Social Systems Research Institute

University of Wisconsin - Madison.

OPERATIONAL CHARACTERISTICS OF MAXIMUM SCORE ESTIMATION

Charles F. Manski
T. Scott Thompson

8519

This research was supported under National Science Foundation Grant SES-8319335 and by a grant from the University of Wisconsin Graduate School. Computational facilities were provided by the Center for Demography and Ecology of the University of Wisconsin.

December 1985

ABSTRACT

This paper reports on the operational characteristics of maximum score estimation of a linear model from binary response data. A series of previous articles have shown that in theory, the maximum score method makes possible binary response analysis under very weak distributional assumptions. Here, we present evidence on the properties of maximum score estimation in practice.

After reviewing the known asymptotic theory of maximum score estimation, the paper describes an algorithm for maximum score estimation and characterizes its performance. Then findings from a Monte Carlo study comparing maximum score and logit maximum likelihood estimation are reported. Finally, the accuracy of bootstrap estimation of maximum score root mean square errors is evaluated.

This paper reports on the operational characteristics of maximum score(MS) estimation of a linear model from binary response data. A series of articles(Manski,1975,1985a,1985b,1985c) have shown that in theory, the maximum score method makes possible binary response analysis under very weak distributional assumptions. Here, we present evidence on the properties of MS estimation in practice.

Section 1 reviews the known asymptotic theory of maximum score estimation. Section 2 describes an algorithm for MS estimation and characterizes its performance. Section 3 reports findings from a monte carlo study comparing MS and logit maximum likelihood(LML) estimation. Section 4 evaluates the accuracy of bootstrap estimation of MS root mean square errors.

1.Theory of MS Estimation

The familiar 'threshold' model of binary response postulates an observable binary indicator y* and a latent scalar variable y such that y*=1 if y≥0 and y*=-1 otherwise. The variable y is itself assumed to be a function of an observable random variable x and an unobservable random variable u. The most common specification is the linear model

$$(1) \quad y = x\beta + u,$$

where $x \in X \subset R^K$, $u \in U \subset R^1$, and where $\beta \in B \subset R^K$ is a parameter vector.

Consider the problem of inference on β from randomly drawn binary response data, that is from observations of (y*,x). From such data, one

can learn the joint distribution of (y^*,x). In particular, one can estimate the conditional response probabilities $Pr(y^*=1|x), x\in X$. These are related to β by

$$(2) \quad Pr(y^*=1|x) = \int 1[x\beta+u\geq 0]dF_{u|x},$$

where $F_{u|x}$ denotes the distribution of u conditional on x. The function $1[\ ]$ takes the value one if the logical condition in the brackets is satisfied and zero otherwise.

Inference on β is possible if one can combine equation (2) with sufficient prior information about $F_{u|x}$. The universal practice in empirical research has been to assume that for some $\gamma\in\Gamma$ and for all $x\in X$,

$$(3) \quad F_{u|x}(*) = \Psi(*|x,\gamma)$$

where Γ is a given finite dimensional parameter space and where $[\Psi(*|x,c), c\in\Gamma]$ is a given family of distribution functions. With few exceptions, empirical work has assumed that $\Psi(*|x,\gamma)$ is the same for all values of x. That is, the disturbances are homoskedastic. The familiar logit and probit models impose the even stronger assumption that the shape of the common distribution is known.

Such sharp distributional assumptions are not necessary. It was shown in Manski(1975,1985a) that β can be identified up to scale and estimated consistently if it is known only that for each $x\in X$, the distribution $F_{u|x}$ has unique median zero. More generally, it suffices that for given $\alpha\in(0,1)$, the α-quantile of $F_{u|x}$ be unique and equal zero. No other restrictions need be imposed on the behavior of the unobservables. Thus, u may be arbitrarily heteroskedastic and the

shapes of the distributions $F_{u|x}, x \in X$ unknown.

To see this, let us restate the zero α-quantile assumption on $F_{u|x}$ in the equivalent form

$$(4) \quad \Pr(y \geq x\beta \mid x) = 1 - \alpha.$$

That is, $x\beta$ is the α-quantile regression of y on x (Koenker and Bassett, 1978). Equation (4) plus the assumed uniqueness of the α-quantile implies that for each $x \in X$,

$$(5) \quad \begin{aligned} x\beta > 0 \quad &\Leftrightarrow \quad \Pr(y^*=1 \mid x) > 1-\alpha \\ x\beta = 0 \quad &\Leftrightarrow \quad \Pr(y^*=1 \mid x) = 1-\alpha \\ x\beta < 0 \quad &\Leftrightarrow \quad \Pr(y^*=1 \mid x) < 1-\alpha. \end{aligned}$$

This relates the parameter β to the estimable response probabilities.

Clearly, (5) cannot identify the scale of β. In general, binary response data do not identify the scale of β unless the scale of the disturbances is fixed. The information about β that is implied by (5) depends on the richness of the distribution of x. For each $b \in R^K$, let

$$(6) \quad X_b = [x \in X : \operatorname{sgn}(xb) \neq \operatorname{sgn}(x\beta)].$$

Given knowledge of the response probabilities, (5) identifies β relative to b if and only if

$$(7) \quad \Pr(x \in X_b) > 0.$$

Let $\| \ \|$ be the Euclidean norm on R^K, let $\beta^* \equiv \beta / \|\beta\|$, and let $B_o \equiv [b \in R^K : b/\|b\| \neq \beta^*]$. Let F_x denote the distribution of x. Manski (1985a,

Lemma 2) gives a sufficient condition on F_x for (5) to identify the normalized parameter β^* relative to all of the elements of B_o. It is enough that the support of F_x not be contained in any proper subspace of R^K and that at least one component of x have an everywhere positive Lebesgue density.

Given identification of β^*, a consistent estimator can be obtained in the following manner. Let the parameter space B be a given subset of $[b \in R^K : \|b\|=1]$ with $\beta^* \in B$. For $b \in B$, define the 'population score'

$$(8) \quad S_\alpha(b) \equiv E[\{y^* - (1-2\alpha)\}\mathrm{sgn}(xb)].$$

It can be shown that as a consequence of (5), the function S_α is maximized on B at β^*. Moreover, if β^* is identified on B, this maximum is unique(Manski,1985a,Lemma 3).

Now consider the sample analog of (8), namely the 'sample score'

$$(9) \quad S_{N\alpha}(b) \equiv \frac{1}{N} \sum_{n=1}^{N} [y_n^* - (1-2\alpha)]\mathrm{sgn}(x_n b)$$

where N is the sample size. It is intuitive that given randomly drawn observations of (y^*,x), maximization of $S_{N\alpha}$ on B should yield a consistent estimate of β^*. This is in fact the case.

Considering the median regression case, Manski(1975) introduced the maximum score estimator and gave a partial proof of consistency. Manski(1985a) generalized the estimator to quantile regressions and gave a complete proof of strong consistency. See Theorem 1 and Corollary 2. That paper also established the large deviations rate of convergence for MS estimation. In particular, Theorem 4 showed that the MS estimate lies outside any fixed neighborhood of β^* with probability that goes to

zero at exponential rate as N->∞.

