SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NY 14853

Technical Report #703

June 1986

ESTIMATING LOGISTIC REGRESSION PROBABILITIES

by

Diane E. Duffy & Thomas J. Santner

I. INTRODUCTION

The standard approach for making inference in binary response regression is based on likelihood methods. For example, regression parameters are estimated by maximum likelihood and hypotheses are tested by likelihood ratio tests. This paper describes alternatives to likelihood based methods. Attention is restricted to estimation although analogous procedures can be developed for other statistical problems.

The observed data are (Y_i, x_i), $i = 1(1)n$ where Y_i are independent Bernoulli (response) random variables and $x_i' = (x_{i1}, \ldots, x_{ip})$ is a vector of p nonstochastic explanatory variables. All vectors in this paper are column vectors and prime denotes transpose. It will be assumed that the $n \times p$ matrix X whose ith row is x_i' has full column rank p and $p \leq n$. The outcome $[Y_i = 1]$ ($[Y_i = 0]$) is referred to as a "success" ("failure") throughout.

We assume the logistic regression model

$$P[Y_i = 1] = \frac{\exp\{x_i'\beta\}}{1 + \exp\{x_i'\beta\}} \equiv \pi_i(\beta) \qquad (1.1)$$

where $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients. This is the most widely applied model for binary regression data. Some of the subject areas where it has been used are epidemiology (Breslow and Day [4]), bioassay (Finney [9]), medicine (Brown [6]), market research (McFadden [16]), and criminology (Larntz [14]). The model's importance is confirmed by the numerous

analogues of linear model extensions that have been developed for it including (i) regression diagnostics (Pregibon [18]), (ii) errors-in-variables (Stefanski and Carroll [21]) and (iii) random effects versions (Pierce and Sands [17]).

This paper considers the problem of estimating the vector of success probabilities $\pi(\beta) = (\pi_1(\beta),\ldots,\pi_n(\beta))$. In many biomedical applications $[Y_i = 1]$ corresponds to the presence of some condition or the success of a treatment. When it is important to discriminate subjects based on the probability of this event, the accurate estimation of π is paramount.

In other applications it may be more appropriate to focus attention on the estimation of β. The logistic regression coefficient β_j has a well-known interpretation as the difference between the log odds ratios of success for two subjects with identical covariate vectors except for their x_j values which differ by one unit. While the problems of estimating $\pi(\beta)$ and β are related, the focus here is on the former. We use the notation $Y \equiv (Y_1,\ldots,Y_n)'$ and $1 + e^{X\beta} \equiv (1 + e^{x_1'\beta},\ldots,1 + e^{x_n'\beta})$ throughout.

The parameter space for the problem is taken to be $\Theta \equiv \{\pi(\beta) : \beta \in \mathbb{R}^p\}$. The action space is $A = \bar{\Theta}$ the closure of Θ under the topology of pointwise convergence of sequences. Thus A is compact which guarantees that certain estimators discussed later exist.

We restrict attention to nonrandomized estimators. However, the usual decision theoretic complete class justification for this restriction does not apply because A is not convex in our

formulation. Changing A to be $[0,1]^n$ would alleviate this problem but would permit estimates outside the model (1.1) which we choose not to allow. Lastly, we take the loss corresponding to action $a \in A$ and state of nature $\pi \in \Theta$ to be squared error loss

$$L(\pi,a) = \|\pi - a\|^2 \equiv \sum_1^n (\pi_i - a_i)^2.$$

There are other appealing losses for this problem and some comments will be made in Section 5 about the performance of the maximum likelihood (mle) and other estimators under weighted squared error loss

$$L_w(\pi,a) = \sum_1^n \frac{(\pi_i - a_i)^2}{\pi_i(1-\pi_i)}. \tag{1.2}$$

An important extension of model (1.1) which will not be carried out here for ease of exposition is to allow polychotomous responses. References will be cited where appropriate which consider this more general case; Duffy [8] contains the details of such a formulation.

The remainder of the paper is organized as follows. Section 2 reviews the mle of π and its properties. Section 3 introduces Bayes and related estimators and Section 4 considers an empirical ridge-type estimator. The paper concludes with results from a simulation study of the small sample properties of the different estimators.

II. MAXIMUM LIKELIHOOD ESTIMATION

The likelihood function corresponding to $y \in \mathcal{Y} \equiv \{0,1\}^n$ is

$$\mathcal{L}(\pi;y) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i};$$

$\mathcal{L}$ is continuous in $\pi \in [0,1]^n$ for all $y \in \mathcal{Y}$. Thus for each $y \in \mathcal{Y}$ there exists $\hat{\pi}^M = \hat{\pi}^M(y) \in A = \overline{\Theta}$ satisfying

$$\mathcal{L}(\hat{\pi}^M;y) = \sup_{\overline{\Theta}} \mathcal{L}(\pi;y). \tag{2.1}$$

Further, $\hat{\pi}^M(y)$ is unique for all $y \in \mathcal{Y}$. There are two cases to analyze: y for which the right hand sup in (2.1) is attained in (i) Θ or (ii) $\overline{\Theta}\backslash\Theta$. One convenient way to differentiate these cases is to first define for $y \in \mathcal{Y}$, $C(y) = \{\beta \in \mathbb{R}^P: (2y_i - 1)x_i'\beta > 0,\ 1 \le i \le n\}$ and $Q(y) = \{\beta \in \mathbb{R}^P: (2y_i - 1)x_i\beta \ge 0,\ 1 \le i \le n\}$. In words, $C(y)$ are those β completely separating y in the sense that $x_i'\beta > (<)\ 0$ for $Y_i = 1$ (0) or equivalently $\pi_i(\beta) > (<)\ 1/2$ for $Y_i = 1$ (0). A similar interpretation holds for $Q(y)$. For any $y \in \mathcal{Y}$, $0 \in Q(y)$ and $0 \notin C(y) \subset Q(y)$ by definition. Following Albert and Anderson [1], we define a partition of $\mathcal{Y} \equiv \mathcal{Y}_C \cup \mathcal{Y}_Q \cup \mathcal{Y}_O$ by $\mathcal{Y}_C = \{y \in \mathcal{Y} : C(y) \neq \emptyset\}$, $\mathcal{Y}_Q = \{y \in \mathcal{Y} : \emptyset = C(y),\ Q(y) \neq \{0\}\}$, and $\mathcal{Y}_O = \{y \in \mathcal{Y} : Q(y) = \{0\}\}$. The sets $\mathcal{Y}_C, \mathcal{Y}_Q$, and $\mathcal{Y}_O$ contain <u>completely separable</u>, <u>quasiseparable</u> and <u>overlapped</u> outcomes, respectively.

The second ingredient required to deduce the uniqueness of $\hat{\pi}^M(y)$ is the observation that Θ is homeomorphic to $\mathbb{R}^P$ via $\pi(\beta)$ and thus the log likelihood can be written as

$$\ell(\beta;y) = \sum_{i=1}^{n} \{y_i x_i'\beta - \ell n(1 + \exp[x_i\beta])\}$$

$$= Y'X\beta - 1_n'\ell n(1 + \exp[X\beta]).$$

where 1_n is a vector of 1's of length n and functions of vector quantities denote componentwise operations. It follows from the strict concavity of $\ell(\beta;y)$ that there exists at most one $\beta \in \mathbb{R}^P$ maximizing $\ell(\beta;y)$.

It is well known (Haberman [11]; Silvapulle [20]; Albert and Anderson [1]; Santner and Duffy, [19]) that $\ell(\beta;y) \to -\infty$ as $\|\beta\| \to \infty \iff y \in \mathcal{Y}_0$ (and thus there exists a unique $\hat{\pi}^M(y) \in \Theta$ in this case).

