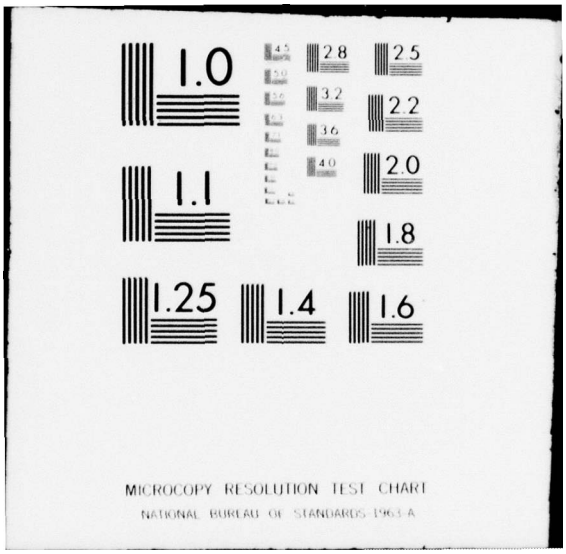
AD-A059 908 MISSOURI UNIV-COLUMBIA DEPT OF STATISTICS F/G 12/1
IDFNTIFICATION PROBLEMS IN THE EXPONFNTIAL FAMILY.(U)
AUG 78 R P KELLEY N00014-75-C-0443
UNCLASSIFIED 78-2 NL

1 OF 1
AD
A059908

END
DATE
FILMED
12-78
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

(14) 78-2

(9) Interim technical rpt.,

LEVEL II

(12)

(6) IDENTIFICATION PROBLEMS IN THE EXPONENTIAL FAMILY

(10) Robert Patrick/Kelley*

(11) Aug 78

ABSTRACT (12) 68 p.

We consider identification problems for members of the exponential family, applying the results to the density of the logistic model for life table data.

For this model we want to know the following:

1. When are the maximum likelihood estimates for the model parameters unique;

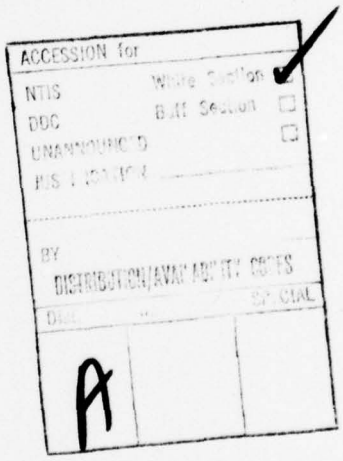
2. What type of inferences may be made if the maximum likelihood estimates are not unique?

Noting the form of the likelihood for the logistic model, question 1 is considered for the exponential family. We obtain an answer for question 1 in this context. Applying this answer to the logistic model, we find that a unique maximum likelihood estimate exists if and only if the density is in one to one correspondence with its parameter space.

To answer question 2, we consider members of the exponential family where the density is not in one to one correspondence with its parameter space. As a guiding example of such a density, we consider the normal linear

(15) * Supported in part by ONR/N00014-75-C-0443

model of less than full rank, discussing the concepts of estimable function and testable hypothesis, which have been developed for this particular case. We then show that the concept of uniform identifiability is a generalization of the concept of an estimable function. Further, through the idea of an identifiable set, we extend the concept of a testable hypothesis. We then apply the resulting theory to the logistic model.



CONTENTS

1. INTRODUCTION

Thompson (1976) introduced a logistic model for covariate effect in the analysis of grouped life times. Maximum likelihood is proposed as a method of estimating the parameters of the model. It is noted that the likelihood function is both concave and differentiable everywhere on the parameter space; thus, a point is a global maximum if and only if it is a solution to the likelihood equations. However, in a numerical example, the likelihood equations are linearly dependent and, to obtain a unique solution, a constraint must be imposed on the parameters. This causes a problem in the application of the logistic model; based on the same data two statisticians might obtain different maximum likelihood estimates for the parameters. One might be able to correct the problem of non-uniqueness by a reparameterization of the model, but, in doing so, the physical meaning associated with the parameters might be lost; thus, it is desirable to know the following:

1. When are the maximum likelihood estimates for the parameter of the model unique;

2. What type of inference may be made if the maximum likelihood estimates are not unique?

Noting the form of the likelihood for the logistic model given in Thompson (1976), we will consider questions

1 and 2 for the exponential family and use the logistic model as an example to illustrate the resulting theory. We first consider background material for the exponential family.

2. SOME PROPERTIES OF THE EXPONENTIAL FAMILY

Let μ be a σ-finite measure on E^p and

$$p(y;\alpha) = \exp(\alpha^T y - \phi(\alpha)) \tag{1}$$

be a probability density function with respect to μ. Families of form (1) are said to be exponential families with natural parameter α, see Lehmann (1959).

Now

$$\int p(y;\alpha)d\mu(y) = 1 ;$$

thus,

$$\phi(\alpha) = \ln \int \exp(\alpha^T y)d\mu(y) , \tag{2}$$

or $\phi(\cdot)$ is the log moment generating function of μ.

Let $\Lambda = \{\alpha | \phi(\alpha) < \infty\}$, then Λ is said to be the natural parameter space of (1). Lehmann shows that Λ is a convex set. For α^0 in Λ^0, the interior of Λ, the moments of (1) can be obtained from (2) by differentiating under the integral sign. In particular, if $Y = (Y_1,\ldots,Y_p)^T$ has density $p(y;\alpha^0)$ then

$$E\ Y_i = \left.\frac{\partial \phi}{\partial \alpha_i}\right|_{\alpha=\alpha_0} ,$$

$$\text{var}\ Y_i = \left.\frac{\partial^2 \phi}{\partial \alpha_i^2}\right|_{\alpha=\alpha_0} ,$$

and

$$\operatorname{cov}(Y_i, Y_j) = \left.\frac{\partial^2 \phi}{\partial \alpha_i \partial \alpha_j}\right|_{\alpha=\alpha_0} .$$

Let $\Sigma(\alpha)$ be the covariance matrix of Y evaluated at α. Then

$$\Sigma(\cdot) = \left(\frac{\partial^2 \phi}{\partial \alpha_i \partial \alpha_j}\right) .$$

A basic property of the exponential family is that the range of Y does not depend on the choice of parameter.

Lemma 1. The support of μ is the support of P_α for all α.

Proof. Let

$$P_\alpha(K) = \int_K p(y;\alpha)\, d\mu(y) .$$

$P_\alpha(K) = 0$ implies $\int_K \exp(\alpha^T y)\, d\mu = 0$; however, $\exp(\alpha^T y) > 0$ for all y. Therefore, $P_\alpha(K) = 0$ if and only if $\mu(K) = 0$.

At this point we need two preliminary results having indirect implications for the exponential family.

Lemma 2. Let Σ be a positive semi-definite matrix, then $\Sigma g = 0$ if and only if $g^T \Sigma g = 0$.

Proof. Σ may be represented as

$$P^T \Delta P ,$$

where P is an orthogonal matrix and Δ is a diagonal matrix whose entries, δ_i, are the eigenvalues of Σ; thus,

$$g^T \Sigma g = g^T P^T \Delta P g = (Pg)^T \Delta (Pg) .$$

Hence, letting $Pg = V$, we have

$$g^T \Sigma g = \sum_{j=1}^{p} \delta_j v_j^2 .$$

Let $g^T \Sigma g = 0$. Now $\delta_j \geq 0$, $j = 1,\ldots,p$, so that $\delta_j v_j = 0$.

$$\Sigma g = P^T \Delta P g = P^T \Delta V = P^T (\delta_1 v_1 \ldots \delta_p v_p)^T = 0 .$$

This proves the "if" part. The converse is obvious.

We will denote the column space of a matrix M (the set of a linear combination of the columns of M) by $Col(M)$ and the rank of M (the number of linearly independent rows in M) by $Rank(M)$.

Lemma 3. Let Y be a random vector with covariance Σ. Writing $L = \bigcup_{i=1}^{k} \{y | b_i^T y = c_i\}$ for the smallest linear manifold containing the support of Y, then

$$Col(\Sigma) = \bigcup_{i=1}^{k} \{y | b_i^T y = 0\} .$$

Proof. Let S be the support of Y and

$$L^* = \bigcup_{i=1}^{k} \{y | b_i^T y = 0\} .$$

We will show that g is perpendicular to L^* if and only if g is perpendicular to $Col(\Sigma)$; thus, due to the uniqueness of the orthogonal complement of a subspace we will have the result.

Let g be perpendicular to L^* . Now, $L = \ell^\circ + L^*$ where ℓ° is a fixed but arbitrary element of L ; thus, $g^T\ell = g^T\ell^\circ$ is constant for all ℓ in L . $g^T\ell$ is constant for all ℓ in L implies $g^T y$ is constant for all y in S ; thus,

$$Var(g^T Y) = g^T \Sigma g = 0$$

and, by Lemma 2, g is perpendicular to $Col(\Sigma)$.

Let g be perpendicular to $Col(\Sigma)$ then $g^T \Sigma g = Var(g^T Y) = 0$. $Var(g^T Y) = 0$ implies there exists a constant c such that $g^T Y = c$ with probability one. By definition, S is the smallest closed set which contains Y with probability one; thus, since $\{g^T y = c\}$ is closed, $S \subseteq \{g^T y = c\}$. Let $S^* = \{y_1 - y_2 \mid y_1, y_2 \in S\}$, then $g^T s = 0$ for all s in S^* . Now, L^* is the set of all linear combinations of elements of S^* ; thus, $g^T \ell^* = 0$ for all ℓ^* in L^* .

A result like Lemma 3 is stated in Jennrich and Moore (1975).

Corollary 1. $\text{Rank}(\Sigma) = p - \text{Rank}((b_1,\ldots,b_k)^T)$.

Corollary 2. $\text{Rank}(\Sigma) < p$ if and only if there exists a vector b such that b^TY is almost surely constant.

Proof. $\text{Rank}(\Sigma) < p$ if and only if $\text{Rank}((b_1,\ldots,b_k)^T) \geq 1$.

Corollary 3. Given a density of form (1) with covariance matrix $\Sigma(\alpha)$, let A be a matrix whose columns form an orthonormal basis for the column space of $\Sigma(\alpha)$. A can be chosen to be independent of the parameters α . Hence the rank and the singularity of $\Sigma(\alpha)$ does not vary with α .

Proof. Apply Lemma 3 then Lemma 1.

We now consider the convexity of $\phi(\cdot)$ on Λ . Results on this topic may also be found in Berk (1972).

Theorem 1. $\phi(\cdot)$ is convex on Λ ; furthermore, $\phi(\cdot)$ is strictly convex on Λ if and only if the covariance matrix, $\Sigma(\alpha)$, of $(Y_1,\ldots,Y_p)^T$ is full rank.

Proof. Let α,α^* be in Λ then for $0 < \lambda < 1$

$$\exp \phi(\lambda\alpha + (1-\lambda)\alpha^*) = \int[\exp(\alpha^T y)]^{\lambda}[\exp(\alpha^{*T} y)]^{1-\lambda} d\mu(y)$$

$$\leq [\int \exp(\alpha^T y)\, d\mu(y)]^{\lambda}[\int \exp(\alpha^{*T} y)\, d\mu(y)]^{1-\lambda}$$

with equality holding if and only if $(\alpha-\alpha^*)^T Y$ is almost surely (μ) constant (see Royden (1968), page 113). Thus, from (2), $\phi(\cdot)$ is convex on Λ.

