmatrix} Q & \Sigma_{12}^{-1} \\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1} \end{bmatrix}, \quad (3.80)$$

where the elements of Q are given by expression (3.44), then

$$i. \beta_{(2)} \beta_{(1)} (\beta_{(2)0}; \beta_{(1)}) = \frac{1}{v_1} \Sigma_{22.1} = \frac{1}{v_1} (\Sigma_{22} - \Sigma_{21} (\Sigma_{11})^{-1} \Sigma_{12}), \quad (3.81)$$

the noncentrality parameter is equal to $\delta/\sqrt{v_1}$ and the test is equivalent to testing δ equal to zero. Again we have two different tests for testing the same hypothesis, $\delta = 0$, both of them satisfying the conditions of Pitman's theorem and from (3.73), (3.74) and (3.81) we have that the asymptotic relative efficiency of the likelihood ratio test to the F test (given by (3.72)) for testing that a subset of the explanatory variables does not have effect on the response is given by expression (3.55) independent of the dimensions of β and $\beta_{(2)}$.

3.3.4 THE RATIO OF LINEAR COMBINATIONS OF THE PARAMETERS

Since only ratios of discriminant functions are unique a hypothesis of the type

$$H: \frac{\beta_i}{\beta_j} = \rho , \quad (3.82)$$

which specifies that the ratio of two coefficients of the discriminant function is equal to a constant, is of interest.

More generally a hypothesis of the form

$$H: \frac{\sum_{j=1}^p m_j \beta_j}{\sum_{k=1}^p l_k \beta_k} = \rho , \quad (3.83)$$

with $\sum_j m_j = \sum_j l_j = 1$; which reproduces (3.82) when $m_i = 1$, $l_j = 1$, $m_k = 0$, for $k \neq i$, and $l_k = 0$ for $k \neq j$, might be of interest.

To test (3.83) is equivalent to testing

$$H: \sum_{j=1}^p (m_j - \rho l_j) \beta_j = 0 . \quad (3.84)$$

Using the fact that asymptotically the estimators of β under both methods are normal, if we call f the vector whose components are $(m_j - \rho l_j)$, we have that, asymptotically, $f'\hat{\beta}$ and $f'b$ are asymptotically normally distributed with mean $f'\beta$ and variances $f'\text{var}(\hat{\beta})f$ and $f'\text{var}(b)f$ respectively. Therefore to test (3.84) we use a normal test, and the asymptotic relative efficiency of the two tests is defined according to Pitman's efficiency as the ratio of the variances evaluated at the null value.

In the case of the linear discriminant function we have that the variance of $f'b$ is

$$f' \left(\frac{1}{\pi_1 \pi_0} + \Delta^2 \right) \Sigma^{-1} f + (f' \Sigma^{-1} \mu)(f' \Sigma^{-1} \mu)' , \quad (3.85)$$

and under the null hypothesis the second term is equal to zero.

In the logistic regression case, we have that $f'\text{var}(\hat{\beta})f$ is equal to $f'A'\text{var}(\hat{\beta}_s)Af$, where A is defined as in section 3.2.3. From the definition of A , we have that under the null value $f'A'\text{var}(\hat{\beta}_s)Af$ is equal to

$$\frac{1}{v_1} f' \Sigma^{-1} f, \quad (3.86)$$

and the relative efficiency of both tests is again given by expression (3.55), independently of the values of ρ and the vectors m and l .

3.4 RELATION OF LINEAR DISCRIMINANT ANALYSIS AND LINEAR REGRESSION

It is well known (Fisher, 1936) that a multiple of Fisher's linear discriminant function coefficient vector can be obtained by fitting a linear equation by least squares using the explanatory variables as independent variables and a binary dependent variable, y , to separate the individuals into two samples.

Different choices of the values of the binary variable give different values of the estimated linear discriminant function, but all of them are multiples of $S^{-1}(\bar{x}_1 - \bar{x}_0)$.

For simplicity we choose the values of the independent binary variable equal to one if the individual is coming from population one and zero otherwise.

If we call Z the $n \times p$ matrix which columns represents the centred values of the explanatory variables and whose rows represent individuals, we have that b_{ls} is the solution of

$$Z'Zb_{ls} = Z'u \quad (3.87)$$

where u represents the centred values of y (which corresponds to the values originally proposed by Fisher, n_0/n and $-n_1/n$).

We can rewrite (3.87) as

$$(S + cdd')b_{ls} = cd, \quad (3.88)$$

where S is the pooled sum of squares defined as $(n - 2)$ times expression (3.5), d is defined as in section 3.2.1 and $c = n_1n_0/n$.

Now from linear regression theory we have that the sum of squares due to regression is $b'_{ls}Z'u = cb'_{ls}d$ and the total sum of squares is equal to $c(n_1(n_0/n)^2 + n_0(n_1/n)^2)$, leaving the residual sum of squares, SS_e equal to

$$SS_e = c(1 - b'_{ls}d). \quad (3.89)$$

Combining (3.88) and (3.89) we have that

$$b_{ls} = c(1 - b'_{ls}d)S^{-1}d = (n - 2)S^{-1}d(SS_e), \quad (3.90)$$

and hence the least squares estimators are equal to $1/SS_e$ times the estimators of the linear discriminant function, when the values of y are chosen to be 1 and 0.

The assumptions in standard regression analysis to test hypotheses are that the independent or explanatory variables are fixed, and the dependent variable has a normal distribution with common variance σ^2 . It is obvious that none of them is satisfied, but nevertheless, the tests used in linear regression are equivalent to the tests we have describe in sections 3.3.1, 3.3.2 and 3.3.3, when the explanatory variables have a normal distribution.

The usual test in regression analysis, for testing that the vector of coefficients is zero, is based on the ratio of the mean square due to the regression and the mean residual sum of squares; in other words we use the statistic

$$\frac{n - p - 1}{p} \cdot \frac{c(b'_{ls}d)}{c(1 - b'_{ls}d)}, \quad (3.91)$$

which, from (3.90), corresponds to

$$\frac{n - p - 1}{p(n - 2)}cd'S^{-1}d, \quad (3.92)$$

which is the F test based in Hotelling's statistic, discussed in section 3.3.2.

To test that a subset of the vector of parameters of dimension q is equal to zero the usual test in regression analysis is the ratio of the difference of the residual sum of squares calculated after omitting the last q variables and the residual sum of squares obtained from the full model to the residual sum of squares obtained from the full model divided by their degrees of freedom; when the independent variables are normally distributed this statistic has an F distribution with q and $n - p - 1$ degrees of freedom. Noting that

$$SS_{\text{Reg}} = cd'S^{-1}d(SS_e) = c - SS_e, \quad (3.93)$$

we have that

$$SS_{e_p} = \frac{c}{(n - 2) + cD_p^2}, \quad (3.94)$$

where the subscript on D and SS_e emphasizes that the Mahalanobis' distance is calculated from the p explanatory variables. Similarly we obtain the value of SS_{e_k} , and therefore the standard test in regression analysis can be written as expression (3.72), which is the test suggested in section 3.3.3.