When $y \in \mathcal{Y}_C \cup \mathcal{Y}_Q$, $\mathcal{L}(\pi;y) < \sup\{\mathcal{L}(\pi;y) : \pi \in \overline{\Theta}\}$ for every $\pi \in \Theta$. The continuity of $\mathcal{L}(\pi;y)$ in π and the strict concavity of $\ell(\beta;y)$ imply that there is a unique $\hat{\pi}^M \in \overline{\Theta}\backslash\Theta$. A bit more can be stated about the character of $\hat{\pi}^M(y)$. If $y \in \mathcal{Y}_C$ then $\hat{\pi}^M(y) = y$ and $\sup\{\mathcal{L}(\pi;y) : \pi \in \overline{\Theta}\} = 1$. To see this choose $\beta \in C(y)$, then $k\beta \in C(y)$ for $k > 0$, $\pi_i(k\beta) \to 1$ (0) for $y_i = 1$ (0) and $\mathcal{L}(\pi(k\beta);y) \to 1$ as $k \to \infty$. When $y \in \mathcal{Y}_Q$ then $\hat{\pi}_i^M(y) = y_i$ for some components, $0 < \hat{\pi}_i^M(y) < 1$ for other components, and $0 < \sup\{\mathcal{L}(\pi;y) : \pi \in \overline{\Theta}\} < 1$. The proof of this is more complicated and only those elements required for later work will be introduced (see Santner and Duffy [19] for full

details). It is possible to choose (i) $\beta^1 \in \mathbb{R}^P$ which yields a maximal set M of indices i, $1 \le i \le n$ for which $x_i'\beta > (<)\ 0$ as $y_i = 1\ (0)$; and (ii) $\beta^2 \in \mathbb{R}^P$ which satisfies $\ell_R(\beta^2;y) = \sup\{\ell_R(\beta;y) : \beta \in \mathbb{R}^P\}$ where

$$\ell_R(\beta;y) = \sum_{i \notin M} \left\{ y_i x_i'\beta - \ell n(1 + \exp[x_i'\beta]) \right\}$$

is a logistic log likelihood for observations $\{Y_i : i \notin M\}$. For $i \in M$, $\pi_i(k\beta^1 + \beta^2) \to 1\ (0)$ as $y_i = 1\ (0)$ and $k \to \infty$ while

$$\begin{aligned} \sup_{\mathbb{R}^P} \ell(\beta;y) &\ge \lim_{k\to\infty} \ell(k\beta^1 + \beta^2;y) \\ &= \ell_R(\beta^2;y) \\ &= \sup_{\mathbb{R}^p} \ell_R(\beta;y) \ge \sup_{\mathbb{R}^p} \ell(\beta;y). \end{aligned}$$

Thus $\hat{\pi}^M(y) = y_i$ for $i \in M$, $0 < \hat{\pi}_i^M(y) = \pi_i(\beta^2) < 1$ for $i \notin M$, and $0 < \sup\{\ell(\hat{\pi};y) : \hat{\pi} \in \overline{\Theta}\} < 1$.

An immediate by-product of the concavity of $\ell(\beta;y)$ over $\beta \in \mathbb{R}^P$ is that for $y \in \mathcal{Y}_0$ we have $\hat{\pi}^M = \hat{\pi}(\beta^M)$ where β^M is the (unique) solution of

$$\nabla\ell(\beta^M;y) = X'(y - \hat{\pi}(\beta^M)) = 0. \qquad (2.2)$$

Equation (2.2) is the analogue of the normal equations; $y - \hat{\pi}^M$ is orthogonal to the column space of X. Furthermore the mle $\hat{\pi}^M$ solves (2.2) in general.

Theorem 2.1. For any $y \in \mathcal{Y}$ the mle $\hat{\pi}^M(y)$ satisfies

$$X'(y - \hat{\pi}^M(y)) = 0. \tag{2.3}$$

Proof. The previous paragraph shows for $y \in \mathcal{Y}_O$ $\hat{\pi}^M$ is in Θ and satisfies (2.3). When $y \in \mathcal{Y}_C$, $\hat{\pi}^M(y) = y$ trivially satisfies (2.3). Lastly if $y \in \mathcal{Y}_Q$ then

$$\begin{aligned} X'y &= \sum_{i \in M} x_i' y_i + \sum_{i \notin M} x_i' y_i \\ &= \sum_{i \in M} x_i' \pi_i^M + \sum_{i \notin M} x_i' y_i \\ &= \sum_{i \in M} x_i' \pi_i^M + \sum_{i \notin M} x_i \pi_i^M \end{aligned}$$

where the last equality holds since $\nabla \ell_R(\beta^2; y) = 0$.■

In addition to its normal equation interpretation, $\hat{\pi}^M(y)$ satisfies the invariance property

$$\hat{\pi}^M(y) = 1 - \hat{\pi}^M(1 - y), \; y \in \mathcal{Y}. \tag{2.4}$$

To prove (2.4) it suffices to show that $1 - \hat{\pi}^M(y) \in \overline{\Theta}$ for all $y \in \mathcal{Y}$ since the uniqueness of $\hat{\pi}^M(y)$ implies

$$\begin{aligned} & X'(y - \hat{\pi}^M(y)) = 0 \\ \Leftrightarrow\; & X'(1 - y - 1 + \hat{\pi}^M(y)) = 0 \\ \Rightarrow\; & \hat{\pi}^M(1 - y) = 1 - \hat{\pi}^M(y). \end{aligned}$$

For $y \in \mathcal{Y}_O$, $\hat{\pi}^M(y) = \hat{\pi}(\beta^M) \Rightarrow 1 - \hat{\pi}^M(y) = 1 - \hat{\pi}(\beta^M) = \hat{\pi}(-\beta^M) \in \Theta$. Else, $y \in \mathcal{Y}_C \cup \mathcal{Y}_Q$, $\hat{\pi}^M(y) \in \overline{\Theta} \backslash \Theta$ and there exists a sequence

$\{\beta^k\} \subset \mathbb{R}^p$ with $\hat{\pi}^M(y) = \lim_{k\to\infty} \pi(\beta^k)$. Then $1 - \hat{\pi}^M(y) = \lim_{k\to\infty} (1 - \pi(\beta^k)) = \lim_{k\to\infty} \pi(-\beta^k) \in \overline{\Theta}$ where the convergence of $\pi(-\beta^k)$ is an easy consequence of the convergence of $\pi(\beta^k)$.

Remark 2.1. The mle $\hat{\pi}^M$ in the multinomial response case satisfies a similar gradient interpretation and invariance property.■

We next study the risk of $\hat{\pi}^M$, beginning with a simple example that illustrates some features of the risk curve. Some comments are made about the behavior of the risk in the general case and we conclude with a summary of its admissibility properties.

Example 2.1. Suppose $Y_1,\ldots,Y_n$ are iid Bernoulli observations and, in an abuse of notation, let $\pi = P[Y_i = 1]$ be the success probability for each Y_i. Assuming $\pi \in (0,1)$, define $\beta \equiv \text{logit}(\pi) \in \mathbb{R}^1$. The mle is $\hat{\pi}^M(y) = \bar{y}$ and its risk is $R(\pi,\bar{Y}) = \pi(1 - \pi)$. Thus $\hat{\pi}^M$ has minimal risk at the extreme points $\pi = 0$ and 1 and maximum risk at $\pi = 1/2$.■

In the general case it is convenient to regard the risk as a function of $\beta \in \mathbb{R}^p$ and write

$$R(\beta,\hat{\pi}^M) = \sum_{y} \left\{ \|\pi(\beta) - \hat{\pi}^M(y)\|^2 \exp\left[\sum_{i=1}^{n} y_i x_i'\beta\right]/D(\beta) \right\} \qquad (2.5)$$

where $D(\beta) = \pi_{i=1}^{n} \{1 + \exp[x_i'\beta]\}$.

Note that $R(\beta;\hat{\pi}^M) = R(-\beta,\hat{\pi}^M)\ \forall\ \beta \in \mathbb{R}^p$ by the invariance (2.4) and the invariance of the underlying distribution when $y \to 1 - y$ and $\beta \to -\beta$.

The value $\beta = 0$ corresponds to $\hat{\pi}_i(\beta) \equiv 1/2$ with risk $R(0,\hat{\pi}^M) = \sum_{\mathcal{Y}} \|1_n(1/2) - \hat{\pi}^M(y)\|^2/2^n$. Here 1_n is a vector of 1's of length n. We show that the risk always has a stationary point at $\beta = 0$ as in Example 2.1. Calculation gives

$$\frac{\partial R(\beta,\hat{\pi}^M)}{\partial \beta_j} = \frac{1}{D(\beta)} \sum_{\mathcal{Y}} \exp\left[\sum_{i=1}^n y_i x_i'\beta\right] \left\{[2(\pi(\beta) - \hat{\pi}^M(y))'\Lambda\xi_j] + \|\pi(\beta) - \hat{\pi}^M(y)\|^2[(y - \pi(\beta))'\xi_j]\right\},\ 1 \le j \le p$$

where Λ is an n x n diagonal matrix with ith diagonal element $\pi_i(\beta)[1 - \pi_i(\beta)]$ and $\xi_1,\ldots,\xi_p$ are the columns of X. Substituting $\xi_j'y = \xi_j'\hat{\pi}^M(y)$ from Theorem 2.1, $\beta = 0$ and rearranging terms shows

$$\frac{\partial R(0,\hat{\pi}^M)}{\partial \beta_j} = \frac{1}{2^n} \sum_{\mathcal{Y}} \left\{\frac{\xi_j'1_n}{4} - \frac{y'\xi_j}{2} + \|(\tfrac{1}{2})1_n - \hat{\pi}^M(y)\|^2(y - (\tfrac{1}{2})1_n)'\xi_j\right\}$$

$$= \frac{1}{2^n}\left\{\sum_{\mathcal{Y}} \|(\tfrac{1}{2})1_n - \hat{\pi}^M(y)\|^2 y'\xi_j - \left[\frac{\xi_j'1_n}{2}\right] \sum_{\mathcal{Y}} \|(\tfrac{1}{2})1_n - \hat{\pi}^M(y)\|^2\right\} \quad (2.6)$$

Decomposing $\mathcal{Y} = \mathcal{Y}_1 \cup \mathcal{Y}_0$ where $\mathcal{Y}_i = \{y \in \mathcal{Y} : y_1 = i\}$ and using the equalities $\mathcal{Y}_0 = \{1 - y : y \in \mathcal{Y}_1\}$ and $\|1_n(1/2) - \hat{\pi}^M(y)\|^2 = \|1_n(1/2) - (1_n - \hat{\pi}^M(1 - y))\|^2$ shows that the right hand side of (2.6) is

$$\frac{1}{2^n}\left\{[\xi_j' 1_n - \xi_j' 1_n] \sum_{\mathcal{Z}_1} \|1_n(1/2) - \hat{\pi}^M(y)\|^2\right\} = 0.$$

It is an open question as to whether the stationary point is a (local) maximum for all X.