Now, if $\phi(\cdot)$ is not strictly convex then there exists $b = \alpha - \alpha^*$ with α, α^* in Λ such that $b^T Y$ is almost surely (μ) constant. Conversely, suppose there exists a nonzero vector b such that $b^T Y$ is almost surely (μ) constant. Pick α^* in Λ and let $\alpha = \alpha^* + b$. From (2), $\phi(\alpha) < \infty$ and α is in Λ. Now, $(\alpha - \alpha^*)^T Y$ is almost surely (μ) constant, so $\phi(\cdot)$ is not strictly convex. Thus $\phi(\cdot)$ is not strictly convex if and only if there exists a nonzero vector b such that $b^T Y$ is almost surely (μ) constant. The Theorem follows from Corollary 2.

Corollary 4. Let $\ell(\cdot)$ be the log likelihood of a sample of size n taken from (1) then $\ell(\cdot)$ is concave on Λ ; furthermore, $\ell(\cdot)$ is strictly concave on Λ if and only if $\dot{\phi}(\alpha)$ is full rank.

Proof. Let $y^1 \ldots y^n$ be a random sample from (1), then

$$\ell(\alpha) = \sum_{i=1}^{n} \alpha^T y^i - n\,\phi(\alpha) \ ;$$

thus

$$\ell(\lambda\alpha + (1-\lambda)\alpha^*) - (\lambda\ell(\alpha) + (1-\lambda)\ell(\alpha^*))$$

$$= -n[\phi(\lambda\alpha + (1-\lambda)\alpha^*) - (\lambda\phi(\alpha) + (1-\lambda)\phi(\alpha^*))] \ .$$

Corollary 5. Let $\hat{\alpha}$ maximize $\ell(\cdot)$ on Λ then, assuming $\ddagger(\alpha)$ is full rank, $\hat{\alpha}$ is unique.

Proof. Λ is a convex set; thus, if there exists $\hat{\alpha}_1, \hat{\alpha}_2$ which maximize $\ell(\cdot)$ then $a\hat{\alpha}_1 + (1-a)\hat{\alpha}_2$ is in Λ $(0 < a < 1)$, and from Corollary 4

$$\ell(a\hat{\alpha}_1 + (1-a)\hat{\alpha}_2) > a\ell(\hat{\alpha}_1) + (1-a)\ell(\hat{\alpha}_2) = \ell(\hat{\alpha}_1) ,$$

which is a contradiction.

Lemma 4 (Berk (1972)). $\phi(\cdot)$ is strictly convex on Λ if and only if $p(\cdot;\alpha_1) = p(\cdot;\alpha_2)$ implies $\alpha_1 = \alpha_2$ for any α_1, α_2 in Λ.

Clearly $\ddagger(\alpha)$ need not be of full rank; however, in the following we will show that if $\ddagger(\alpha)$ has rank $r < p$ then we may, by suitable transformation, obtain a family of form (1) with r parameters and covariance matrix of full rank. This fact was mentioned in Berk (1972).

Theorem 2. Given a density, $p(y;\alpha)$, of form (1) with covariance matrix of less than full rank, and the matrix A of Corollary 3, the transformed variables $Z = A^T Y$ again have a density of form (1) but with natural parameter $B = A^1\alpha$ and covariance matrix of full rank. Further,

$$p(y;\alpha) = f_Z(A^T y; A^T \alpha) .$$

Proof. Suppose $\dot{\phi}(\alpha)$ has rank $r < p$. From Lemmas 1 and 3

$$b_i^T Y = c_i \ , \quad i = 1,\ldots,p-r \ ,$$

almost surely (μ).

Let $B = (b_1,\ldots,b_{p-r})^T$ and $C = (c_1,\ldots,c_{p-r})^T$. We may write

$$\begin{aligned} \alpha^T Y &= \alpha^T (A:B)(A:B)^T Y \\ &= \alpha^T AA^T Y + \alpha^T BB^T Y \\ &= \alpha^T AA^T Y + \alpha^T BC \ , \end{aligned}$$

almost surely μ. Now, from (2)

$$\phi(\alpha) = \alpha^T BC + \phi(AA^T\alpha) \ .$$

Substituting in (1) we get

$$p(y;\alpha) = \exp(\alpha^T AA^T y - \phi(AA^T\alpha)) \ .$$

Let $Z = A^T Y$ and μ^* be the measure defined by

$$\mu^*(B) = \mu(\{y | A^T y \in B\})$$

for all r-dimensional Borel sets B. Now, by the change of variables theorem, see for example Lehmann (1959),

$$\int_{\{y \mid A^T y \in B\}} \exp(\alpha^T A A^T y - \phi(AA^T\alpha)) d\mu(y)$$

$$= \int_B \exp(\alpha^T A z - \phi(AA^T\alpha)) d\mu^*(z)$$

$$= \int_B \exp(\beta^T z - \phi^*(\beta)) d\mu^*(z)$$

where $\beta = A^T\alpha$ and $\phi^*(\beta) = \phi(A\beta)$. Hence $\exp(\beta^T z - \phi^*(\beta))$ is the density of Z with respect to μ^*. This density is a member of the exponential family with natural parameter β.

The covariance matrix of Z is $A^T \Sigma A$ which is of full rank.

Next, let

$$g(x;\theta) = \exp[\sum_{k=1}^{m} \eta_k(\theta)\psi_k(x) - Q(\theta)] \tag{3}$$

be a density, with respect to a σ-finite measure μ, where $\theta \in \Theta \subseteq E^p$. Such densities are said to be members of the exponential family. The following Theorem, a statement of which may be found in Berk (1972), relates the exponential family and the exponential family with natural parameterization.

Theorem 3. $Y = \psi(X)$ has density

$$f_Y(y;\eta(\theta)) = \exp[\sum_{k=1}^{m} \eta_k(\theta) y_k - \phi(\eta(\theta))] ,$$

a member of the exponential family with natural parameter $\eta(\theta)$. Further,

$$g(x;\theta) = f_Y(\psi(x);\eta(\theta)) \ .$$

Proof. Consider the transformation $Y_k = \psi_k(X)$; $k = 1,\ldots,m$ and let $\xi(A) = \mu\{x|\psi(x) \in A\}$. By the change of variables theorem,

$$\int_{\psi^{-1}(A)} \exp\Big(\sum_{k=1}^{m} \eta_k(\theta)\psi_k(x) - Q(\theta)\Big)d\mu(x)$$

$$= \int_A \exp\Big(\sum_{k=1}^{m} \eta_k(\theta)y_k - Q(\theta)\Big)d\xi(y) \ ;$$

and

$$\exp\Big(\sum_{k=1}^{m} \eta_k(\theta)y_k - Q(\theta)\Big)$$

is the density of $Y_1,\ldots,Y_m$ with respect to the measure ξ.

Now

$$\int \exp\Big(\sum_{k=1}^{m} \eta_k(\theta)y_k - Q(\theta)\Big)d\xi(y) = 1$$

implies $Q(\theta) = \phi(\eta(\theta))$ where $\phi(\cdot)$ is the log moment generating function of ξ . Substituting in the density of Y , and then in (3), we obtain the theorem.

Corollary 6. Let A be a matrix whose columns form an orthonormal basis for the column space of the covariance matrix of $\psi(X)$. $Z = A^T\psi(X)$ has density

$f_Z(z;\beta) = \exp(\beta^T z - \phi(\beta))$, a member of the exponential family with natural parameter $\beta = A^T\eta(\theta)$. The covariance matrix of Z has full rank. Further

$$g(x;\theta) = f_Z(A^T\psi(x);A^T\eta(\theta)) \ .$$

Proof. Apply Theorem 3, then Theorem 2.

In summary, applying Theorem 3 and then Corollary 4, the log likelihood for a random sample taken from a family of form (3) may be written as $\ell(\eta(\theta))$ where $\ell(\cdot)$ is concave. Also, from Corollary 4, $\ell(\cdot)$ is strictly concave if and only if the covariance matrix of $Y = \psi(X)$ is full rank. A density of form (3) will be said to be in canonical form if the covariance matrix of $\psi(X)$ is of full rank.

If the density of X is not in canonical form then from Corollary 6 we may write the log likelihood of the sample as $\ell(A^T\eta(\theta))$ where $\ell(\cdot)$ is strictly concave.

3. UNIQUENESS OF MAXIMUM LIKELIHOOD IN THE EXPONENTIAL FAMILY

Let us now consider the maximum likelihood problem for families of the form (3). Suppose that $\hat{\theta}$ is a maximum likelihood estimate for θ. When will $\hat{\theta}$ be unique? Huzurbazar (1949) demonstrates the uniqueness of $\hat{\theta}$ in families of the form (3) where $m = p$. However, from the following example, we see that $m = p$ is not necessary for the uniqueness of $\hat{\theta}$.

Example 1 (Charnes, et. al. (1975)). We wish to model the effect of radiation on bacteria in suspension. For each radiation does level several dilutions will be placed on petri dishes and the number of resulting colonies counted.

Let

X_{i1} = concentration of bacteria in suspension,

X_{i2} = radiation dose,

n_i = number of dilutions observed at the i^{th} dose level

Y_{ij} = number of colonies counted for the j^{th} dilution.

We assume that $Y_{ij} (j = 1 \ldots n_i,\ i = 1 \ldots N)$ are independent Poisson distributed random variables, Y_{ij} having expected value

$$\theta_1 X_{i1} \exp(-\theta_2 X_{i2}) .$$

The parameter θ_1 represents the number of colonies forming, per unit volume of suspension, when no radiation is present; θ_2 describes the radiation sensitivity of the bacteria.

We will estimate the parameters by maximum likelihood. The likelihood for a sample $Y_{ij}(j = 1 \ldots n_i,\ i = 1 \ldots N$ is proportional to

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} [Y_{ij}(\ln(\theta_1 X_{i1}) - \theta_2 X_{i2}) - \theta_1 X_{i1} \exp(-\theta_2 X_{i2})] ,$$

or, employing the dot notation of the analysis of variance,

$$(4) \qquad \sum_{i=1}^{N} [y_{i\cdot}(\ln(\theta_1 X_{i1}) - \theta_2 X_{i2}) - n_i \theta_1 X_{i1} \exp(-\theta_2 X_{i2})] .$$

Now, this likelihood is one from a family of the form (3) where $m = N$ and $p = 2$. If (4) has a maximum is it unique? The likelihood equations for (4) are

$$(5) \qquad \sum_{i=1}^{N} [y_{i\cdot}\theta_1^{-1} - n_i X_{i1} \exp(-\theta_2 X_{i2})] = 0 ,$$

and

$$(6) \qquad \sum_{i=1}^{N} [-y_{i\cdot} X_{i2} + n_i \theta_1 X_{i1} X_{i2} \exp(-\theta_2 X_{i2})] = 0 .$$

From (5)

$$(7) \qquad \hat{\theta}_1 = \sum_{i=1}^{N} y_{i\cdot} \Big/ \sum_{i=1}^{N} n_i X_{i1} \exp(-\hat{\theta}_2 X_{i2}) ;$$

substituting into (6) we have

$$(8) \qquad \frac{\sum_{i=1}^{N} n_i X_{i1} \exp(-\hat{\theta}_2 X_{i2})}{\sum_{i=1}^{N} n_i X_{i1} X_{i2} \exp(-\hat{\theta}_2 X_{i2})} = \frac{\sum_{i=1}^{N} y_{i\cdot}}{\sum_{i=1}^{N} y_{i\cdot} X_{i2}} \quad .$$

Let

$$g(\theta_2) = \frac{\sum_{i=1}^{N} n_i X_{i1} \exp(-\theta_2 X_{i2})}{\sum_{i=1}^{N} n_i X_{i1} X_{i2} \exp(-\theta_2 X_{i2})}$$

then

$$g'(\theta_2) =$$

$$\frac{-\left(\sum_{i=1}^{N} n_i X_{i1} X_{i2} \exp(-\theta_2 X_{i2})\right)^2 + \sum_{i=1}^{N} n_i X_{i1} \exp(-\theta_2 X_{i2}) \sum_{i=1}^{N} n_i X_{i1} X_{i2}^2 \exp(-\theta_2 X_{i2})}{\left(\sum_{i=1}^{N} n_i X_{i1} X_{i2} \exp(-\theta_2 X_{i2})\right)^2} \quad .$$

Letting $a_i = \sqrt{n_i X_{i1} \exp(-\theta_2 X_{i2})}$ and $b_i = X_{i2} a_i$ then by Schwarz inequality

$$\left(\sum_{i=1}^{N} a_i b_i\right)^2 \leq \sum_{i=1}^{N} a_i^2 \sum_{i=1}^{N} b_i^2$$

with equality holding only if $\frac{b_i}{a_i} = X_{i2}$ is constant for all i ; thus, assuming X_{i2} is not constant in i ,

$g'(\theta_2) > 0$. Therefore, if a solution to the likelihood equations exists, it is unique.

