TWO SAMPLE COMPARISONS WITH MIXED DISCRETE AND CONTINUOUS VARIANTS

by

WEN XU, B.S., M.S.

A DISSERTATION

IN

MATHEMATICS

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for
the Degree of

DOCTOR OF PHILOSOPHY

Approved

August, 1996

ACKNOWLEDGMENTS

I would like to express my appreciation to Dr. Clyde Martin for his constant encouragement and guidance. I am also grateful to Dr. Frits Ruymgaart, Dr. Ronald Anderson, Dr. Song Yang and Dr. Shan Sun for serving on my committee.

A word of thanks goes to all my friends in Lubbock for their helpful suggestions. I am also indebted to my friends in Alabama Quality Assurance Foundation.

My deepest gratitude and love go to my family: my father Tsui Kuen-Faat, my mother Lam Ling and my brother Tsui Pui for their love, encouragement and advice.

CONTENTS

ABSTRACT

The comparisons of means, dispersions or distributions are well understood and researched for univariate investigations. Much literature is available for the design and implementation of studies involving multivariate normal samples. Most of investigations into multivariate, non-normal comparisons still remain open.

This dissertation presents what is available and provides two simple methods to transform the multivariate two sample comparison into the well known chi-square test or logistic model. The methodologies and techniques will be examined by simulations and utilized to analyze the data from Texas Tech Health Science Center on the screening of Alzheimer's disease.

LIST OF TABLES

CHAPTER I

INTRODUCTION

This dissertation will attempt to test whether two samples have the same means, variances and distributions. The problem is hereafter referred to as two sample comparisons. Numerous applications of this problem exist such as comparing the quality of two different brands, comparing two medical treatments, and comparing the effectiveness of two drugs for a certain disease.

We classify the two sample comparisons into two groups: univariate two sample comparisons, which we know much about, and multivariate two sample comparisons, which we do not know much about. In univariate cases, we could almost say the problem is solved when the samples are normally distributed. There are many results for the unknown distribution cases such as a run test, the Mann-Whitney-Wilcoxon test and others. In the multi-variate case, the treatments of normally distributed samples have been developed by others. Currently, discretely distributed samples and non-normally continuous samples are gaining more attention because of their practical applications. In this dissertation, we will present a brief literature review, then we will provide two methods of transferring the two sample comparisons into χ^2 test or the regression significance test. We will check on significance levels and power by simulation and then apply the methods to a real data set.

Our interests are particularly focused on the comparisons of multivariate two samples with two different types of variables, namely continuous and discrete variables. Apparently, there is not much literature on this subject, yet we found this study necessary because our research projects with the Texas Tech Health Science Center are related to this problem. For example, the goal of the intensive care unit (ICU) project is to assess the outcomes and costs of various groups of babies that have entered the neonatal intensive care units. Each group contains certain variables that could be continuous or discrete, that is, we are trying to distinguish complicated multivariate samples. Another example is the Alzheimer's disease project. The project analyzes and compares several

groups of people with common histories, which might, hopefully, lead to the discovery of the underlying causes of Alzheimer's disease. The data contains many variables with mixture continuous and discrete types. The ventilation treatment comparison project is an additional example where the multivariate two sample comparisons are needed. The purpose of the ventilation study is to test the difference between two ventilation treatments: conventional ventilation (CV) and high frequency oscillatory ventilation (HFOV). The null hypothesis is that there is no difference between the effects of the two treatments in sixteen different aspects. Some of the sixteen variables such as the amount of oxygen needed or the amount of surfactant doses needed are continuous. Other variables are discrete. Those include the incidence of survival or the incidence of chronic lung disease.

This dissertation is composed of seven chapters. In the second chapter, we will present methods for the two sample problems in univariate settings under different conditions. It will be a summary of the univariate two sample problems. From the third chapter through the seventh chapter, we will concentrate on the multivariate two sample problems. Specifically, in the third chapter, we will present the solutions for problems in which the samples are multivariate normally distributed. The fourth chapter will discuss the results of the multivariate nonparametric two sample problem. The fifth chapter will present the maximum likelihood quadratic exponential model on multivariate binary samples, then we will consider a way to compare the samples by the use of a chi-square test. In the sixth chapter, we will study mixed continuous and discrete multivariate two sample comparisons. We will first present the generalized estimating equation method, and then discuss the method of transferring two sample comparisons to logistic regression tests. In the seventh and final chapter, we will inspect the chi-square test used in Chapter V, and the logistic regression test shown in Chapter VI by simulations. Then we will apply the methods with real data regarding Alzheimer's disease.

CHAPTER II

UNIVARIATE MODELS

Univariate two sample comparisons are the problems of comparing two samples with only one variable. The problem occurs frequently in practice. For example, people would like to know which medication stops their cold faster. This is the univariate type problem because there is only one quantity, the time necessary to stop a cold. In this chapter, we will present the results under certain common conditions. First, we will consider a case in which the variable has a normal distribution. Then we will look at the discrete variable cases. Finally, we will present some results in a situation in which the distribution of the variable is not known.

2.1 Normally Distributed Random Variables

Suppose we have two samples: $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$. x_i's are assumed to be independently and identically distributed (*i.i.d.*) random copies of a variable with a normal distribution $N(\mu, \sigma^2)$. y_i's are *i.i.d.* with normal $N(\nu, \tau^2)$. To compare whether these two samples have the same distribution, we only need to test the difference of the means and the difference of variances.

It is reasonable to perform the test for the variances first, and the hypotheses are:

$$H_0: \sigma^2 = \tau^2, \quad vs \quad H_1: \sigma^2 \neq \tau^2.$$

Here we use s_x^2 and s_y^2 to estimate the variances σ^2 and τ^2, where s_x^2 and s_y^2 are defined to be:

$$s_x^2 = \sum (x_i - \bar{x})^2 / (n-1),$$
$$s_y^2 = \sum (y_i - \bar{y})^2 / (m-1).$$

The test statistic is:

$$F_0 = \frac{s_x^2}{s_y^2},$$

where F_0 has F-distribution with degrees of freedom: $n-1$ and $m-1$ under the null

hypothesis H_0. Hence, for a given significance level α, we will reject H_0 if :

$$F_0 > F_{\alpha/2,n-1,m-1} \quad \text{or}$$

$$F_0 < F_{1-\alpha/2,n-1,m-1}.$$

After testing for the variances, if we do not reject H_0, we then perform the test for the means. The hypotheses of testing means are :

$$H_0: \mu = \nu, \quad vs \quad H_1: \mu \neq \nu.$$

Since $\bar{x} \sim N(\mu, \sigma^2/n)$, $\bar{y} \sim N(\nu, \tau^2/m)$, and $(\bar{x} - \bar{y}) \sim N(\mu - \nu, \sigma^2/n + \tau^2/m)$. We use sample variances s_x^2 and s_y^2 to substitute σ^2 and τ^2.

We define:

$$s_p^2 = [(n-1)s_x^2 + (m-1)s_y^2]/(n+m-2),$$

as the estimate for the common variance.

The test statistic for the two sample t-test is:

$$t_0 = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}},$$

where t_0 has the t-distribution with $df = n+m-2$ degrees of freedom under the null hypothesis H_0. For a given significance level α, we will reject H_0 if $|t_0| > t_{\alpha/2,m+n-2}$. We will conclude that the two samples have the same distribution if we do not reject the null hypotheses of both of the tests.

Even in the case of rejecting the common variance hypothesis H_0: $\sigma^2 = \tau^2$ in the first test, people might still be interested in testing the means. There exits an approximated t-test in this situation. The test statistic is:

$$t_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}.$$

Here t_0 has approximated t-distribution the degrees of freedom:

$$df = \frac{(\frac{s_x^2}{n} + \frac{s_y^2}{m})^2}{\frac{(s_x^2/n)^2}{n-1} + \frac{(s_y^2/m)^2}{m-1}}.$$

We will reject the hypothesis H_0: $\mu = \nu$ if $|t_0| > t_{\alpha/2,df}$.

2.2 Binary Variables

In this section, we consider the tests between two samples with univariate binary variables. A binary variable is a random variable taking only two values: *0* and *1*. We choose to discuss a binary variable instead of other discrete random variables because it occurs quite often in practice, especially in medical applications, and it is a very typical discrete random variable.

Two samples $\{x_1, x_2, ..., x_n\}$ and $\{y_1, y_2, ..., y_m\}$ are from two binary variables. That is, x_i's are *i.i.d.* with binary distribution $B(1, p_1)$, y_i's are *i.i.d.* with distribution $B(1, p_2)$. Here B stands for binomial distribution, p_1 is the probability of x variable taking value *1*, and p_2 is the probability of y taking *1*.

Let us first consider the normal approximation test. To compare whether these two samples have the same distribution, in this case, we only need to test:

$$H_0: p_1 = p_2, \quad vs \quad H_1: p_1 \neq p_2.$$

Since: $x_1, x_2, ..., x_n \sim B(1, p_1)$, and $y_1, y_2, ..., y_m \sim B(1, p_2)$, we know that:

$$\sum x_i \sim B(n, p_1), \quad \sum y_i \sim B(m, p_2).$$

Then,

$$\frac{\bar{x} - p_1}{\sqrt{p_1(1-p_1)/n}} \to N(0,1), \qquad \frac{\bar{y} - p_2}{\sqrt{p_2(1-p_2)/m}} \to N(0,1),$$

for large n and m.

Under the null hypothesis H_0: $p_1 = p_2$, the following statistic is approximately normally distributed for large sample sizes n and m. Let $cp = (\sum x + \sum y)/(n+m)$, and $s_p^2 = cp(1-cp)$. The test statistic is:

$$z_0 = \frac{\bar{x} - \bar{y}}{s_p\sqrt{(1/n) + (1/m)}}.$$

We reject H_0: $p_1 = p_2$ if $|z_0| > z_{\alpha/2}$.