We next consider the behavior of $R(\beta,\hat{\pi}^M)$ for "extreme" $\pi(\beta)$ or more precisely as $\|\beta\| \to \infty$. In brief, the risk $R(\beta,\pi^M)$ converges to zero in those β directions for which all components of $\pi(\beta)$ converge to 0 or 1 as in Example 2.1, and the risk converges to a positive value in β directions where $\pi_i(\beta) = 1/2$ (i.e. $x_i'\beta = 0$) for one or more indices.

To make this precise, decompose the unit sphere $\mathcal{B} = \{\beta \in \mathbb{R}^P : \|\beta\| = 1\}$ into $\mathcal{B}_C = \{\beta \in \mathcal{B} : \beta \in C(y)$ for some $y \in \mathcal{Y}\}$ and $\mathcal{B}_Q = \{\beta \in \mathcal{B} : \beta \notin C(y)$ for any $y \in \mathcal{Y}\}$; then $\mathcal{B} = \mathcal{B}_C \cup \mathcal{B}_Q$ with $\mathcal{B}_C \cap \mathcal{B}_Q = \emptyset$. The latter is obvious, while the former follows because $X\beta \neq 0$ for any $\beta \in \mathcal{B}$, hence $\beta \in \mathcal{B}_C$ if $x_i'\beta \neq 0$ for all i, $1 \leq i \leq n$ (take $y_i = 1$ (0) as $x_i'\beta > (<) 0$) and $\beta \in \mathcal{B}_Q$ otherwise (take $y_i = 1$ or 0 or arbitrary as $x_i'\beta >$ or $<$ or $= 0$). Geometrically $\mathcal{B}_C$ is the intersection of the unit sphere with a union of cones and $\mathcal{B}_Q$ is the intersection of the unit sphere with a union of subspaces.

Theorem 2.2.

(i) For any $\beta \in \mathcal{B}_C$, $\lim_{k\to\infty} R(k\beta, \hat{\pi}^M) = 0$

(ii) For any $\beta \in \mathcal{B}_Q$, $0 < \lim_{k\to\infty} R(k\beta,\hat{\pi}^M) < Z/4$ where Z is the number of indices i, $1 \leq i \leq n$, for which $x_i'\beta = 0$.

Proof. The same argument works for both $\beta \in \mathcal{B}_C$ and $\beta \in \mathcal{B}_Q$ although the final details differ. Fix $\beta \in \mathcal{B}$ and let $S = \{i : x_i'\beta \neq 0\}$ be the components of $X\beta$ $(\pi(\beta))$ separated by 0 (1/2), $Z = n - |S|$ is the number of components of $X\beta$ $(\pi(\beta))$ which are zero (1/2), and $\mathcal{Y}_{\neq} \equiv \{y \in \mathcal{Y} : (2y_i - 1)x_i'\beta \geq 0, 1 \leq i \leq n\}$ are those y such that the corresponding $\pi(\beta)$ correctly separates all components according to the criteria $\pi_i(\beta) \geq (\leq)\ 1/2$ for $y_i = 1$ (0). Here $|A|$ denotes the cardinality of the set A. Then $Z = (<)\ n$ according as $\beta \in \mathcal{B}_C$ $(\mathcal{B}_Q)$ and $|\mathcal{Y}_{\neq}| = 2^Z$. As $k \to \infty$,

$$f(y|k\beta) \equiv \prod_{i=1}^{n} \{\pi_i(k\beta)\}^{y_i}\{1 - \pi_i(k\beta)\}^{1-y_i}$$
$$\to \begin{cases} 1/2^Z, & y \in \mathcal{Y}_{\neq} \\ 0, & y \notin \mathcal{Y}_{\neq}. \end{cases} \tag{2.7}$$

Furthermore for $y \in \mathcal{Y}_{\neq}$ and $i \in S(\beta)$, $\pi_i(k\beta) \to y_i = \hat{\pi}_i^M(y)$. Hence

$$R(k\beta, \hat{\pi}^M) \to \sum_{\mathcal{Y}_{\neq}} \frac{1}{2^Z} \sum_{i \notin S(\beta)} (\tfrac{1}{2} - \hat{\pi}_i^M(y))^2.$$

When $\beta \in \mathcal{B}_C$ the right hand side is zero as $S(\beta) = \{1, \ldots, n\}$. When $\beta \in \mathcal{B}_Q$ the right hand side is bounded above by

$$\sum_{\mathcal{Y}_{\neq}} \frac{1}{2^Z} \sum_{i \notin S(\beta)} \frac{1}{4} \leq \frac{Z}{4}.$$

That the right hand side is positive can be argued by contradiction.■

Example 2.2. Suppose $n = 5$, $p = 2$ and

$$X = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 0 & 2 \\ -2 & 1 \\ 1 & 1 \end{bmatrix}.$$

Figure 2.1 is a 3-dimensional plot of $(\beta_1,\beta_2,R(\beta,\hat{\pi}^M))$ for β_1,β_2 each satisfying $-4(1/3)4$. Observe that $R(0,\hat{\pi}^M)$ is the (global) maximum and that the risk decreases as $\|\beta\|$ increases. In this example $B_Q = \{(0,\pm1), (\pm1,0), (\pm\sqrt{2}/2,\mp\sqrt{2}/2), (\pm1/\sqrt{5},\pm2/\sqrt{5}), (\pm\sqrt{2}/2,\pm\sqrt{2}/2)\}$. The right hand side of Figure 2.1 clearly shows 3 ridges in $R(\beta,\hat{\pi}^M)$ corresponding to the $(\sqrt{2}/2,\sqrt{2}/2)$, $(1,0)$ and $(\sqrt{2}/2,-\sqrt{2}/2)$ directions, and the front view shows 4 ridges corresponding to $(\sqrt{2}/2,-\sqrt{2}/2)$, $(0,1)$, $(-1/\sqrt{5},-2/\sqrt{5})$, $(-\sqrt{2}/2,-\sqrt{2}/2)$. In all other directions $\beta \notin B_Q$, $R(k\beta,\hat{\pi}^M) \to 0$ as $k \to \infty$.■

We conclude our discussion of the small sample risk properties of the mle by considering its admissibility. It is a textbook exercise (Berger [2] p. 165) to show $\hat{\pi}^M$ is admissible in the simple case of Example 2.1. Johnson [12] or Gutmann [10] can be used to establish the admissibility when the Y_i are from p independent Bernoulli populations. For general (Y,X) Duffy [8] establishes admissibility of $\hat{\pi}^M$ in specific problems by showing that it is unique totally Bayes (Brown [5]). However regardless of the admissibility of $\hat{\pi}^M$, the preceeding discussion shows that its risk is only small for β large in norm. Thus in the spirit of the work of Bishop et al. [3] on the estimation of

multinomial probabilities, the remainder of this paper develops alternative estimators which perform better than $\hat{\pi}^M$ over the central portion of the parameter space.

III. BAYES AND RELATED ESTIMATORS

The natural starting point to develop estimators with lower risk over the central portion of β-space is to consider Bayes estimators with respect to priors putting mass near the origin. Suppose $\beta_1,\ldots,\beta_p$ have iid $N(0,\sigma^2)$ priors and denote the corresponding Bayes estimator $\hat{\pi}^B = \hat{\pi}^B(y)$. The prior mean and variance of 0 and σ^2 imply that $\hat{\pi}^B$ will have lower risk than $\hat{\pi}^M$ in a region near the origin and that σ^2 will control the amount of improvement at (near) $\beta = 0$ and the size of the region where $\hat{\pi}^B$ beats $\hat{\pi}^M$.