Subsequently, it has been shown that if the observations are generated by response-based rather than random sampling, then an appropriately weighted version of the MS estimator is strongly consistent(Manski,1985b). In another paper, the problem of inference from binary response panel data with arbitrary person-specific effects has been studied. It has been found that a version of the MS estimator is strongly consistent if, for each person, the distribution of u is time-stationary(Manski,1985c).

Given these consistency results, one would like to enrich the theory of MS estimation by characterizing MS sampling behavior. Thus far, however, no exact or asymptotic sampling theory has been developed. In particular, the local rate of convergence has not been determined. Nor has the existence of a limiting distribution been proved. The monte carlo analysis of Section 3 provides experimental evidence which may help guide the development of sampling theory.

2.Computation of Maximum Score Estimates

In order to facilitate the computation of maximum score estimates in practice, we have written a user-oriented computer program. This FORTRAN 77 program embodies a much refined version of the algorithm described in Manski(1975). To date, it has been implemented on the VAX 11-780 computer and on the IBM Personal Computer. Subject to machine-specific limitations, the program accomodates any sample size and number of explanatory variables. It provides point MS estimates for any quantile regression, not just the median. An optional post-processor

characterizes the set of parameter values on which the score is maximized. Another option offers bootstrap estimation of root mean square errors.

The MS algorithm searches for the maximum of the sample score function in a manner that differs from the approaches to numerical optimization commonly applied in econometric work. To understand why, we need to compare the MS problem with the optimization problems familiar to econometricians.

In most econometric work, the criterion function is smooth. Given an initial parameter estimate, derivatives of the criterion function provide local steepness and curvature information that help one to select a direction in which to search for a better estimate. Most often, evaluation of the criterion function is time consuming and computation time is proportional to the number of function evaluations. This makes it expensive to conduct anything but a crude search in the chosen direction. As a consequence of these characteristics, the algorithms used in practice are generally good at selecting a direction in which to move from the initial estimate. But they usually have limited capacity for optimizing the criterion function along this direction. Some popular algorithms do not search at all. Instead, they take a step of predetermined length in the chosen direction.

Maximum score estimation differs from the usual econometric optimization problem in two important respects. First, the sample score (9) is not smooth but rather a step function. So local analysis provides no information useful in selecting a fruitful direction for search from an initial estimate. This makes maximization of the score function more difficult than the usual optimization problem.

In a second respect though, MS estimation is easier than the usual

problem. In contrast to other criterion functions, the sample score can be evaluated all along a given search direction in little more time than that required for evaluation at a single point. Moreover, evaluation at a single point can be performed very quickly. It follows that given a search direction, one can maximize the score function along this direction with little effort.

To see why this is so, let the parameter space B be the unit hypersphere in R^K. Let $b^o \in B$ be an initial parameter estimate and let $c \in B$ be a given vector orthogonal to b^o. For $\lambda \in [0,2\pi)$, define

$$(10) \quad b(\lambda) \equiv \cos(\lambda)b^o + \sin(\lambda)c.$$

We shall show how to evaluate the score function on the great circle connecting b^o and c, that is on the set of points $\{b(\lambda), \lambda \in [0,2\pi)\}$.

First observe that it suffices to consider values λ in the interval $[0,\pi)$. This follows from the fact that for all $\lambda \in [0,\pi)$, $\cos(\lambda+\pi)=-\cos(\lambda)$ and $\sin(\lambda+\pi)=-\sin(\lambda)$. Hence $b(\lambda+\pi)=-b(\lambda)$. By the definition (9) of the sample score function,

$$(11) \quad S_{N\alpha}[b(\lambda+\pi)] = \frac{1}{N}\sum_{n=1}^{N}[y_n^* - (1-2\alpha)]\operatorname{sgn}[x_n' b(\lambda+\pi)] = -S_{N\alpha}[b(\lambda)].$$

So knowledge of the score at $b(\lambda)$ implies knowledge at $b(\lambda+\pi)$.

Next, fix n and consider the expression

$$(12) \quad x_n' b(\lambda) = \cos(\lambda)x_n' b^o + \sin(\lambda)x_n' c$$

as a function on $\lambda \in [0,\pi)$. In the usual case where $x_n' b^o \neq 0$, there exists

a unique $\lambda_n \in (0,\pi)$ solving the equation $x_n'b(\lambda_n)=0$ and partitioning the interval $[0,\pi)$ into two sub-intervals. In particular, if $x_n'b^o>0$, then $x_n'b(\lambda)>0$ for $\lambda\in[0,\lambda_n)$ and $x_n'b(\lambda)<0$ for $\lambda\in(\lambda_n,\pi)$. If $x_n'b^o<0$, then $x_n'b(\lambda)<0$ for $\lambda\in[0,\lambda_n)$ and $x_n'b(\lambda)>0$ for $\lambda\in(\lambda_n,\pi)$. So the score for observation n varies over $\lambda\in[0,\pi)$ as follows:

$$(13a)\quad [y_n^*-(1-2\alpha)]\,\mathrm{sgn}[x_n'b(\lambda)] \;=\; [y_n^*-(1-2\alpha)]\,\mathrm{sgn}[x_n'b^o]\,\mathrm{sgn}(\lambda_n-\lambda).$$

It may occur that $x_n'b^o=0$. If so, $\lambda_n=0$ solves the equation $x_n'b(\lambda_n)=0$ and $\mathrm{sgn}[x_n'b(\lambda)]=\mathrm{sgn}(x_n'c)$ for all $\lambda\in(0,\pi)$. In this case then, the score varies with λ as follows:

$$(13b)\quad [y_n^*-(1-2\alpha)]\,\mathrm{sgn}[x_n'b(\lambda)] \;=\; [y_n^*-(1-2\alpha)]\,\mathrm{sgn}[-x_n'c]\,\mathrm{sgn}(\lambda_n-\lambda).$$

Now let the observations be ordered so that $\lambda_n \le \lambda_{n+1}$, $n=1,\ldots,N$. Equations (13a)-(13b) imply that the interval $[0,\pi)$ is partitioned into 2N+1 sub-intervals, on each of which the sample score is constant. These are $[0,\lambda_1),[\lambda_1],(\lambda_1,\lambda_2),[\lambda_2],(\lambda_2,\lambda_3),\ldots,[\lambda_N],(\lambda_N,\pi)$. It follows that to evaluate the score on the entire interval $[0,\pi)$, we need only compute $(x_n'b^o, x_n'c; n=1,\ldots,N)$ and solve for the values $(\lambda_1,\ldots,\lambda_N)$.

To summarize, given an initial MS estimate, it is difficult to identify a good direction in which to search for a better estimate. But it is easy to maximize the score along any given direction. These characteristics make it undesirable to design an MS algorithm in the traditional manner, which emphasizes selection of a search direction and downplays optimization along this direction. Rather, it seems advantageous to reverse the traditional priorities.