We could also perform the χ^2- test for the hypotheses. The test statistic is:

$$Q = \frac{(\sum x - n\hat{p})^2}{n\hat{p}} + \frac{(\sum y - m\hat{p})^2}{m\hat{p}},$$

where $\hat{p} = (\sum x + \sum y)/(n + m)$ is the estimate for the common probability under the null hypothesis. The test statistic has the χ^2-distribution with *1* degree of freedom. It is also possible to use nonparametric methods which we will present in the next section.

2.3 Nonparametric Tests

When the samples have unknown distributions, we will test whether the two samples have the same distribution by utilizing nonparametric methods. Let us discuss the Kolmogorov-Smirnov test first. The test hypotheses are:

$$H_0: F(X)=G(Y) \quad vs \quad H_1: F(X) \neq G(Y),$$

where $F(X)$ and $G(Y)$ are the cumulative distribution functions for the random variables X and Y. We define the empirical cumulative distribution functions of the two samples as:

$\hat{F}(x) = 1/n$ (number of observations in sample *1* that are $\leq x$), and

$\hat{G}(x) = 1/m$ (number of observations in sample *2* that are $\leq x$).

We then define a quantity M as $M = \max|\hat{F}(x)-\hat{G}(x)|$. If n and m are large, the two distribution functions would be considered as significantly different at the α significance level if $M\sqrt{(nm)/(n+m)} > \kappa_\alpha$, where the value κ_α can be found in the table in Fleiss (1986).

If we are only interested in the locations of the two samples, the Mann-Whitney-Wilcoxon test seems to be more powerful than the Kolmogorov-Smirnov test. The hypotheses for this test are:

H_0: the locations of the two samples are equal, and

H_1: the locations are different.

Here, the location could be mean value or median. If the sample sizes are large (both greater than *20*), the Mann-Whitney-Wilcoxon test can be performed by the following procedure:

1. Rank all $n+m$ observations in the pooled sample. Assign the average rank for each of the tied values, for example, if z_3 and z_4 are tied, then they are both ranked *3.5* each.
2. Let $\bar{R}_1$ be the mean of the ranks of the n observations in sample *1*, while $\bar{R}_2$ is the mean of the ranks of the m observations in sample *2*.
3. If there are no ties, then the test statistic satisfies:

$$\chi_0^2 = \frac{12nm(\bar{R}_1 - \bar{R}_2)^2}{(n+m)^2(n+m+1)} \sim \chi 2(1).$$

We reject the null hypothesis if $\chi_0^2 > \chi^2_{1,\alpha}$.

4. If the T values are tied, with *t1* ties at the first tied value, *t2* ties at the second tied value, and so forth, we can adjust the above test statistic as:

$$\chi^2 = \frac{12nm(\bar{R}_1 - \bar{R}_2)^2}{(n+m)^2(n+m+1)f} \sim \chi^2(1),$$

where

$$f = 1 - \frac{\sum ti\ (ti-1)(ti+1)}{(n+m)(n+m-1)(m+n+1)}.$$

We have presented the methods of handling the two sample problem in univariate situations. We will move on to the multivariate two sample comparisons in the next chapter.

CHAPTER III

MULTIVARIATE NORMAL MODEL

For multivariate two sample comparisons, there are not many results available except for multivariate normal cases. Let $X_1,...,X_n$, $X_i=(x_{1i},...,x_{pi})$, be a sample of a p-variate multivariate normal variable with distribution $MVN(\mu, \Sigma)$. $Y_1,...,Y_m$ form a sample of p-variate from $MVN(v, \Omega)$. μ, v are the mean vectors. Σ and Ω are the covariance matrices of the two samples. Since the mean vector and covariance matrix specify a multivariate normal distribution completely, testing the equality of the distributions is equivalent to testing the equality of both the means and the covariance matrices.

Consider the likelihood ratio test for the variance matrices first. The test hypotheses are: $H_0: \Sigma=\Omega$ vs $H_1: \Sigma\neq\Omega$.

Let S_1, S_2 be the estimated covariance matrices for the samples, where

$$S_1=1/v_1[\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'],$$

$$S_2=1/v_2[\sum_{i=1}^{m}(y_i-\bar{y})(y_i-\bar{y})'],$$

$v_1=n-1$, $v_2=m-1$. Suppose S is the pooled covariance matrix based on $N=v_1+v_2$ $n+m-2$ degrees of freedom. Then the test statistic is:

$$M = N\ log|S| - \sum_{i=1}^{2} v_i\ log\ |S_i|.$$

M is asymptotically distributed as χ^2 with $f_1=1\ 2[p(p+1)]$ degrees of freedom. A better test is adjusted as:

$$M \sim \chi 2(f_1) / (1 - D_1),$$

$$D_1 = \frac{2p^2 + 3p - 1}{6(p+1)}\left(\sum_{i=1}^{2} 1/v_i - 1/N\right).$$

We reject the null hypothesis H_0 if $M > \chi^2_{\alpha}(f_1)/(1 - D_1)$.

If we do not reject the null hypothesis in the covariance matrix test, we could test the means assuming that $\Sigma = \Omega$ is true. The test hypothesis is:

$$H_0: u = v, \quad vs \quad H_1: u \neq v.$$

The log likelihood function can be written as:

$$l(\mu, v, \Sigma, \Omega) = -\frac{1}{2}\left(n \log|2\pi\Sigma| + n\, tr\Sigma^{-1}(C + [\bar{x}-\mu][\bar{x}-\mu]') + m \log|2\pi\Omega| + m\, tr\Omega^{-1}(D + [\bar{y}-v][\bar{y}-v]')\right),$$

where

$$C = A/n = 1/n\left[\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'\right],$$

$$D = A_2/m = 1/m\left[\sum_{i=1}^{m}(y_j - \bar{y})(y_j - \bar{y})'\right].$$

C and D are the maximum likelihood estimates of Σ, Ω respectively. $\bar{x}$ and $\bar{y}$ are the maximum likelihood estimates of μ and v. Under the null hypothesis H_0 and the assumption $\Sigma = \Omega$, the maximum likelihood estimate of the pooled mean and covariance matrix are:

$$\overline{xy} = \frac{1}{n+m}\left(\sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i\right),$$

$$\frac{1}{n+m}T = \frac{1}{n+m}\left[\sum_{i=1}^{n}(x_i - \overline{xy})(x_i - \overline{xy})' + \sum_{i=1}^{m}(y_i - \overline{xy})(y_i - \overline{xy})'\right].$$

Let $H = n(\bar{x} - \overline{xy})(\bar{x} - \overline{xy})' + m(\bar{y} - \overline{xy})(\bar{y} - \overline{xy})'$, $E = T - H$. Hence, H is a $p \times p$ sum of squares and product matrix based on q degrees of freedom, where q is the rank of the matrix. E is the residual matrix with $n+m-2-p$ degrees of freedom after fitting all the

parameters by maximum likelihood estimates. Let $l_1,...,l_p$ be the eigenvalues of ET^{-1}, and $c_1,...,c_p$ be the eigenvalues of HE^{-1}. Then we know that $l_i=c_i\ (1+c_i)$.

There are various test statistics based on these invariant eigenvalues; however, we will only present four popular tests.

1. Wilks's criterion: Let $W = \prod_{i=1}^{p}(1 - l_i) = |E| / |H + E|$, then,

$$- [n + m - 2 - \frac{1}{2}(p + q + 1)]\ log(W) \sim a\chi 2([pq]) ,$$

where a is a multiplier tabulated by Schatzoff (1966).

2. Two sample Hotelling's T^2 test: $T^2 = 1 + \frac{nm}{n + m}(\bar{x} - \bar{y})' E^{-1} (\bar{x} - \bar{y})$ is the test statistic. It is similar to the univariate t^2 test.

$$T^2 \sim \frac{(n + m - 2)p}{n + m - p - 1}F(p,\ n + m - p - 1).$$

We reject the null hypothesis H_0 if

$$T^2 > \frac{(n+m-2)p}{n+m-p-1} F_\alpha (p,\ n+m-p-1).$$

3. The Lawley-Hotelling trace method: Let $H1 = \sum_{i=1}^{p} c_i = tr(HE^{-1})$. Approximately,

$$\frac{n^* H1}{p^2 q} \sim F(pq,\ n^*) ,$$

where $n^*=p(n+m-p-q-3)+2$.

4. The Pillai trace: Let the test statistic be $D1 = \sum_{i=1}^{p} l_i = tr\ H(H + E)^{-1}$, then the approximate test is:

$$\frac{p(n + m - 2 - q)D1}{pq(p - D1)} \sim F(pq,\ p(n + m - 2 - q)) .$$

We can see that the multivariate two sample problem is much more complicated than the univariate cases even when the distributions are multivariate normal. In the next chapter, we will consider the nonparametric tests.

CHAPTER IV

MULTIVARIATE NONPARAMETRIC MODELS

Nonparametric models for multivariate two sample problems are difficult. They involve deep mathematics and probability theories. We need to make some assumptions in order to use nonparametric methods. The basic assumptions are: the distribution functions have to be absolutely continuous to avoid ties on the ranks; the two samples are from the same unknown distribution family; and large sample sizes help to develop asymptotic tests. In this chapter, we will first consider the location problem, then we will look at the covariance matrices.