Let $H(\cdot|\sigma^2)$ denote the corresponding prior measure on $\mathbb{R}^P$. Then $\hat{\pi}^B(y)$ minimizes

$$\int_{\mathbb{R}^P} \|\pi(\beta) - \pi\|^2 P[Y = y|\pi(\beta)]H(d\beta|\sigma^2)$$

over $\pi \in A$. Unfortunately $\hat{\pi}^B$ is computationally impractical as it requires a minimization over $\mathbb{R}^P$ of an objective function which is itself an integral over $\mathbb{R}^P$. We turn instead to the related Bayesian maximum likelihood estimator $\hat{\pi}^R = \hat{\pi}^R(y)$ which is the mode of the posterior distribution.

Formally $\hat{\pi}^R(y)$ is defined by $\hat{\pi}^R(y) = \hat{\pi}(\beta^R)$ for any $\beta^R \in \mathbb{R}^P$ satisfying

$$\ell_p(\beta^R;y) = \sup_{\mathbb{R}^p} \ell_p(\beta;y) \qquad (3.1)$$

where

$$\ell_p(\beta;y) = y'X\beta - 1_n'\ell n(1 + \exp[X\beta]) - \|\beta\|^2/2\sigma^2.$$

The function $\ell_p(\beta;y)$ is proportional to the posterior density of β given $Y = y$. The estimator $\hat{\pi}^R(y)$ exists and is unique for all $y \in \mathcal{Y}$ since $\ell_p(\beta;y)$ is strictly concave in β and $\ell_p(\beta;y) \to -\infty$ as $\|\beta\| \to \infty$.

In the special case when $\ell_p(\beta;y)$ is symmetric about the origin then $\beta^R = 0$ since $\ell_p(0;y) \geq (\ell_p(\beta;y) + \ell_p(-\beta;y))/2 = \ell_p(\beta;y)$ $\forall$ $\beta \in \mathbb{R}^P$. In this case $\beta^R = 0$ is also the mean of the posterior distribution. In general the strict concavity of $\ell_p(\beta;y)$ means that $\hat{\pi}^R(y)$ can be calculated by any straightforward iterative procedure, Newton Raphson for example, which solves

$$\nabla\ell_p(\beta;y) = X'(y - \hat{\pi}(\beta)) - \beta/\sigma^2 = 0 \qquad (3.2)$$

for $\beta = \beta^R$. Computationally $\hat{\pi}^R(y)$ requires the same amount of work as $\hat{\pi}^M(y)$. In addition, the symmetry $\ell_p(\beta;y) = \ell_p(-\beta;1 - y)$ implies that $\hat{\pi}^R$ satisfies the invariance

$$\hat{\pi}^R(y) = 1 - \hat{\pi}^R(1 - y),\ y \in \mathcal{Y}. \qquad (3.3)$$

Hence the risk of $\hat{\pi}^R$ is symmetric about the origin; i.e. $R(\beta,\hat{\pi}^R) = R(-\beta,\hat{\pi}^R)$ $\forall$ $\beta \in \mathbb{R}^P$.

The posterior mode can be viewed as a ridge-type estimator in that it is a restricted mle which prevents the selection of β "too large" in norm. To formally state this characterization, we explicitly indicate the dependence of β^R on σ as $\beta^R(\sigma)$.

Theorem 3.1. The estimate $\beta^R(\sigma)$ is the solution of problem $\mathcal{R}$,

$$\mathcal{R} \qquad \begin{aligned} &\text{maximize } \ell(\beta;y) \\ &\text{subject to} \\ &\|\beta\|^2 \leq K \equiv \|\beta^R(\sigma)\|^2. \end{aligned}$$

Proof. Denote the La Grangian of $\mathcal{R}$ by

$$T(y;\beta) = \ell(\beta;y) + u(K - \|\beta\|^2).$$

Then $(\bar{\beta},\bar{u}) \in \mathbb{R}^{P+1}$ maximizes $T(y,\beta)$ if

$$\begin{aligned} \frac{\partial T}{\partial \beta} = X'(y - \pi(\bar{\beta})) - 2\bar{u}\bar{\beta} &= 0 \\ \bar{u}(K - \|\bar{\beta}\|^2) &= 0 \\ \bar{u} &\geq 0 \\ \|\bar{\beta}\|^2 &\leq K. \end{aligned} \tag{3.4}$$

The point $(\bar{u},\bar{\beta}) = (1/2\sigma^2,\beta^R(\sigma))$ satisfies (3.4), thus $\beta^R(\sigma)$ solves $\mathcal{R}$.■

We consider the behavior of the risk function of $\hat{\pi}^R(y)$. First, like $\hat{\pi}^M$, $R(\beta,\hat{\pi}^R)$ has a stationary point at $\beta = 0$. This can be seen by performing analogous computations to those in Section 2 to show that

$$\frac{\partial R(0,\hat{\pi}^R)}{\partial \beta_j} = \frac{1}{2^n} \sum_y \left\{ \frac{1}{2} [\pi(0) - \hat{\pi}^R(y)]' \xi_j + \|\pi(0) - \hat{\pi}^R(y)\|^2 [y - \pi(0)]' \xi_j \right\} = 0,$$

where ξ_j is defined in Section 2. The required substitution in this case is $\xi_j'y = \xi_j'\hat{\pi}^R(y) + \beta_j^R/\sigma^2$.

We conjecture that the nature of this stationary point (local minimum, local maximum) depends on σ^2. Qualitatively we expect $\hat{\pi}^R(y)$ to be strongly pulled toward 1/2 as the prior precision increases ($\sigma^2 \to 0$) and to behave like the mle as the prior becomes more diffuse ($\sigma^2 \to \infty$). This intuitive argument suggests that $\beta = 0$ is a local minimum for small σ^2 and a local maximum for large σ^2. We now analyze the effect of σ^2 in $\hat{\pi}^R$ in a rigorous manner.

The Implicit Function Theorem guarantees the map $\beta^R(\sigma) : (0,\infty) \to \mathbb{R}^P$ is a continuous function of σ. Thus $\beta^R(\sigma) \to 0$ as $\sigma \to 0$ since $\ell_p(\beta;y) \to -\infty$ as $\sigma \to 0$ for $\beta \neq 0$ and hence $\hat{\pi}^R(y) \to 1/2$ as $\sigma \to 0$ for all y. Similarly $\ell_p(\beta;y) \to \ell(\beta;y)$ as $\sigma \to \infty$ hence, for $y \in \mathcal{Y}_0$, $\beta^R(\sigma) \to \beta^M$, and for all $y \in \mathcal{Y}$, $\hat{\pi}^R(y) \to \hat{\pi}^M(y)$. The above arguments show that the path traced in $\mathbb{R}^P$ by $\{\beta^R(\sigma) : 0 < \sigma < \infty\}$ emanates from the origin. One more aspect of this path is described below.

Theorem 3.2. The norm $\|\beta^R(\sigma)\|^2$ is nondecreasing in σ.

Proof. Let $\ell_p(\beta;y,\sigma)$ explicitly denote the dependence of $\ell_p(\cdot)$ on σ^2. Then for all $\beta \in \mathbb{R}^P$ and $0 < \sigma_1 < \sigma_2 < \infty$,

$$\ell_p(\beta;y,\sigma_1) = \ell_p(\beta;y,\sigma_2) + \frac{\|\beta\|^2}{2}\left[\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right] \leq \ell_p(\beta;y,\sigma_2)$$

$$\Rightarrow \sup_{\mathbb{R}^P} \ell_p(\beta;y,\sigma_1) \leq \sup_{\mathbb{R}^P} \ell_p(\beta;y,\sigma_2);$$

equivalently

$$\sup_{\beta:\|\beta\|^2\le K_1} \ell(\beta;y) \le \sup_{\beta:\|\beta\|^2\le K_2} \ell(\beta;y)$$

where $K_i = \|\beta^R(\sigma_i)\|^2$. This implies
$\{\beta : \|\beta\|^2 \le K_1\} \subset \{\beta : \|\beta\|^2 \le K_2\}$ which implies
$\|\beta^R(\sigma_1)\|^2 = K_1 \le K_2 = \|\beta^R(\sigma_2)\|^2$ and completes the proof.■

Example 2.2. (continued) Figure 3.1 plots the risk of $\hat{\pi}^R$ for $\sigma = 1$ over the same (β_1,β_2) region as in Figure 2.1. Figure 3.2 is a plot of the indicator function of β for which $R(\beta,\hat{\pi}^R) \le R(\beta,\hat{\pi}^M)$ which facilitates comparisons of the two risk surfaces. The mode estimator $\hat{\pi}^R$ has lower risk than $\hat{\pi}^M$ over a large central portion of the β-space. Comparisons (not shown) with the risk of $\hat{\pi}^B$ shows two features. First, $\hat{\pi}^B$ has lower risk than $\hat{\pi}^R$ over a small central region about the origin. This indicates that $\hat{\pi}^B$ pulls the estimate of $\hat{\pi}$ more strongly to $1_n(1/2)$ than does $\hat{\pi}^R$. Second, $\hat{\pi}^R$ beats the mle $\hat{\pi}^M$ over a region roughly twice as large as the region over which $\hat{\pi}^B$ beats the mle.■

As for normal theory ridge estimators there is no universally agreed upon method of choosing the unspecified shrinkage parameter. The next section proposes an estimator based on a data dependent choice of σ^2.