Our MS algorithm does not attempt to select a single, best search

direction from the initial estimate. Instead, it selects a set of K-1 orthogonal directions and searches recursively along all of these. To be specific, let $c^1, c^2, \ldots, c^{K-1}$ be a selected orthonormal set of vectors in R^K, all being orthogonal to the initial estimate b^o. The algorithm first maximizes the score function on the great circle connecting b^o with c^1. As explained above, this maximization is easy to perform. In general, the maximum score is attained on one or more intervals of positive length on the great circle. The algorithm chooses the midpoint of a maximizing interval to be the new initial estimate. Then it maximizes the score on the great circle connecting the revised initial estimate with c^2. The process repeats until all K-1 directions have been searched. To begin the next iteration, the algorithm selects a new set of orthonormal vectors $c^1, \ldots, c^{K-1}$.

Iterations are continued until a trial maximum is achieved, that is until an orthogonal set of great circle searches yields no increase in the value of the score function. But a given set of great circle searches may miss a direction of increase of the score function. Moreover, even if the trial maximum is a true local maximum, it need not be the global maximum. So one cannot be certain that the trial maximum is the actual maximum.

For these reasons, upon achieving a trial maximum the algorithm conducts a user-determined number of 'end-game' searches. Each end-game search is an iteration in which a new set of orthonormal search directions is drawn by a randomization mechanism. Random drawing of directions ensures that if iterations were continued indefinitely, with probability one all regions of the parameter space would be searched eventually. If the specified number of end-game searches produce no increase in the score function, the trial maximum is declared the actual

maximum. Otherwise, search resumes.

Table 1 gives benchmarks for computation time per iteration for various values of the sample size N and number of explanatory variables K. These benchmarks are for the VAX11-780 version of the MS program, compiled and executed under version 4.1 of the VMS operating system.

Table 1: CPU Time Benchmarks(Seconds Per Iteration)

	Number of Variables(K)		
Sample Size(N)	2	6	11
100	0.04	0.18	0.43
1000	0.32	1.74	3.94
10000	4.03	24.20	53.59

When K=2, there is only one search direction orthogonal to the initial estimate. So the algorithm always finds the score function maximum in one iteration. Clearly, the computation of MS estimates in this case is extremely fast.

When K>2, the number of regular and end-game iterations required to achieve the maximum depends on the configuration of the data and on one's initial parameter value. Hence, we cannot give precise benchmarks for total CPU time. The program offers a simple data-driven method for generating an initial value. Using this method, we have found that in moderate size problems(i.e.K≤6,N≤1000), a trial maximum is usually found in five to ten iterations. For one to be confident that the trial maximum is the actual maximum, ten to twenty end-game iterations should be performed. So convergence is usually attained in under one minute of CPU time, excluding program overhead. Of course, solution of larger problems may require much more time.

3.A Monte Carlo Study of MS and LML Estimation

To assess the usefulness of MS estimation, we need to compare its performance with that of methods routinely applied in empirical work. The obvious standard of comparison is maximum likelihood estimation of the homoskedastic logit model.

For purposes of this discussion, we maintain the assumption that $x\beta$ is the median regression of y on x and that F_x is sufficiently rich for the median regression restriction to identify β^*. But we do not necessarily assume that the distributions $F_{u|x}, x \in X$ are identical or logistic. So MS estimation of β^* is consistent but LML estimation may not be.

An empirical researcher who is not confident that the disturbances are homoskedastic or logistic should prefer MS estimation of β^* to LML, provided that the sample is large enough. We would like to know whether the asymptotic advantage of MS estimation is retained in samples of moderate size. In the absence of knowledge of the exact sampling distributions of MS and LML estimates, we address this question through Monte Carlo experimentation.

3.1. Design of the Experiments

The experiments reported here all concern the estimation of a simple median regression. That is, $\alpha=0.5$, $K=2$, and $x=(1,z)$, where $z \in Z \subset R^1$ is a scalar random variable with distribution F_z. In this case, $\beta=(\beta_1,\beta_2) \in R^2$ and β^* is a point on the unit circle. Observe that β^* has an equivalent representation as a scalar parameter, namely the angle between β^* and

the β_1 axis. We denote this angle by λ^* and take it as the parameter of interest. In the experiments, we set $\beta_1=\beta_2=1$. So the value of the angular parameter is $\lambda^*=\pi/4\underset{\sim}{=}.79$.

We have performed experiments with a variety of specifications for $[F_{u|z}, z\in Z]$. Let v be a scalar random variable whose distribution G has unique median zero. Let h(z) be a positive-valued function of z. Now let

$$(14) \qquad y = 1 + z + h(z)v.$$

Then the median regression of y on (1,z) is 1+z for all scaling functions h(*) and distributions G. But the response probability Pr(y*=1|z) varies with h(*) and G. In particular,

$$(15) \qquad \Pr(y^*=1|z) = \int 1\left[\frac{1+z}{h(z)} + v \geq 0\right] dG.$$

Equation (15) and the fact that G has median zero imply that for all h(*), $1+z>0 \iff \Pr(y^*=1|z)>.5$. On the other hand, Pr(y*=1|z) is not necessarily monotonic in z. Rather, it increases with (1+z)/h(z).

An experiment was specified by fixing $[h(*), G, F_z, N]$. Given a specification, data for each trial were generated by drawing N pseudo-random realizations of (z,v), converting these into values of y via (14), and applying the indicator function $y^*=1[y\geq 0]$. On each trial, the MS estimate of λ^* was obtained by the circle search algorithm described in Section 2. The LML estimate of (β_1,β_2) was computed using a modified Newton-Raphson method. The angle subtended by the estimate was then computed to obtain the LML estimate of λ^*. In each experiment, 1000 trials were performed.

We shall discuss six sets of experiments. In all of these, the distribution G is normalized to have median zero and variance five. In the first four sets of experiments, G is logistic and the disturbances have different scaling function h(*). These are

$$(16) \quad h_1(z) = \exp(z-1)$$

$$(17) \quad h_2(z) = |1+z|/\sqrt{2} \quad \text{for } z \neq -1$$
$$h_2(z) = 1 \quad \text{for } z=-1$$

$$(18) \quad h_3(z) = 2.07\exp(-|1+z|).$$

$$(19) \quad h_4(z) = 1.$$

The remaining two sets of experiments assume homoskedastic disturbances with nonlogistic distributions, specifically t(3) and uniform.

Note that the scaling functions $h_1(*)$, $h_2(*)$, $h_3(*)$, and $h_4(*)$ are all normalized so as to satisfy the condition $\int h(z)^2 d\Phi = 1$, where Φ is the standard normal distribution. This normalization and the variance normalization of G were made in an attempt to enhance comparability of experimental results across the different experiments. It must be admitted, however, that the appropriate norm for this purpose is not clear.

Observe that each of the scaling functions given in (16)-(19) implies a different form for (1+z)/h(z), hence a different shape for the response probability as a function of z. In particular, $(1+z)/h_1(z) = (1+z)e^{1-z}$ and has derivative $-ze^{1-z}$. So $(1+z)/h_1(z)$ increases with z in the region $z<0$, is maximized at $z=0$, and then decreases with z in the region $z>0$. It follows that $P(y^*=1|z)$ rises from zero at $z=-\infty$ until it attains a maximum value of $\int 1[e+v>0]dG$ at $z=0$. As z increases from 0 to ∞, $P(y^*=1|z)$ falls to .5.