4.1 Two Sample Multivariate Location Models

Let $X_1,\ldots, X_n$ and $Y_1,\ldots, Y_m$ be samples from p-variate distributions with $c.d.f.$s: $F(t_1,\ldots,t_p)$ and $F(t_1-\Delta_1,\ldots,t_p-\Delta_p)$, respectively, where $X_i=(x_{1i},\ldots,x_{pi})'$. Assume that F is absolutely continuous with absolutely continuous marginal distributions. $\Delta=(\Delta_1,\ldots,\Delta_p)'$ is the amount of location shift from X distribution to Y distribution.

The hypotheses of testing the locations are:

$$H_0: \Delta=0, \quad vs \quad H_1: \Delta\neq 0.$$

We consider the following two tests:

1. The multivariate version of the Mann-Whitney-Wilcoxon test statistic is

$$U^* = N^{-1}U'(\hat{V})^{-1}U = U'(N\hat{V})^{-1}U$$

where

$$U=(u_1,\ldots,u_p)',$$

$$u_i = \sum_{t=1}^{m}[\frac{R_{it}}{N+1} - 1/2],$$

and R_{it} is the rank of Y_{it} in the combined sample of the ith component. Let $N=n+m$ be the size of the combined sample. $\hat{V}$ is the estimated asymptotic covariance matrix with the elements:

$$\hat{v}_{ii} = var(U_i/\sqrt{N}) = \frac{mn}{12N(N+1)},$$

$$\hat{v}_{ij} = \frac{nm}{N^2(N-1)(N+1)^2}\left[\sum_{t=1}^{N} R_{it}R_{jt} - N(N+1)^2/4\right].$$

Under H_0: $\Delta=0$, U^* is asymptotically χ^2 with p degrees of freedom. Hence, we reject H_0 at the significance level α if $U^* \geq \chi^2_\alpha(p)$.

2. Mood's test: For $i=1,...,p$, let $M_i=\#(Y_{it} > \hat{c}_i) - m/2$, $t=1,...,m$, where $\hat{c}_i$ is the median of the combined sample in the ith component. Let $M'=(M_1,...,M_p)$. The test is :

$$M^* = M'(N\hat{V})^{-1}M \sim \chi^2(p),$$

where $\hat{V} = ((\hat{v}_{ij}))_{i,j=1,..p}$, $\hat{v}_{ii} = \frac{mn}{4m(N-1)}$, $\hat{v}_{ij} = \frac{mn}{N(N-1)}(N_{11}/(N-1) - 1/4)$, where N_{11} is the number of pairs, and the ith and jth components are both positive in the combined sample. We reject H_0, at significance level α if $M^* \geq \chi^2_\alpha(p)$.

4.2 Two Sample Multivariate Covariance Models

In this section, we study the problem of testing the homogeneity of covariance matrices by using rank order statistics. Let $F^{(1)}$ be the distribution of the first sample, and $F^{(2)}$ be the distribution of the second sample. Assume that the distributions are completely specified by location vectors and covariance matrices.

First, we consider the equality of covariance matrices, assuming the identity of locations. The hypotheses are: H_0: $F^{(1)}=F^{(2)}$ *vs* H_1: $F^{(1)} \neq F^{(2)}$, which implies that $\Sigma^{(1)}=\Sigma^{(2)}$ assuming the same locations. let's rewrite the notation in order to present the rank matrices more clearly. Let $X_\alpha^{(1)}$, $\alpha=1,...,n_1$, be the first sample, $X_\alpha^{(2)}$, $\alpha=1,...,n_2$, be the second sample, $N=n_1+n_2$. Pool the N observations into a combined sample: $Z_N' = (X_1^{(1)},...,X_{n_2}^{(2)})$. The rank matrix of the combined sample is:

$$R_N = \begin{bmatrix} R_{11}^{(1)} & \cdots & R_{1n_1}^{(1)} & \cdots & R_{1n_2}^{(2)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_{p1}^{(1)} & \cdots & R_{pn_1}^{(1)} & \cdots & R_{pn_2}^{(2)} \end{bmatrix}$$

Define a score function $E_{N,\alpha}^{(i)} = J_{N(i)}(\alpha/(N+1))$ $\alpha=1,...,N$, $i=1,...,p$, where

$$J_{N(i)}(\alpha/(N+1)) = [12/(N^2-1)]^{1/2}(\alpha-(N+1)/2),$$

where $\alpha=1,...,N$, $i=1,...,p$. The corresponding score matrix is:

$$E_N = \begin{bmatrix} E^{(1)}_{N,R_{11}^{(1)}} & \cdots & E^{(1)}_{N,R_{1n_1}^{(1)}} & \cdots & E^{(p)}_{N,R_{1n_2}^{(2)}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ E^{(p)}_{N,R_{p1}^{(1)}} & \cdots & E^{(p)}_{N,R_{pn_1}^{(2)}} & \cdots & E^{(p)}_{N,R_{pn_2}^{(2)}} \end{bmatrix}$$

Define

$$S_{N,ij}^{(k)} = \frac{1}{n_k-1}[\sum_{\alpha=1}^{n_k} E_{N,R_{i\alpha}^{(k)}}^{(i)} E_{N,R_{j\alpha}^{(k)}}^{(j)} - \frac{1}{n_k}(\sum_{\alpha=1}^{n_k} E_{N,R_{i\alpha}^{(k)}}^{(i)})(\sum_{\alpha=1}^{n} E_{N,R_{j\alpha}^{(k)}}^{(j)})], \; i \le j=1,...,p; \; k=1,2.$$

Then define

$$S_{N,ij}^{(*)} = \frac{1}{N-1}[\sum_{k=1}^{2}\sum_{\alpha=1}^{n_k} E_{N,R_{i\alpha}^{(k)}}^{(i)} E_{N,R_{j\alpha}^{(k)}}^{(j)} - N\,\bar{E}_N^{(i)}\bar{E}_N^{(j)}],$$

where

$$\bar{E}_N^{(i)} = \frac{1}{N}[\sum_{k=1}^{2}\sum_{\alpha=1}^{n_k} E_{N,R_{i\alpha}^{(k)}}^{(i)}], \; i=1,...,p.$$

Let

$$S_N^{(k)} = \{S_{N,ij}^{(k)}, \; i \le j=1,...,p\},$$

$$S_N^{(*)} = \{S_{N,ij}^{(*)}, \; i \le j=1,...,p\}.$$

Then define

$$v_{ij,i'j'}(R_N) = \frac{1}{N}\sum_{k=1}^{2}\sum_{\alpha=1}^{n_k} E_{N,R_{i\alpha}^{(k)}}^{(i)} E_{N,R_{j\alpha}^{(k)}}^{(j)} E_{N,R_{i'\alpha}^{(k)}}^{(i')} E_{N,R_{j'\alpha}^{(k)}}^{(j')} - S_{N,ij}^{*} S_{N,i'j'}^{*}.$$

Let $V_N(R_N)=((v_{ij,i'j'}(R_N))$, and the test statistic is:

$$\delta_N = \sum_{k=1}^{2} n_k [S_N^{(k)} - S_N^{*}] V_N^{-1}(R_N)[S_N^{(k)} - S_N^{*}]'$$

For large samples, δ_N approximately goes to $\chi^2\,(p(p+1)\,2)$. For a given significance level ϵ, we reject H_0 if $\delta_N \geq \chi^2_{\epsilon}(p(p+1)/2)$.

The above test is based on the assumption that the locations of the distributions are equal. If we do not have that assumption, we can consider the locations and covariance matrices simultaneously. Let

$$U_{N,ij}^{(k)} = \frac{1}{n_k - 1}\sum_{\alpha=1}^{n_k} E_{N,R_{i\alpha}}^{(i)(k)}\, E_{N,R_{j\alpha}}^{(j)(k)}\,, \quad k=1,2;\ i\leq j=1,\ldots,p.$$

Define

$$Z(R_N)=((Z_{ij,i'}(R_N)))_{i\leq j,\ i'=1,\ldots,p},$$

where

$$z_{iji'}(R_N) = \frac{1}{N}\sum_{k=1}^{2}\sum_{\alpha=1}^{n_k} E_{N,R_{i\alpha}}^{(i)(k)}\, E_{N,R_{j\alpha}}^{(j)(k)}\, E_{N,R_{i'\alpha}}^{(i')(k)} - \bar{E}_N^{(i')}\, v_{ij}(R_N),$$

and

$$W(R_N)=((w_{iji'j'}(R_N)))_{i\leq j,\,i'\leq j'=1,\ldots,p},$$

where

$$w_{iji'j'}(R_N) = \frac{1}{N}\sum_{k=1}^{2}\sum_{\alpha=1}^{n_k} E_{N,R_{i\alpha}}^{(i)(k)}\, E_{N,R_{j\alpha}}^{(j)(k)}\, E_{N,R_{i'\alpha}}^{(i')(k)}\, E_{N,R_{j'\alpha}}^{(j')(k)} - v_{ij}(R_N)v_{ij}(R_N).$$

Let the test statistic be :

$$L_N^* = \sum_{k=1}^{2} n_k Q_k\, [W^*((R_N))]^{-1} Q_k\,,$$

where

$$W^*(R_N) = \begin{bmatrix} V(R_N) & Z(R_N) \\ Z'(R_N) & W(R_N) \end{bmatrix},$$

$$Q'_k=(T_N^{(k)'}, U_N^{(k)'})\,,$$

$$T_N^{(k)}=(T_{N,1}^{(k)},\ldots,T_{N,p}^{(k)})$$

and

$$T_{N,i}^{(k)} = \frac{1}{n_k}\sum_{\alpha=1}^{n_k} E_{N,R_{i\alpha}}^{(i)(k)}\,, \quad i=1,\ldots,p.$$

$$U_N^{(k)}=(U_{N,ij}^{(k)},\ 1\leq i\leq j\leq p)\ k=1,\ldots,c.$$

For large samples, $L_N^* \sim \chi^2(p(p+3))$.