IV. EMPIRICAL MODE ESTIMATOR

Section III showed that $\hat{\pi}^R$ acts like $\hat{\pi}^M$ for large σ^2 and hence has small risk for $\|\beta\|$ large and "extreme" $\pi(\beta)$; for small σ^2, $\hat{\pi}^R$ pulls toward $1_n(1/2)$ and hence has small risk for $\|\beta\|$ small and "central" $\pi(\beta)$. This section proposes an empirical mode estimator $\hat{\pi}^E$ based on data-dependent choice of σ^2. The goal is to estimate σ^2 to be large when $\pi(\beta)$ is extreme and to be small when $\pi(\beta)$ is central.

The standard empirical Bayes approach to this problem is to choose σ^2 to maximize the marginal distribution of Y given σ^2 which is given by

$$m(y|\sigma^2) \equiv \int_{\mathbb{R}^P} f(y|\beta)H(d\beta|\sigma).$$

There do not exist closed form expressions for the resulting estimator of σ^2. An attractive iterative scheme for maximizing $m(y|\sigma^2)$ over $\sigma^2 \in (0,\infty)$ is the EM algorithm (Dempster, Laird, and Rubin [7]). We consider (Y,β) as the "complete data" and Y as the "incomplete data." The joint density of (Y,β) is an exponential family with sufficient statistic $t(y,\beta) = \|\beta\|^2$. Hence, given a current guess σ_k^2 of σ^2, the next cycle of the EM algorithm is

E-step: Estimate $t = \|\beta\|^2$ by

$$t_k = E[\|\beta\|^2 | y, \sigma_k^2]$$

$$= \int_{R^P} \|\beta\|^2 f(y|\beta) H(d\beta|\sigma_k^2)/m(y|\sigma_k^2).$$

M-step: Set $\sigma_{k+1}^2 = t_k/m$.

The EM algorithm as above is not computationally practical because of the severe demands of the E-step. Following Leonard [15] and Laird [13], we propose an alternative algorithm obtained by approximating the conditional distribution of β given y and σ^2 with a p-variate normal distribution having mean $\beta^R = \beta^R(\sigma^2)$ and covariance matrix $\Sigma^R = \Sigma^R(\beta^R, \sigma^2)$. We take β^R to be the mode of the true conditional distribution of β given y and σ^2; equivalently β^R maximizes $\ell_p(\beta; y)$ or solves (3.2). We choose

$$\Sigma^R = X' \Lambda X + (\beta^R)(\beta^R)'/\sigma^4 \qquad (4.1)$$

where Λ is $n \times n$ diagonal with ith diagonal element $\hat{\pi}_i(\beta^R)[1 - \hat{\pi}_i(\beta^R)]$. Under mild regularity conditions Σ^R is the asymptotic covariance matrix of β^R (Duffy [8]).

With this approximation the EM algorithm becomes:

E-step: Estimate $t = \|\beta\|^2$ by

$$t_k = E[\|\beta\|^2 | y, \sigma_k^2]$$

$$= (\beta^R(\sigma_k^2))'(\beta^R(\sigma_k^2)) + \mathrm{Tr}\{[\Sigma^R(\beta^R(\sigma_k^2), \sigma_k^2)]^{-1}\}. \qquad (4.2)$$

M-step: Set $\sigma_{k+1}^2 = t_k/M$.

Here $Tr(A)$ denotes the trace of A. If the algorithm given in (4.2) converges to an estimate $\hat{\sigma}^2$ of σ^2, then the empirical mode estimator is defined as $\hat{\pi}^E(y) = \hat{\pi}(\beta^R(\hat{\sigma}^2))$.

Remark 4.1. An alternative choice of Σ^R considered by Laird [13] is $\Sigma^R = [-\nabla^2 \ell_R(\beta^R;y)]^{-1}$. With this choice the approximating normal distribution and the true conditional distribution of β have the same curvature at β^R.■

The example below indicates that $\hat{\pi}^E$ achieves its goal of mimicking the risk behavior of $\hat{\pi}^M$ or $\hat{\pi}^R$ for central and extreme β, respectively.

Example 2.2. (continued) For each $y \in \mathcal{Y}$, $\hat{\pi}^E(y)$ was computed by (4.2). The estimates $\hat{\sigma}^2(y)$ lie in $[.79,1.51]$ with $\hat{\sigma}^2 > 1$ if and only if $y \in \mathcal{Y}_C$. Note that in this example $|\mathcal{Y}_C| = 10$, $|\mathcal{Y}_0| = 22$ and $\mathcal{Y}_Q = \emptyset$. Figure 4.1 is an indicator plot of β with $R(\beta,\hat{\pi}^E) < R(\beta,\hat{\pi}^M)$ which shows that $\hat{\pi}^E$ has lower risk than the mle $\hat{\pi}^M$ over a larger portion of the β-space than does $\hat{\pi}^R$. The improvement occurs at the extreme portions of β-space where $\hat{\pi}^E$ is better able to protect against extreme probabilities. Figure 4.2 is an indicator comparing the risk of $\hat{\pi}^E$ and $\hat{\pi}^R$. It confirms that $\hat{\pi}^E$ is precisely doing better than the mode estimator $\hat{\pi}^R$ for the extreme points in β-space.■

Several comments should be made about our experience with the approximate EM algorithm. It was implemented with starting value $\sigma_0^2 = 1$ and terminated when either $|\sigma_k^2 - \sigma_{k+1}^2| \leq \varepsilon$ or an upper bound B on the number of cycles was achieved. The E-step

maximization was always performed iteratively with initial value $\beta^0 \equiv 0$. Empirical evidence suggests both that the algorithm is well-behaved in that small to moderate changes in the starting values do not alter its convergence, and that $\beta^R(\sigma^2)$ and $\hat{\pi}^E$ are insensitive to small changes in σ^2. Thus in the simulation study reported in Section 5, the values $\varepsilon = .01$ and $B = 25$ were selected. The algorithm converged in over 99% of the problems and generally in far fewer than 25 cycles.

V. EMPIRICAL SMALL SAMPLE STUDY

This section reports results of a pilot simulation study analyzing the small sample risk properties of the estimators described in Sections II-IV. In an effort to assess the sensitivity of the results with respect to square error loss and the parameter estimated, we also estimated the risk with respect to the weighted squared error loss on π given in (1.2) and the squared error loss on β,

$$L_\beta(\beta,\hat{\beta}) = \|\beta - \hat{\beta}\|^2.$$

The risk was estimated for 8 random problems corresponding to the combinations of (n,m) = (sample size, dimension of β): (10,2), (30,2), (30,4), (30,6), (50,2), (50,4), (50,6), and (50,8). For each (n,m) a random matrix X with entries independently and identically distributed according to the uniform distribution over $[-1,+1]$ was generated, and two β_0 values were selected. This yielded 16 test problems. The value of $\|\beta_0\|$ is of particular importance as it determined the degree of centrality of the components of the associated $\pi(\beta_0)$. Specifically, the two chosen β_0 values gave (i) all $\pi_i(\beta_0) \in [1/4,3/4]$, $1 \le i \le n$, i.e., a central case, and (ii) $\pi_i(\beta_0) \in [1/4,3/4]$ for 50% of the indices i, $1 \le i \le n$, i.e., a more extreme case. The former $\pi(\beta_0)$ is more favorable to estimators which pull toward 1/2 such as $\hat{\pi}^B$ and $\hat{\pi}^R$ while the latter is more favorable to $\hat{\pi}^M$.

One hundred and fifty data sets were generated from the model (1.1) for each test problem. For each data set the estimators $\hat{\pi}^M$, $\hat{\pi}^B$, $\hat{\pi}^R$, and $\hat{\pi}^E$ were computed and their average loss calculated. Pilot runs with 300 replications showed the average loss values to be quite stable after 150 data sets.