In the second case, $(1+z)/h_2(z) = \sqrt{2}\,\mathrm{sgn}(1+z)$ so the response

probability is a step function. In particular, $P(y^*=1|z)=\int 1[-\sqrt{2}+v>0]dG$ for all $z<-1$, $P(y^*=1|z)=.5$ at $z=-1$, and $P(y^*=1|z)=\int 1[\sqrt{2}+v>0]dG$ for all $z>-1$.

Finally, $(1+z)/h_3(z) = (1+z)e^{|1+z|}/2.07$. Here, $P(y^*=1|z)$ rises continuously from 0 to 1 as z increases from $-\infty$ to ∞. In the vicinity of $z=-1$, the expression $(1+z)/h_3(z)$ is well approximated by $(1+z)/2.07$. Hence, for z near -1, $P(y^*=1|z)$ behaves like a logit probability function. As $z\to-\infty$ and $z\to\infty$, however, $P(y^*=1|z)$ approaches its asymptotes more rapidly than do logit probabilities.

Tables 2 through 7 present bias, standard deviation(SDEV) and root mean square error(RMSE) findings for MS and LML estimation from the six sets of experiments. The tables also give asymptotic bias(A.Bias) findings for LML estimation. Asymptotic biases are determined by solving numerically the limiting form of the LML problem, namely

$$(20)\quad \max_{b\in R^2} \int P(y^*=1|z)\log\frac{\exp(b_1+b_2z)}{1+\exp(b_1+b_2z)} + P(y^*=-1|z)\log\frac{1}{1+\exp(b_1+b_2z)}\,dF_z.$$

The top part of each table holds the sample size fixed at N = 200 and explores the effect on estimation of changing the centering of F_z. In particular, we conduct experiments with $F_z = N(\mu,1)$ and μ varying over the interval $[-2,1]$. The bottom part of each table holds $F_z = N(0,1)$ and explores the effect of sample size. We report experiments with N=25,50,100,200,400. The rows for N=200 reproduce results from the top half of the tables.

3.2. Findings on LML Estimation

EFFECT OF HETEROSKEDASTICITY ON BIAS: Our discussion of the effect of heteroskedasticity on the response probabilities suggests that the bias of LML estimation should vary with both h(*) and μ. Tables 2 through 5 show that this variation is substantial. In all the tables, the experimentally determined LML finite sample biases are very close to the corresponding asymptotic biases. Hence, it suffices to examine the asymptotic biases.

Table 2 reveals that under $h_1(*)$, the LML bias decreases from the small positive value .06 when μ=-2 to the large negative value -.91 when μ=1. It is easy to explain this pattern. When the probability mass of z is concentrated in the region z<<0, the LML estimate of λ* should be positively biased as $(1+z)/h_1(z)$ increases more rapidly than z when z is small. But when the mass is concentrated at larger values of z, we see negative bias as $(1+z)/h_1(z)$ rises increasingly slowly and then decreases with z. When μ=1, the concentration of mass in the region z>0 makes LML estimation so negatively biassed as to imply that λ* equals $\pi/4-.91 \approx -.12$. Thus, LML estimation leads one to infer that the slope β_2 of the median regression is negative rather than positive.

Table 3 indicates that under $h_2(*)$, the variation in bias with μ is similarly large. Here, the LML bias decreases from .17 when μ=-2 to -.64 when μ=1. To explain this pattern, recall that under $h_2(*)$, the response probability is a step function. In particular, P(y*=1|z) is a constant less than .5 for z<-1 and a constant greater than .5 for z>-1. As $\mu \to -\infty$, the probability mass of z becomes increasingly concentrated in the region z<-1. So the plim of the LML estimate of β_1 approaches a negative constant, that of β_2 approaches zero, and that of λ* approaches

π. Given that the true value of λ^* is $\pi/4$, the limiting asymptotic bias as $\mu \to -\infty$ is $3\pi/4$. In contrast, as $\mu \to \infty$, the mass of z becomes concentrated in the region $z > -1$. Hence, the plim of the LML estimate of β_1 approaches a positive number, that of β_2 again approaches zero, and that of λ^* approaches zero. So the limiting asymptotic bias as $\mu \to \infty$ is $-\pi/4$.

Table 4 demonstrates that not all forms of heteroskedasticity are so damaging to LML estimation. Here, the LML bias increases from -.03 as $\mu=-2$ to .16 at $\mu=1$. To understand this pattern, observe that $h_3(*)$ differs from $h_1(*)$ and $h_2(*)$ in two respects. First, the range of $h_3(*)$ is $(0, 2.07]$ while $h_1(*)$ and $h_2(*)$ range over $(0, \infty)$. So $h_3(*)$ represents a more moderate form of hetroskedasticity than do $h_1(*)$ and $h_2(*)$. We should therefore expect that the LML biases under $h_3(*)$ are more moderate than under $h_1(*)$ and $h_2(*)$. Second, as we noted earlier, the response probabilities under $h_3(*)$ are logit-like for z near -1. It follows that as long as μ is not too far from -1, the LML bias should be particularly small. This is precisely what Table 4 shows. As μ moves toward -1, the LML bias essentially vanishes. On the other hand, noticeable biases do emerge as μ moves far enough away from -1.

As expected, the LML bias vanishes in Table 5. When the disturbances are homoskedastic, LML estimation is consistent.

EFFECT OF DISTRIBUTION SHAPE ON BIAS: The experiments of Tables 6 and 7 indicate that the bias of the LML estimator is minimal when the disturbances are homoskedastic t(3) and minimal to moderate when they are homoskedastic uniform. We have conducted other experiments, not reported here, with homoskedastic disturbances drawn from other nonlogistic distribution. In none of these experiments was the LML bias

large in magnitude.

It is not surprising that a failure of G to be logistic has less severe consequences for LML estimation than does a failure of h(*) to be flat. It follows from (15) that if the disturbances are homoskedastic, then Pr(y*=1|z) is an increasing function of z irrespective of the shape of G. This fact limits the damage that misspecification of G can do.

Ruud(1983) proved that in some contexts, misspecification of G produces no bias at all. In particular, assume that F_x is normal and that the model (1) has an intercept, so that

$$(21) \quad y = \beta_1 + \sum_{k=2}^{K} \beta_k x_k + v.$$

Then maximum likelihood estimation with G misspecified yields consistent estimates of the scale normed slope parameters $(\beta_2,\ldots,\beta_K)/\|\beta_2,\ldots,\beta_K\|$. Ruud did not address the question of bias in estimation of β^*, which involves the value of the intercept. The asymptotic bias results of Tables 6 and 7 show that his result does not extend to this problem.

Observe that in Table 6, as previously, the experimentally determined LML finite sample biases are very close to the corresponding asymptotic biases. In Table 7, however, the exact sample biases deviate systematically from the asymptotic biases as μ decreases. In particular, at μ=-2 the asymptotic bias is -.02 but the finite sample bias for N=200 is .06. To investigate this discrepancy, we conducted experiments at μ=-2 with larger sample sizes and found that as N increases, the finite sample bias moves slowly towards the asymptotic bias. Specifically, for samples of size N=400 and N=800, the finite sample biases are .01 and -.00 respectively.