There are other methods dealing with the nonparametric multivariate two sample comparisons, such as the projection pursuit method which is based on selecting important low dimensional projections by iteratively maximizing an appropriate projection.

Multivariate generalizations of the Wald-Wolfowttz and Smirnov two sample tests use minimal spanning trees in graph theory. The details can be found in the reference section of this dissertation.

CHAPTER V

MULTIVARIATE BINARY MODEL

There is growing interest in the statistical and medical literature concerning the analysis of correlated binary data. This type of data arises when repeated measures of a binary response variable are taken over time or when two or more measurements are taken at one time on the same individual. The two sample comparisons associated with multivariate binary responses are useful in medical treatment comparisons since many outcomes of treatments are binary.

We first study the two sample comparisons in a different prospect. We introduce a dummy variable which is used to indicate which sample an observation is from. For two sample cases, we assign the values *0* or *1* to the dummy variable. Then we consider the two sample comparison problem as a regression problem. We consider the comparison problem to be a regression problem because there are regression techniques which can be used for significance tests.

In this chapter, we consider a model which can be used to test the identity of the means: the maximum likelihood quadratic exponential model by Zhao and Prentice (1990). A general model, the generalized estimation equation (GEE) model by Zeger and Liang (1986), is also suitable in this situation. However, it is more appropriate to present the GEE model in the next chapter.

Consider the combined sample of the two samples: $y_1,...,y_K$, where $K=n+m$, and $y_k'=(y_{k1},...,y_{kp})$ is a p-variate binary response vector. We can express the distribution of y_k as a quadratic exponential distribution:

$$pr(\ y_k\ ) = \Delta_k^{-1}\ exp\ [y_k'\theta_k + w_k'\lambda_k + c_k(y_k\ )] \qquad (5.1),$$

where $w'_k=(y_{k1}y_{k2},..., y_{k1}y_{k3},..., y_{k2}y_{k3},...,)$. The parameters are: $\theta'_k=(\theta_{k1},..., \theta_{kp})$, and $\lambda'_k=(\lambda_{k12}, \lambda_{k13},..., \lambda_{k23},...)$. $\Delta_k=\Delta_k(\theta_k, \lambda_k)$ is a normalization constant which is defined by: $\Delta_k = \sum exp[y_k'\theta_k + w_k'\lambda_k + c_k(y_k\ )]$. $c_k(.)$ is a shape function which can be expressed as a linear combination of the products of some elements of y_k.

Let $\mu_k=E(y_k)$ be a function of $x'_k\beta$: $\mu_k=\mu_k(x'_k\beta)$, where x_k is the covariate, and $x_k=(1,1)$ if y_k is from the first sample, $x_k=(1,0)$ if y_k is from the second sample. $\beta=(\beta_0\beta_1)'$ is a $2\times p$ matrix where $\beta_0=(\beta_{01},\ldots,\beta_{0p})$, $\beta_1=(\beta_{11},\ldots,\beta_{1p})$. The null hypothesis H_0:the two populations have the same mean, is equivalent to H_0: $\beta_1=0$ if the function $\mu_k(.)$ is one to one.

Let $\sigma_k=(\sigma_{k12},\sigma_{k13},\ldots,\sigma_{k(p-1)p})=\sigma_k(\beta,\alpha)$ be the pairwise covariance, then μ_k and σ_k depend on the parameters β and α. Here we need to transform the parameters (θ_k,λ_k) to $(\mu_k(.),\sigma_k(.))$. Let $\eta_k=((\eta_{kij}))_{i=j=1,\ldots,p}$ where $\eta_{kij}=\sigma_{kij}+\mu_{ki}\mu_{kj}$. Then by equation (5.1),

$$\mu_k=\sum y_k p(y_k)=\sum y_k \exp[y'_k\theta_k+w'_k\lambda_k+c_k(y_k)]\Delta_k^{-1},$$

$$\eta_k=\sum w_k p(y_k)=\sum y_k \exp[y'_k\theta_k+w'_k\lambda_k+c_k(y_k)]\Delta_k^{-1}.$$

By differentiating these two equations, we get the Jacobian of the transformation from (θ_k,λ_k) to $(\mu_k(.),\sigma_k(.))$ is $\tilde{V}_k = cov(y'_k, w'_k)$. The log likelihood of $p(y_k)$ is:

$$l_k=y'_k\theta_k+w'_k\lambda_k+c_k(y_k)-\log\Delta_k.$$

Then, we get the score function:

$$K^{-\frac{1}{2}}\sum_{k=1}^{K} D'_k V_k^{-1} f_k = 0,$$

where

$$D_k = \begin{bmatrix} \partial\mu_k/\partial\beta & 0 \\ \partial\sigma_k/\partial\beta & \partial\sigma_k/\partial\alpha \end{bmatrix},$$

$$V_k = \begin{bmatrix} var(y_k) & cov(y_k,s_k) \\ cov(s_k,y_k) & var(s_k) \end{bmatrix},$$

$$f_k = \begin{bmatrix} y_k-\mu_k \\ s_k-\sigma_k \end{bmatrix},$$

where $s'_k=(s_{k12},s_{k13},\ldots,s_{k23},\ldots)$, $s_{kij}=(y_{ki}-\mu_{ki})(y_{kj}-\mu_{kj})$. $\hat{\beta}$ and $\hat{\alpha}$ are pseudo-maximum likelihood estimators, and $(K^{-\frac{1}{2}}(\hat{\beta}-\beta), K^{-\frac{1}{2}}(\hat{\alpha}-\alpha))$ has an asymptotic normal distribution with mean 0 and consistent asymptotic covariance matrix:

$$K^{-1} W^{-1} \left(\sum_{k=1}^{K} D'_k V_k^{-1} f_k f'_k V_k^{-1} D_k\right) W^{-1},$$

where

$$W=K^{-1}\sum(D'_k V_k^{-1} D_k).$$

The null hypothesis for the population means is equivalent to H_0: $\beta_1=0$, and $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\beta_1})$ approximately, here $\sigma^2_{\beta_1}$ is the submatrix of

$$K^{-1} W^{-1} \left(\sum_{k=1}^{K} D_k' V_k^{-1} f_k f_k' V_k^{-1} D_k\right) W^{-1}.$$

The iterative method of estimating $(\hat{\beta}, \hat{\alpha})$ is developed by using the Newton-Raphson iteration, for the initial value (β_0, α_0),

$$(\beta_w', \alpha_w') = (\beta_0', \alpha_0') + (\sum D_{k0}' V_{k0}^{-1} D_{k0})^{-1}(\sum D_{k0}' V_{k0}^{-1} f_{k0}),$$

here V_{k0} is the V_k evaluated at (β_0, α_0).

We may apply the easier multinormial χ^2 test for the multivariate binary vectors. Let the p-variate binary two samples be: $\{y'_1,...,y'_{n1}\}$ and $\{y''_1,...,y''_{n2}\}$, where $y_k^i=(y_{k1}^i,...,y_{kp}^i)'$. Since the y_{kj}^i s are binary variables, we define a function:

$$y_k^i \Rightarrow t = 1+\sum_{j=1}^{p} y_{kj}^i \, 2^{(j-1)}.$$

That is , each distinct y_k^i defines an integer number t uniquely. Thus, vectors y_k^i s can be expressed as *i.i.d.* multinormial distributed variables taking values from *1* to c, where c is the largest t computed by either sample, and $c \le 2^p$. We can transfer the samples $\{y_k^i\}$ as $\{z_1^i,...,z_c^i\}$, where z_j^i is the number of the vectors in the sample i maps to value j by the function. Let $p_j^i = pr$ (vectors from population i map to value j). We want to test whether the two samples are from the same population. Thus, the hypotheses are:

$$H_0: p_j' = p_j'' \text{ , for all } j=1,...,c. \quad vs \quad H_1: \text{ not all } p_j\text{s are equal.}$$

If we have large sample sizes $n1$ and $n2$, we can use the χ^2 test:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{c} \frac{(z_j^i - n_i p_j^i)^2}{n_i p_j^i}$$

which is approximately distributed by χ^2 $(2c-2)$ under the null hypothesis. Since we don't know the true p_j^is, we need to estimate them by $\hat{p}_j^i = z_j^i / n_i$ under H_0, $\hat{p}_j^i = (z_j'+z_j'') / (n_1+n_2)$. Those $c-1$ estimations cost us $c-1$ degrees of freedom. The test

statistic is:

$$\chi_0^2 = \sum_{i=1}^{2}\sum_{j=1}^{c}\frac{[z_j^{\,i} - n_i((z_j'+z_j'')/(n_1+n_2))]^2}{n_i\,[(z_j'+z_j'')/(n_1+n_2)]} \sim \chi^2(c-1).$$

We reject H_0 if $\chi_0^2 > \chi_\alpha^2(c-1)$.

This method can be extended to a problem with any other categorical variables by introducing $t-1$ binary variables to represent a categorical variable with t categories. In chapter seven, we will check the significance level and the power of this method by simulation and apply it to a data set.

CHAPTER VI

MODELS FOR MIXED DISCRETE AND CONTINUOUS DATA

We now consider the two sample comparisons in which the samples are from populations with some variables which are discretely distributed and some which are continuously distributed as regression problems.