Summaries of typical results for 4 of the 16 problems are presented in Table 4.1. The problems presented are for the two choices of β_0 for each of $(n,m) = (10,2)$ and $(50,6)$. Table 4.1 gives the estimated risk values R, R_w and R_β for L, L_w and L_β, respectively. The corresponding estimators are given in increasing order with respect to R and denoted by M,B,R or E for $\hat{\pi}^M$, $\hat{\pi}^B$, $\hat{\pi}^R$ or $\hat{\pi}^E$, respectively. Whenever $\hat{\beta}^M$ did not exist, L_β was not computed. We also give the observed mean and standard deviation of the estimate $\hat{\sigma}^2$ computed by (4.2).

Note first that $\hat{\pi}^M$ has the highest risk in 10 out of the 12 combinations of problem size by choice of β_0 by loss clearly indicating that many settings are better served by an alternate technique. In the problems shown both $\hat{\pi}^R$ and $\hat{\pi}^E$ perform well but in more extreme situations studied, $\hat{\pi}^E$ has markedly lower risk. The 3 different risk functions are, with one exception, identical in their ranking of the estimators, hence improvements over maximum likelihood are not restricted to loss L. Finally, we observe that the mean of $\hat{\sigma}^2$ increases with the degree of extremeness of $\pi(\beta)$ as desired. This phenomenon held throughout the entire simulation study and is the reason why $\hat{\pi}^E$ has competitive risk properties across all sizes and states of nature considered.

Table 4.1

Simulated Risk Values

$(n,m) = (10,2)$

Every $\pi_i(\beta_0) \in [1/4,3/4]$.				50% of $\pi_i(\beta_0) \in [1/4,3/4]$.			
Estimator	R	R_w	R_β	Estimator	R	R_w	R_β
B	.132	.63	.65	M	.276	1.97	13.25*
R	.134	.64	.67	E	.281	2.74	4.97
E	.210	.97	1.96	R	.325	4.14	6.35
M	.348	1.58	5.30*	B	.346	4.54	6.87
* based on 140 reps $\hat{\sigma}^2$: 1.53 (.655)				* based on 98 reps $\hat{\sigma}^2$: 1.99 (.710)			

$(n,m) = (50,6)$

Every $\pi_i(\beta_0) \in [1/4,3/4]$.				50% of $\pi_i(\beta_0) \in [1/4,3/4]$.			
Estimator	R	R_w	R_β	Estimator	R	R_w	R_β
E	.68	3.13	.77	R	.75	4.45	1.24
R	.83	3.73	1.04	E	.79	4.79	1.33
M	1.33	5.90	2.36	M	1.24	6.76	3.65 *
$\hat{\sigma}^2$: .674 (.113)				* based on 149 reps $\hat{\sigma}^2$: .921 (.194)			

In conclusion, the mode and empiric mode estimators have reasonable small sample risk properties and are computationally competitive with the maximum likelihood estimator. Duffy [8] establishes some asymptotic properties $\hat{\pi}^R$ and $\hat{\pi}^E$. Further study of the properties of these estimators is currently in progress.

ACKNOWLEDGMENT

The second author's research was partially supported by the Biomechanical Engineering Program at Cornell University and by the U.S. Army Research Office through the Mathematical Science Institute at Cornell University.

REFERENCES

[1] Albert, A. and Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*, 1-10.

[2] Berger, J. (1980). *Statistical Decision Theory: Foundations, Concepts and Models*. Springer-Verlag, New York.

[3] Bishop, Y., Fienberg, S. and Holland, P. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

[4] Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research. Vol. 1: The Analysis of Case-Control Studies*. IARC Scientific Publications No. 32, International Agency for Research on Cancer, Lyon, France.

[5] Brown, L.D. (1981). A complete class theorem for statistical problems with finite sample spaces. *Ann. Statist. 9*, 1289-1300.

[6] Brown, W.B. (1980). Prediction analyses for binary data in Miller, R., et al. eds. *Biostatistics Casebook*. New York: John Wiley and Sons.

[7] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Royal Statist. Soc. B. 39*, 1-38.

[8] Duffy, D.E. (1986). *Alternative Methods of Estimation in Logistic Regression*. Ph.D. thesis, Cornell University, Ithaca.

[9] Finney, D.J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika 34*, 320-334.

[10] Gutmann, S. (1982). Stein's paradox is impossible in problems with finite sample space. *Ann. Statist. 10*, 1017-1020.

[11] Haberman, S.J. (1974). *The Analysis of Frequency Data*. Chicago, IL: The University of Chicago Press.

[12] Johnson, B.M. (1971). On the admissible estimators for certain fixed sample binomial problems. *Ann. Math. Statist. 42*, 1579-1587.

[13] Laird, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika 65*, 581-590.

[14] Larntz, K. (1980). Linear logistic models for the parole decision making problem in Fienberg, S.E. and A.J. Reiss, Jr., eds., *Indicators of Crime and Criminal Justice: Quantitative Studies*, U.S. Government Printing Office, Washington, D.C.

[15] Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *J. Royal Statist. Soc. B. 37*, 23-37.

[16] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior in P. Zarembka, ed., *Frontiers in Econometrics*. New York: Academic Press.

[17] Pierce, D.A., and Sands, B.R. (1975). Extra-Bernoulli variation in binary data. Department of Statistics Technical Report 46, Oregon State University, Eugene.

[18] Pregibon, D. (1981). Logistic regression diagnostics. *Ann. of Statist. 9*, 705-724.

[19] Santner, T.J. and Duffy, D.E. (1986). A Note on Albert and Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. To appear in *Biometrika*.

[20] Silvapulle, M.J. (1981). On the existence of maximum likelihood estimators for the binomial response model. *J. Royal Statist. Soc. B. 43*, 310-313.

[21] Stefanski, L.A., and Carroll, R.J. (1985). Covariate measurement error in logistic regression. *Ann. Statist. 13*, 1375-1351.

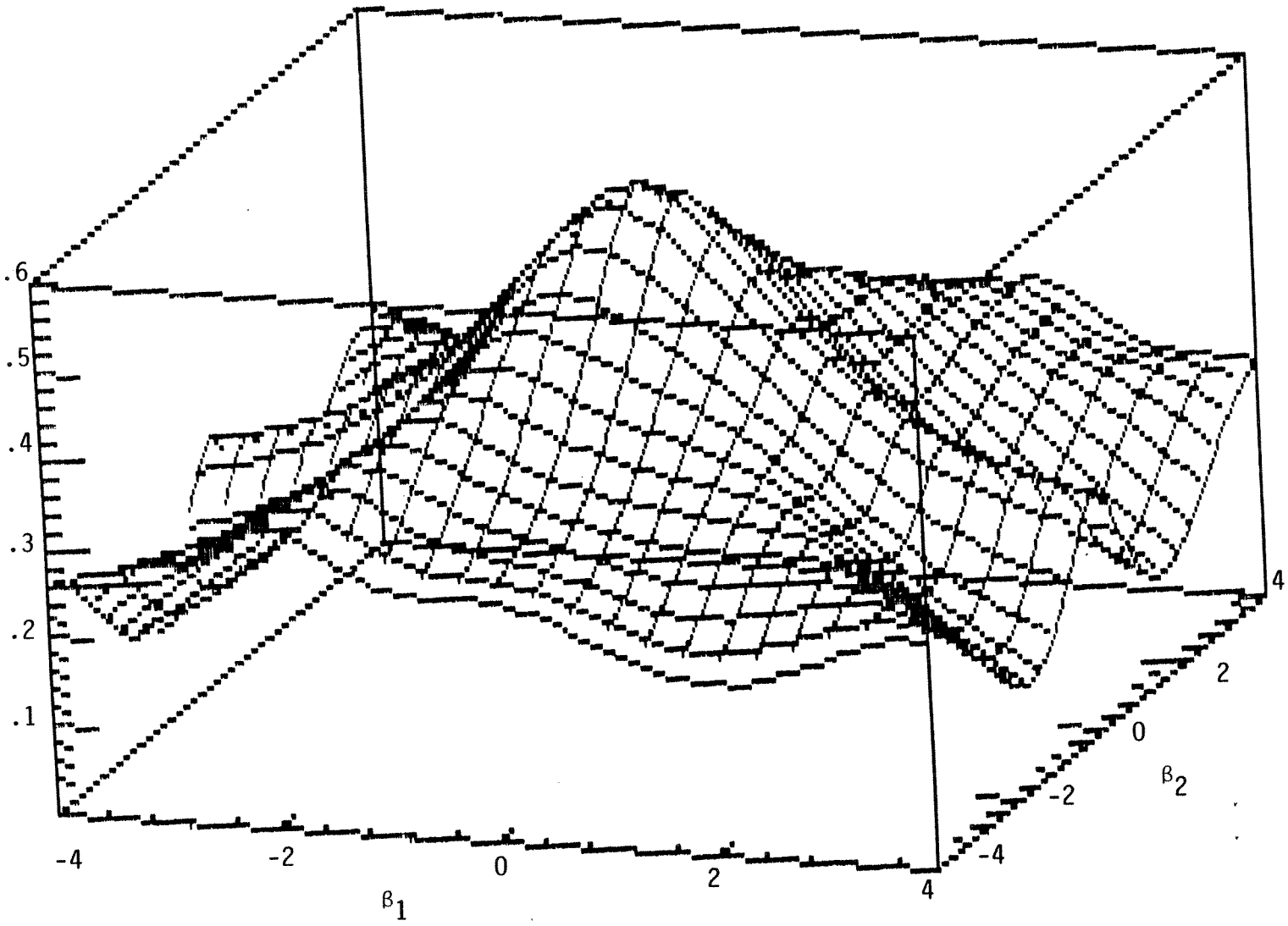

Figure 2.1

Plot of $((\beta_1,\beta_2), R(\beta,\hat{\pi}^M))$ for $\beta_1 = -4(1/3)4$ and $\beta_2 = -4(1/3)4$ in Example 2.2.