Given a solution to the likelihood equations we will show that it is a local, and thus unique global, maximum by showing that the Hessian of the likelihood is negative definite when evaluated at the solution. A matrix is positive definite if and only if its principle minors are positive (see Nobel (1969), page 395). The principle minors of the negative of the Hessian of the likelihood evaluated at the solution of the likelihood equations are

$$\sum_{i=1}^{N} y_{i.} \hat{\theta}_1^{-2}$$

and

$$\sum_{i=1}^{N} y_{i.} \hat{\theta}_1^{-1} \sum_{i=1}^{N} n_i X_{i1} X_{i2}^2 \exp(-\hat{\theta}_2 X_{i2})$$

$$- \left(\sum_{i=1}^{N} n_i X_{i1} X_{i2} \exp(-\hat{\theta}_2 X_{i2}) \right)^2 .$$

The first principle minor is greater than zero and by replacing $\hat{\theta}_1$ by the expression given by (7) the second is equal to

$$\sum_{i=1}^{N} n_i X_{i1} \exp(-\hat{\theta}_2 X_{i2}) \sum_{i=1}^{N} n_i X_{i1} X_{i2}^2 \exp(-\hat{\theta}_2 X_{i2})$$

$$- \left(\sum_{i=1}^{N} n_i X_{i1} X_{i2} \exp(-\hat{\theta}_2 X_{i2}) \right)^2$$

which is also greater than zero; thus the Hessian of the likelihood evaluated at the solution of the likelihood equations is negative definite.

In summary we have shown that given a solution to the likelihood equations (5) and (6) this solution is a unique global maximum of the likelihood (4).

To demonstrate the existence of solutions in a numerical example we consider the following data:

i	x_{i1}	x_{i2}	n_i	y_{ij}					
1	1	0	6	299	283	280	246	264	254
2	1	1	2	169	184				
3	2	2	5	179	224	188	202	194	
4	4	3	5	233	261	229	286	264	
5	10	4	4	401	410	356	388		
6	4	5	5	157	146	134	161	159	

Using a search technique to solve (8) we obtain $\hat{\theta}_2 = .4459$ and, by evaluating (7) at $\hat{\theta}_2$, we have $\hat{\theta}_1 = 256.9$.

Now, returning to the general discussion, suppose $\hat{\theta}$ is a maximum likelihood estimate. The likelihood for a sample of size n taken from a population of form (3) is $\ell(\eta(\theta))$, where

$$\ell(\cdot) = \sum_{i=1}^{n} \ln f_Y(\psi(x_i);\cdot) .$$

From this we observe:

Theorem 4. $\hat{\theta}$ is unique if and only if

(i) $\eta(\hat{\theta})$ is the unique maximum of $\ell(\cdot)$ on the range of $\eta(\cdot)$, and

(ii) there exists no other θ such that $\eta(\theta) = \eta(\hat{\theta})$.

The following example illustrates the use of Theorem 4 in the logistic model of Thompson (1976).

Example 2 (Logistic Life Study Model). The log likelihood of the logistic life study model is

$$L(\beta,\eta) = \sum_{j=1}^{n} \sum_{V_j \cup S_j} [y_{ij}(z_{ij}\beta + \eta_j) - \ln(1 + \exp(z_{ij}\beta + \eta_j))]$$

where z_{ij}^T is a vector of variables for the i^{th} individual in the j^{th} time interval, S_j is the set of survivors of the j^{th} time interval, V_j is the set of failures in the j^{th} time interval, y_{ij} equals 1 if the i^{th} individual is in V_j and 0 if the i^{th} individual is in S_j, and $(\beta^T : \eta_1, \ldots, \eta_m)^T$ is a vector of unknown parameters to be estimated.

Now, $L(\beta,\eta)$ is the log likelihood of a density which is a member of the exponential family and in this case the function $\eta(\cdot)$ is the linear function determined by $z_{ij}\beta + \eta_j$, $i \in V_j \cup S_j$, $j = 1,\ldots,m$ and

$$\ell(\alpha) = \sum_{j=1}^{m} \sum_{V_j \cup S_j} [y_{ij}\alpha_{ij} - \ln(1 + \exp(\alpha_{ij}))] .$$

For the logistic model, hypothesis (i) of Theorem 4 holds since we are maximizing a strictly concave function over a linear manifold; and, hypothesis (ii) becomes necessary and sufficient for a maximum likelihood estimate, $(\hat{\beta}^T : \hat{\eta}_1 \ldots \hat{\eta}_m)$, to be unique.

Thus, $(\hat{\beta}^T : \hat{\eta}_1 \ldots \hat{\eta}_m)$ will be a unique maximum likelihood estimate if and only if the matrix

$$Z = \left[\begin{array}{c:c:c} & & \\ z_{11} & 1 0 \ . \ . \ . \ 0 & \\ \vdots & \vdots & \\ z_{n_1 1} & 1 0 \ . \ . \ . \ 0 & \\ \hdashline & & \\ \vdots & \vdots & \\ \hdashline & & \\ z_{1k} & 0 \ldots 0 1 0 \ldots 0 & \\ \vdots & \vdots & \\ z_{n_k k} & 0 \ldots 0 1 0 \ldots 0 & \\ \hdashline & & \\ \vdots & \vdots & \\ \hdashline & & \\ z_{1\ell} & 0 \ . \ . \ . \ 0 1 & \\ \vdots & \vdots & \\ z_{n_\ell \ell} & 0 \ . \ . \ . \ 0 1 & \end{array}\right]$$

is full rank.

Though Theorem 4 was stated with the exponential family in mind, we may apply it to any problem where the density is a composite function. In the following example, we consider a modified logistic model, applying Theorem 4 to obtain conditions for the uniqueness of maximum likelihood estimates:

Example 3 (Modified Logistic Life Study). In Thompson (1976) items censored in an interval were considered to be not at risk in the interval; thus, no contribution to the likelihood was obtained from the interval in which an individual was censored. Thompson (1977) considers a modification of the logistic model to obtain information from the interval in which censoring occurred.

The log likelihood of the modified logistic model is

$$L(\beta,\eta) = \sum_{j=1}^{m} \Big[\sum_{V_j \cup S_j} y_{ij}(z_{ij}\beta+\eta_j) - \sum_{V_j \cup S_j} \ell n(1+\exp(z_{ij}\beta+\eta_j)) - 2^{-1} \sum_{L_j} \ell n(1+\exp(z_{ij}\beta+\eta_j)) \Big] ,$$

where L_j is the set of individuals censored in the j^{th} interval. Now,

$$L(\beta,\eta) = \ell(z_{ij}\beta+\eta_j) \; ; \; i \in V_j \cup S_j \cup L_j \; , \quad j = 1,\ldots,n)$$

where

$$\ell(\alpha_{ij}) = \sum_{j=1}^{m} \Big[\sum_{V_j \cup S_j} y_{ij}\alpha_{ij} - \sum_{V_j \cup S_j} \ell n(1+\exp\alpha_{ij}) - 2^{-1} \sum_{L_j} \ell n(1+\exp\alpha_{ij}) \Big]$$

thus, if $\ell(\cdot)$ is strictly concave then by applying the same reasoning as in Example 2, a maximum likelihood estimate will be unique if and only if the matrix Z is full rank. To show that $\ell(\cdot)$ is strictly concave note that

$$\frac{\partial^2 \ell(\alpha)}{\partial \alpha_{k\ell} \partial \alpha_{jk}} = \begin{cases} 0 & i \neq k \text{ or } j \neq \ell \\ -\left(\frac{1}{1+\exp\alpha_{ij}}\right)^2 & i = k, \quad j = \ell \\ & i \quad V_j \quad S_j \\ -2^{-1}\left(\frac{1}{1+\exp\alpha_{ij}}\right)^2 & i = k, \quad j = \ell \\ & i \in L_j \end{cases}$$

Thus, the matrix of second order partials of $\ell(\cdot)$ is negative definite, and hence (see Roberts and Varberg (1973) page 103), $\ell(\cdot)$ is strictly concave.

Conditions (i) and (ii) of Theorem 4 are difficult to verify when working with a particular problem; therefore, it is desirable to find conditions sufficient for both (i) and (ii) which are more tractable.

Consider (i). A necessary and sufficient condition for $\eta(\hat{\theta})$ to be unique is that

$$\eta(\Theta) \cap \{\alpha \mid \ell(\alpha) \geq \ell(\eta(\hat{\theta}))\} = \{\eta(\hat{\theta})\} .$$

In the case that $g(x;\theta)$ is in canonical form and $\nabla\ell(\eta(\hat{\theta}))$ exists, a sufficient condition for (i) may be established using the following well known fact (see, for example, Roberts and Varberg (1973), page 98):

Lemma 5. Let $\ell(\cdot)$ be strictly concave on Λ and differentiable at α_1 then

$$\ell(\alpha_2) \geq \ell(\alpha_1) \quad (\alpha_1 \neq \alpha_2)$$

implies

$$(\alpha_2 - \alpha_1)^T \nabla\ell(\alpha_1) > 0 .$$

Theorem 5. If $(\eta(\theta) - \eta(\hat{\theta}))^T \nabla\ell(\eta(\hat{\theta})) \leq 0$ for all θ then $\eta(\hat{\theta})$ is unique.

Proof. By assumption

$$\eta(\Theta) \subseteq \{\alpha | (\alpha - \eta(\hat{\theta}))^T \nabla\ell(\eta(\hat{\theta})) \leq 0\}$$

and from Lemma 5

$$\{\alpha | \ell(\alpha) \geq \ell(\eta(\hat{\theta})) \text{ and } \alpha \neq \eta(\hat{\theta})\} \subseteq$$

$$\{\alpha | (\alpha - \eta(\hat{\theta}))^T \nabla\ell(\eta(\hat{\theta})) > 0\} ;$$

Thus, $\eta(\Theta) \cap \{\alpha | \ell(\alpha) \geq \ell(\eta(\hat{\theta}))\} = \{\eta(\hat{\theta})\}$.