When the responses are continuous and assumed to be approximately normal, there is a general class of linear models that are suitable for analysis. However, when the responses are not near normal, there are few analogues to perform the analysis. Recently, the generalized estimating equation (GEE) approach by Zeger and Liang (1986) has been studied and applied extensively. The generalized estimating equation method uses a working generalized linear model for the marginal distribution of the outcome variables. Instead of specifying the joint distribution of the outcome variables, the GEE method introduces estimating equations that give consistent estimates of the regression parameters and of their variances under weak assumptions about the joint distribution.

Consider the combined sample of the two samples: $y_1,...,y_K$ where $K=n+m$, and $y_k'=(y_{k1},...,y_{kp})$ is a p-variate mixed discrete and continuous response vector. x_k is the covariate, and $x_k=(1,1)$ if y_k is from the first sample, and $x_k=(1,0)$ if y_k is from the second sample. $\beta=(\beta_0\beta_1)'$ is a $2\times p$ matrix where $\beta_0=(\beta_{01},...,\beta_{0p})$, $\beta_1=(\beta_{11},...,\beta_{1p})$. The null hypothesis H_0: the two populations have the same mean, is equivalent to H_0: $\beta_1=0$ if the function $\mu_k(.)$ is one to one. Assume the marginal density of y_{ij} is:

$$f(\ y_{ij}\ ) = exp\ [(y_{ij}\theta_{ij} - a(\theta_{ij}\ ) + b(y_{ij}\ ))\phi],$$

where $\theta_{ij}=h(\eta_{ij})$, $\eta_{ij}=x_{ij}\beta$. Here, $h(.)$ is called the link function.

Under independent working assumptions, which we call the independent estimating equation (IEE) method, $\hat{\beta}_I$ is the solution of the score equation:

$$U_I(\beta)=\Sigma(x_i^T\Delta_i s_i)=0,$$

where $\Delta_i=diag(d\theta_{ij}/d\eta_{ij})$ is a $p\times p$ matrix. $s_i=y_i-E(y_{ij})$ is a $p\times 1$ vector for the ith observation defined for each i, while the $p\times p$ diagonal matrix $A_i=diag(var(y_{ij}))$ is the covariance matrix of y_i. By calculation, we get:

$$E(y_{ij})=a'(\theta_{ij}) \text{ and } var(y_{ij})=a''(\theta_{ij})\ \phi.$$

Theorem 1: The estimate $\hat{\beta}_I$ *of* β *is consistent and* $K^{\frac{1}{2}}(\hat{\beta}_I-\beta)$ *is asymptotically multivariate Gaussian as* $K \to \infty$ *with mean=0 and covariance matrix* V_1 *given by:*

$$V_1 = \lim_{k\to\infty} K(\sum_{i=1}^{K} x_i^T \Delta_i A_i \Delta_{ix_i})^{-1}(\sum_{i=1}^{K} x_i^T \Delta_i \ cov(y_i)\Delta_i x_i)(\sum_{i=1}^{K} x_i^T \Delta_i A_i \Delta_{ix_i})^{-1}$$
$$= \lim_{k\to\infty} K[H_1(\beta)]^{-1}H_2(\beta)[H_1(\beta)]^{-1}$$

$$\hat{var}\ (\hat{\beta}_I) = [H_1(\hat{\beta}_I)]^{-1}([\sum_{i=1}^{K} x_i^T \Delta_i\ s_i\ s_i^T\ \Delta_i\ x_i]_{\hat{\beta}_I})\ [H_i(\hat{\beta}_I)]^{-1}.$$

Under H_0: $\beta 1=0$, which is equivalent to test the identity of the means, $K^{1/2}\hat{\beta}_1 \sim MVN(0, \Sigma_{\beta 1})$, where $\Sigma_{\beta 1}$ is the submatrix, which is related to $\hat{\beta}_1$, of $\hat{var}(\hat{\beta}_I)$.

If we take the correlation into account, we can increase the efficiency of the test. Let $R(\alpha)$ be a $p\times p$ symmetric matrix which fulfills the requirements of being a correlation matrix, α is an $s\times 1$ vector which fully characterizes $R(\alpha)$, which is called a "working" correlation matrix. Define $V_i=A_i^{1/2}R(\alpha)A_i^{1/2}/\phi$, $V_i=cov(Y_i)$ if $R(\alpha)$ is indeed the true correlation matrix for the Y'_is. The generalized estimate equations are:

$$\sum_{i=1}^{K} D_i^T V_i^{-1} s_i = 0, \qquad (6.1)$$

where

$$D_i = \frac{d(a_i(\theta))}{d\beta} = A_i \Delta_i x_i ,$$

and

$$\Delta_i = diag(d\theta_{ij}/d\eta_{ij})$$

are $p\times p$ matrices.

A $K^{1/2}$-consistent estimate of α when β and ϕ are known as $K^{1/2}(\hat{\alpha}-\alpha) = O_p(1)$, given $\hat{\alpha} = \hat{\alpha}(y, \beta, \phi)$. Let $\hat{\phi} = \hat{\phi}(y, \beta)$, a consistent estimate when β is known. Then the equation *(6.1)* becomes

$$\sum_{i=1}^{K} U_i\ [\beta,\ \hat{\alpha}(\beta,\ \hat{\phi}(\beta))] = 0 \qquad (6.2)$$

where $V_i=(\beta, \alpha)=D_i^T V_i^{-1} s_i$, $\hat{\beta}_G$ is defined to be the solution of equation *(6.2)*. For large sample sizes, the following theorem is true.

Theorem 2: Under mild regularity conditions and given that:

(a) $\hat{\alpha}$ *is* $K^{1/2}$*-consistent given* β *and* ϕ*;*

(b) $\hat{\phi}$ *is* $K^{1/2}$*-consistent given* β*; and*

(c) $|\frac{\partial \hat{\alpha}(\beta, \phi)}{\partial \phi}| \leq H(y, \beta)$ *which is* $O_p(1)$*,*

Then $K^{1/2}(\hat{\beta}_G - \beta)$ *is asymptotically multivariate Gaussian with mean 0 and covariance* V_G *given by:*

$$V_G = \lim_{k \to \infty} K(\sum_{i=1}^{K} D_i^T V_i^{-1} D_i)^{-1} [\sum_{i=1}^{K} D_i^T V_i^{-1} cov(y_i) V_i^{-1} D_i] (\sum_{i=1}^{K} D_i^T V_i^{-1} D_i)^{-1}.$$

The variance estimate $\hat{V}_G$ of $\hat{\beta}_G$ can be obtained by replacing $cov(y_i)$ by $s_i s_i^T$ (note: $s_i = y_i - a'_i(\theta)$) and replacing β, α, ϕ by their estimates, in the expression V_G. The consistency of $\hat{\beta}_G$ and $\hat{V}_G$ depends on the specification of the mean, not on the correct choice of $R(\alpha)$. Under H_0: $\beta 1 = 0$, which is equivalent to test the identity of the means, $K^{1/2} \hat{\beta}_1 \sim MVN(0, \Sigma_{\beta 1})$, where $\Sigma_{\beta 1}$ is the submatrix , which related to $\hat{\beta}_1$, of $\hat{var}(\hat{\beta}_G)$.

There is an iterative procedure to compute $\hat{\beta}_G$

$$\hat{\beta}_{j+1} = \hat{\beta}_j - [\sum_{i=1}^{K} D_i^T(\hat{\beta}_j) \tilde{V}_i^{-1}(\hat{\beta}_j) D_j(\hat{\beta}_j)]^{-1} [\sum_{i=1}^{K} D_i^T(\hat{\beta}_j) \tilde{V}_i^{-1}(\hat{\beta}_j) s_j(\hat{\beta}_j)]$$

where $\tilde{V}_i(\beta) = V_i [\beta, \hat{\alpha}(\beta, \hat{\phi}(\beta))]$. Define $D = (D_1^T, \ldots, D_K^T)$, $s = (s_1^T, \ldots, s_K^T)$, $\tilde{V}$ be $pK \times pK$ block diagonal matrix with V''_is as the diagonal elements.

Estimators of α and ϕ: at a given iteration, α and ϕ can be estimated by Pearson residuals:

$$\hat{r}_{it} = [y_{it} - a'(\hat{\theta}_{it})] / [a''(\hat{\theta}_{it})]^{1/2},$$

where $\hat{\theta}_{it}$ depends on the current value for β. We estimate ϕ by

$$\hat{\phi}^{-1} = \sum_{i=1}^{K} \sum_{j=1}^{p} \hat{r}_{ij}^2 / (N - p),$$

where $N = Kp$. Since α depends on $R(\alpha)$, we need to estimate $R(\alpha)$ before we can estimate α. The general approach is to estimate α by a simple function of

$$\hat{R}_{uv} = \sum_{i=1}^{K} (\hat{r}_{iu}\ \hat{r}_{iv}) \ / \ (N - p).$$

In the GEE method, we need to correctly specify the link function which may be difficult if there are both continuous and discrete variables in the samples. On the other hand, the GEE method seems too general and complicated for two sample comparisons. There are a few alternative ways to study the mixture continuous and discrete variables: We may approximate discrete distributions by some continuous distributions. We may approximate the continuous distributions by truncating them to discrete distributions. We think it is better to study the discrete part and the continuous part separately. We take the construction of the discrete and continuous variables into account. If the two kinds of variables are independent, we can write their joint distribution as the product of the two multivariate distributions. Otherwise, there are two reasonable approaches to construct them: The binary variables depend on the continuous variables. It arises naturally in many situations, for example, blood pressure and cholesterol may affect the chances of a heart attack. Continuous variables depend on discrete variables, for example, a person's physical measurements such as height and weight may depend on his/her sex and race.

If we test continuous and discrete variables separately, in most of the cases, we have to assume the independence between continuous and discrete variables because it is hard to know the conditional distributions between them.