3.5 SIMULATION STUDIES

In order to examine the behaviour of the tests when the sample sizes are small we simulate data and apply the tests to them and calculate the power of the tests in different situations. Obviously the power of the tests depends on the sample size and the distance from the null value to the actual value of the parameters; it is obtained as

one minus the proportion of times that the test is accepted when the null hypothesis is false.

We generate data using the NAG subroutine for generating normal data; in all the cases we generate 1000 samples, and in all the cases we take, without loss of generality, the covariance matrix to be equal to the identity, so that $\Delta^2 = \beta'\beta$.

In the case of the hypothesis $H: \beta = 0$ we generate data from bivariate and trivariate normal distributions, for a given Δ we obtain the values of β_i^2 as Δ^2 divided by two or three, and from there for a fixed value of the mean vector in one population (usually zero, but any other choice would not change the results) we obtain the value of the mean vector in the other population.

We generate data for values of Δ from .5 to 3.5, but in all the cases the power of Hotelling's test statistic and the likelihood ratio test when Δ is equal to 3.5 is one; when Δ is equal to 3 in most of the cases the estimated power of both tests is one and in the few it is not equal to one it is equal in both tests, so that the estimated relative efficiency is one.

We generate data for two different values of the total sample size and for each one of these for three different values of the sample sizes in each group. The values for the total sample sizes are 20, 40 and for each one of these the values of the sample sizes in the groups are $(n/2, n/2)$, $(n/4, 3n/4)$, $(n/10, 9n/10)$. Note that if we generate data for a bigger sample size, because both tests are α consistent, the powers approach one and therefore the relative efficiency is one.

The results in terms of Δ , π_1 and the different sample sizes are shown in tables 3.2.1 and 3.2.2. When the Mahalanobis distance is equal to .5 and n is equal to 20 the powers of both tests are less than .26 for all values of π_1 considered in the study; this is also true for $\Delta^2 = 1$ and π_1 equal to .1 when $n = 20$.

Table 3.2.1.

Ratio of the powers of the likelihood ratio test based on the logistic regression estimators and the Hotelling's test statistic, for testing the hypothesis $H: \beta = 0$, when the explanatory variables are bivariate normal, based on 1000 simulated samples.

Δ^2	.5	1	1.5	2	2.5
n = 20					
$\pi_1 = 0.5$	1.000	1.529	1.625	1.167	1.034
$\pi_1 = 0.25$	1.181	1.700	1.500	1.208	1.000
$\pi_1 = 0.1$	1.571	3.000	2.143	2.200	1.208
n = 40					
$\pi_1 = 0.5$	1.037	1.000	1.035	1.034	1.000
$\pi_1 = 0.25$	1.150	1.028	1.333	1.074	1.000
$\pi_1 = 0.1$	1.083	1.111	2.125	1.412	1.000

β is the vector whose components are all equal to the square root of $\Delta^2/2$.

The standard error of the estimated relative efficiency is less than 1×10^{-4} .

For the trivariate normal distribution the powers of both tests are bigger than the powers for the bivariate normal; particularly, when Δ is equal to or less than 1.

In general, for both cases, the estimated relative efficiency for small sample sizes is very close to one for any fixed alternative investigated. However, for n equal to 20 the powers are not very big, especially when the alternative is close to the null value.

Table 3.2.2.
Ratio of the powers of the likelihood ratio test based on the logistic regression estimators and the Hotelling's test statistic, when the explanatory variables are coming from a trivariate normal, for testing the hypothesis $H: \beta = 0$, based on 1000 simulated samples.

Δ^2	.5	1	1.5	2	2.5
$n = 20$					
$\pi_1 = 0.5$	1.500	1.133	1.117	1.000	1.000
$\pi_1 = 0.25$	1.500	2.167	2.143	1.000	1.000
$\pi_1 = 0.1$	1.750	1.800	2.285	1.333	1.136
$n = 40$					
$\pi_1 = 0.5$	1.500	1.000	1.000	1.000	1.000
$\pi_1 = 0.25$	1.000	1.500	1.071	1.105	1.036
$\pi_1 = 0.1$	3.136	1.764	1.041	1.000	1.000

β is the vector whose components are all equal to the square root of $\Delta^2/3$.

The standard error of the estimated relative efficiency is less than 1×10^{-1} .

In the case of testing that a subset of the parameters, of dimension q , is equal to zero, to compare the power of the conditional F test defined in section 3.3.2 with the likelihood ratio test based on the logistic regression estimators, we simulate samples from a trivariate normal distribution. First we test the hypothesis $H: \beta_1 = 0$ (note that these tests in this particular case are equivalent to the normal test) where the parameters are calculated in the same way as described for Hotelling's test (i.e., the values of the β are obtained as the square root of Δ^2 divided by three, and from there we obtain the values of μ_1 having fixed μ_0); the results are shown in table 3.3.1.

Table 3.3.1.

Ratio of the powers of the likelihood ratio test and the conditional F test for testing that one of the parameters is zero when the observations are coming from a trivariate normal and the parameters are equal.

Δ^2	.5	1	1.5	2	2.5	3	3.5	4
n = 20								
$\pi_1 = 0.5$	1.577	1.286	1.317	1.358	1.275	1.247	1.333	1.288
$\pi_1 = 0.25$	1.548	1.727	1.385	1.185	1.520	1.480	1.273	1.279
$\pi_1 = 0.1$	2.333	2.600	2.057	2.000	1.808	1.315	1.193	1.250
n = 40								
$\pi_1 = 0.5$	1.175	1.195	1.193	1.064	1.050	1.023	1.051	0.973
$\pi_1 = 0.25$	1.162	1.147	1.142	1.202	.985	1.096	.988	.977
$\pi_1 = 0.1$	1.321	1.657	1.339	1.197	1.213	1.151	1.079	.982

β is the vector whose components are all equal to the square root of $\Delta^2/3$.

The standard error of the estimated relative efficiency is in between .17 and .21 for any value of π_1 and Δ^2 .

In a second simulation, the parameter whose null value is zero is calculated as the square root of Δ^2 divided by four and the other two are each the square root of half of three quarters of Δ^2 , and from there we obtain the values of the mean of the distribution. The results are shown in table 3.3.2.

For a third simulation, the parameter whose null value is zero is calculated as the square root of Δ^2 divided by ten and the other two are each the square root of half of nine tenth of Δ^2 . The results are shown in table 3.3.3.

When n is equal to 20 the logistic regression estimators are more efficient than the linear discriminant estimators, for all the alternative values we use, but the power of both tests is very small (the maximum power in the case Δ^2 equal to 4 and π_1 equal to .5 is .44 for the logistic regression estimators and .34 for the linear discriminant estimators) and so the relative efficiency is not a good indicator of the performance of the test. This is also true for Δ^2 equal to or less than 2.5 in the case of n = 40.

Table 3.3.2.

Ratio of the powers of the likelihood ratio test and the conditional F test for testing that one of the parameters is zero when the observations are coming from a trivariate normal.