We will illustrate Theorem 5 using Example 1. The covariance matrix of $Y_{1.},\ldots,Y_{N.}$ is a matrix with diagonal terms $\theta_1 X_{i1} \exp(-\theta_2 X_{i2})$ and off diagonal terms

zero; thus, provided X_{i1} is not zero for any i, the density of $Y_{1.},\ldots,Y_{N.}$ is in canonical form. Now,

$$\eta(\theta) = (\ell n(\theta_1 X_{11}) - \theta_2 X_{12},\ldots,\ell n(\theta_1 X_{N1}) - \theta_2 X_{N2})^T$$

and

$$\ell(\alpha) = \sum_{i=1}^{N} (y_{i.}\alpha_i - n_i \exp \alpha_i) \ ;$$

thus,

$$(\eta(\theta) - \eta(\hat{\theta}))^T = (\ell n(\frac{\theta_1}{\hat{\theta}_1}) + (\hat{\theta}_2 - \theta_2)X_{12},\ldots,\ell n(\frac{\theta_1}{\hat{\theta}_1}) + (\hat{\theta}_2 - \theta_2)X_{N2})$$

and

$$\nabla\ell(\eta(\hat{\theta})) = (y_{1.} - n_1\hat{\theta}_1 X_{11} \exp(-\hat{\theta}_2 X_{12}),\ldots,y_{N.} - n_N\hat{\theta}_1 X_{N1} \exp(-\hat{\theta}_2 X_{N2}))^T \ .$$

Therefore

$$(\eta(\theta) - \eta(\hat{\theta}))^T \nabla\ell(\eta(\hat{\theta})) =$$

$$\ell n(\frac{\theta_1}{\hat{\theta}_1}) \sum_{i=1}^{N} (y_{i.} - n_i\hat{\theta}_1 X_{i1} \exp(-\hat{\theta}_2 X_{i2})) +$$

$$(\hat{\theta}_2 - \theta_2) \sum_{i=1}^{N} X_{i2}(y_{i.} - n_i\hat{\theta}_1 X_{i1} \exp(-\hat{\theta}_2 X_{i2})) \ .$$

Now, from (5)

$$\sum_{i=1}^{N} (y_{i.} - n_i \hat{\theta}_1 X_{i1} \exp(-\hat{\theta}_2 X_{i2})) = 0 ,$$

and from (7) and (8)

$$\sum_{i=1}^{N} X_{i2}(y_{i.} - \hat{\theta}_1 n_i X_{i1} \exp(-\hat{\theta}_2 X_{i2})) = 0 ;$$

thus, from Theorem 5 $\eta(\hat{\theta})$ is unique.

An assumption stronger than (ii) is that $\eta(\cdot)$ be one to one, this is the case in Example 1.

In fact, $\eta(\theta^*) = \eta(\theta)$ implies

$$\theta_1^* X_{i1} \exp(-\theta_2^* X_{i2}) = \theta_1 X_{i1} \exp(-\theta_2 X_{i2}) , \quad i = 1,\ldots,N$$

or

$$\theta_1^*/\theta_1 \exp((\theta_2 - \theta_2^*) X_{i2}) = 1 , \quad i = 1,\ldots,N .$$

Thus, assuming X_{i2} is not constant in i, $\theta_1^* = \theta_1$ and $\theta_2^* = \theta_2$.

In summary, through application of Theorem 5 and by showing that $\eta(\cdot)$ is one to one, we have shown that hypothesis (i) and (ii) of Theorem 4 hold; thus, the maximum likelihood estimate for the parameters in Example 1 is unique.

4. ESTIMABLE FUNCTIONS AND TESTABLE HYPOTHESES FOR THE NORMAL LINEAR MODEL

From Example 2 we see that the likelihood in Thompson (1976) will admit a unique maximum likelihood estimate if and only if the linear transformation determined by the matrix of covariates is one to one. Therefore, to look at inference problems for the logistic model when the maximum likelihood estimates are not unique, we will consider members of the exponential family in which $\eta(\cdot)$ is not a one to one function. In this case, problems of identification, as discussed in Koopmans and Reiersøl (1950), arise.

Before we discuss the identification problem in general, let us consider another example of a member of the exponential family where $\eta(\cdot)$ is not one to one -- the normal linear model of less than full rank.

In the normal linear model we have an $n \times 1$ dimensional random vector, Y, which we express as

$$Y = X\beta + \varepsilon ,$$

where X is an $n \times p$ matrix of known values, β is a $p \times 1$ vector of unknown parameters, and ε is an $n \times 1$ vector of errors distributed as a multivariate normal with mean 0 and variance $\sigma^2 I$. The log likelihood for y is

$$-\frac{n}{2} \ell n\, \sigma^2 - \frac{1}{2} \frac{(y-X\beta)^T (y-X\beta)}{\sigma^2}$$

or

$$\frac{y^T X\beta}{\sigma^2} - \frac{y^T y}{2\sigma^2} - \frac{\beta^T X^T X\beta}{2\sigma^2} - \frac{n \ln \sigma^2}{2} \quad ;$$

thus, letting

$$\eta(\beta,\sigma^2) = (\beta^T X^T/\sigma^2 \vdots -\frac{1}{2\sigma^2})^T \ ,$$

$$\psi(y) = (y^T \vdots y^T y)$$

and

$$Q(\beta,\sigma^2) = -\frac{\beta^T X^T X\beta}{2\sigma^2} - \frac{n}{2} \ln \sigma^2 \ ,$$

we may write the likelihood for y as

$$\eta(\beta,\sigma^2)^T \psi(y) - Q(\beta,\sigma^2) \ .$$

Now, $\eta(\beta_*,\sigma_*^2) = \eta(\beta,\sigma^2)$ if and only if $\sigma_*^2 = \sigma^2$ and $X(\beta_* - \beta) = 0$; thus, $\eta(\cdot,\cdot)$ is one to one if and only if X is full rank. Therefore, the normal linear model of less than full rank is a member of the exponential family for which $\eta(\cdot)$ is not one to one.

In the normal linear model the concepts of estimable functions and testable hypotheses are introduced to remedy problems caused by X being less than full rank.

We will denote the null space of the matrix M (the set of solutions to $Mx = 0$) by $\text{Null}(M)$ and the row space

of M (the set of all linear combinations of the rows of M) by $\text{Row}(M)$.

An estimable function is defined as follows: A linear function of β is estimable if and only if there is a linear function of Y which is an unbiased estimate of it. From this definition we have the following:

Theorem 6. $\lambda^T\beta$ is estimable if and only if $\lambda = X^T r$ for some r .

Proof. If $\lambda^T\beta$ is estimable then there exists a vector r such that $E(r^TY) = r^TX\beta = \lambda^T\beta$ for all ; thus, $\lambda = X^Tr$.

Conversely, $\lambda = X^Tr$ implies $E(r^TY) = \lambda^T\beta$.

We may restate Theorem 6 as $\lambda^T\beta$ is estimable if and only if λ^T is in $\text{Row}(X)$.

Now $\hat{\beta}$ is a maximum likelihood estimate for β if and only if $\hat{\beta}$ solves the normal equations,

$$X^TX\beta = X^TY .$$

One important property of estimable functions is given by the following result:

Theorem 7. $\lambda^T\beta$ is estimable if and only if $\lambda^T\hat{\beta}$ is constant for all $\hat{\beta}$ maximizing the likelihood.

Proof. Rao (1965, page 181).

Another important property of estimable functions concerns tests of hypotheses. A hypothesis, H, stating that β is in $S = \{\beta | \lambda_i^T \beta = m_i,\ i = 1,\ldots,\ell\}$ is called testable if and only if $\lambda_i^T \beta$ is estimable for each i. Without loss of generality we will assume that the λ_i's are linearly independent.

Searle (1971) discusses testable hypotheses, showing that the sum of squares error under a nontestable hypothesis, where all $\lambda_i^T \beta$ are not estimable, is equal to the sum of squares due to error. Let ~H be the hypothesis stating that β is in S^c. Seely (1977) shows the intersection of the sets of expected values under the null H and the alternative ~H is empty if and only if H is testable. The following is a version of Seely's result.

Theorem 8.

$$XS \cap XS^c = \phi$$

if and only if $\lambda_i^T \beta$ is estimable for each i.

Proof. Suppose for some i, $\lambda_i^T \beta$ is not estimable then, from Theorem 6, λ_i is not in Row(X); thus, there exists a β^* in Null(X) for which $\lambda_i^T \beta^* \neq 0$. Let β_1 be in S then $\beta_1 + \beta^*$ is in S^c; however,

$$X(\beta_1 + \beta^*) = X\beta_1 + X\beta^* = X\beta_1 .$$

Therefore,

$$XS \cap XS^c \neq \phi \ .$$

Conversely, if $XS \cap XS^c \neq \phi$ then there exists β_1 in S and β_2 in S^c such that $X\beta_1 = X\beta_2$; thus, $\beta_2 - \beta_1$ is in $\text{Null}(X)$.

Now, for at least one i , $\lambda_i^T\beta_1 \neq \lambda_i^T\beta_2$; thus, $\lambda_i^T(\beta_2 - \beta_1) \neq 0$. Therefore, λ_i is not in $\text{Row}(X)$ and, from Theorem 6, $\lambda_i^T\beta$ is not estimable.

The next theorem gives a more exact relationship between XS and XS^c .

Theorem 9. If $\lambda_i^T\beta$ is not estimable for at least one i then $XS \subseteq XS^c = \text{Col}(X)$. Furthermore, if $\lambda_i^T\beta$ is not estimable for any i then $XS = XS^c$.

Proof. Suppose for some i that $\lambda_i^T\beta$ is not estimable, then from Theorem 6, λ_i is not in $\text{Row}(X)$; thus, there is a δ in $\text{Null}(X)$ such that $\lambda_i^T\delta \neq 0$.

Let β be in S then

$$\lambda_i^T(\beta+\delta) = \lambda_i^T\beta + \lambda_i^T\delta = m_i + \lambda_i^T\delta \neq m_i \ ;$$

thus, $\beta+\delta$ is in S^c . Now,

$$X(\beta+\delta) = X\beta + X\delta = X\beta \ ;$$

hence, $X\beta$ is in XS^c . Therefore, $XS \subseteq XS^c$.

Now, $\text{Col}(X) = XS \cup XS^C = XS^C$.

Suppose $\lambda_i^T\beta$ is not estimable for any i , then, from Theorem 6, for each i , λ_i^T is linearly independent of the rows of X . Now, since the vectors λ_i , $i = 1,\ldots,\ell$, are linearly independent, the equations in the variable n ,

$$\lambda_i^T n = m_i - \lambda_i^T\beta \ , \quad i = 1,\ldots,\ell$$

and

$$Xn = 0$$

have at least one solution for all β . Hence, for β^* in S^C there exists n^* such that $\beta^* + n^*$ is in S and

$$X(\beta^* + n^*) = X\beta^* \ .$$

Therefore, $XS^C \subseteq XS$ and the theorem follows.

We may extend the results of Theorem 8 to hypotheses involving inequality constraints.

Theorem 10. Let

$$S = \{\beta \mid \lambda_i^T\beta = m_i, \ i = 1,\ldots,s \ \text{ and } \ \lambda_i^T\beta \geq m_i, \ i = s+1,\ldots,e\}$$

then $XS \cap XS^C = \phi$ if and only if $\lambda_i^T\beta$ is estimable for each i .

Proof. If $\lambda_i^T\beta$ is not estimable for some i, then, as in the proof of Theorem 9, there exists a δ in Null(X) such that $\lambda_i^T\delta \neq \phi$. Let β^* be in S then we may find a real number r such that $\beta^* + r\delta$ is in S^c. Now,

$$X(\beta^* + r\delta) = X\beta^* + rX\delta = X\beta^* ;$$

thus, $XS \cap XS^c \neq \phi$. The proof of the converse is the same as that of the converse of Theorem 8.

5. IDENTIFIABLE PARAMETRIC FUNCTIONS

5.1 Definitions and Properties

In the discussion of the normal linear model the concept of an estimable function was used to solve some of the problems associated with X being less than full rank. To generalize this concept to functions of the parameter of an exponential family member where $\eta(\cdot)$ is not one to one, we considered the concept of identifiability. (See Theorem 17.)