We want to find a simple and practical way to compare two samples with mixed continuous and discrete variables without making too many assumptions. Regression analysis has become a natural choice for this purpose because there are few restrictions on independent variables in regression. Therefore, we could consider the mixed variables in the two samples as independent variables in regression and create a dependent variable to identify the two groups. Regression analyses procedures are available in any statistical software and are simple to use.

Let us first compare the means of two samples. Suppose $\{x_1, x_2, \ldots, x_{n+m}\}'$ is the combined sample of two k-variate samples, where $x_i=(x_{i1}, x_{i2},\ldots,x_{ik})$. We can generate a variable y such that $y_i=1$ if x_i belongs to the first sample, and $y_i=0$ if x_i is in the second sample. Consider the logistic regression model:

$$Pr(y_i=1) = F(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) .$$

Where $F(.)$ is the logistic distribution function, that is, $F(x)=1/(1+exp(-x))$.

We are interested in the joint significance of all the independent variables. The test statistic for the joint significance is : $-2Log(l) = -2\sum_{i=1}^{n+m} log(\hat{p}_i)/(n+m)$, where the estimate $\hat{p}_i$ of $p_i=Pr(y_i=1)$ is obtained by replacing the regression coefficients by their maximum likelihood estimates. Under the null hypothesis, H_0: $\beta_1=\beta_2=\ldots=\beta_k=0$ (which means the joint effects of the covariates are not significant.), $-2log(L) \sim \chi^2(k)$ approximately. We reject H_0 if $-2Log(L) > \chi^2_\alpha(k)$. Since the hypothesis H_0 here implies that the covariates $(x_{.1},x_{.2},\ldots,x_{.k})$ do not affect the values of y significantly, it makes sense to say that the means of the two samples are not statistically different under H_0.

Comments:

1. This method can be used to compare the means of k-variate samples with a mixture of any kind of variables.
2. This method works when the covariates $(x_{.1},x_{.2},\ldots,x_{.k})$ are mildly correlated because we are only interested in the simultaneous effects. If two covariates are highly correlated, we will just delete one of them. In fact, we should compute the pairwise correlation to drop some variables before we perform the logistic regression analysis.
3. The logistic regression is not likely to have a big R^2 value since the dependent variate y is generated independently. The $y's$ indicate which sample the covariates belong to.
4. The method could be used to compare the means of more than two samples with the mean of a controlled sample.
5. The method could be used on the comparisons of the covariance matrices by modifying the covariates as $(ss_{i11},ss_{i12},\ldots,ss_{i22},\ldots,ss_{ikk})$ where $s_{ijl} = (x_{ij}-\bar{x}_{.j})(x_{il}-\bar{x}_{.l})$. The generations of the s_{ijl}'s can be done easily with Stata software.

In the next chapter, we will give the simulation results on the significance level and power for both of the χ^2 method and the logistic regression methods. We will then have a complete statistical analysis using the Alzheimer's disease data.

CHAPTER VII
SIMULATIONS AND APPLICATIONS

The χ^2 test for multivariate binary variables in chapter five and the logistic regression method for testing the means or covariance matrices of mixed continuous and discrete variables in chapter six are examined in this chapter. Our goal is to analyze data regarding Alzheimer' disease, hence, we set u simulations with a situation similar to the data.

7.1 Simulations

We would like to know the significance level and power for these two methods. First, we generate a set of 4-dimensional multivariate binary variables using the method developed by Emrich and Piedmonte (1991), then we perform the χ^2 test and the logistic method on the data set. Next, we add two more continuous variables into the sampling data set, and perform the logistic regression test.

The procedure to find the significance level is:

1. We generate two random samples with different sizes. The variables in these samples follow the same distribution.
2. Perform the tests and save the value of the test statistic into a data set.
3. Repeat step 1 and 2, *100* times, then count the number of the test statistic values which are greater than the critical value $\chi^2_{0.05}(df)$. This count is, in fact, the number of disagreements based on the test and the true situations. Thus, the significance level is the count divided by *100*.
4. We could repeat the above steps a couple of times and get several significance levels.

The idea of finding the power is similar to the above procedure except for a change in step 1. To find the power, we want to generate two random samples with different distributions. The power is the count divided by *100*. Here the count is interpreted as the number of the agreements based on the test and on the truth.

Test the distributions of two samples with *4*-variate correlated binary variables: suppose (x_1, x_2, x_3, x_4) are the variables, and we generate two samples with sizes *181* and *191*. Define $p'_i=Pr(x'_i=1)$, $p''_i=Pr(x''_i=1)$, where x'_i , x''_i are the values of the *i*th variable from sample *1* and sample *2*, respectively. Let δ_{ij} be the correlation of the *i*th and the *j*th variables. The results are listed as the following:

a. $p'_1=p''_1=0.154$, $p'_2=p''_2=0.086$, $p'_3=p''_3=0.346$, $p'_4=p''_4=0.475$.

$\delta_{12}=0$, $\delta_{13}=-0.103$, $\delta_{14}=-0.017$, $\delta_{23}=-0.1$, $\delta_{24}=-0.12$, $\delta_{34}=0.05$.

After running the procedure for significance level 5 times by the χ^2-test and the logistic regression method, we get the significance levels:

χ^2-test	0.05	0.05	0.06	0.04	0.05
logistic	0.09	0.02	0.11	0.05	0.05

b. If we change probabilities in the second sample slightly, but keep the correlations, we find:

$p'_1=0.154$, $p'_2=0.086$, $p'_3=0.346$, $p'_4=0.475$

$p''_1=0.254$, $p''_2=0.076$, $p''_3=0.330$, $p''_4=0.475$.

$\delta_{12}=0$, $\delta_{13}=-0.103$, $\delta_{14}=-0.017$, $\delta_{23}=-0.1$, $\delta_{24}=-0.12$, $\delta_{34}=0.05$.

We will get the powers of those methods as:

χ^2-test	0.28	0.22	0.22	0.26	0.22
logistic	0.41	0.43	0.44	0.44	0.45

From these two tables, we see that the tests are quite significant, but not very powerful for small changes of the distributions.

c. From the following table, we see the power will improve greatly if we change the probabilities in the second sample.

$p'_1=0.154$, $p'_2=0.086$, $p'_3=0.346$, $p'_4=0.475$

$p''_1=0.3$, $p''_2=0.086$, $p''_3=0.346$, $p''_4=0.275$.

$\delta_{12}=0$, $\delta_{13}=-0.103$, $\delta_{14}=-0.017$, $\delta_{23}=-0.1$, $\delta_{24}=-0.12$, $\delta_{34}=0.05$.

The powers in this case are:

χ^2-test	0.96	0.92	0.96	0.86	0.96
logistic	1.00	0.99	0.96	0.99	0.99

From these three tables, it seems that the logistic regression method is slightly better than the χ^2 method. In the next three cases, we add two continuous variables into the samples, and consider only the logistic regression method.

d. Suppose $(x_1, x_2, x_3, x_4, y_1, y_2)$ are the variables in the samples, where (x_1, x_2, x_3, x_4) have exact the same distribution as in case (a), $y_1 \sim N(0,1)$, $y_2 \sim N(1,4)$, moreover, (x_1, x_2, x_3, x_4), y_1, y_2 are independent. We run the procedure for the significance levels 4 times:

Logistic	0.07	0.03	0.09	0.04

e. We keep the same distribution of the first sample, and the same distribution of (x_1, x_2, x_3, x_4) in the second sample. The distributions of y_1, y_2 are changed as the following and the powers are:

$y_1 \sim N(2,4)$, $y_2 \sim N(3,9)$	1.00	1.00
$y_1 \sim N(2,4)$, $y_2 \sim N(1,9)$	1.00	1.00
$y_1 \sim N(0,1)$, $y_2 \sim N(2,4)$	0.93	0.95

The first row in the table shows that if the means and the variances of both y_1, y_2 are different in the two samples, we reject the hypothesis of two samples having the same means 100 times out of 100 times. In the second row, we keep the same mean of y_2 in both samples. In the third row, the only difference between the two samples is that y_2 has different means. The result is still satisfactory.

f. Both samples have the same distribution. (x_1, x_2, x_3, x_4) have the same distribution as in case (a), $y_1 \sim EXP(1\ 2)$, $y_2 \sim N(1,4)$. The significance levels are:

Logistic regression	*0.05*	*0.07*	*0.09*

g. The first sample has the same distribution as in (f). In the second sample, (x_1, x_2, x_3, x_4) and y_2 have the same distribution as in the first sample. In the first row of the following table, y_1 in the second sample is distributed as $N(1,9)$. In the second row, y_1 is distributed as $EXP(1\ 3)$, and the powers are:

$y_1 \sim N(1,9)$	*0.81*	*0.79*	*0.71*
$y_1 \sim EXP(1\ 3)$	*0.78*	*0.70*	*0.79*

In the next two tables, we consider the situation in which y_1, and y_2 are correlated.

h. Suppose (x_1, x_2, x_3, x_4) in both samples have the same distribution as in case (a), $(y_1, y_2)'$ are multinormial with mean $(0,0)'$ and correlation matrix $\Sigma = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$ in both samples. The significance levels are:

Logistic	*0.01*	*0.03*	*0.06*	*0.04*	*0.07*

i. We change the mean vector of $(y_1, y_2)'$ to $(0,1)'$ in the second sample , and keep all other information the same as in case (h). The powers are:

Logistic	*1.00*	*1.00*	*1.00*

Although it is impossible to exam the methods in all the combinations, from the results of the above simulations, we believe we can make very reliable decisions by using these methods in the similar settings. It is appropriate to use these methods to analyze the Alzheimer's disease data.