Δ^2	.5	1	1.5	2	2.5	3	3.5	4
n = 20								
$\pi_1 = 0.5$	1.629	1.245	1.257	1.256	1.219	1.112	1.204	1.270
$\pi_1 = 0.25$	1.709	1.552	1.333	1.357	1.200	1.395	1.230	1.195
$\pi_1 = 0.1$	2.157	2.344	1.750	2.272	1.693	1.311	1.222	1.200
n = 40								
$\pi_1 = 0.5$	1.190	1.135	1.059	1.050	1.032	1.026	1.005	0.995
$\pi_1 = 0.25$	1.256	1.058	1.064	1.093	1.013	1.059	1.062	0.952
$\pi_1 = 0.1$	1.267	1.631	1.410	1.137	1.175	1.194	1.081	0.952

The parameter we are testing is the smallest of the three, the square root of $\Delta^2/4$.

The standard error of the estimated relative efficiency is in between .17 and .21 for any value of π_1 and Δ^2 .

Table 3.3.3.

Ratio of the powers of the likelihood ratio test and the conditional F test for testing that one of the parameters is zero when the observations are coming from a trivariate normal.

Δ^2	.5	1	1.5	2	2.5	3	3.5	4
n = 20								
$\pi_1 = 0.5$	1.363	1.318	1.538	1.375	1.340	1.063	1.142	1.196
$\pi_1 = 0.25$	1.375	1.470	1.535	1.333	1.022	1.406	1.125	1.191
$\pi_1 = 0.1$	2.214	1.842	1.555	1.454	1.500	1.392	1.354	1.111
n = 40								
$\pi_1 = 0.5$	1.214	1.363	1.032	1.057	0.951	1.078	0.977	0.976
$\pi_1 = 0.25$	1.173	1.113	0.964	1.014	0.970	1.027	1.000	1.027
$\pi_1 = 0.1$	0.818	1.500	1.461	1.800	1.100	0.913	0.923	0.963

The parameter we are testing is the smallest of the three, the square root of $\Delta^2/10$.

The standard error of the estimated relative efficiency is in between .17 and .21 for any value of π_1 and Δ^2 .

3.6 ROBUSTNESS

3.6.1 INTRODUCTION

Because of the relative abundance of theoretical results in the normal case, there is a tendency to regard distributions as normal unless otherwise proven or the evidence for not doing so is very clear, or to find a transformation to reach normality, and to use the standard normal theory whenever possible. The question is whether such assumptions of normality are likely to be seriously misleading.

In this context there are two departures from the assumptions to consider; first, that the observations come from multivariate normal distributions with different means and unequal covariance matrices and secondly that the distribution is not normal, but that the observations come from a continuous multivariate distribution with longer or shorter tails than the normal with the same covariance matrix and different means.

Obviously another case would be a mixture of the two cases mentioned, i.e., different covariance matrices in the populations and the distribution not being a multivariate normal but this case is not considered in the present work.

We have to consider the effect of the departures from the assumptions, first on the estimation of the parameters and therefore on the estimation of the risk of being in one or other of the two groups, and then if this effect is not significant the effect of these departures on the tests of hypotheses related to the parameters.

In the particular case of hypothesis testing problems we call a test robust if it is α insensitive to changes in the distributional form, with α the level of significance, i.e., insensitive to departures from the assumptions which underlie it. Additional changes in

the distributional form that do not have an effect on the significance level can have an effect on the power of the test.

3.6.2 DEPARTURES FROM THE ASSUMPTION OF EQUAL COVARIANCE MATRIX

In the case of unequal covariance matrices the linear logistic model no longer hold. If we apply Bayes' theorem to the posterior probability the logarithm of the ratio of the conditional probability of being in group one to the conditional probability of being in group two is not a linear function of the observations but a quadratic one, given by

$$\begin{aligned} \log\left(\frac{P(y = 1 | x)}{P(y = 0 | x)}\right) &= \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{\mu_1 \Sigma_1^{-1} \mu_1}{2} + \frac{\mu_0 \Sigma_0^{-1} \mu_0}{2} + \frac{1}{2} \log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) \\ &+ (\Sigma_1^{-1} \mu_1 - \Sigma_0^{-1} \mu_0)' x + \frac{1}{2} x' (\Sigma_0^{-1} - \Sigma_1^{-1}) x ; \end{aligned} \quad (3.95)$$

obviously the closer Σ_1 and Σ_0 are the less effect the departure from the assumption of equal covariance matrix has in the estimation of the parameters, but when this departure is significant, we have that any test used to test that a subset of the parameters is equal to zero does not measure the effect of the variables associated with that particular set of parameters. In other words, if we want to see the effect of x_i on the response in the presence of the other variables, to test that β_i is equal to zero does not mean that x_i has no effect on the response, because the logarithm of the ratio of the conditional

probabilities of being in the different groups is not linear in x , and the effect of the variables is not in one to one association with the parameters. Nevertheless the conditional likelihood ratio test is still valid for testing that the parameters of the logistic model are equal to a certain value, because it is based on the conditional distribution and asymptotically the test has a chi square distribution, central under the hypothesis β (or a subset of it) equal to zero.

We are going to consider the situation where the observations are normally distributed with mean $\mu_i = ((-1)^{i-1} \Delta/2, 0, \dots, 0)$ and variance $I(1 + (-1)^{i-1} \delta)$, for small δ , with $i = 1, 0$ depending on the population from which they are coming. Note that this is equivalent to having a normal distribution with mean μ_i and variance $\Sigma(1 + (-1)^{i-1} \delta)$ and applying the transformation discussed in section 3.2.2.

In this particular case expression (3.95) reduces to

$$\log\left(\frac{\pi_1}{\pi_0}\right) + \Delta x + (p/2) \log(1 - 2\delta) + \delta x^2 + \delta \frac{\Delta^2}{4}, \quad (3.96)$$

and as δ tends to zero the model reduces to the standard case and, because the expected value of x^2 is finite, as δ tends to zero the expected value of (3.96) tends to the expected value of the logarithm of the conditional probabilities in the equal covariance matrices case.

The power of the likelihood ratio test that the last k parameters are equal to zero against local alternatives ($\beta_{(2)a} = \beta_{(2)0} + \psi/\sqrt{n}$, where n is the sample size and ψ small) is given by the non centrality parameter

$$(\beta_{(2)a} - \beta_{(2)0})' i_{\beta_{(2)}\beta_{(1)}}(\beta_{(2)0}; \beta_{(1)})(\beta_{(2)a} - \beta_{(2)0}), \quad (3.97)$$

where $i_{\beta_{(2)}\beta_{(1)}}(\beta_{(2)0}; \beta_{(1)})$ represents the covariance matrix of the limiting distribution of