Let Y be a sample with density $f(y;\theta)$. From Koopmans and Reiersøl (1950) a function $h(\cdot)$ of θ will be called identifiable at θ_0 if $f(\cdot;\theta) = f(\cdot;\theta_0)$ implies $h(\theta) = h(\theta_0)$. The significance of identifiability is as follows: Suppose that an observation Y is produced according to some member of the class of densities $f(\cdot;\theta)$, $\theta \in \Theta$. From Y we wish to make an inference about the true θ, say θ_0. The characteristic of θ, in which we are interested, is $h(\theta)$. If $h(\cdot)$ is not identifiable at θ_0 then there exists θ' such that $f(\cdot;\theta') = f(\cdot;\theta_0)$ but $h(\theta') \neq h(\theta_0)$. Thus, even if we could infer the density perfectly, we could still not discriminate between $h(\theta')$ and $h(\theta_0)$.

Theorem 11. If $f(\cdot;\theta_0) = f(\cdot;\theta_1)$ then h is identifiable at θ_0 if and only if it is identifiable at θ_1.

Proof. Suppose $h(\cdot)$ is identifiable at θ_0. By definition, $h(\theta_1) = h(\theta_0)$, and $f(\cdot;\theta) = f(\cdot;\theta_0)$ implies $h(\theta) = h(\theta_0)$. Hence $f(\cdot;\theta) = f(\cdot;\theta_1)$ implies $h(\theta) = h(\theta_1)$.

Theorem 12. In the special case that $Y = (X_1,\ldots,X_n)$ is a random sample from a density $f_X(x;\theta)$ then $h(\cdot)$ is identifiable at θ_0 if $f_X(\cdot;\theta) = f_X(\cdot;\theta_0)$ implies $h(\theta) = h(\theta_0)$.

Proof. Since $f(x;\theta) = \prod_{i=1}^{n} f_X(x_i;\theta)$, then $f(\cdot;\theta) = f(\cdot;\theta_0)$ is equivalent to $f_X(\cdot;\theta) = f_X(\cdot;\theta_0)$.

From Theorem 12, if we are observing a random sample from some density then the set of functions identifiable at θ_0 is the same for all sample sizes and we may check $h(\cdot)$ for identifiability at θ_0, for any particular sample size, by checking at sample size one.

Let $\hat{\theta}$ be a maximum likelihood estimate for θ then, from Zehna (1966), $h(\hat{\theta})$ is a maximum likelihood estimate for $h(\theta)$.

Theorem 13. $h(\hat{\theta})$ is a unique maximum likelihood estimate for $h(\theta)$ only if $h(\cdot)$ is identifiable at $\hat{\theta}$.

Proof. If $h(\hat{\theta})$ is unique then $h(\cdot)$ is constant on $\{\theta|f(y;\theta) = f(y;\hat{\theta})\}$. Now, $\{\theta|f(\cdot;\theta) = f(\cdot;\hat{\theta})\} \subseteq \{\theta|f(y;\theta) = f(y;\hat{\theta})\}$; thus, $h(\cdot)$ is constant on $\{\theta|f(\cdot;\theta) = f(\cdot;\hat{\theta})\}$, so $h(\cdot)$ is identifiable at $\hat{\theta}$.

Now we return to $g(x;\theta)$, a density of form (3).

Theorem 14. Assuming $\eta(\hat{\theta})$ is unique, $h(\hat{\theta})$ is a unique maximum likelihood estimate for $h(\theta)$ if and only if $h(\cdot)$ is identifiable at $\hat{\theta}$.

Proof. From Theorem 3,

$$f(x;\theta) = \prod_{i=1}^{n} g(x_i;\theta) = \prod_{i=1}^{n} f_Y[\psi(x_i);\eta(\theta)]$$

so that $\eta(\theta) = \eta(\hat{\theta})$ implies $f(\cdot;\theta) = f(\cdot;\hat{\theta})$. Therefore, if h is identifiable at $\hat{\theta}$, then $\eta(\theta) = \eta(\hat{\theta})$ implies $h(\theta) = h(\hat{\theta})$. The converse is given by Theorem 13.

5.2 Uniformly Identifiable Parametric Functions

Let $g(x;\theta)$ be a density of form (3).

Lemma 6. Assuming $g(x;\theta)$ is in canonical form, then $g(\cdot;\theta) = g(\cdot;\theta_0)$ if and only if $\eta(\theta) = \eta(\theta_0)$.

Proof. From Theorem 3, $g(x;\theta) = f_Y(\psi(x);\eta(\theta))$. Since $g(x;\theta)$ is in canonical form, from Theorem 1, the function $\phi(\cdot)$ for $f_Y(y;\alpha)$ is strictly convex on $\{\alpha|\phi(\alpha) < \infty\}$. Thus, from Lemma 4, $g(\cdot;\theta) = g(\cdot;\theta_0)$ if and only if $\eta(\theta) = \eta(\theta_0)$.

Theorem 15. Assuming $g(x;\theta)$ to be in canonical form, then $h(\cdot)$ is identifiable at θ_0 if and only if $\eta(\theta) = \eta(\theta_0)$ implies $h(\theta) = h(\theta_0)$.

Proof. The result follows from Lemma 6 and the definition of identifiable at θ_0 .

The rest of this section depends on Theorem 15, so we will restrict our attention to densities of form (3) in canonical form.

Corollary 7. Let $\eta(\theta_1) = \eta(\theta_0)$ then $h(\cdot)$ is identifiable at θ_0 if and only if $h(\cdot)$ is identifiable at θ_1 .

This follows from Theorems 11 and 15.

Koopmans and Reiersøl (1950) call $h(\cdot)$ uniformly identifiable if $h(\cdot)$ is identifiable at θ_0 for all θ_0 in Θ. If $h(\cdot)$ is not uniformly identifiable then the set

$$\Theta_h = \{\theta | h(\cdot) \text{ is identifiable at } \theta\}$$

is important.

Theorem 16. For α in $\eta(\Theta)$ let $r(\alpha) = h(\theta)$, where $\eta(\theta) = \alpha$, then $r(\cdot)$ is a function from $\eta(\Theta_h)$ to $h(\Theta_h)$; that is, $h(\theta) = r(\eta(\theta))$ for θ in Θ_h .

Proof. $r(\cdot)$ is a function from $\eta(\Theta_h)$ to $h(\Theta_h)$ since, if $\alpha = \eta(\theta_0) = \eta(\theta_1)$ then, from Theorem 15, $r(\alpha) = h(\theta_0) = h(\theta_1)$.

Corollary 8. $h(\cdot)$ is uniformly identifiable if and only if there exists a function $r(\cdot)$ such that $h(\theta) = r(\eta(\theta))$ for all θ in Θ.

Proof. If $h(\cdot)$ is uniformly identifiable then $\Theta_h = \Theta$; thus, from Theorem 16, there exists a function $r(\cdot)$ such that $h(\theta) = r(\eta(\theta))$ for all θ in Θ.

Conversely, if $h(\theta) = r(\eta(\theta))$ for all θ in Θ, then from Theorem 15 $h(\cdot)$ is uniformly identifiable.

Corollary 9. In the case that $\eta(\theta) = M\theta$ for a matrix M, $h(\theta) = \lambda^T\theta$ for some vector θ and $\Theta = E^P$, $h(\cdot)$ is identifiable at θ_0 if and only if $h(\cdot)$ is uniformly identifiable.

Proof. Suppose $h(\cdot)$ is identifiable at θ_0, then from Theorem 15, $M\theta = M\theta_0$ implies $\lambda^T\theta = \lambda^T\theta_0$.

Let $M\theta = 0$ then $M(\theta + \theta_0) = M\theta_0$; thus, $\lambda(\theta + \theta_0) = \lambda\theta_0$, so $\lambda\theta = 0$.

Thus, λ^T is perpendicular to Null(M), so λ^T is in Row(M). Therefore

$$\lambda^T\theta = r^TM\theta$$

for some vector r and, from Corollary 8, $n(\theta) = \lambda^T\theta$ is uniformly identifiable. The converse is a special case.

Theorem 17. In the normal linear model $\lambda^T\beta$ is estimable if and only if it is uniformly identifiable.

Proof. Suppose $\lambda^T\beta$ is estimable, then from Theorem 6,

$$\lambda^T\beta = r^TX\beta$$

for all β; thus, from Corollary 8, $\lambda^T\beta$ is uniformly identifiable.

Conversely if $\lambda^T\beta$ is uniformly identifiable then, again applying Corollary 8, there exists a function $r(\cdot)$ such that

$$\lambda^T\beta = r(X\beta)$$

for all β. Now, $\lambda^T\beta$ being linear implies $r(\cdot)$ must be linear; thus, there exists r such that

$$\lambda^T\beta = r^TX\beta \ ,$$

and from Theorem 6 $\lambda^T\beta$ is estimable.

This result was obtained in Reiersøl (1963) by a different method.

5.3 Comparison of Uniformly Identifiable Functions and Those Possessing an Unbiased Estimate

From Theorem 17 we see that the concept of uniform identifiability is one possible generalization of the concept of an estimable function. Another generalization is

those functions having an unbiased estimate.

Let Y be a sample with density $f(y;\theta)$.

Theorem 18. If $u(\theta)$ has an unbiased estimate, then $u(\cdot)$ is uniformly identifiable.

Proof. There exists a function $z(\cdot)$ such that

$$u(\theta) = \int z(y) f(y;\theta) du(y)$$

for all θ; thus, $f(\cdot;\theta_0) = f(\cdot;\theta_1)$ implies $u(\theta_0) = u(\theta_1)$.

In the following example we look at a density where there is a function which is uniformly identifiable, but does not have an unbiased estimate.

Example 4. Let Y_1, Y_2 be independent, binary random variables such that

$$P(Y_1=1) = \frac{\exp(\beta_1+\beta_2+\beta_3)}{1+\exp(\beta_1+\beta_2+\beta_3)}$$

and

$$P(Y_2=1) = \frac{\exp(\beta_1+\beta_2-\beta_3)}{1+\exp(\beta_1+\beta_2-\beta_3)} .$$

Letting $X_1 = (1,1,1)$, $X_2 = (1,1,-1)$ and $\beta = (\beta_1,\beta_2,\beta_3)^T$, we may write the density of Y_1, Y_2 as

$$\exp\, y_1 X_1\beta + y_2 X_2\beta - \ell n(1+\exp(X_1\beta)) - \ell n(1+\exp(X_2\beta)) ,$$

which is of form (3), in canonical form.

Now, $\eta(\cdot)$ is linear with coefficient matrix $X = (X_1^T X_2^T)^T$. So, from Corollary 8, the function

$$\lambda_0^T\beta = (\tfrac{1}{2},-\tfrac{1}{2})X\beta = \beta_3$$

is uniformly identifiable.

We will show that $\lambda_0^T\beta$ does not possess an unbiased estimate. Assume there exists a function $z(\cdot,\cdot)$ such that $\lambda_0^T\beta$ is equal to the expected value of $z(Y_1,Y_2)$ for all β.

Now,

$$\lim_{\beta_3\to\infty} \frac{\exp(y_1X_1\beta)}{1+\exp(X_1\beta)} = \begin{cases} 1 & y_1 = 1 \\ 0 & y_1 = 0 \end{cases}$$

and

$$\lim_{\beta_3\to\infty} \frac{\exp(y_2X_2\beta)}{1+\exp(X_2\beta)} = \begin{cases} 0 & y_2 = 1 \\ 1 & y_2 = 0 \end{cases} ;$$

thus

$$\lim_{\beta_3\to\infty} E(z(Y_1,Y_2)) = \lim_{\beta_3\to\infty} \Sigma\, z(y_1,y_2) \frac{\exp(y_1X_1\beta)}{1+\exp(X_1\beta)} \frac{\exp(y_2X_2\beta)}{1+\exp(X_2\beta)}$$

$$= z(1,0) ,$$

where the summation is over the sample space. However, the expected value of $z(Y_1,Y_2)$ is β_3 for all β, so $z(1,0) = \lim_{\beta\to\infty} \beta_3 = +\infty$. Now, if $z(1,0) = \infty$ then the

expected value of $z(Y_1,Y_2)$ is also ∞ for all β, which is a contradiction.