STANDARD DEVIATION OF LML ESTIMATION: We shall discuss only briefly the experimentally determined standard deviations for LML estimation. The tables show that in samples of size 200, the SDEV varies moderately with h(*), μ, and G. In particular, it ranges over the interval [.04,.19] in Table 2, [.08,.38] in Table 3, [.06,.14] in Table 4, [.10,.23] in Table 5, [.08,.20] in Table 6, and [.16,.27] in Table 7. In each case, the SDEV is minimized at μ=-1 or μ=-.05. These findings are unsurprising but we have made no attempt to explain them in detail.

The tables also show that for fixed h(*) and μ, the SDEV varies predictably with sample size. It is well known that $\sqrt{N}$ times the difference between the LML estimate and its plim converges to a limiting normal distribution centered on zero. This suggests that the finite sample SDEVs should decrease at rate $\sqrt{N}$. In all the tables, we find exactly this pattern.

3.3. Findings on MS Estimation

In all the experiments reported here, MS estimation is consistent. We hope to learn from the experiments how h(*), G, μ, and N affect the precision of MS estimation, as measured by RMSE.

Formally, RMSE is determined by both bias and SDEV. But the experiments consistently show that the RMSE of MS estimation is dominated by its SDEV. In essentially every experiment conducted, the reported MS SDEV is at least ninety percent of the value of the reported RMSE. Given this, our discussion will focus more heavily on the SDEV of MS estimation than on its bias.

To help interpret the findings, we first pose two hypotheses based on heuristic reasoning. These are:

THE CERTAIN RESPONSE HYPOTHESIS: All else equal, the precision of MS estimation should increase the closer the response probabilities $P(y^*=1|x)$ are to 0 and 1.■

The argument is simple. Given an observation (y^*,x), consider the score $\{y^*-(1-2\alpha)\}sgn(xb)$ evaluated at β^*. The score is positive if $y^*=sgn(x\beta^*)$ and is negative otherwise. Assume that $x\beta^*>0$. By (5), $Prob(y^*=sgn(x\beta^*)|x] = P(y^*=1|x) > 1-\alpha$. So the probability of a positive score at β^* increases as $P(y^*=1|x)$ approaches one. Symmetrically, when $x\beta^*<0$ the probability of a positive score at β^* is $1-P(y^*=1|x)$, which increases as $P(y^*=1|x)$ approaches zero. It follows that the score evaluated at β^* should tend to be higher the more certain is the response y^* given x. A similar argument shows that the score evaluated at $b\neq\beta^*$ should tend to be lower the more certain are the responses. Thus, one should expect increasing precision of the maximum score estimate as the responses become more certain.

THE THRESHOLD HYPOTHESIS: Let $\nu>0$. All else equal, the precision of MS estimation should deteriorate as $Prob(|x\beta^*|<\nu)\to 0$. Moreover, the closer $x\beta^*$ is to zero, the more informative x should be about β^*.■

It is known that a quantile regression assumption can identify β^* only if F_x gives positive probability to the event that $x\beta^*$ is near the threshold value zero. Formally, the requirement is that for every $\nu>0$, $Prob(|x\beta^*|<\nu)>0$. See Manski(1985), Lemma 1. If the condition fails, β^* is locally unidentified. That is, there exist b close to β^* such that for all $x\in X$, $sgn(xb)=sgn(x\beta^*)$. It seems reasonable to extrapolate from this identification condition and hypothesize that all else equal, the precision of MS estimation should be low if there is low probability of drawing x such that $x\beta^*$ is near zero. It also seems reasonable to think

that observations with $x\beta^*$ near zero are the most informative in estimation, as they can distinguish β^* from the most other elements of the parameter space.

EFFECT OF μ: With this as background, let us fix N, G, and h(*) and see how the precision of MS estimation varies with μ. Examining the tables, we find that the precision of MS estimation is consistently highest in the vicinity of $\mu=-1$. In each of Tables 2 through 6, $\mu=-1$ minimizes both the absolute bias and the standard deviation of MS estimation. In Table 7, bias and SDEV are minimized at $\mu=-.5$. In all the tables, the experimentally determined biases and SDEVs tend to increase monotonically as μ moves away from the neighborhood of -1.

This clear pattern of results seems to provide strong evidence for the threshold hypothesis. When $\mu=-1$, the probability mass of 1+z is most heavily concentrated in the neighborhood of zero. As μ moves away from -1, the mass of 1+z in the neighborhood of zero falls. It is of some practical interest to observe that in all the tables, the MS precision remains relatively close to its optimal value when μ is in the interval $[-2,0]$. Precision deteriorates seriously only at $\mu=1$, where fully .975 of the probability mass of 1+z is to the right of 0. This suggests that successful MS estimation requires only a modest concentration of $x\beta^*$ data in the neighborhood of zero.

EFFECT OF h(*): Now fix N, G, and μ and let us examine how precision varies with h(*). Comparison of Tables 2 through 5 shows that for essentially all values of μ, the most precise estimates occur under $h_2(*)$. In every case, $h_2(*)$ gives the smallest absolute bias. In every case but $\mu=.5$, $h_2(*)$ gives the smallest SDEV. In contrast, we

find that the least precise estimates essentially always occur under $h_3(*)$. For every value of μ, $h_3(*)$ gives the largest absolute bias. For all μ except $\mu=1$, $h_3(*)$ gives the largest SDEV.

The estimates under $h_4(*)$ always seem to be slightly more precise than those under $h_3(*)$. The most distinguishing feature of the $h_1(*)$ results is their range. For $\mu\in[-2,.5]$, $h_1(*)$ yields estimates that are almost or as precise as those of $h_2(*)$. But when $\mu=1$, $h_1(*)$ gives the largest SDEV and RMSE in all our experiments. Thus, the effect of μ on precision is more marked for $h_1(*)$ than for the other heteroskedasticity functions.

These findings are plausibly explained by combining the certain response hypothesis with other ideas. By the certain response hypothesis, MS estimation is most precise when $P(y*=1|z)$ tends to remain far from .5. Equivalently, estimation is most precise when $(1+z)/h(z)$ tends to remain far from 0. Taken alone, this hypothesis cannot explain our findings because the relative distance from 0 of $(1+z)/h_1(z)$, $(1+z)/h_2(z)$, $(1+z)/h_3(z)$, and $(1+z)/h_4(z)$ varies with z. So the certain response hypothesis has predictive power only if we can specify the relevant region of z values.

Clearly, the relevant region should move with μ. But the threshold hypothesis states that the most informative z values in MS estimation are those with z near -1. So it is reasonable to speculate that as long as μ is not too far from -1, MS estimation should be most precise when $h(*)$ is such that $(1+z)/h(z)$ is furthest from 0 for z near -1.