β , given by n^{-1} times

$$i. \beta_{(2)\beta_{(1)}}(\beta_{(2)0}; \beta_{(1)}) = \frac{1}{v_1'} \Sigma_{22.1} = \frac{1}{v_1'} (\Sigma_{22} - \Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}) , \quad (3.98)$$

and v_1' is approximately (ignoring terms of order δ^2) given by

$$\begin{aligned} & \frac{\pi_1 \pi_0 e^{-\Delta^2/8}}{\sqrt{2\pi}} \left\{ \int_{-\infty}^{\infty} \frac{e^{-(x^2 + \delta\Delta x)/2} (\pi_1 e^{(\Delta x + \delta x^2 + \Delta^2/4)/2} + \pi_0 e^{-(\Delta x + \delta x^2 + \Delta^2/4)/2})}{(\pi_1 e^{(\Delta x + \delta x^2 + \Delta^2/4 + p\delta)/2} + \pi_0 e^{-(\Delta x + \delta x^2 + \Delta^2/4 + p\delta)/2})^2} dx \right. \\ & \left. + \delta \int_{-\infty}^{\infty} \frac{e^{-(x^2 + \delta\Delta x)/2} (\pi_0 e^{-(\Delta x + \delta x^2 + \Delta^2/4)/2} - \pi_1 e^{(\Delta x + \delta x^2 + \Delta^2/4)/2})}{(\pi_1 e^{(\Delta x + \delta x^2 + \Delta^2/4 + p\delta)/2} + \pi_0 e^{-(\Delta x + \delta x^2 + \Delta^2/4 + p\delta)/2})^2} dx \right\} . \quad (3.99) \end{aligned}$$

In the case of the linear discriminant function, it was constructed under the assumption of equal covariance matrix and this matrix is estimated with the pooled estimator S . This estimator is used as well in the test statistics for testing that the parameters are equal to zero. The distribution of $(n - 2)$ times the estimator is no longer Wishart; we have that

$$(n - 2)S = (n_1 - 1)S_1 + (n_0 - 1)S_0, \quad (3.100)$$

where S_i is Wishart distributed $W(\Sigma_i; n_i - 1, p)$. Therefore we have that

$$E((n - 2)S) = (n_1 - 1)\Sigma_1 + (n_0 - 1)\Sigma_0, \quad (3.101)$$

and assuming that Σ_i is equal to $\Sigma(1 + (-1)^{i-1}\delta)$, we have that

$$E((n - 2)S) \approx (n - 2) \Sigma \{ 1 + \delta(\pi_1 - \pi_0) \}. \quad (3.102)$$

On the other hand we have that the covariance matrix of $\bar{x}_1 - \bar{x}_0$ is

$$\left(\frac{1}{n_1} + \frac{1}{n_0}\right) \Sigma\{1 + \delta(\pi_0 - \pi_1)\} . \quad (3.103)$$

Therefore in the case of the equal sample sizes ($n_1 = n_0$) we could assume that $(n - 2)S$ is Wishart distributed $W(\Sigma\{1 + \delta(\pi_1 - \pi_0)\}; (n - 2), p)$, and the standard theory for testing that the discriminant coefficients are equal zero can be applied. For instance Hotelling's test statistic is a function of $(\bar{x}_1 - \bar{x}_0)'(\Sigma_1/n_0 + \Sigma_0/n_1)^{-1}(\bar{x}_1 - \bar{x}_0)$ and the statistic $(\bar{x}_1 - \bar{x}_0)'(\Sigma_1/n_1 + \Sigma_0/n_0)^{-1}(\bar{x}_1 - \bar{x}_0)$ has a chi square distribution with noncentrality parameter given by $(\mu_1 - \mu_0)'(\Sigma_1/n_1 + \Sigma_0/n_0)^{-1}(\mu_1 - \mu_0)$, so when n_1 is equal to n_0 , we can apply Hotelling's test statistic without affecting the level of significance or power of the test. The same result can be extended to the case of the conditional F test for testing that a subset of the parameters is equal to zero.

Ito & Schull (1964) study the robustness of Hotelling's test statistic to departures from the assumption of equal covariance matrix. The approach they take to the problem is to find constants c and f such that the distribution of the statistic can be approximated as c times a chi square with f degrees of freedom, where f and c are adjusted to give the correct two first moments of the distribution.

In the case of two samples it reduces to finding the positive characteristic roots of $\Sigma_1(\Sigma_0)^{-1}$; then the significance level and the power of the test are function of these roots. In the particular situation where Σ_1 is equal to $\Sigma(1 + (-1)^{i-1}\delta)$ these roots are $(1 + 2\delta)$; when δ is equal to zero the roots are equal to one and the case reduces to the equal covariance matrix case and obviously the power and significance level are not affected. When the sample sizes are equal the power and significance level are also not affected. This approach can be extended to the case of the conditional F test for testing that a subset of the parameters is equal to zero. Nevertheless we can no longer compare

the power of this test with the likelihood ratio test based on the logistic regression estimators because the significance level might be different.

Hotelling's test was constructed for testing hypotheses about the mean of normally distributed random variables, but because of the form of the linear discriminant function it can be used to test hypotheses about the linear discriminant coefficients. To see that this is still valid in the case of unequal covariance matrices, we see that the inverse of the pooled estimator of the covariance matrix is given by

$$S^{-1} = \frac{n-2}{n_1 n_0} \left(\frac{n_1-1}{n_1 n_0} S_1 + \frac{n_0-1}{n_1 n_0} S_0 \right)^{-1}. \quad (3.104)$$

If we assume that the sample sizes in both groups are big we have that S_i is approximately equal to Σ_i , and then

$$S^{-1} = \Sigma^{-1} \{1 + \delta(\pi_0 - \pi_1)\}. \quad (3.105)$$

Therefore to test that the means of the populations are equal is equivalent to testing that the linear discriminant function coefficients are zero.

We note (from expression (3.102)) that the expected value of the pooled sample covariance matrix can be written as

$$E(S) = \Sigma \{1 + \delta(\pi_0 - \pi_1)\} \{1 + 2\delta(\pi_1 - \pi_0)\}, \quad (3.106)$$

so if we define the linear discriminant function coefficient estimators as

$$(1 + 2\delta(\pi_0 - \pi_1))(S^*)^{-1}(\bar{x}_1 - \bar{x}_0), \quad (3.107)$$

where $S^* = S(1 + 2\delta(\pi_0 - \pi_1))$, having assumed that $(n - 2)S$ is Wishart distributed $W(\Sigma\{1 + \delta(\pi_1 - \pi_0)\}; n - 2, p)$, we have that $(n - 2)S^*$ is Wishart distributed $W(\Sigma^* = \Sigma\{1 + \delta(\pi_0 - \pi_1)\}; n - 2, p)$, independent of $\bar{x}_1 - \bar{x}_0$ which (from expression (3.103)) has covariance equal to $(1/n_1 + 1/n_0)\Sigma^*$. Hence we can calculate the mean and covariance of the linear discriminant function coefficients as in section 3.2.1 replacing Σ by Σ^* . Then the mean of the linear discriminant function is given (ignoring terms of order δ^2 and assuming S^* Wishart distributed) by

$$\frac{(1 + \delta(\pi_0 - \pi_1))(n - 2)}{(n - p - 3)} \Sigma^{-1}(\mu_1 - \mu_0), \quad (3.108)$$

and the variance is

$$\begin{aligned} (1 + 3\delta(\pi_0 - \pi_1))C \left\{ \left(\frac{n(n - 3)}{n_1 n_0} + \Delta^2 \right) \Sigma^{-1} + \Sigma^{-1} \mu \mu' \Sigma^{-1} \frac{n - p - 1}{n - p - 3} \right\} \\ + \delta(\pi_1 - \pi_0)C \left\{ \Delta^2 \Sigma^{-1} + \Sigma^{-1} \mu \mu' \Sigma^{-1} \frac{n - p - 1}{n - p - 3} \right\}, \end{aligned} \quad (3.109)$$

where C is equal to $(n - 2)^2 / ((n - p - 2)(n - p - 3)(n - p - 5))$.