5.4 Examples

Example 5 (Logistic Life Study Model - Continued). In example 2 we saw that a unique maximum likelihood estimate for $(\beta^T:\eta_1,\ldots,\eta_m)^T$ exists if and only if matrix Z is full rank. Assuming that Z is not full rank, we wish to consider the class of uniformly identifiable functions.

From Corollary 8 a function, $h(\cdot)$, is uniformly identifiable if and only if there exists a function $r(\cdot)$ such that

$$h[(\beta^T:\eta_1,\ldots,\eta_m)^T] = r(Z(\beta^T:\eta_1,\ldots,\eta_m)^T)$$

or, for differentiable $r(\cdot)$,

$$\nabla h[(\beta^T:\eta_1,\ldots,\eta_m)^T] = Z^T \nabla r(Z(\beta^T:\eta_1,\ldots,\eta_m)^T) ;$$

thus, for linear $h(\cdot)$, $h(\cdot)$ is uniformly identifiable if and only if for some vector r

$$\lambda = Z^T r$$

where $\lambda = \nabla h[(\beta^T:\eta_1,\ldots,\eta_m)^T]$. In other words the class of linear uniformly identifiable functions is that class of functions whose gradients are in $\mathrm{Row}(Z)$.

Let λ_i, $i = 1,\ldots,s$, span the space orthogonal to the row space of Z. The functions $\lambda_i^T(\beta^T:\eta_1,\ldots,\eta_m)^T$,

$i = 1,\ldots,s$ are useful in the computation of a maximum likelihood estimate. As noted in Example 2, there exists a unique $\hat{\eta}$ in the range of $Z(\beta^T : \eta_1,\ldots,\eta_m)^T$ which maximizes the likelihood; thus, a maximum likelihood estimate for $(\beta^T : \eta_1,\ldots,\eta_m)^T$ can be found by solving

$$Z(\beta^T : \eta_1,\ldots,\eta_m)^T = \hat{\eta} .$$

If we solve these equations under the restriction $\lambda_i^T(\beta^T : \eta_1,\ldots,\eta_m)^T = 0$, $i = 1,\ldots,s$, then we have a full rank system of equations, and thus, a unique solution.

The following are examples of densities of form (3) where $\eta(\cdot)$ is not a one to one function, and, like the logistic model, satisfy hypothesis (i) of Theorem 4 for all samples sizes.

Example 6 (Retrospective Study), Cox (1970).

We might like to estimate the conditional probability of getting cancer given a person smokes minus the conditional probability of getting cancer given a person does not smoke. The ideal way to do this would be to take a sample of both smokers and nonsmokers, follow the state of their health for a number of years, and then, check the group to see how many develop lung cancer. This is called a prospective study. In practice this can be a long and expensive process; thus, another method, called a retrospective study, is sometimes used.

In a retrospective study we take a group of lung cancer patients and a control group and check to see whether or not they smoked. We can express both studies diagramatically as follows:

Prospective Study

	non-smokers $u=0$	smokers $u=1$
no cancer $w=0$	$P(w=0\|u=0) = \frac{\pi_{00}}{\pi_{00}+\pi_{01}}$	$P(w=0\|u=1) = \frac{\pi_{10}}{\pi_{10}+\pi_{11}}$
cancer $w=1$	$P(w=1\|u=0) = \frac{\pi_{01}}{\pi_{00}+\pi_{01}}$	$P(w=1\|u=1) = \frac{\pi_{11}}{\pi_{10}+\pi_{11}}$

Retrospective Study

	no cancer $w=0$	cancer $w=1$
non-smokers $u=0$	$P(u=0\|w=0) = \frac{\pi_{00}}{\pi_{00}+\pi_{10}}$	$P(u=0\|w=1) = \frac{\pi_{01}}{\pi_{01}+\pi_{11}}$
smokers $u=1$	$P(u=1\|w=0) = \frac{\pi_{10}}{\pi_{00}+\pi_{10}}$	$P(u=1\|w=1) = \frac{\pi_{11}}{\pi_{01}+\pi_{11}}$

where $\pi_{ij} = P(u=i, w=j)$, $i,j = 0,1$. The parameter space is $\{\pi = (\pi_{00},\pi_{01},\pi_{10},\pi_{11}) \mid \pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} = 1 \ , \ 0 \le \pi_{ij} \le 1\}$; however, due to the nature of the data collection methods we may obtain meaningful estimates only for the conditional probabilities in either study.

Henceforth we confine attention to the retrospective study. Let $p_1 = P(u=0|w=0)$, $p_2 = P(u=0|w=1)$ and r_{ij} the number of observations in the ij cell, then the log likelihood for a retrospective sample is

$$\ell(p_1,p_2) = r_{00} \ln p_1 + r_{10} \ln (1-p_1) + r_{01} \ln p_2 + r_{11} \ln (1-p_2) \ .$$

Now, let

$$\eta_1(\pi) = \pi_{00}/\pi_{00} + \pi_{10}$$

and

$$\eta_2(\pi) = \pi_{01}/\pi_{01} + \pi_{11} \ .$$

The range of $\eta(\cdot) = (\eta_1(\cdot),\eta_2(\cdot))^T$ is the unit square. $\ell(\eta_1(\cdot),\eta_2(\cdot))$ is the log likelihood of the sample as a function of the parameters $(\pi_{00},\pi_{01},\pi_{10},\pi_{11})$. The unique maximum of $\ell(\cdot,\cdot)$ on the range of $\eta(\cdot)$ is $(r_{00}/(r_{00} + r_{10}),\ r_{01}/(r_{01} + r_{11}))$; thus hypothesis (i) of Theorem 4 is satisfied. However, $\eta(\cdot)$ is a many to one function for each value of its range; in fact,

$$\eta(\pi) = \eta(\pi^*)$$

if and only if $\pi_{10}/\pi_{00} = \pi^*_{10}/\pi^*_{00}$ and $\pi_{11}/\pi_{01} = \pi^*_{11}/\pi^*_{01}$; and thus, we will not be able to obtain a unique maximum likelihood estimate for π.

We might wish to ask what parametric functions are uniformly identifiable? Consider

$$h_1(\pi) = \pi_{11}/(\pi_{10} + \pi_{11}) - \pi_{01}/(\pi_{00} + \pi_{01})$$

and

$$h_2(\pi) = \ell n(\pi_{11}/\pi_{10}) - \ell n(\pi_{01}/\pi_{00})$$

$$= \ell n \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}} \quad .$$

Now, $h_1(\pi)$ is the difference of the conditional probability of cancer given a person smokes and the conditional probability of cancer given a person does not smoke, while $h_2(\pi)$ is the difference of the log odds of the two conditional probabilities.

First,

$$h_2(\pi) = \ell n\left(\frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}\right) = \ell n\left(\frac{\eta_1(\pi)}{1-\eta_1(\pi)}\right) - \ell n\left(\frac{\eta_2(\pi)}{1-\eta_2(\pi)}\right) ;$$

thus, from Corollary 8, $h_2(\cdot)$ is uniformly identifiable.

On the other hand, we will show that $\Theta_{h_1} = \{\pi | h_1(\pi) = 0\}$ is a proper subset of Θ, and hence, h_1 is not uniformly identifiable. With this objective in mind suppose $h_1(\pi_0) = 0$ and write

$$h_1(\pi) = \frac{\pi_{01}}{\pi_{01} + \pi_{00}\gamma(\pi)} - \frac{\pi_{01}}{\pi_{01} + \pi_{00}} ,$$

where

$$\gamma(\pi) = \frac{1 - \eta_1(\pi)}{\eta_1(\pi)} \frac{\eta_2(\pi)}{1 - \eta_2(\pi)} .$$

Then, $h_1(\pi) = 0$ if and only if $\gamma(\pi) = 1$. From Corollary 8, $\gamma(\cdot)$ is uniformly identifiable; thus from Theorem 15, if $\eta(\pi) = \eta(\pi_0)$ then $\gamma(\pi) = \gamma(\pi_0)$ and $h_1(\pi) = h_1(\pi_0)$. Thus h_1 is identifiable at π_0.

Now, suppose $h_1(\pi_1) \neq 0$. We want to show that π_1 is not in Θ_{h_1}. Let $\pi_1 = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ and $\pi_c = (c\pi_{00}, \pi_{01}, c\pi_{10}, \pi_{11})$, then $\eta(\pi_c) = \eta(\pi_1)$ and $\gamma(\pi_c) = \gamma(\pi_1)$. Now, $h_1(\pi_1) = h_1(\pi_c)$ says

$$\frac{\pi_{01}}{\pi_{01} + c\pi_{00}\gamma(\pi_1)} - \frac{\pi_{01}}{\pi_{01} + c\pi_{00}} = \frac{\pi_{01}}{\pi_{01} + \pi_{00}\gamma(\pi_1)} - \frac{\pi_{01}}{\pi_{01} + \pi_{00}} ,$$

$$\frac{c(1 - \gamma(\pi_1))}{(\pi_{01} + c\pi_{00}\gamma(\pi_1))(\pi_{01} + c\pi_{00})} = \frac{1 - \gamma(\pi_1)}{(\pi_{01} + \pi_{00}\gamma(\pi_1))(\pi_{01} + \pi_{00})} ,$$

and, since $h_1(\pi_1) \neq 0$, $\gamma(\pi_1) \neq 1$, so

$$(\pi_{01} + c\pi_{00}\gamma(\pi_1))(\pi_{01} + c\pi_{00}) = c(\pi_{01} + \pi_{00}\gamma(\pi_1))(\pi_{01} + \pi_{00}) .$$

Thus, if $h_1(\pi_1) = h_1(\pi_c)$, then

$$\pi_{01}^2 - c(\pi_{01}^2 + \pi_{00}^2\gamma(\pi_1)) + c^2\pi_{00}^2\gamma(\pi_1) = 0 .$$

But this last equation cannot hold for all c $(0 < c < 1)$, so there exists some c' such that $\eta(\pi_1) = \eta(\pi_{c'})$ but $h_1(\pi_1) \neq h_1(\pi_{c'})$; thus, from Theorem 15, $h_1(\cdot)$ is not identifiable at π_1 .

Example 7 (The Projectile Example). In all the preceding examples the data has been discrete. In this example we consider a problem involving continuous data.

Let X_i $(i = 1,\ldots,n)$ be independent and identically distributed observations of the distance traveled by a projectile fired at elevation θ and initial velocity v . Assuming no air resistance, the distance traveled is given by $v^2 \sin 2\theta/g$, where g is the gravitational constant.

Suppose the X_i's are exponentially distributed with mean $v^2 \sin 2\theta/g$.

Except for an additive constant, the log likelihood can be written as $\ell(\eta(v,\theta))$ where

$$\ell(\lambda) = \left(\sum_{i=1}^{n} x_i\right)\lambda - n \ell n \lambda ,$$

and

$$\eta(v,\theta) = g/v^2 \sin 2\theta .$$

The parameter space is $\{(v,\theta) \mid \theta$ is in $(0,\pi/2)$ and v is in $(0,\infty)\}$.