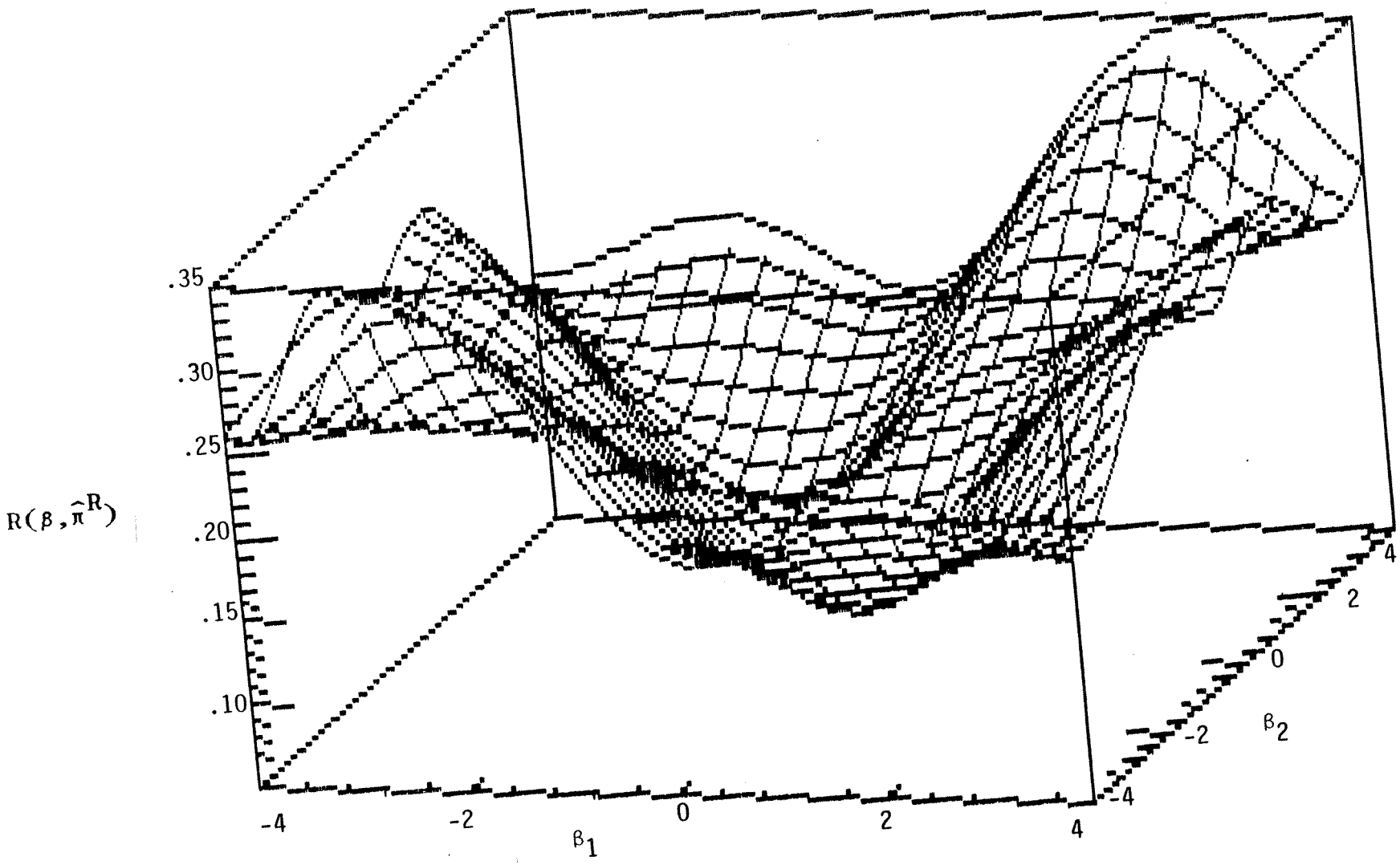

Figure 3.1

Plot of $((\beta_1,\beta_2), R(\beta,\hat{\pi}^R))$ when $\sigma^2 = 1$ for $\beta_1 = -4(1/3)4$ and $\beta_2 = -4(1/3)4$ in Example 2.2.

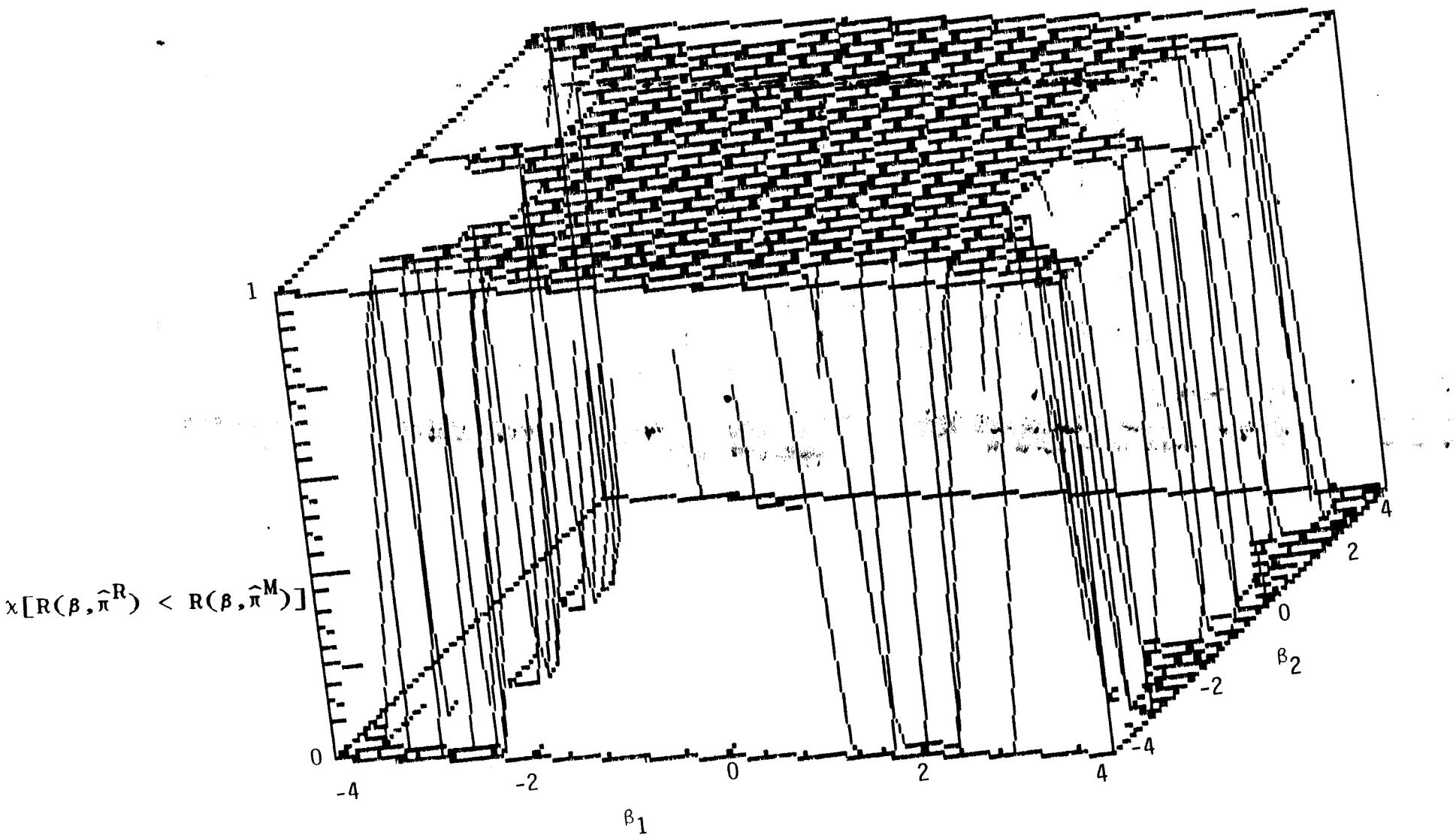

Figure 3.2

Indicator plot of $((\beta_1,\beta_2), \chi[R(\beta,\hat{\pi}^R) \leq R(\beta,\hat{\pi}^M)])$ when $\sigma^2 = 1$ for $\beta_1 = -4(1/3)4$ and $\beta_2 = -4(1/3)4$ in Example 2.2.