It is very easy to see, following the same idea as in section 3.2.1, in the equal covariance case, that asymptotically this estimator is normally distributed with mean $(1 + \delta(\pi_0 - \pi_1)) \Sigma^{-1}(\mu_1 - \mu_0)$ and variance

$$(1 + 3\delta(\pi_0 - \pi_1)) \left\{ \left(\frac{1}{\pi_1 \pi_0} + \Delta^2 \right) \Sigma^{-1} + \Sigma^{-1} \mu \mu' \Sigma^{-1} \right\} + \delta(\pi_0 - \pi_1) \{ \Delta^2 \Sigma^{-1} + \Sigma^{-1} \mu \mu' \Sigma^{-1} \} \quad (3.110)$$

and when β_1 is equal to zero this expression reduces to

$$(1 + 5\delta(\pi_0 - \pi_1))\left(\frac{1}{\pi_1\pi_0} + \Delta^2\right)\Sigma^{-1} + \delta(\pi_0 - \pi_1)\Delta^2\Sigma^{-1}. \quad (3.111)$$

Hence as an approximation for the asymptotic relative efficiency we could use Pitman's theorem, noting, from expression (3.96), that the expected value of the logistic regression estimator of β_1 is $\beta_1(1 + O(\delta))$, where the term of order δ is, in absolute value, equal to or smaller than the absolute value of $\delta/2(\pi_1 - \pi_0)$. (To justify this we notice that the general form of the estimator of α is $a + bx + cx^2$ and the general form of the estimator of β_1 is $a_1/x + b_1 + c_1x$, so, as x tends to zero, we notice that a_1 must be zero and α tends to a ; then if we assume that the contribution of the term δx^2 corresponds only to β we have that c is equal to zero and the expected value of β is $\Delta(1 + \delta/2(\pi_1 - \pi_0))$. Another way of justifying this is by replacing the values of α and β by $\{\log(\pi_1/\pi_0) + (p/2)\log(1-2\delta) + \delta\Delta^2/4\}$ and $\Delta(1 + \delta/2(\pi_1 - \pi_0))$ and evaluating the expected value of the first derivatives, which we find that to be of order δ). The expected value of the linear discriminant function estimator is β_1 times $(1 + \delta(\pi_0 - \pi_1))$, and we have that the asymptotic relative efficiency in terms of π_1 , π_0 , Δ , δ and p , assuming that the expected value of the logistic regression estimator is equal to $\beta_1(1 + \delta/2(\pi_1 - \pi_0))$, can be approximately given by

$$v_1' \left\{ (1 + 2\delta(\pi_0 - \pi_1))\left(\frac{1}{\pi_1\pi_0} + \Delta^2\right) + 2\delta(\pi_1 - \pi_0)\Delta^2 \right\}. \quad (3.112)$$

The results for some values of π_1 , Δ , δ and p are shown in tables 3.4.1 and 3.4.2. We can see that the relative efficiency in this case (different covariance matrices) is not symmetrical with respect to the values of π_1 and π_0 , because population one has bigger variance than population zero. The logistic regression estimators are more efficient when the population with less variance is bigger, i.e., π_1 is less than .5

and Δ is smaller than or equal to 1.5.

Table 3.4.1.

Asymptotic relative efficiency of the normal test based on the logistic regression estimators to the normal test based on the linear discriminant function estimators, when the explanatory variables are bivariate normal and the covariance matrices are different.

$p = 2$

Δ	.5	1	1.5	2	2.5	3	3.5	4
$\pi_1 = 0.1$								
$\delta = 0.05$	1.176	1.128	1.031	0.892	0.728	0.559	0.404	0.274
$\delta = 0.10$	1.314	1.239	1.113	0.950	0.770	0.589	0.424	0.287
$\delta = 0.20$	1.396	1.284	1.130	0.952	0.766	0.585	0.423	0.288
$\pi_1 = 0.3$								
$\delta = 0.05$	1.085	1.062	1.013	0.925	0.798	0.644	0.485	0.340
$\delta = 0.10$	1.154	1.127	1.073	0.979	0.844	0.682	0.515	0.362
$\delta = 0.20$	1.232	1.222	1.177	1.084	0.943	0.769	0.586	0.417
$\pi_1 = 0.5$								
$\delta = 0.05$	0.996	0.995	0.970	0.903	0.791	0.646	0.491	0.346
$\delta = 0.10$	0.983	0.993	0.977	0.916	0.805	0.660	0.503	0.357
$\delta = 0.20$	0.934	0.983	1.003	0.964	0.865	0.721	0.558	0.401
$\pi_1 = 0.7$								
$\delta = 0.05$	0.900	0.904	0.879	0.812	0.706	0.573	0.433	0.304
$\delta = 0.10$	0.792	0.810	0.802	0.750	0.657	0.537	0.408	0.288
$\delta = 0.20$	0.562	0.600	0.623	0.606	0.544	0.453	0.350	0.251
$\pi_1 = 0.9$								
$\delta = 0.05$	0.797	0.795	0.759	0.680	0.568	0.444	0.324	0.221
$\delta = 0.10$	0.586	0.594	0.581	0.532	0.451	0.356	0.261	0.179
$\delta = 0.20$	0.122	0.179	0.179	0.175	0.160	0.133	0.101	0.046

Because the variance of the logistic regression estimators depends on p , we calculate the asymptotic relative efficiency for different values of p ; from the tables we can see that as p increases the relative efficiency of the logistic regression estimators increases when π_1 is smaller than .5 (with the exception of π_1 equal to .3 and Δ less than 1).

Table 3.4.2.
Asymptotic relative efficiency of the normal test based on the logistic regression estimators to the normal test based on the linear discriminant function estimators, when the explanatory variables are ten dimensional multivariate normal and the covariance matrices are different.
 $p = 10$

Δ	.5	1	1.5	2	2.5	3	3.5	4
$\pi_1 = 0.1$								
$\delta = 0.05$	1.548	1.433	1.261	1.056	0.840	0.634	0.451	0.302
$\delta = 0.10$	2.118	1.897	1.620	1.327	1.041	0.777	0.549	0.367
$\delta = 0.20$	2.542	2.401	2.135	1.800	1.442	1.097	0.789	0.536
$\pi_1 = 0.3$								
$\delta = 0.05$	1.191	1.156	1.094	0.992	0.851	0.684	0.514	0.340
$\delta = 0.10$	1.246	1.253	1.218	1.126	0.979	0.796	0.602	0.425
$\delta = 0.20$	0.909	1.095	1.289	1.370	1.309	1.137	0.905	0.664
$\pi_1 = 0.5$								
$\delta = 0.05$	0.941	0.963	0.958	0.905	0.780	0.658	0.502	0.356
$\delta = 0.10$	0.787	0.867	0.924	0.917	0.839	0.708	0.550	0.396
$\delta = 0.20$	0.404	0.529	0.725	0.898	0.958	0.898	0.754	0.574
$\pi_1 = 0.7$								
$\delta = 0.05$	0.741	0.779	0.792	0.758	0.676	0.558	0.428	0.304
$\delta = 0.10$	0.492	0.559	0.627	0.649	0.612	0.526	0.415	0.302
$\delta = 0.20$	0.168	0.211	0.295	0.401	0.468	0.467	0.408	0.320
$\pi_1 = 0.9$								
$\delta = 0.05$	0.573	0.593	0.600	0.569	0.498	0.403	0.302	0.210
$\delta = 0.10$	0.292	0.314	0.346	0.364	0.348	0.301	0.236	0.170
$\delta = 0.20$	0.041	0.043	0.049	0.061	0.073	0.074	0.063	0.047