$\ell(\cdot)$ has a unique maximum on $(0,\infty)$, the range of $\eta(v,\theta)$, in fact

$$\hat{\lambda} = n/\sum_{i=1}^{n} x_i .$$

The function $\eta(\cdot,\cdot)$ is not one to one, so there is no unique maximum likelihood estimate of (v,θ) .

Let us now consider the first component of the terminal velocity of the projectile, a function of the parameter which might be of interest. Thus, $h_3(v,\theta) = v \cos\theta$. We show that $h_3(\cdot)$ is not uniformly identifiable by proving $\Theta_{h_3} \subseteq \{(v,\theta) \mid v > 0,\ \theta = \pi/4\}$.

Let (v^*,θ^*) be in Θ_{h_3} , we first show that

$$\{(v,\theta) \mid \eta(v,\theta) = \eta(v^*,\theta^*)\} = \{(v^*,\theta^*)\} .$$

Suppose $\eta(v',\theta') = \eta(v^*,\theta^*)$, then from Theorem 15, since (v^*,θ^*) is in Θ_{h_3} , $h_3(v',\theta') = h_3(v^*,\theta^*)$. Thus,

$$(v')^2 \sin 2\theta' = (v^*)^2 \sin 2\theta^*$$

and

$$v' \cos \theta' = v^* \cos \theta^* \ .$$

Now,

$$\left(\frac{v'}{v^*}\right)^2 = \frac{\sin 2\theta^*}{\sin 2\theta'} = \frac{\sin \theta^* \cos \theta^*}{\sin \theta' \cos \theta'}$$

and

$$\frac{v'}{v^*} = \frac{\cos \theta^*}{\cos \theta'} \ ,$$

which implies

$$\frac{v'}{v^*} = \frac{\sin \theta^*}{\sin \theta'} = \frac{\cos \theta^*}{\cos \theta'} \ ,$$

and

$$\sin(\theta^* - \theta') = \sin \theta^* \cos \theta' - \cos \theta^* \sin \theta' = 0 \ ,$$

or, since θ' and θ^* are in $(0,\pi/2)$, $\theta' = \theta^*$. Also, $v' = v^*$. Therefore, $\{(v,\theta) \mid \eta(v,\theta) = \eta(v^*,\theta^*)\} \subseteq \{(v^*,\theta^*)\}$. The reverse containment is obvious.

But this equivalence of sets implies that $\theta^* = \pi/4$, since if $\theta^* < \pi/4$, let $\theta_1 = \pi/4 + (\pi/4 - \theta^*)$ and $v^* = v_1$. Hence $\eta(v^*,\theta^*) = \eta(v_1,\theta_1)$ but $\theta^* \neq \theta_1$. A similar argument holds if $\theta^* > \pi/4$.

6. IDENTIFIABLE SETS AND TESTABLE HYPOTHESES

6.1 Identifiable Sets

In Chapter 5 we considered the problem of making inferences in a family of form (3), where $\eta(\cdot)$ is not one to one. As an example, we looked at the normal linear model of less than full rank, considering the concepts of estimable functions and testable hypotheses which have been developed for this particular case. We then showed that the uniform identifiability of Koopmans and Reiersøl (1950) is a generalization of the concept of estimable functions.

In this chapter we generalize the concept of testable hypotheses.

Let Y be a sample with density $f(y;\theta)$. A subset S of Θ is called identifiable if and only if θ_0 in S and $f(\cdot;\theta) = f(\cdot;\theta_0)$ implies θ in S. We now consider some basic properties of identifiable sets.

Let $I_S(\cdot)$ be the indicator function of S ($I_S(\theta) = 1$ for θ in S, 0 elsewhere).

Lemma 7. S is identifiable if and only if $I_S(\cdot)$ is uniformly identifiable.

Proof. The statement "$I_S(\theta_0) = 1$ and $f(\cdot;\theta) = f(\cdot;\theta_0)$ implies $I_S(\theta) = 1$" is equivalent to "$f(\cdot;\theta) = f(\cdot;\theta_0)$ implies $I_S(\theta) = I_S(\theta_0)$."

As an immediate consequence of Lemma 7, certain results for identifiable functions also hold for identifiable sets. In particular, if Y is a random sample from some density then, by applying Lemma 7 and then Theorem 12, the collection of identifiable sets is the same for all sample sizes, and we may check for identifiablity of a particular set at a particular sample size by checking at sample size 1.

Now, for m in the range $h(\cdot)$, let $S_m = \{\theta | h(\theta) = m\}$.

Lemma 8. S_m is identifiable if and only if $h(\cdot)$ is identifiable at θ for all θ in S_m.

Proof. The statement "$\theta_0 \in S_m$ and $f(\cdot;\theta) = f(\cdot;\theta_0)$ implies $\theta \in S_m$" is equivalent to "$f(\cdot;\theta) = f(\cdot;\theta_0)$ implies $h(\theta) = h(\theta_0)$, for all $\theta_0 \in S_m$."

Corollary 10. S_m is identifiable for all m in $h(\Theta)$ if and only if $h(\cdot)$ is uniformly identifiable.

Example 8 (Retrospective Study - Continued). In Example 6 it was shown that $h_1(\cdot)$ is not uniformly identifiable. It was also shown that

$$\Theta_{h_1} = \{\pi | h_1(\pi) = 0\} .$$

Thus, for $h_1(\cdot)$, S_m is identifiable for $m = 0$ but is not identifiable for any other value of m.

6.2 Identifiable Sets and the Exponential Family

Let $g(x;\theta)$ be a density of form (3) and S be a subset of Θ. Throughout this section we will restrict our attention to $g(x;\theta)$ in canonical form.

Theorem 19. Suppose $g(x;\theta)$ is in canonical form, then S is identifiable if and only if θ_0 in S and $\eta(\theta) = \eta(\theta_0)$ implies θ is in S.

Proof. Apply Lemma 6 and the definition of "S is identifiable."

Corollary 11. S is identifiable if and only if $\eta(S) \cap \eta(S^c) = \phi$.

Proof. The statement "$\eta(S) \cap \eta(S^c) = \phi$" is equivalent to "$\theta_0$ in S and $\eta(\theta) = \eta(\theta_0)$ implies θ is in S."

Let G be a subset of $\eta(\Theta)$ then $\eta^{-1}(G)$ will denote $\{\theta | \eta(\theta) \text{ is in } G\}$.

Theorem 20. Suppose S is identifiable and R is not, then

$$\eta^{-1}(\eta(S \cap R)) = S \cap \eta^{-1}(\eta(R)) .$$

Proof. The statement "$\eta^{-1}(\eta(S \cap R)) \subseteq S \cap \eta^{-1}(\eta(R))$" is equivalent to "$\eta(\theta) = \eta(\theta_0)$ and θ_0 in $S \cap R$ implies θ is in $S \cap \eta^{-1}(\eta(R))$", which is true from Theorem 19 and

the definition of $\eta^{-1}(\eta(R))$.

The statement "$\eta^{-1}(\eta(S \cap R)) \supseteq S \cap \eta^{-1}(\eta(R))$" is equivalent to "$\eta(\theta) = \eta(\theta_0)$, θ in S and θ_0 in R implies θ in $\eta^{-1}(\eta(S \cap R))$", which is true from Theorem 19 and the definition of $\eta^{-1}(\eta(S \cap R))$.

Theorem 21. Let I be the collection of identifiable subsets of Θ . I is closed under intersection, union and complement.

Proof. Let S and S_λ , λ in Δ , be in I.

First, $\bigcap_{\lambda\in\Delta} S_\lambda$ is in I . Suppose θ_0 is in $\bigcap_{\lambda\in\Delta} S_\lambda$ and $\eta(\theta) = \eta(\theta_0)$. Then, for all λ in Δ , θ_0 is in S_λ and $\eta(\theta) = \eta(\theta_0)$. Thus, from Theorem 19, θ is in $\bigcap_{\lambda\in\Delta} S_\lambda$ and $\bigcap_{\lambda\in\Delta} S_\lambda$ is identifiable.

Second, $\bigcup_{\lambda\in\Delta} S_\lambda$ is in I . Suppose θ_0 is in $\bigcup_{\lambda\in\Delta} S_\lambda$ and $\eta(\theta) = \eta(\theta_0)$. Then, for some λ in Δ , θ_0 is in S_λ and $\eta(\theta) = \eta(\theta_0)$; thus, from Theorem 19, θ is in $\bigcup_{\lambda\in\Delta} S_\lambda$ and $\bigcup_{\lambda\in\Delta} S_\lambda$ is identifiable.

Finally, as a direct consequence of the definition of "S is identifiable" we see that S^c is identifiable.

Let $h_i(\cdot)$, $i = 1,\ldots,\ell$, be functions of θ . Using Theorem 21 we may extend the results of Lemma 8 as follows.

Theorem 22. Let $S_{m_i} = \{\theta | h_i(\theta) = m_i\}$, $i = 1,\ldots,\ell$. $\bigcap_{i=1}^{\ell} S_{m_i}$ is identifiable for all $(m_1,\ldots,m_\ell)$, m_i in $h_i(\theta)$, if and only if $h_i(\cdot)$ is uniformly identifiable for all i .

Proof. Suppose $\bigcap_{i=1}^{\ell} S_{m_i}$ is identifiable for all $(m_1,\ldots,m_\ell)$. We show that, for every i, S_{m_i} is identifiable for all m_i in $h_i(\theta)$; thus, from Corollary 10, $h_i(\cdot)$ is uniformly identifiable for all i . Let $1 \le j \le \ell$, m_j' be a fixed value of m_j , and $\Delta_j = \{(m_1,\ldots,m_n) | m_j = m_j'\}$. Now, since $\bigcap_{i=1}^{\ell} S_{m_i}$ is identifiable for all $(m_1,\ldots,m_\ell)$, from Theorem 21,

$$\bigcup_{(m_1,\ldots,m_\ell)\in\Delta_j} \bigcap_{i=1}^{\ell} S_{m_i}$$

is identifiable. However,

$$\bigcup_{(m_1,\ldots,m_\ell)\in\Delta_j} \bigcap_{i=1}^{\ell} S_{m_i} = S_{m_j'} \cap \left(\bigcap_{\substack{i=1 \\ i\neq j}}^{\ell} \left(\bigcup_{m_i \in h_i(\theta)} S_{m_i} \right) \right) ,$$

$$= S_{m_j'} ,$$

since, for any i , $\bigcup_{m_i \in h_i(\theta)} S_{m_i} = \theta$. So, $S_{m_j'}$ is identifiable.

Conversely, suppose $h_i(\cdot)$ is uniformly identifiable for all i. Then, from Corollary 10 and Theorem 21, $\bigcap_{i=1}^{n} S_{m_i}$ is identifiable for all $(m_1,\ldots,m_n)$.

6.3 Generalizing Testable Hypotheses

Let $S = \{\beta \mid \lambda_i^T\beta = m_i,\ i = 1,\ldots,\ell\}$ where β and σ^2 are the parameters of a normal linear model. Now, $\eta(S) = \{(\beta^T X^T/\sigma^2 : -1/2\sigma^2) \mid \beta \text{ is in } S,\ \sigma^2 > 0\}$; thus, $\eta(S) \cap \eta(S^C) = \phi$ if and only if $XS \cap XS^C = \phi$. Therefore, from Theorem 8 and Corollary 11, S is identifiable if and only if $\lambda_i^T\beta$ is estimable for all i, so the hypothesis H, stating that β is in S, is testable if and only if S is identifiable.

We generalize the concept of a testable hypothesis as follows: Let Y be a sample with density $f(y;\theta)$. The hypothesis H, θ in S, is <u>testable</u> if and only if S is identifiable.

