Maximum Score Methods

by Robert P. Sherman

In a seminal paper, Manski (1975) introduces the Maximum Score Estimator (MSE) of the structural parameters of a multinomial choice model and proves consistency without assuming knowledge of the distribution of the error terms in the model. As such, the MSE is the first instance of a semiparametric estimator of a limited dependent variable model in the econometrics literature.

Maximum score estimation of the parameters of a binary choice model has received the most attention in the literature. Manski (1975) covers this model, but Manski (1985) focuses on it. The key assumption that Manski (1985) makes is that the latent variable underlying the observed binary data satisfies a linear α-quantile regression specification. (He focuses on the linear median regression case, where $\alpha = 0.5$.) This is perhaps an under-appreciated fact about maximum score estimation in the binary choice setting. If the latent variable were observed, then classical quantile regression estimation (Koenker and Bassett, 1978), using the latent data, would estimate, albeit more efficiently, the same regression parameters that would be estimated by maximum score estimation using the binary data. In short, the estimands would be the same for these two estimation procedures.

Assuming that the underlying latent variable satisfies a linear α-quantile regression specification is equivalent to assuming that the regression parameters in the linear model do not depend on the regressors and that the error term in the model has zero α-quantile conditional on the regressors. Under these assumptions, Manski (1985) proves strong consistency of the MSE. The zero conditional α-quantile assumption does not require the existence of any error

moments and allows heteroscedastic errors of an unknown form. This flexibility is in contrast to many semiparametric estimators of comparable structural parameters for the binary choice model. Many of these latter estimators require the existence of error moments and most require more restrictive assumptions governing the relation of errors to regressors. See Powell (1994) for more.

The weak zero conditional α-quantile assumption comes at a price, however. Extrapolation power is limited: off the observed support of the regressors it is not possible to identify the conditional probability of the choice of interest, but only whether this probability is above or below $1 - \alpha$. See Manski (1995, pp.149-150). There are also disadvantages associated with the estimation procedure. The maximum score criterion function is essentially a sum of indicator functions of parameters. This lack of smoothness precludes using standard optimization routines to compute the MSE. Moreover, Kim and Pollard (1990) show that this type of discontinuity leads to a convergence rate of $n^{-1/3}$ rather than the $n^{-1/2}$ convergence rate attained by most semiparametric estimators of parameters in this model. In addition, Kim and Pollard (1990) show that the MSE has a nonstandard limiting distribution. The properties of this distribution are largely unknown, making asymptotic inference problematic. Also, Abrevaya and Huang (2005) prove that the bootstrapped MSE is an inconsistent estimator of the parameters of interest, seemingly precluding bootstrap inference.

To repair some of these shortcomings, Horowitz (1992) develops a smoothed MSE (SMSE) for the linear median regression case. Standard optimization routines can be used to compute this estimator. In addition, the SMSE converges at a faster rate than the MSE and has a normal limit law allowing first order asymptotic inference. Horowitz (2002) proves that the bootstrapped SMSE provides asymptotic refinements and in various simulations demonstrates

the superiority of bootstrap tests over first order asymptotic tests. Kordas (2005) generalizes Horowitz' (1992) SMSE to cover all α-quantiles.

In the next section, we present the multinomial choice model under random utility maximization as well as some intuition behind maximum score estimation in this context. We then discuss the relation between maximum score estimation in the binary response model and quantile regression. Next, we present Kim and Pollard's (1990) intuitive argument for the nonstandard rate of convergence of the MSE in the binary model. Finally, we discuss Horowitz' (1992) method of smoothing the MSE.

The Random Utility Maximization Model of Choice and the MSE

Manski (1975) developed the MSE for the multinomial choice model in the context of random utility maximization. Suppose the ith individual in a sample of size n from a population of interest must make exactly one of J choices, where $J \geq 2$.

For $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, J\}$, let U_{ij} denote the utility to individual i of making choice j. Assume the structural form $U_{ij} = X_{ij}'\beta + \epsilon_{ij}$ where X_{ij} is an observable $m \times 1$ vector of explanatory variables, β is a unknown $m \times 1$ parameter vector, and ϵ_{ij} is an unobservable random disturbance. (A more general set-up can be accommodated. For example, there can be a different parameter vector associated with each choice.)

The utilities associated with the choices an individual faces are latent, or unobservable. However, an individual's choice is observable. Suppose we adopt the maximum utility model of choice: if individual i makes choice j then $U_{ij} > U_{ik}$ for all $k \neq j$. For any event E, define the indicator function $\{E\} = 1$ if E occurs and 0 otherwise. Define

$$Y_{ij} = \{U_{ij} > U_{ik}, \text{ for all } k \neq j\} = \{X_{ij}'\beta + \epsilon_{ij} > X_{ik}'\beta + \epsilon_{ik}, \text{ for all } k \neq j\}. \tag{1}$$

If choice j has maximum utility, then $Y_{ij} = 1$. Otherwise, $Y_{ij} = 0$. Thus, for each individual i, we observe X_{ij}, $j = 1, 2, \ldots, J$ and Y_{ij}, $j = 1, 2, \ldots, J$.