In conclusion we can say that when π_1 , the probability of the population with larger variance, is bigger than or equal to .5 the logistic regression estimators are less efficient asymptotically relative to the linear discriminant function estimators and even more, they are relatively less efficient than in the case of equal covariance matrices. That the logistic regression estimators are significantly more efficient than the linear discriminant function estimators when population two is much bigger than population

one, Δ^2 is smaller than 1.5 the efficiency increases as p and the departures from equal covariance increase.

Obviously a better approximation and the possibility to investigate bigger departures from the standard case can be obtained using the terms of order δ^2 in the approximations.

3.6.3 DEPARTURES FROM THE ASSUMPTION OF NORMALITY

Alternative distributional models to the multivariate normal distribution are provided by the class of elliptical distributions. The use of this class of distributions in robustness studies is of interest because the class contains long-tailed and short-tailed distributions relative to the normal and the contours of equal density for all the members of the family have the same elliptical shape as the normal. As earlier studies on robustness suggest (see for instance Bartlett (1935), Geary(1947)) tests on population means (which the tests on linear discriminant function coefficients essentially are) are rather insensitive to departures from normality in terms of null distributions and significance levels. Even more, for some tests, including the Hotelling's T^2 test, it is possible to show (see Kariya (1981a)) that the null distribution does not depend on the normal distribution of the observations but on the invariance of the distribution under the transformation Px , for any P in the space of the orthogonal matrices. We shall see this using properties of the elliptical distribution family and therefore we start by reviewing the properties of that family.

3.6.3.1 ELLIPTICAL DISTRIBUTIONS

A random vector X is said to have an elliptical distribution with parameters μ and V if its density function is of the form

$$c_p |V|^{-1/2} h((x - \mu)'V^{-1}(x - \mu)), \quad (3.113)$$

for some function h , where V is positive definite.

Obviously the constant c_p can be absorbed into the function h , but writing the distribution in the form given by (3.112) h is independent of p . If X has an elliptical distribution we write that X is $E_p(\mu, V)$ which does not specify a particular elliptical distribution but only that the distribution belongs to the class of elliptical distributions or $E_p^h(\mu, V)$, which specifies the particular distribution. The characteristic function of an elliptical distribution is

$$\phi_h(t) = \exp(it'\mu)\psi_h(t'Vt), \quad (3.114)$$

for some function ψ_h . Provided they exist the mean and covariance of an elliptical distribution are given by μ and $\Sigma = kV$, where k is given by $-2\psi_h'(0)$.

Many properties of elliptical distributions have been studied by Keller(1970)and they are summarized in Muirhead (1982,pp32-40).

The class of elliptical distributions includes the multivariate normal, multivariate Student's t , Cauchy and mixtures of normal distributions, among others. For the normal distribution we have that $h(u)=\exp(-u/2)$, $c_p=(2\pi)^{-1/2}$, $V=\Sigma$ and $\psi_h(u)=\exp(-u/2)$.

Any $E_p^h(\mu, V)$ random variable can be reduced to an $E_p^h(0, I_p)$, random

variable, by the linear transformation

$$Z = P^{-1}(X - \mu), \quad (3.115)$$

where P is a non singular $p \times p$ matrix such that $P'V^{-1}P = I_p$ and $E_p^h(0, I_p)$ represents a spherical distribution, which is characterized by the fact that if Z has a spherical distribution HZ has the same distribution for all $p \times p$ orthogonal matrices H .

Anderson & Fang (1982c) prove that if X has an elliptical distribution, where h is a decreasing and differentiable function the maximum likelihood estimators of μ and Σ obtained from a sample of size n are $\hat{\mu} = \bar{x}$ and $\hat{\Sigma} = \lambda_{\max} S$, where $\lambda_{\max}$ is the value which maximizes the function $\lambda^{-np/2} h(p/\lambda)$ (in the normal case $\lambda_{\max} = 1/n$). They also find the joint distribution of $\bar{x}$ and S , and the marginal distributions of $\bar{x}$ and S .

If x_i , $i = 1, \dots, n$ are distributed as $E_p^h(\mu, V)$ the distribution of $\bar{x}$ is an elliptical distribution $E_p^h(\mu, (1/n)V)$, and the distribution of S , writing m instead of $n-1$ for simplicity is given by

$$\frac{\pi^{mp/2 - p(p-1)/4}}{\prod_{i=1}^p \Gamma(\frac{m-i-1}{2})} |S|^{(m-p-1)/2} |\Sigma|^{-m/2} c_{p_n} h(\text{tr} \Sigma^{-1} S), \quad (3.116)$$

which in the normal case reduces to the Wishart distribution $W(\Sigma; n-1, p)$.

The density of S^{-1} is given by (see Anderson & Fang (1982b))

$$\frac{\pi^{mp/2 - p(p-1)/4}}{\prod_{i=1}^p \Gamma(\frac{m-i+1}{2})} |\Psi|^{m/2} |S^{-1}|^{-(m+p+1)/2} c_{p_n} h(\text{tr} \Psi S), \quad (3.117)$$

which in the normal case reduces to the inverse Wishart distribution, where Ψ is equal to Σ^{-1} .

Campanis, Huang & Simons(1981) prove that if x is an elliptical distributed random variable $E_p(\mu, V)$ then $E((x - \mu)'(x - \mu)) = (ER^2/n)\Sigma$, where ER^2 depends on the particular elliptical distribution. Based on this result Anderson & Fang(1982c) show that the unbiased estimator of Σ is $np/((n-1)ER^2)S$; depending on the particular distribution S might be asymptotically unbiased for Σ .

Mitchell & Krzanowski(1981) show that the Mahalanobis' distance is an appropriate measure of the distance between two elliptical distributions and is given by $c_h(\mu_1 - \mu_0)\Sigma^{-1}(\mu_1 - \mu_0)$, where c_h is a constant depending on the particular elliptical distribution and is the inverse of the asymptotic efficiency of the sample mean as an estimator of μ .

When the explanatory variables follow an elliptical distribution with different means and the same covariance matrix (or the same matrix V), the linear discriminant function coefficients are a multiple of the maximum likelihood estimators, whereas the logistic regression model in general no longer holds, in the sense that the logarithm of the ratio of the conditional probabilities is not a linear function of x . Therefore we expect the linear discriminant function estimators to be more efficient in estimating the effect of a particular variable on the response.

In particular in the case of testing hypotheses about the parameters, Kariya(1981b) shows that any test for testing that a subset of the means of two normal distributed population with the same covariance matrix are equal is robust when the true distribution is elliptical.