This reasoning successfully explains our findings for $\mu\in[-2,.5]$. To see this, observe that at $z=-1$, $(1+z)/h(z)=0$ for all specifications of $h(*)$. But the behavior of $(1+z)/h(z)$ in a neighborhood of $z=-1$ differs across specifications. In particular, $(1+z)/h_2(z)$ is a step function at

$z=-1$ while $(1+z)/h_1(z)$, $(1+z)/h_3(z)$, and $(1+z)/h_4(z)$ are differentiable in z. Evaluated at $z=-1$, the derivative of $(1+z)/h_1(z)$ with respect to z is e^2, that of $(1+z)/h_3(z)$ is 1/2.07, and that of $(1+z)/h_4(z)$ is 1. Thus, focussing on the neighborhood of $z=-1$, the certain response hypothesis predicts that precision should be highest under $h_2(*)$, followed by $h_1(*)$, $h_4(*)$, and finally $h_3(*)$. This is precisely what we find for $\mu\in[-2,.5]$.

At $\mu=1$, the ranking of precisions changes, $h_1(*)$ now giving the least precise estimates. This deviation from the previous pattern is presumably due to the fact that $(1+z)/h_1(z)$ is a decreasing function for $z>0$. In contrast, $(1+z)/h(z)$ is either constant of increasing for the other scaling functions. When $\mu=1$, much of the probability mass of z is in the region where $(1+z)/h(z)$ is closer to zero under $h_1(*)$ than it is under the other scaling functions.

EFFECT OF G: Comparison of Tables 5 and 6 shows that the behavior of MS estimation is broadly similar when G is logistic and t(3). In contrast, the MS estimates are always less precise when G is uniform. We shall not attempt to draw conclusions from this pattern.

EFFECT OF N: It remains to consider how the precision of MS estimation varies with N. The consistency of MS implies that as $N\to\infty$, both bias and SDEV should fall to zero. In fact, this happens in all of the experiments. But the available theory does not predict the rates at which bias and SDEV converge to zero.

The experiments provide only rough evidence about the rate of convergence of the MS bias. The absence of a clear relationship of bias to sample size may reflect underlying structural complexity. It seems

more likely, however, that the culprit is the small magnitude of the MS biases. Because the biases are close to zero, sampling and rounding errors can easily muddy rate of convergence calculations.

At any rate, inspection of the tables shows that in every case, the experimentally estimated bias changes little if at all as N increases from 25 to 50 and then falls as N increases from 50 to 400. We are not able to determine whether the rate of decrease of the bias is a simple function of N. Nor is it certain that the bias falls at the same rate in all six sets of experiments.

The experiments do reveal one very important pattern. That is, the MS biases always fall at least as rapidly and appear to fall more rapidly than do the corresponding MS SDEVs. This implies that the local asymptotic behavior of MS estimation may be determined either by SDEV alone(if bias falls more rapidly than does SDEV) or by both bias and SDEV(if bias and SDEV fall at the same rate). The experiments favor the former hypothesis but do not seem to us conclusive.

Now let us examine the behavior of the MS SDEV as a function of sample size. Here the experimental evidence is clear but rather unconventional.

To interpret the outcomes in a succinct manner, we assume that in each of the six sets of experiments, the simple model

$$(22) \quad \sigma(N) * N^{r} = e^{c}$$

holds. Here, $\sigma(N)$ is the MS SDEV in samples of size N and (c,r) are parameters specific to the experimental specification. In the context of this model, N^{r} is the rate of convergence of the SDEV. To estimate (c,r), we rewrite (22) in the equivalent form

(23) $\log[\sigma(N)] = c - r*\log(N)$

and use the experimentally estimated values of $\log[\sigma(N)]$ on samples of size N=50,100,200,400 to fit equation (23) by least squares. The results are given in Table 8.

Table 8: Least Squares Fits of (c,r)

Experiment of Table	Parameter c Estimate	Std.Err.	Parameter r Estimate	Std.Err.	R^2
2	-0.740	(0.160)	0.383	(0.032)	.986
3	2.486	(0.341)	1.014	(0.068)	.991
4	-0.204	(0.114)	0.240	(0.023)	.982
5	0.385	(0.016)	0.360	(0.003)	.999
6	0.165	(0.101)	0.359	(0.020)	.994
7	0.891	(0.089)	0.403	(0.018)	.996

Examination of the table shows that the six estimates of the model (23) all fit the experimental data remarkably well. The reported standard errors for the estimates of r are uniformly less than ten percent the magnitudes of the estimates. And the R^2 values are always above .98.

The estimates of r indicate three distinct rates of convergence for the MS SDEV. In the heteroskedastic experiment of Table 2 and in all of the experiments with homoskedastic disturbances(Tables 5, 6, and 7), the estimate of r is in the narrow band [0.36,0.40]. This suggests strongly a common rate of convergence that is faster than $N^{1/3}$ but slower than $N^{1/2}$. The rate of convergence under the heteroskedastic specifications of Tables 3 and 4 are clearly different. In the case of Table 3, the rate is almost certainly N^1. In the experiment of Table 4, it is in the neighborhood of $N^{1/4}$.

The estimates in Table 8 do not use the experimental data for samples of size of N=25. Given that the MS bias does not change between samples of size 25 and 50, we felt that N=25 may be too small a sample size for the behavior of $\sigma(N)$ to have stabilized. Nevertheless, one would like to know whether inclusion of the data for N=25 changes the picture drawn by Table 8. So we have re-estimated (c,r) utilizing these data.

We find that when the data for N=25 are included, the estimates of c in Table 8 change substantially. On the other hand, the estimates of r remain essentially unchanged in every case but one. The exception is the experiment of Table 2, where the estimate of r becomes 0.585, its standard error becomes 0.118 and the R^2 drops to .89. In the five other cases, the estimates of r remain within 0.03 of the corresponding estimates given in Table 8, the reported standard errors remain very small, and the R^2 values stay above .98. Thus, inclusion of the data for N=25 has no effect on our rate of convergence conclusions for the experiments of Tables 3 through 7. In the case of Table 2, it appears that the behavior of the MS SDEV has not yet stabilized at N=25.

Overall, our experimental findings on the rate of convergence of the MS bias and SDEV send a very strong signal. That is, the local asymptotic behavior of MS estimation is unlike that of the estimators of classical statistics. In general, the bias of a classical estimator converges at rate N and its SDEV at rate $\sqrt{N}$. In contrast, The rate of convergence of the MS SDEV appears to vary with the scaling function h(*). The rate is slower than $\sqrt{N}$ in every experiment except that of Table 3, where it is faster.

In retrospect, it is not entirely surprising that the MS estimate converges most quickly under the scaling function $h_2(*)$ and least quickly under $h_3(*)$. Under $h_2(*)$, the response probabilities are a

discontinuous function of z, the discontinuity being precisely at the threshold value z=-1. Thus, observations with z near the threshold value are extremely informative. Under all the other scaling functions, the response probabilities increase smoothly with z so observations near the threshold value are less valuable. In particular, under $h_3(*)$ the value of (1+z)/h(z) increases most slowly for z near the threshold. Apparently, the behavior of the response probabilities near the threshold affects not only the magnitude of the MS SDEV but also its rate of convergence.

It is of interest to note that the experimental results corroborate an asymptotic theoretical conclusion of Chamberlain(1984). There it was found that under the assumptions of Manski(1985), there exists no 'regular' estimator of b*, that is no estimator that converges at rate $\sqrt{N}$ and has certain uniformity properties.