7.2 Applications on the Alzheimer's Disease Data

The data set we are going to analyze was collected by the Texas Tech Health Science Center. It is a phone screen database on Alzheimer's disease. The original data set contains more than five hundred records, and each record contains more than eighty

variables. There are many missing values. We first deleted the variables or records with too many missing values, then we selected the variables we are interested in. After these preliminary operations, there are four hundred and sixty two records with twenty two variables remaining in the data set to be analyzed.

7.2.1 Data Descriptions

The variable which we use to identify the groups is the diagnostic variable: *dx*. It is a categorical variable with three levels, *dx=1* means the diagnosis is probable Alzheimer's disease; *dx=2* means diagnosis as possible Alzheimer's; and *dx=3* means other diagnosis. We can consider these three groups as three samples. By comparing these three samples, we hope to find some variables related to the disease.
The variables are listed in the following table (Table 7.1).

7.2.2 Statistical Analyses

We begin with a multinomial regression analysis to obtain some basic flavors of the data structure. The results are listed in Table 7.2.

When we compare samples *dx=1* and *dx=2* , we find the variables of head injury history, alcohol problems and drug study to be relatively more significant than the other variables. Although the variable of family history of memory problems does not appear to be very significant here, we want to include it for further study because people tend to assume that Alzheimer's is a genetic disease. The comparison of groups *dx=2* and *dx=3* suggests that the continuous variables of age, years of education, and years on medication are relatively significant. Since *200* out of *462* values of the years of education variable are missing, we have to give up the variable. Thus, we selected four binary variables: eadhx, alcohl, study and famhx, plus two continuous variables: age and yrsdur for further statistical investigations.

Table 7.1 Variable Descriptions

variable	label	type	value description
obs	observation number	integer	
study	drug study	binary	study=1: qualify for drug study.
age	age at screening	continuous	
sex	patient's gender	binary	sex=1: female.
educ	years of education	continuous	
hand	handedness	binary	hand=1: left handed
total	total scores of daily living test	integer	total$\in$(4,16), small is good
depr	problem of depression	binary	depr=1: yes
delus	delusions	binary	delus=1: yes
halluc	hallucinations	binary	halluc=1: yes
mothx	history of emotional disturbance	binary	mothx=1: yes
eadhx	history of head injury	binary	mothx=1: yes
sleep	problem with sleeping	binary	sleep=1: yes
alcohl	history of alcohol abuse	binary	alcohl=1: yes
drugs	history of drug abuse	binary	drugs=1: yes
la0g	language comprehen. problem	binary	laog=1: yes
famhx	hist. of memory problem in family	binary	famhx=1: yes
meds	hist. of severe medical conditions	codes	medical codes
sym1	first symptom	codes	medical codes
yrsdur	years of using medications	continuous	
cond	significant medical conditions	codes	medical codes

Table 7.2 Multinomial Regression Analysis

Multinomial regression				*Number of obs = 186*	
				chi2(34) = 88.02	
				Prob > chi2 = 0.0000	
Log Likelihood = -109.88579				*Pseudo R2 = 0.2860*	
dx	*Coef.*	*Std. Err.*	*z*	*P>|z|*	*[95% Conf. Interval]*
		dx=1			
sex	*.0062302*	*.381957*	*0.016*	*0.987*	*-.7423918* *.7548522*
hand	*.9739503*	*.7694733*	*1.266*	*0.206*	*-.5341896* *2.48209*
depr	*-.4086533*	*.3826778*	*-1.068*	*0.286*	*-1.158688* *.3413813*
delus	*.035545*	*.3847654*	*0.092*	*0.926*	*-.7185814* *.7896714*
halluc	*.1989002*	*.4792495*	*0.415*	*0.678*	*-.7404117* *1.138212*
mothx	*-.9242813*	*1.331225*	*-0.694*	*0.487*	*-3.533435* *1.684872*
eadhx	*-2.942516*	*.8078878*	*-3.642*	*0.000*	*-4.525947* *-1.359086*
sleep	*.4525199*	*.4223096*	*1.072*	*0.284*	*-.3751916* *1.280231*
alcohl	*-2.074928*	*.9138164*	*-2.271*	*0.023*	*-3.865976* *-.2838813*
drugs	*-46.09925*	.	.	.	. .
la0g	*-.673064*	*.4509499*	*-1.493*	*0.136*	*-1.55691* *.2107816*
famhx	*-.5236103*	*.378315*	*-1.384*	*0.166*	*-1.265094* *.2178734*
study	*.7158623*	*.3874543*	*1.848*	*0.065*	*-.0435342* *1.475259*
age	*-.03562*	*.0249966*	*-1.425*	*0.154*	*-.0846125* *.0133725*
educ	*-.0083808*	*.0281283*	*-0.298*	*0.766*	*-.0635112* *.0467496*
total	*.0229239*	*.0726392*	*0.316*	*0.752*	*-.1194463* *.1652942*
yrsdur	*.0020098*	*.057359*	*0.035*	*0.972*	*-.1104118* *.1144313*
_cons	*3.036969*	*2.033939*	*1.493*	*0.135*	*-.9494784* *7.023416*

Table 7.2 Continued

dx	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
			dx=3			
sex	-2.483205	1.89367	-1.311	0.190	-6.19473	1.22832
hand	-43.60364	.	.	.	.	.
depr	2.481919	1.717776	1.445	0.149	-.8848589	5.848698
delus	-1.898772	1.896415	-1.001	0.317	-5.615677	1.818133
halluc	-.0810019	3.119235	-0.026	0.979	-6.19459	6.032586
mothx	-2.450484	2.997904	-0.817	0.414	-8.326267	3.4253
eadhx	-4.142503	3.031537	-1.366	0.172	-10.08421	1.7992
sleep	-.0446563	1.910008	-0.023	0.981	-3.788202	3.69889
alcohl	1.288457	2.754188	0.468	0.640	-4.109653	6.686567
drugs	-35.78311	.	.	.	.	.
la0g	2.854184	2.543891	1.122	0.262	-2.131752	7.84012
famhx	-1.82556	1.954876	-0.934	0.350	-5.657046	2.005926
study	-1.628805	1.754883	-0.928	0.353	-5.068313	1.810703
age	-.3708888	.167339	-2.216	0.027	-.6988671	-.0429104
educ	-1.084987	.575298	-1.886	0.059	-2.21255	.0425767
total	-1.913934	1.171971	-1.633	0.102	-4.210954	.3830865
yrsdur	.6357236	.3811333	1.668	0.095	-.111284	1.382731
_cons	40.80538	18.45198	2.211	0.027	4.640152	76.9706
(Outcome dx==2 is the comparison group)						

Let us study the continuous variables first. We would like to know whether we could assume they are normal or not, so we perform the Shapiro-Wilk normality test.

Shapiro-Wilk W test for age and years on medications					
Variable	*Obs*	*W*	*V*	*z*	*Pr > z*
age	*459*	*0.96701*	*10.276*	*5.579*	*0.00000*
yrsdur	*402*	*0.82849*	*47.427*	*9.185*	*0.00000*

The test showed that both of these are nearly normal. We hope to transform them to approximately normally distributed variables, thus, we perform the Box-Cox transformation.

Transform: (variable^L-1) /L ⇒ new variable		
variable	*L*	*new variable*
age	*3.3913*	*bcage*
yrdsdur	*0.0764*	*bcyrs*

We then perform the normality test on the new variables, and obtain:

Shapiro-Wilk W test for normal data					
Variable	*Obs*	*W*	*V*	*z*	*Pr > z*
bcage	*459*	*0.99495*	*1.574*	*1.086*	*0.13881*
bcyrs	*402*	*0.99392*	*1.682*	*1.238*	*0.10789*

We can not reject the normal assumption. Therefore, we could consider these two new variables to be normally distributed. From now on we will study the variables bcage and bcyrs instead of age and yrsdur.

We compute the correlations between the six selected variables. We first compute the correlation matrix for the discrete variables, Table 7.3, then compute the correlations for all the variables, Table 7.4.

Table 7.3 Correlation Matrix of the Discrete Variables

Correlation of: eadhx alcohl study famhx (obs=372) (symmetry)				
	eadhx	*alcohl*	*study*	*famhx*
eadhx	*1.0000*			
alcohl	*0.0091*	*1.0000*		
study	*-0.1031*	*-0.0998*	*1.0000*	
famhx	*-0.0165*	*-0.1168*	*0.0477*	*1.0000*

Table 7.4 Correlation Matrix of All Selected Variables

Correlations of: bcage bcyrs eadhx alcohl study famhx (obs=332)						
	bcage	*bcyrs*	*eadhx*	*alcohl*	*study*	*famhx*
bcage	*1.0000*					
bcyrs	*0.0307*	*1.0000*				
eadhx	*-0.0142*	*0.1013*	*1.0000*			
alcohl	*-0.1678*	*0.0078*	*0.0313*	*1.0000*		
study	*-0.0370*	*-0.2064*	*-0.1101*	*-0.0900*	*1.0000*	
famhx	*-0.0760*	*-0.0038*	*-0.0431*	*-0.1241*	*0.0431*	*1.0000*

Some values in these two tables are different because these two tables are computed by using a different number of observations, and more values are missing in the set which contained more variables. The variables are not highly correlated.

7.2.3 χ^2 and Logistic Tests for Binary Variables

Samples tested in this section are composed of four binary variables, namely eadhx, alcohl, study, and famhx.

a. The sample with dx=1 is randomly divided into two sub-samples. χ^2 and logistic tests are conducted to find out whether the two sub-samples are the same. In the

χ^2 test, there are *13* out of *16* cells are occupied by the data. Hence, *df=12*. The χ^2 value is computed from the data:

chi(12)=11.253$<\chi^2_{0.05}(12)=21.026$.