As in the normal case the tests used in regression analysis can be used to test the discriminant function coefficients and the null distribution of the tests is unchanged when the explanatory variables come from an elliptical distribution. Therefore the level of significance of the test is the same for any distribution in the class of elliptical distributions.

Anderson & Fang(1982a) show, in a very elegant way, that if x is an elliptically distributed random variable $E_p(0, V)$ with $P(x = 0) = 0$ and A is an $n \times n$

symmetric matrix, then the distribution of $x'Ax$ is $G_2(k/2; (n - k)/2, h)$ if and only if the rank of A is k and $A\Sigma A = A$ (in particular if $\Sigma = I$ this is true if and only if $A = A^2$), where if $z_1, \dots, z_{m-1}$ are distributed according to $G_m(\alpha_1, \dots, \alpha_{m-1}; \alpha_m; h)$ the distribution of $z_1, \dots, z_{m-1}$ is written as

$$\frac{\Gamma(\frac{n}{2})}{\prod_{i=1}^m \Gamma(\alpha_i)} \prod_{i=1}^{m-1} z_i^{\alpha_i-1} \int_{\sqrt{\sum_{i=1}^m z_i}}^{\infty} r^{-(n-2)} \left\{ r^2 - \sum_{i=1}^{m-1} z_i \right\}^{\alpha_m-1} dF(r), \quad (3.118)$$

where $dF(r)$ depends on the particular distribution and its directly related to h .

They prove that if x is an elliptically distributed random variable $E_p(0, I)$, with $P(x = 0) = 0$ and A and B are $n \times n$ symmetric matrices such that $A^2 = A$ and $B^2 = B$ with rank of A equal to k and rank of B equal to q satisfying $AB = 0$ then the ratio of the quadratic forms $(x'Ax, x'Bx)$ divided by their ranks has an F distribution with degrees of freedom equal to the ranks of the matrices, independent of the particular elliptical distribution.

Consider the linear model $y = \alpha + X\beta + e$, where y , e and β represent $n \times 1$ vectors and X represents an $n \times p$ matrix, with the error, e , assumed to have an elliptical distribution $E_p^h(0, V)$. It is easy to see that y is elliptically distributed $E_p^h(X\beta, V)$; the least squares estimator b of β is given by $(X'X)^{-1}X'y$ and the distribution of b is $E_p^h(\beta, (X'X)^{-1}X'VX(X'X)^{-1})$. In particular, if V is equal to $\sigma^2 I$, the sum of squared residuals is given by $e'(I_p - X(X'X)^{-1}X')e = e'Ae$ and is distributed as $\sigma^2 G_2((n-p-1)/2; p/2; h)$.

Now to test the hypothesis that a subset of the vector of parameters of dimension q is equal to zero, we calculate the estimators under the condition $\beta_{(2)} = 0$, and obtain the sum of squared residuals under this model. The difference of the sum of squared residuals in the restricted model and the full model might be expressed as $e'Be$, where $B = X(X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}H(X'X)^{-1}X'$ and H is a matrix which

ones in the q diagonal elements corresponding to the parameters we want to test and zeros anywhere else. Then $e'Be$ is distributed as $\sigma^2 G_2(q/2; (n - q)/2; h)$ and it is easy to see that $A^2 = A$, $B^2 = B$ and $AB = 0$. Therefore the ratio of the quadratic forms divided by the rank of the matrices is distributed as an F with q and $n - p - 1$ degrees of freedom, independently of the particular elliptical distribution of x .

In the case of the logistic regression model applying Bayes theorem, we have that

$$\log\left(\frac{P(y = 1 | x)}{P(y = 0 | x)}\right) = \log\left(\frac{h(x - \mu_1)V^{-1}(x - \mu_1)}{h(x - \mu_2)V^{-1}(x - \mu_2)}\right), \quad (3.119)$$

which is not in general linear in x . Again the parameters of the logistic model are calculated from the conditional likelihood therefore the (conditional) likelihood ratio test is still valid and chi square distributed, with degrees of freedom equal to the dimension of the vector of parameters that are tested.

Kay and Little (1987) extend the idea of the logistic regression model to consider that the log ratio of the conditional distributions of Y given x is not a linear function of x but of some function of x , $g(x)$. Then the logistic regression is appropriate if we use $g(x)$ instead of x as the vector of explanatory variables. They find the transformations for the case x is a member of the exponential family. To find the appropriate transformation in multivariate cases could be complicated; however in the case of elliptical distributions, we can always transform $(x - \mu_1)'V^{-1}(x - \mu_1)$ to a function which depends only on x_1 and Δ , the Mahalanobis distance, and then find the transformation of x_1 needed.

Obviously the logistic regression estimators are less efficient than the maximum likelihood estimators, because, as in the normal case the logistic regression estimators are obtained maximizing the first term in expression (3.7), ignoring the information contained in the second term. This loss of efficiency depends on the particular elliptical distribution and it is not very easy to obtain. In the case when the explanatory variables

have an e-contaminated normal distribution, it is not difficult to show that the asymptotic efficiency of the logistic regression estimators relative to the maximum likelihood estimators is the same as in the case of normally distributed explanatory variables .

3.7 CONCLUSIONS

The linear discriminant function estimators are more efficient than the logistic regression estimators in testing hypotheses about a subset of the parameters when the explanatory variables, or the set of properties, are normally distributed. The loss of efficiency could be as big as two thirds when the distance between the two populations (Mahalanobis' distance) is bigger than four, but if the distance is equal to or smaller than two the efficiency is over 85%. Nevertheless, as the simulation studies suggest, when the sample sizes are small and an alternative hypothesis is fixed the logistic regression estimators could be more efficient.

In testing hypotheses about the whole vector of parameters the logistic regression estimators are asymptotically fully efficient relative to the linear discriminant function estimators; this is true even for small sample sizes.

The linear discriminant function estimators are more efficient than the logistic regression estimators when the explanatory variables are elliptically distributed. When the covariance matrix is different in the two populations the tests based on the logistic regression estimators are still valid and chi square distributed, whereas the tests based on the linear discriminant function estimators might have a different distribution and the level of significance of the tests might change. Nevertheless if the difference of the covariance matrices is not very big and the Mahalanobis distance is bigger than two the

linear discriminant function estimators are more efficient. The logistic regression estimators are more efficient if the proportion of individuals from the population with smaller variance is much bigger in the total sample.

However, this loss of efficiency is based on the assumed functional form of the conditional distribution of x in the two populations and in many applications these assumptions would be unreasonable.

If both approaches are applicable, and the logistic regression is effectively linear, logistic regression assumed less, in that given the linear regression the form of the distribution of x within sub-populations is irrelevant. On the other hand, if multivariate normality or some other specific distributional form can be assumed, then the discriminant approach is more efficient.