To see the importance of this definition, let ~H (read not H) state that θ is in S^C. If S is not identifiable then there is a θ_0 in S and a θ_1 in S^C such that $f(\cdot;\theta_0) = f(\cdot;\theta_1)$; thus, the set of distributions under H and ~H are not disjoint. This means that based on observed Y we cannot say whether H or ~H is true.

Certain hypothesis testing results for the normal linear model will extend to densities $g(x;\theta)$ of form (3) in canonical form.

Searle (1971) shows that in the normal linear model the sum of squares used in testing a hypotheses with some estimable components and some non-estimable components is the same as the sum of squares for testing that same hypothesis but with the non-estimable components deleted. We may extend Searle's result in the following way.

Suppose that S and R are both subsets of θ, S being identifiable and R not. Let H state that θ is in $S \cap R$ and H' state that θ is in $S \cap \eta^{-1}(\eta(R))$.

Theorem 24. The maximum likelihood ratio statistic testing H versus $\sim H$ is the same as testing H' versus $\sim H'$.

Proof. Let $x_1,\ldots,x_n$ be an observed sample from $g(x;\theta)$. The maximum likelihood ratio statistic testing H versus $\sim H$ is

$$\frac{\sup_{\theta \in S \cap R} \prod_{i=1}^{n} g(x_i;\theta)}{\sup_{\theta \in \Theta} \prod_{i=1}^{n} g(x_i;\theta)} .$$

From Theorems 3 and 20,

$$\sup_{\theta \in S \cap R} \prod_{i=1}^{n} g(x_i;\theta) = \sup_{\theta \in S \cap R} \prod_{i=1}^{n} f_Y(\psi(x_i);\eta(\theta))$$

$$= \sup_{\theta \in \eta^{-1}\eta(S \cap R)} \prod_{i=1}^{n} f_Y(\psi(x_i);\eta(\theta))$$

$$= \sup_{\theta \in S \cap \eta^{-1}(\eta(R))} \prod_{i=1}^{n} g(x_i;\theta)$$

which is the numerator of the maximum likelihood ratio statistic testing H' versus $\sim H'$. The denominator is the same for testing both hypotheses.

To see that Theorem 24 is truly a generalization of Searle (1971), let $S = \{\beta | \lambda_i^T \beta = c_i,\ i = 1,\ldots,j\}$ and $R = \{\beta | \lambda_i^T \beta = c_i,\ i = j+1,\ldots,\ell\}$ where $\lambda_i^T \beta$ is estimable for $i = 1,\ldots,j$, and not estimable for $i = j+1,\ldots,\ell$. From Theorem 24 the sum of squares for H versus $\sim H$ is the same as the sum of squares for H' versus $\sim H'$. Now,

$$\eta^{-1}(\eta(R)) = \{(\beta^T : \sigma^2)^T \mid \beta \text{ is in } X^{-1}(X(R)) \text{ and } \sigma^2 > 0\}.$$

From Theorem 9, $X^{-1}(X(R)) = E^P$; thus, $\eta^{-1}(\eta(R)) = \Theta$, so $S \cap \eta^{-1}(\eta(R)) = S$. Thus, the sum of squares for testing β in $S \cap R$ is the same as the sum of squares for testing β in S.

Example 9 (Logistic Life Study Model - Continued).

Thompson (1976) discusses the use of covariates in the analysis of life table data, introducing a logistic model for the conditional probability of failure in a time interval given survival to the beginning of the interval. An example is given using the following data:

Table 1

Times of Remission (weeks) of Leukemia Patients

(Gehan (1965), from Freireich et. al.)

Sample 0 (drug 6-MP)	6*,6,6,6,7,9*,10*,10,11*,13,16,17* 19*,20*,22,23,25*,32*,32*,34*,35*
Sample 1 (control)	1,1,2,2,3,4,4,5,5,8,8,8,8,11,11, 12,12,15,17,22,23

*Censored

Here, the covariate effect is containment in sample 0 or sample 1. The conditional probability that individual i fails in interval j given survival to the beginning of the interval is represented as

$$\frac{\exp(z_{ij0}\beta_0 + z_{ij1}\beta_1 + \eta_j)}{1 + \exp(z_{ij0}\beta_0 + z_{ij1}\beta_1 + \eta_j)}$$

where, $z_{ijk} = 1$ if the i^{th} individual is in the k^{th} class, 0 otherwise. β_0 and β_1 are control and treatment effects respectively and η_j is the effect of the j^{th} interval.

We wish to test the hypothesis H^* , $\beta_0 = \beta_1$, of equal drug and control effect. Let $S = \{\beta | \beta_0 = \beta_1\}$, here $\beta = (\beta_0, \beta_1, \eta_1, \ldots, \eta_\ell)^T$. From Lemma 8, S is identifiable if and only if $\beta_0 - \beta_1$ is identifiable at β for all β in S . The density of the logistic life study model is of form (3), in canonical form, $\eta(\beta) = Z\beta$, and $\Theta = E^{\ell+2}$, thus, from Corollary 9, $\beta_0 - \beta_1$ is identifiable at β if and only if $\beta_0 - \beta_1$ is uniformly identifiable. From Corollary 8, $\beta_0 - \beta_1$ is uniformly identifiable if and only if $(1,-1,0,\ldots,0)$ is in $\text{Row}(Z)$. This will be the case if there is any interval with at least one member from each class at risk in the interval. The data in Table 1 shows 21 members from class 0 and 21 members from class 1 at risk in the first interval. Thus H^* is testable.

We show that the likelihood ratio test of H^* is the same as that of H , $\beta_0 = \beta_1 = 0$. Write

$$\{\beta | \beta_0 = \beta_1 = 0\} = \{\beta | \beta_0 - \beta_1 = 0\} \cap \{\beta | \beta_0 + \beta_1 = 0\} .$$

Now $(1,1,0,\ldots,0)$ is not in $\text{Row}(Z)$, $\beta_0 + \beta_1$ is not uniformly identifiable, and $\{\beta | \beta_0 + \beta_1 = 0\}$ is not identifiable. From Theorem 24, the likelihood ratio test of H is the same as that of H' , β is in $\{\beta | \beta_0 - \beta_1 = 0\} \cap Z^{-1}(Z\{\beta | \beta_0 + \beta_1 = 0\})$. But $Z^{-1}(Z\{\beta | \beta_0 + \beta_1 = 0\}) = E^{\ell+2}$ so $H' = H = H^*$.

We prove this last by showing that $Z\{\beta|\beta_0+\beta_1\neq 0\} \subseteq Z\{\beta|\beta_0+\beta_1=0\}$. Let β^* be such that $\beta_0^*+\beta_1^*\neq 0$. There is $\beta^\dagger$, with $\beta_0^\dagger+\beta_1^\dagger=0$, such that $Z\beta^\dagger = Z\beta^*$. Consider the equations, in the variable β ,

$$\beta_0+\beta_1 = -\beta_0^*-\beta_1^*$$

and

$$Z\beta = 0 .$$

These equations have a solution, β', since $(1,1,0,\ldots,0)$ is not in $\mathrm{Row}(Z)$. $\beta^\dagger = \beta^*+\beta'$.

In summary, we have shown that $H^*,\beta_2 = \beta_1$ is testable and that its likelihood ratio test is the same as that for H , $\beta_0 = \beta_1 = 0$.

7. CONCLUSION

We have considered problems of identification which arise in making inference about the exponential family when the density is not in one to one correspondence with the parameter space. Such problems logically precede all questions of inference. Using data we cannot hope to distinguish between two parametric values corresponding to the same density.

One can assume this problem away by a reparameterization, but in doing so, the physical meaning associated with the parameters might be lost.

As a guiding example we considered the normal linear model of less than full rank discussing the concepts of estimable function and testable hypothesis. Many of the classic properties proved there have been extended to the general exponential family through the ideas of uniformly identifiable function and identifiable set.

These general ideas are illustrated with several numerical and computational examples: i) a Poisson model for the analysis of some data on the survival of bacteria after radiation, ii) a logistic life study model, iii) analysis of a retrospective study of cancer and smoking and iv) a physical example involving terminal velocity of a projectile. It is found that some parametric

questions simply cannot be answered from data, for the data contains no information about them, and sometimes two questions cannot be distinguished from one another using data. Other parametric questions can reasonably be asked and answered in a data analysis sense.

BIBLIOGRAPHY

Berk, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models, Annals of Mathematical Statistics, Vol. 43, pp. 193-204.

Charnes, A., Frome, E. L., Yu, P. L. (1975). The Equivalence of Iterative Weighted Least Squares and Maximum Likelihood Estimates in the Exponential Family. Center for Cybernetic Studies, The University of Texas, Austin. Research Report number 237.

Cox, D. R. (1970). The Analysis of Binary Data. London, Methuen.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. Biometrika, Vol. 52, pp. 203-224.

Huzurbazar, V. S. (1949). On a property of distributions admitting sufficient statistics. Biometrika, Vol. 36, p. 71.

Jennrich, R. I., Moore, R. H. (1975). Maximum Likelihood Estimation and Nonlinear Least Squares. Health Sciences Computing Facility, University of California, Los Angeles. Technical Report 9.

Koopmans, T. C. and Reiersøl, O. (1950). The Identification of Structural Characteristics. The Annals of Mathematical Studies, Vol. 21, pp. 165-181.

Lehmann, E. L. (1959). Testing Statistical Hypotheses, Wiley.

Nobel, R. (1969). Applied Linear Algebra, Prentice-Hall.

Rao, C. R. (1965). Linear Statistical Inference and its Application. Wiley.

Reiersøl, O. (1963). Identifiability, Estimability, Phenorestricting Specifications and Zero Lagrange Multipliers in the Analysis of Variance. Skandinavish Aktuarietidskrift, Vol. 46, pp. 131-142.

Roberts, A. W. and Varberg, D. E. (1973). Convex Functions, Academic Press.

Royden, H. L. (1968). Real Analysis, Macmillian.

Searle, S. R. (1971). Linear Models, Wiley.

Seely, J. (1977). Estimability and the Linear Hypothesis, The American Statistician, Vol. 31, #3, p. 121.

Thompson, W. A., Jr. (1976). On the Treatment of Grouped Observations in Life Studies. Department of Statistics, University of Missouri-Columbia. Technical Report number 57.

Thompson, W. A., Jr. (1977). On the Treatment of Grouped Observations in Life Studies, Biometrics, Vol. 33, pp. 463-470.

Zehna, P. W. (1966). Invariance of Maximum Likelihood Estimation. Annals of Mathematical Statistics, Vol. 37, p. 744.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 78-2	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Identification Problems in the Exponential Family		5. TYPE OF REPORT & PERIOD COVERED Interim Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert Patrick Kelley		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C0443 (NR042 282)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics University of Missouri-Columbia		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE AUG 1978 13. NUMBER OF PAGES 67
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of the document is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Exponential family, identifiable, estimable, testable

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We consider identification problems for members of the exponential family, applying the results to the density of the logistic model for life table data.

For this model, we want to know the following:

1. When are the maximum likelihood estimates for the model parameters unique; and

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. 2. What type of inferences may be made if the maximum likelihood estimates are not unique?

Noting the form of the likelihood for the logistic model, question 1 is considered for the exponential family. We obtain an answer for question 1 in this context. Applying this answer to the logistic model, we find that a unique maximum likelihood estimate exists if and only if the density is in one to one correspondence with its parameter space.

To answer question 2, we consider members of the exponential family where the density is not in one to one correspondence with its parameter space. As a guiding example of such a density, we consider the normal linear model of less than full rank, discussing the concepts of estimable function and testable hypothesis, which have been developed for this particular case. We then show that the concept of uniform identifiability is a generalization of the concept of an estimable function. Further, through the idea of an identifiable set, we extend the concept of a testable hypothesis. We then apply the resulting theory to the logistic model.