Finally, one should be careful not to extrapolate from these experiments to more general settings involving additional explanatory variables. In some nonclassical estimation problems, the rate of convergence of an estimate necessarily falls as the number of explanatory variables rises. For example, this is so in nonparametric regression. Our experiments only concern the case K=2. We cannot exclude the possibility that when K>2, MS estimates converge at slower rates than those found here.

3.4. Implications for Practice

If estimators are ranked by RMSE, then the ranking of MS and LML in our experiments varies with the scaling function h(*) but not with the distribution shape G. Tables 2 and 3 show that in the presence of two relatively serious forms of heteroskedasticity, MS dominates LML at all values of μ and N where experiments were conducted. In the case of $h_1(*)$, the LML RMSE is 1.1 to 5.2 times as high as the corresponding MS RMSE. Under $h_2(*)$, the LML RMSE ranges from 1.5 to 9.4 times as high. It is of interest to note that the dominance of MS does not derive solely from the lower bias of MS estimation. In all the $h_2(*)$ experiments and most of the $h_1(*)$ ones, MS estimation has a lower SDEV than does LML.

Tables 4 through 7 indicate that if the disturbances are homoskedastic or mildly heteroskedastic, then LML dominates MS. Under $h_3(*)$, the MS RMSE ranges from 1.7 to 3.8 times as high as that for LML. In the three homoskedastic settings studied, the MS RMSE is 1.2 to 2.0 times as high.

What implications can one draw for practice? It would seem that an empirical researcher can rely on LML estimation if he is confident that the distributions $F_{u|x}$ do not vary substantially with x. Apparently, he need not worry seriously about the effects of nonlogistic shape.

A researcher who feels less comfortable with the assumption of homoskedasticity should use MS estimation as a check on LML results. The LML results might be accepted if the two estimates are close. If they differ markedly, it would be prudent to discard the LML estimates and to base inference on the MS results.

A researcher who believes that the distributions $F_{u|x}$ may vary

substantially with x should put no trust in LML estimates. Provided that he is confident that $x\beta$ is a quantile regression of y on x, he can rely on MS estimation. Our experiments indicate that the consistency of MS estimation is complemented by reasonably good behavior in small samples, whatever the actual pattern of heteroskedasticity.

The above conclusions are, of course, tentative. The experiments reported here have held fixed numerous factors potentially important to a comprehensive comparison of MS and LML estimation. For example, we have maintained the assumption that $x\beta$ is the median regression of y on x, rather than some other quantile regression. We have assumed throughout that F_x is normal with fixed variance. We have not explored the possible interactive effect on estimation of heteroskedasticity and nonlogistic shape of the distributions $F_{u|x}, x \in X$. And we have considered only the simple setting in which x has two components, one of which is a constant. It would clearly be of interest to conduct experiments which investigate the consequences of varying factors that we have held fixed.

4. Bootstrap Estimation of the Precision of MS Estimates

The user of MS estimation would like to have some measure of the precision of his estimates. Lacking an analytical approximation to the MS sampling distribution, we decided to explore bootstrap estimation of the RMSE of MS estimates. The available theory of the bootstrap is not sufficiently general to show that this application of the bootstrap is proper. But the bootstrap has intuitive appeal and has been found successful elsewhere. This motivated us to conduct some experiments.

In each experiment, we performed 1000 trials as in Section 3. In

each trial, we computed the MS estimate and then applied the bootstrap to the sample data. By definition, the bootstrap RMSE is the MS RMSE that would obtain if the empirical distribution of (y*,x) were the population distribution and if the computed MS estimate were β^*.

We estimated the bootstrap RMSE in the usual way by monte carlo experimentation on the empirical distribution. On each trial, we drew 50 pseudo-random samples of N observations from the multinomial empirical distribution of (y*,x). On each such pseudo-sample, we computed the maximum score estimate. We then estimated the bootstrap RMSE by the sample mean square deviation of the 50 pseudo-MS estimates.

Table 9 gives the estimated mean and standard deviation of the estimated bootstrap RMSE for twenty four of the specifications examined in Tables 2 through 5. The column headed 'True RMSE' repeats the experimentally determined RMSEs presented in those tables.

We would have liked to examine the bootstrap performance on all the specifications of Tables 2 through 7. The computational burden of experiments to evaluate bootstrap performance is severe, however. To see why, it suffices to point out that each experiment in Table 9 consists of 1000 outer trials and 50 inner(bootstrap) trials. Thus, to obtain the findings reported in Table 9 required computation of $24 \times 1000 \times 50 = 1{,}200{,}000$ MS estimates.

Examination of Table 9 shows that the bootstrap performs uniformly well in the experiments with scaling functions $h_1(*)$, $h_3(*)$, and $h_4(*)$. In these experiments, the mean of the bootstrap estimates of the RMSE is always very close to the experimentally determined value. Thus, the bootstrap appears to provide an unbiased estimate of the RMSE. In most experiments, the standard error of the bootstrap estimates is less than half the mean. In all experiments it is less than the mean.

Now consider the experiments with the scaling function $h_2(*)$. Here, the bootstrap clearly does not behave as nicely. The mean of the bootstrap estimates of the RMSE is consistently higher than that the true RMSE. Moreover, the standard error of the bootstrap estimates are of the same magnitudes as the means. The difference between $h_2(*)$ and the other scaling functions is, of course, that the response probabilities are a step function of z under $h_2(*)$ and vary smoothly otherwise. Apparently, discontinuity in the response probabilities hinders the performance of the bootstrap.

Overall, we find these results encouraging. In practice, one would usually expect the response probabilities to vary smoothly with z, as under $h_1(*)$, $h_3(*)$, or $h_4(*)$, rather than in the step function manner of $h_2(*)$. Assuming this, our experiments suggest that the bootstrap provides enough information about the precision of MS estimates to make the bootstrap a useful tool in empirical research. A conservative approach would be to double the bootstrap estimate of the MS RMSE and to take this as an upper bound on the value of the true RMSE.

References

Chamberlain, G.(1984), "Asymptotic Efficiency in Semiparametric Models with Censoring", unpublished.

Koenker, R. and Bassett, G.(1978), "Regression Quantiles", Econometrica, 46, 33-50.

Manski, C.(1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice", Journal of Econometrics, 3, 205-228.

Manski, C.(1985a), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator", Journal of Econometrics, 27, 313-333.

Manski, C.(1985b), "Semiparametric Analysis of Binary Response from Response-Based Samples", Journal of Econometrics, forthcoming.

Manski, C.(1985c), "Semiparametric Analysis of Random Effects Linear Models from Binary Response Data".