REFERENCES

- Altham, P.M.E.**(1978). ' Two generalization of the binomial distribution '. *Applied Statistics*, **27**,162-167.
- Amemiya, T. and Powell, J.L.**(1983). ' A comparison of the logit model and normal discriminant analysis when the independent variables are binary '. In *Studies in Econometrics, Time Series and Multivariate Statistics*, pp3-30, eds. Karlin, Amemiya and Goodman.
- Anderson, T.W. and Fang, K.-T.**(1982a). ' Distributions of quadratic forms and Cochran's theorem for elliptically contoured distributions and their applications. *Technical Report No. 53*. Department of Statistics, Stanford University.
- Anderson, T.W. and Fang, K.-T.**(1982b). ' On the theory of multivariate elliptically contoured distributions and their applications '. *Technical Report No. 54*. Department of Statistics, Stanford University.
- Anderson, T.W. and Fang, K.-T.**(1982b). ' Maximum likelihood estimators and likelihood ratio criteria for multivariate elliptically contoured distributions '. *Technical Report No. 1*. Department of Statistics, Stanford University.
- Bartlett, M.S.**(1935). ' The effect of non-normality on the t -distribution '. *Proceedings of the Cambridge Philosophical Society*, **31**, 223-231.
- Bartlett, M.S.**(1936). ' Some notes on insecticide tests in the laboratory and in the field. *Journal of the Royal Statistical*, **3**, Suppl. 185-194.
- Blomqvist, N.**(1950). ' On a measure of dependence between two random variables'. *Annals of Mathematical Statistics*, **21**, 593-600.
- Byth, K. and McLachlan, G.**(1980). ' Logistic regression compared to normal discrimination for non-normal populations '. *Australian Journal of Statistics*, **22**, 188-196.

- Cambanis, S., Huang, S. and Simons, G.**(1981). ‘ On the theory of elliptically contoured distributions ’. *Journal of Multivariate Analysis*, **11**, 368-385.
- Cochran, W.G.**(1943). ‘ Analysis of variance for percentages based on unequal numbers ’. *Journal of the American Statistical Association*, **38**, 287-301.
- Cox, D.R. and Hinkley, D.V.**(1974). *Theoretical Statistics*. Chapman and Hall. London.
- Cox, D.R. and Lewis, P.A.W.**(1966). *The Statistical Analysis of a Series of Events*. Methuen. London.
- Crowder, M.J.**(1978). ‘ Beta-binomial ANOVA for proportions ’. *Applied Statistic*, **27**, 34-37.
- Das Gupta, S.**(1968). ‘ Some aspects of discrimination function coefficients ’. *Sankhya*, A, **28**, 387-400.
- Efron, B.** (1975). ‘ The efficiency of logistic regression compared to discriminant analysis ’. *Journal of the American Statistical Association*, **70**, 892-897.
- Firth, D.**(1987). ‘ On the efficiency of quasiliikelihood estimation ’. *Biometrika*, **74**, 233-245.
- Fisher, R.A.**(1936). ‘ The use of multiple measurement in taxonomic problems ’. *Annals of Eugenics*, **7**, 179-188.
- Geary, R.C.**(1947). ‘ Testing for normality ’. *Biometrika*, **34**, 209-242.
- Griffiths, D.A.**(1973). ‘ Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease ’. *Biometrics*, **29**, 637-648.
- Hill, J.R. and Tsai C.**(1988). ‘ Calculating the efficiency of maximum quasiliikelihood estimation ’. *Applied Statistics*, **37**, 219-230.
- Hotelling, H.**(1931). ‘ The generalization of Student's ratio ’. *Annals of Mathematical Statistics*, **2**, 360-378.
- Im, S. and Gianda, D.**(1988). ‘ Mixed models for binomial data with an application to lamb mortality ’. *Applied Statistics*, **37**, 196-204.

- Ito, K. and Schull, W.J.**(1964). ‘ On the robustness of the T^2 test in multivariate analysis of variance when variance-covariance matrices are not equal ’. *Biometrika*, **51**, 71-82.
- Kariya, T.**(1981a). ‘ A robustness property of Hotelling's T^2 test ’. *Annals of Statistics*, **9**, 210-213.
- Kariya, T.**(1981a). ‘ Robustness of multivariate tests ’. *Annals of Statistics*, **9**, 1267-1275.
- Kay, R. and Little S.**(1987). ‘ Transformations of the explanatory variables in the logistic regression model for binary data ’. *Biometrika*, **74**, 495-501.
- Keller, D.**(1970). ‘Distribution theory of spherical distributions and a location-scale parameter generalization ’. *Sankhya, A*, **32**, 419-430.
- Kleinman, J.C.**(1973). ‘ Proportions with extraneous variance: single and independent samples ’. *Journal of the American Statistical Association*, **62**, 46-54.
- Kshirsagar, A.M.**(1972). *Multivariate Analysis*. Dekker, New York.
- Mahalanobis, P.C.**(1936). ‘On the generalized distance in statistic ’. *Proceedings of the National Institute of Science India, A*, **2**, 49-55.
- McCullagh, P.**(1983). ‘Quasilikelihood functions’. *Annals of Statistics*, **11**, 59-67.
- McCullagh, P. and Nelder J.A.**(1983). *Generalized linear models*. Chapman and Hall, London.
- Mitchell A.F.S. and Krzanowsky, W.J.**(1985). ‘ The Mahalanobis distance and elliptical distributions ’. *Biometrika*, **72**, 464-467.
- Muirhead, R.B.**(1982). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, New York.
- Nelder, J.A. and Pregibon, D.**(1987). ‘ An extended quasilikelihood function ’. *Biometrika*, **74**, 221-232.
- Noether, G. E.**(1955). ‘ On a theorem of Pitman ’. *Annals of Mathematical Statistics*, **26**, 64-68.

- Ochi, Y. and Prentice R.L.**(1984). ' Likelihood inference in a correlated probit regression model '. *Biometrika*, **71**, 531-543.
- O'Neill, T.J.**(1978). The general distribution of the error rate of a classification procedure with applications to the logistic regression discrimination '. *Journal of the American Statistical Association*, **75**, 154-160.
- Pierce, D.A. and Sands, B.R.**(1975) ' Extra-bernoulli variation on binary data '. *Technical Report 46*. Department of Statistics, Oregon State University.
- Potthoff, R.F. and Whittinghill, M.**(1966). ' Testing for homogeneity '. *Biometrika*, **53**, 167-182.
- Prentice, R.L.**(1986). ' Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariates measurements errors, *Journal of the American Statistical Association*, **81**, 321-327.
- Press, S.J. and Wilson, S.**(1978). 'Choosing between logistic regression and discriminant analysis '. *Journal of the American Statistical Association*, **73**, 699-705.
- Rao, C.R.**(1946). ' Tests with discriminant functions in multivariate analysis '. *Sankhya*, **7**, 407-414.
- Rao, C.R.**(1948). ' Tests of significance in multivariate analysis '. *Biometrika*, **35**, 58-79.
- Wedderburn, R.W.M.**(1974). ' Quasilikelihood function, generalized linear models and the Gauss-Newton method '. *Biometrika*, **61**, 439-447.
- Williams, D.A.**(1975). ' The analysis of binary responses from toxicological experiments involving reproduction and teratogeneity '. *Biometrics*, **31**, 949-952.
- Williams, D.A.**(1982). ' Extra-binomial variation in logistic linear models '. *Applied Statistics*, **31**, 144-148.