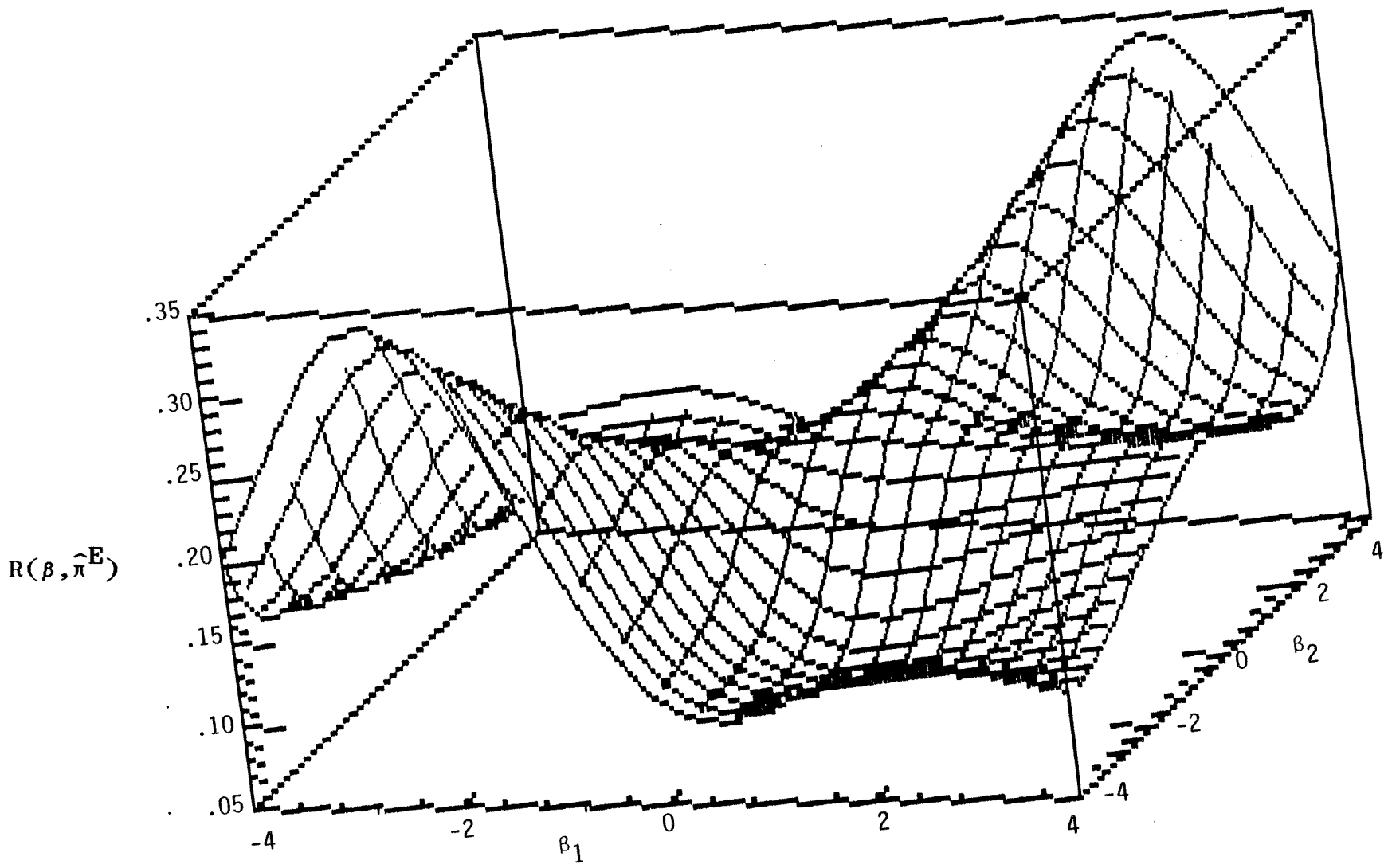

Figure 4.1

Plot of $((\beta_1,\beta_2), R(\beta,\hat{\pi}^E))$ for $\beta_1 = -4(1/3)4$ and $\beta_2 = -4(1/3)4$ in Example 2.2.

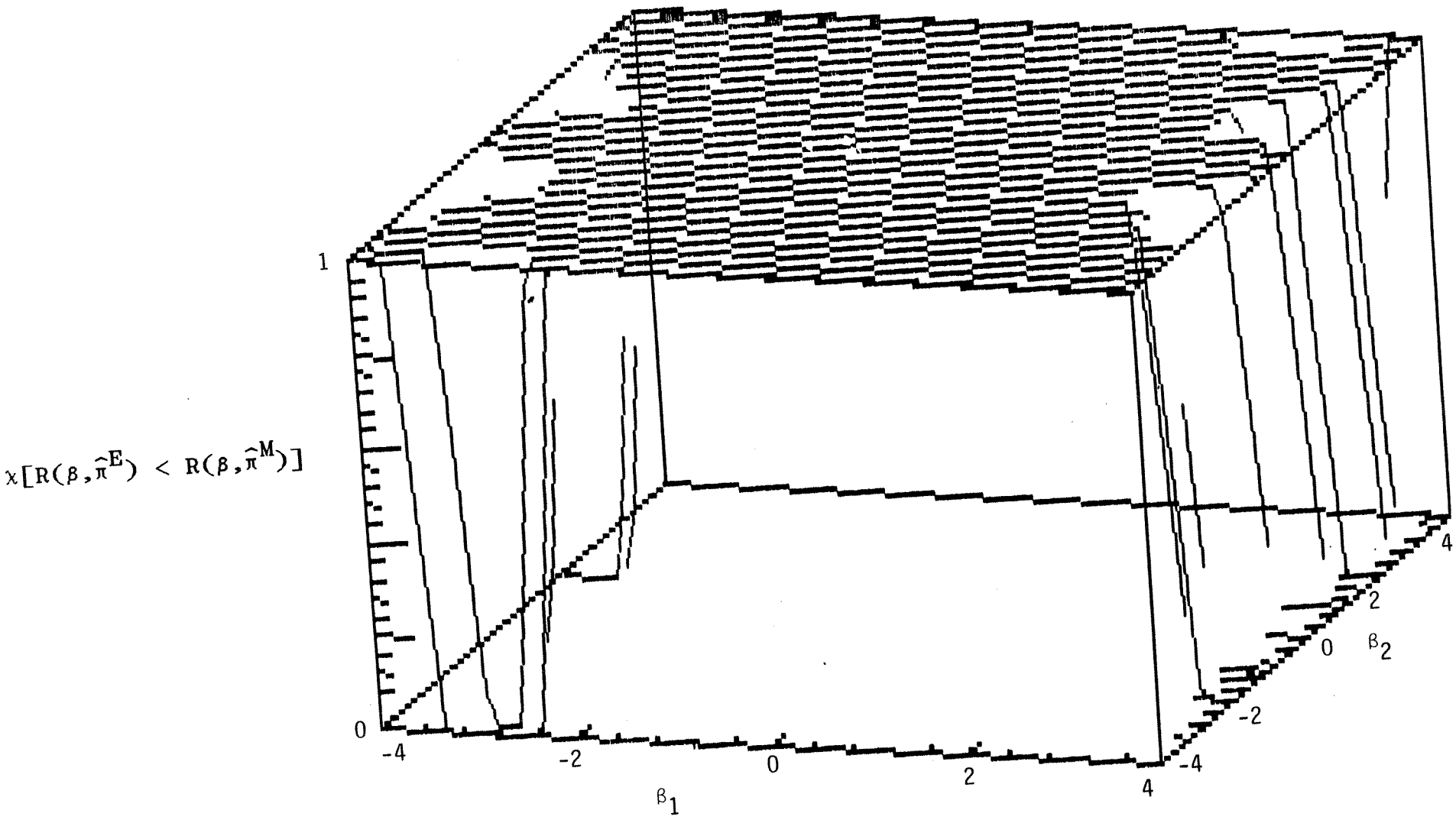

Figure 4.2

Indicator plot of $((\beta_1,\beta_2), \chi[R(\beta,\hat{\pi}^E) < R(\beta,\hat{\pi}^M)])$ for $\beta_1 = -4(1/3)4$ and $\beta_2 = -4(1/3)4$ in Example 2.2.