TABLE 2: Experiments With Heteroskedastic Logistic Disturbances
$h_1(z) \equiv \exp(z-1)$

N = 200

	Maximum Score			Logit Maximum Likelihood			
μ	Bias	SDEV	RMSE	Bias	SDEV	RMSE	A.Bias
-2.0	.03	.06	.07	.06	.05	.08	.06
-1.5	.02	.06	.06	.08	.04	.09	.07
-1.0	.02	.05	.05	.09	.04	.10	.08
-0.5	.02	.05	.06	.08	.05	.09	.08
0.0	.02	.06	.06	-.05	.10	.11	-.05
0.5	.03	.09	.10	-.48	.19	.52	-.49
1.0	-.09	.40	.42	-.89	.13	.90	-.91

$\mu = 0$

	Maximum Score			Logit Maximum Likelihood		
N	Bias	SDEV	RMSE	Bias	SDEV	RMSE
25	.03	.28	.28	-.08	.36	.37
50	.05	.11	.12	-.07	.22	.23
100	.03	.08	.09	-.06	.15	.16
200	.02	.06	.06	-.05	.10	.11
400	.01	.05	.05	-.05	.07	.09

TABLE 3: Experiments With Heteroskedastic Logistic Disturbances
$h_2(z) \equiv |1+z|/\sqrt{2}$

N = 200

	Maximum Score			Logit Maximum Likelihood			
μ	Bias	SDEV	RMSE	Bias	SDEV	RMSE	A.Bias
-2.0	.01	.05	.05	.28	.38	.47	.17
-1.5	.01	.04	.04	.03	.11	.11	.01
-1.0	.01	.03	.03	.01	.08	.08	.00
-0.5	.01	.03	.03	-.01	.09	.09	-.01
0.0	.01	.05	.05	-.13	.15	.20	-.12
0.5	.01	.14	.14	-.38	.21	.43	-.39
1.0	-.01	.29	.29	-.60	.21	.64	-.64

$\mu = 0$

	Maximum Score			Logit Maximum Likelihood		
N	Bias	SDEV	RMSE	Bias	SDEV	RMSE
25	.02	.47	.47	-.11	.51	.52
50	.03	.24	.24	-.13	.32	.35
100	.02	.11	.12	-.13	.23	.26
200	.01	.05	.05	-.13	.15	.20
400	.00	.03	.03	-.13	.11	.17

TABLE 4: Experiments With Heteroskedastic Logistic Disturbances
$h_3(z) = 2.07\exp(-|1+z|)$

N = 200

	Maximum Score			Logit Maximum Likelihood			
μ	Bias	SDEV	RMSE	Bias	SDEV	RMSE	A.Bias
-2.0	.06	.26	.26	-.02	.08	.08	-.03
-1.5	.08	.24	.25	.00	.07	.07	-.00
-1.0	.06	.22	.23	.00	.06	.06	-.00
-0.5	.07	.22	.23	.01	.07	.07	.00
0.0	.09	.23	.25	.03	.08	.09	.03
0.5	.12	.26	.29	.09	.11	.14	.09
1.0	.17	.33	.37	.17	.14	.22	.16

$\mu = 0$

	Maximum Score			Logit Maximum Likelihood		
N	Bias	SDEV	RMSE	Bias	SDEV	RMSE
25	.18	.41	.45	.05	.25	.26
50	.16	.31	.35	.05	.17	.18
100	.12	.28	.30	.04	.12	.13
200	.09	.23	.25	.03	.08	.09
400	.05	.19	.20	.03	.06	.07

TABLE 5. Experiments With Homoskedastic Logistic Disturbances
$h_4(z) = 1$

N = 200

	Maximum Score			Logit Maximum Likelihood			
μ	Bias	SDEV	RMSE	Bias	SDEV	RMSE	A.Bias
-2.0	.06	.25	.26	.03	.17	.17	0
-1.5	.07	.23	.24	.02	.12	.12	0
-1.0	.05	.20	.20	.01	.10	.10	0
-0.5	.06	.20	.21	.00	.11	.11	0
0.0	.07	.22	.23	-.00	.14	.14	0
0.5	.10	.26	.28	.00	.17	.17	0
1.0	.15	.32	.36	.01	.23	.23	0

$\mu = 0$

	Maximum Score			Logit Maximum Likelihood		
N	Bias	SDEV	RMSE	Bias	SDEV	RMSE
25	.17	.51	.54	.02	.42	.42
50	.14	.36	.38	.01	.27	.27
100	.10	.28	.30	.00	.20	.20
200	.07	.22	.23	-.00	.14	.14
400	.03	.17	.17	.00	.10	.10

TABLE 6. Experiments With Homoskedastic t(3) Disturbances
$h_4(z) = 1$

N = 200

	Maximum Score			Logit Maximum Likelihood			
μ	Bias	SDEV	RMSE	Bias	SDEV	RMSE	A.Bias
-2.0	.05	.20	.21	.03	.13	.13	.01
-1.5	.05	.17	.17	.01	.09	.09	.00
-1.0	.04	.15	.16	.01	.08	.08	.00
-0.5	.05	.16	.16	.00	.08	.08	.00
0.0	.05	.17	.18	-.00	.11	.11	-.00
0.5	.08	.21	.23	-.01	.15	.15	-.01
1.0	.11	.30	.32	-.04	.20	.20	-.05

$\mu = 0$

	Maximum Score			Logit Maximum Likelihood		
N	Bias	SDEV	RMSE	Bias	SDEV	RMSE
25	.12	.41	.43	-.00	.33	.33
50	.11	.29	.31	-.00	.21	.21
100	.09	.23	.25	.00	.15	.15
200	.05	.17	.18	-.00	.11	.11
400	.04	.14	.14	-.01	.07	.07

TABLE 7. Experiments With Homoskedastic Uniform Disturbances
$h_4(z) = 1$

N = 200

	Maximum Score			Logit Maximum Likelihood			
μ	Bias	SDEV	RMSE	Bias	SDEV	RMSE	A.Bias
-2.0	.11	.35	.37	.06	.27	.28	-.02
-1.5	.09	.32	.33	.05	.24	.25	-.00
-1.0	.10	.30	.31	.03	.17	.18	-.00
-0.5	.08	.28	.29	.01	.16	.16	.00
0.0	.11	.28	.31	.02	.19	.19	.02
0.5	.15	.33	.36	.08	.22	.23	.07
1.0	.23	.39	.46	.18	.27	.32	.18

$\mu = 0$

	Maximum Score			Logit Maximum Likelihood		
N	Bias	SDEV	RMSE	Bias	SDEV	RMSE
25	.18	.73	.75	.04	.65	.65
50	.18	.50	.53	.03	.40	.40
100	.15	.39	.42	.03	.27	.27
200	.11	.28	.31	.02	.19	.19
400	.07	.22	.23	.02	.13	.13

TABLE 9. Bootstrap Estimation of Root Mean Square Error

Experiment			True RMSE	Estimate of RMSE	
h(*)	μ	N		Mean	SDEV
$h_1(*)$	−2	100	.12	.11	.09
		200	.07	.07	.04
	−1	100	.07	.08	.04
		200	.05	.06	.02
	0	100	.09	.10	.07
		200	.06	.07	.03
$h_2(*)$	−2	100	.22	.34	.35
		200	.05	.13	.16
	−1	100	.07	.12	.12
		200	.03	.05	.03
	0	100	.12	.23	.18
		200	.05	.09	.07
$h_3(*)$	−2	100	.31	.30	.16
		200	.26	.26	.14
	−1	100	.27	.27	.12
		200	.23	.23	.10
	0	100	.30	.29	.12
		200	.25	.24	.10
$h_4(*)$	−2	100	.38	.38	.26
		200	.26	.28	.18
	−1	100	.27	.31	.18
		200	.20	.22	.11
	0	100	.30	.33	.15
		200	.23	.25	.10