As a result, the null hypothesis of the same distribution cannot be rejected at a five percent significance level. This is consistent with the origin of the two sub-samples. For they come from the same sample. In the logistic test,

chi2(4) =9.36$<\chi^2_{0.05}(4)=9.488$.

Again, the null hypothesis of the same mean cannot be rejected at the five percent level.

b. The sample with dx=2 is randomly divided into two sub-samples.

In the χ^2 test, *chi(12)=6.1498* $<\chi^2_{0.05}(12)=21.026$.

In the logistic test, *chi2(4)=0.89*$<\chi^2_{0.05}(4)=9.488$.

Neither test rejects the null hypothesis.

c. The sample *with dx=3* is randomly divided into two sub-samples, and the results follow.

In the χ^2 test, *chi(8)=11.7319*$<\chi^2_{0.05}(8)=15.507$.

In the logistic test, *chi2(3) =7.13*$<\chi^2_{0.05}(3)=7.815$.

Because the two sub-samples are from the same distribution, the two tests above show consistent results.

d. In contrast to the tests in a, b, and c, the two samples compared in this section are from different distributions. One of the samples is formed with *dx=1*, and the other with *dx=2*.

The χ^2 test: *chi(13)=47.667*$>\chi^2_{0.05}(13)=22.362$.

The logistic test: *chi2(4) =38.77*$>\chi^2_{0.05}(4)=9.488$.

We reject the null hypothesis in both tests, and conclude that the two samples are significantly different.

e. Finally, the two samples *with dx=1* and *dx=3*, respectively, are compared. The χ^2 test may not be appropriate since the difference of the sample sizes is very large. However, the logistic test is conducted.

$chi2(4) = 37.39 > \chi^2_{0.05}(4) = 9.488$.

Consequently, the null hypothesis is rejected in favor of the alternative hypothesis. We can conclude with a 95 percent confidence level that the two samples are significantly different.

7.2.4 Logistic Tests for Samples with Both Discrete and Continuous Variables

The sampling data in the binary tests above are extended to include six variables, both discrete and continuous. The two additional continuous variables are bcage and bcyrs. The logistic regression method is used to test the differences of sub-samples.

a. The sample with $dx=1$ is randomly divided into two sub-samples. The two sub-samples are then tested using the logistic test. The test statistic is provided below:

$chi2(5) = 10.08 < \chi^2_{0.05}(5) = 11.07$.

Therefore, the null hypothesis of the same mean is rejected at 5 percent level.

b. Similarly, the sample with $dx=2$ is randomly divided into two sub-samples. In the logistic test, $chi2(6) = 1.81 < \chi^2_{0.05}(6) = 12.592$. Hence, the null hypothesis is not rejected at 5 percent level.

c. Also, the sample with $dx=3$ is randomly separated into two sub-samples. The comparisons of the sub-samples reveal the consistent result with the origin of the two sub-samples. In the logistic test, $chi2(4) = 7.26 < \chi^2_{0.05}(4) = 7.815$.

d. The logistic regression results are shown in the following table for the two different samples with $dx=1$ and $dx=2$, respectively.

Logit Estimates	*Number of obs = 311*
chi2(6) = 46.90	*Prob>chi2 =0.0000*
Log Likelihood = -190.36585	*Pseudo R2 = 0.1097*

From the logistic test, $chi2(6) = 46.90 > \chi^2_{0.05}(6) = 12.592$. We reject the null hypothesis, and conclude that the two samples are significantly different.

e. Compare the samples $dx=1$ and $dx=3$. In the logistic test, $chi2(6) = 37.89 > \chi^2_{0.05}(6) = 12.592$.

Consequently, we have a good reason to believe that the two samples are significantly different.

All the results above in the two sample comparisons are as expected. For the two sub-samples that come from the same sample, as shown in a, b, c of both binary and mixed data, the test statistics are all smaller than the critical values. Consequently, all the null hypotheses are not rejected at 5 percent significance level. On the other hand, tests for the samples from different dx values reject all the null hypotheses at 5 percent level. The methods utilized in this study seem to work well in this application.

REFERENCES

Albert, Paul S. and McShane, Lisa M. (1995): *A Generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data,* Biometrics 51, pp. 627-638.

Andersen, Per Kragh and Rønn, Birgitte B. (1995): *A Nonparametric Test for Comparing Two Samples Where All Observations are Eight Left- or Right-Censored,* Biometrics 51, pp. 323-329.

Barnhart, Huiman X. and Sampson, Allan R. (1995): *Multiple Population Models for Multivariate Random Length Data-with Applications in Clinical Trials,* Biometrics 51, pp. 195-204.

Buhrman, J. M. and Ruymgaart, F. H. (1981): *An Application of Linearization in Nonparametric Multivariate Analysis,* Sankhy$\bar{a}$: The Indian Journal of Statistics, Vol 43, Series A, Pt. 1, pp. 52-66.

Emrich, Lawrence J. and Piedmonte, Marion R. (1991): *A Method for Generating High-Dimensional Multivariate Binary Variates,* The American Statistician, Vol. 45, No. 4, pp. 302-304.

Fill, James Allen and Johnstone, Iain (1984): *On Projection Pursuit Measures of Multivariate Location and Dispersion, The Annals of Statistics,* Vol. 12, No. 1, pp. 127-141.

Fitzmaurice, Garrett M. (1995): *A Caveat Concerning Independence Estimating Equations With Multivariate Binary Data,* Biometrics 51, pp. 309-317.

Friedman, Jerome H. and Rafsky, Lawrence C. (1979): *Multivariate Generalizations of the Wald-Wolfwitz and Smirnov Two-Sample Tests,* The Annals Statistics, Vol. 7, No. 4, pp. 697-717.

George, E. Olusegun and Bowman, Dale (1995): *A Full Likelihood Procedure for Analyzing Exchangeable Binary Data,* Biometrics 51, pp. 512-523.

Have, Thomas R. Ten And Becker, Mark P. (1995): *Multivariate Contingency Tables and the Analysis of Exchangeability,* Biometrics 51, pp. 1001-1016.

Hettmansperger, Thomas P. (1991): *Statistical Inference Based on Ranks,* Krieger Publishing Company, Florida.

Hogg, R. V. and Graig, A. T. (1978): *Introduction to Mathematical Statistics,* 4th ed. Macmillan, New York.

Krzanowski, W. J. (1975): *Discrimination and Classification Using Both Binary and Continuous Variables,* Journal of the American Statistical Association, Vol 70, No. 352.

Krzanowski, W. J. (1980): *Mixtures of Continuous and Categorical Variables in Discriminant Analysis,* Biometrics 36, pp. 493-499.

Krzanowski, W. J. (1982): *Mixtures of Continuous and Categorical Variables in Discriminant Analysis: A Hypothesis-Testing Approach,* Biometrics 38, pp. 991-1002.

Krzanowski, W. J. and Marriott, F. H. C. (1994): *Multivariate Analysis, Part 1, Distributions, ordination and inference,* John Wiley & Sons Inc. New York.

Lauritzen, S. L. and Wermuth, N. (1989): *Graphical Models for Associations Between Variables, Some of Which Are Qualitative and Some Quantitative,* The Annals of Statistics, Vol. 17, No. 1, pp. 31-57.

Liang, Kung-Yee and Zeger, Scott L. (1986): *Longitudinal Data Analysis Using Generalized Linear Models,* Biometrika 73, 1. 13-22.

Liang, Kung-Yee and Zeger, Scott L. and Qaqish, Bahjat (1992): *Multivariate Regression Analyses for Categorical Data,* J. R. Statist. Soc. B. 54, No. 1. 3-40.

Link, William A. and Sauer, John R. (1995): *Estimation and Confidence Intervals for Empirical Mixing Distributions,* Biometrics 51, 810-821.

Lipsitz, Stuart R. and Fitzmaurice, Garret M. and Sleeper, Lynn and Zhao, L. P. (1995): *Estimation Methods for the Joint Distribution of Repeated Binary Observations,* Biometrics 51, pp. 562-570.

Little, Roderick J. A. and Schluchter, Mark D. (1985): *Maximum Likelihood Estimation For Mixed Continuous and Categorical Data With Missing Values,* Biometrika, 72, 3, pp. 497-512.

Montgomery, Douglas C. (1991): *Design and Analysis of Experiments,* 3rd Ed. John Wiley & Sons.

Moran, M. A. and Murphy, B. J. (1979): *A Closer Look at two Alternative Methods of Statistical Discrimination,* Applied Statistics, 28, No. 3, pp. 223-232.

Puri, M. L. and Sen, P. K. (1971): *Nonparametric Methods in Multivariate Analysis*, John Wiley & Sons, New York.

Schatzoff, (1966): *Exact distribution of Wilks 's likelihood ratio criterion*, Biometrika, 53, pp. 347-358.

Vlachonikolis, I. G. and Marriott, F. H. C. (1982): *Discrimination with Mixed Binary and Continuous Data,* Appl. Statist. 31, No. 1, 23-31.

Zeger, Scott L. and Liang, Kung-Yee (1986): *Longitudinal Data Analysis for Discrete and Continuous Outcomes,* Biometrics 42, pp. 121-130.

Zeger, Scott L. and Liang, Kung-Yee and Albert, Paul S. (1988): *Models for Longitudinal Data: A Generalized Estimating Equation Approach,* Biometrics 44, pp. 1049-1060.

Zhao, Lue Ping and Prentice, Ross L. (1990): *Correlated Binary Regression Using a Quadratic Exponential Model,* Biometrika 77, pp. 642-648.