The traditional approach to estimating β in the multinomial choice model under the assumption of random utility maximization is the method of maximum likelihood in which the errors are iid with a distribution known up to scale. The likelihood function to be maximized has the form

$$\sum_{i=1}^{n} \sum_{j=1}^{J} Y_{ij} \log P\{Y_{ij} = 1 \mid X_{i1}, X_{i2}, \ldots, X_{iJ}, b\}\,.$$

For example, when ϵ_{ij} has the Type 1 extreme-value cdf $F(t) = \exp(-\exp(-t)$, $t \in \mathbb{R}$, McFadden (1974) shows that the likelihood probabilities have the multinomial logit specification $\exp(X_{ij}'b) \left[\sum_{k=1}^{J} \exp(X_{ik}'b)\right]^{-1}$. The corresponding likelihood function is analytic and globally concave. Despite the consequent computational advantages, this specification makes very strong assumptions about the distribution of the errors. The MSE is consistent under much weaker assumptions about the errors. Manski (1975) only assumes that the disturbances ϵ_{ij} are independent and identically distributed (iid) across choices and independent but not necessarily identically distributed across individuals.

Write b for a generic element of the parameter space. It follows trivially from (1) that the infeasible criterion function

$$\sum_{i=1}^{n} \sum_{j=1}^{J} Y_{ij} \{X_{ij}'b + \epsilon_{ij} > X_{ik}'b + \epsilon_{ik}, k \neq j\}$$

attains its maximum value of n at $b = \beta$. Since, for each i, the disturbances ϵ_{ij} are iid variates, this suggests estimating β with the maximizer of the so-called score function

$$\sum_{i=1}^{n} \sum_{j=1}^{J} Y_{ij} \{X_{ij}'b > X_{ik}'b, k \neq j\}\,. \tag{2}$$

A score for a parameter b is the number of correct predictions made by predicting Y_{ij} to be 1 whenever $X_{ij}'b$ exceeds $X_{ik}'b$ for all $k \neq j$. A maximizer of the score function is an MSE of β. The maximizer need not be unique.

The MSE in the Binary Choice Model and Quantile Regression

Now consider the binary model where $J = 2$. Define $Y_i = Y_{i1}$ (implying $Y_{i2} = 1 - Y_i$) and $X_i = X_{i1} - X_{i2}$. Then the score function in (2) reduces to

$$\sum_{i=1}^{n} \left[Y_i \{X_i'b > 0\} + (1 - Y_i)\{X_i'b < 0\} \right] . \tag{3}$$

Substitute $1 - \{X_i'b > 0\}$ for $\{X_i'b < 0\}$ in (3) and expand each summand to see that maximizing (3) is equivalent to maximizing

$$S_n(b) = n^{-1} \sum_{i=1}^{n} (2Y_i - 1)\{X_i'b > 0\} . \tag{4}$$

Note that $Y_i = \{Y_i^* > 0\}$ where $Y_i^* = X_i'\beta + \epsilon_i$ with $\epsilon_i = \epsilon_{i1} - \epsilon_{i2}$. For ease of exposition, write (Y^*, Y, X, ϵ) for $(Y_1^*, Y_1, X_1, \epsilon_1)$ and x for an arbitrary point in the support of X. Thus, $Y = \{Y^* > 0\}$ where $Y^* = X'\beta + \epsilon$.

Before proceeding further, we must consider what interpretation to give to the parameter β in the last paragraph. The interpretation depends on our assumptions. For example, if we assume that β does not depend on x and that for every x, $\mathbb{E}[Y^* \mid x] = x'\beta$, then β is such that the conditional mean of Y^* given $X = x$ is equal to $x'\beta$. However, if we assume that $\text{MED}(Y^* \mid x) = x'\beta$, then β is such that the conditional median of Y^* given $X = x$ is equal to $x'\beta$. In general, the β satisfying conditional mean assumption will be different from the β satisfying the conditional median assumption. Similarly, if we assume that for $\alpha \neq 0.5$, the

conditional α-quantile of Y^* given x is equal to $x'\beta$, then this β will, in general, be different from the β satisfying the conditional median assumption.

With this in mind, for $\alpha \in (0,1)$, write $Q_\alpha(Y^* \mid x)$ for the α-quantile of Y^* given $X = x$. Fix an $\alpha \in (0,1)$ and assume the linear α-quantile regression specification. That is, assume that for each x in the support of X, there exists a unique parameter β_α, depending on α but not on x, such that $Q_\alpha(Y^* \mid x) = x'\beta_\alpha$. It is easy to show that this implies a zero conditional α-quantile restriction on ϵ: $Q_\alpha(\epsilon \mid x) = 0$ for all x. (However, it is also easy to show that the argument does not go the other way!)

For $\alpha \in (0,1)$, define

$$S_n^\alpha(b) = n^{-1} \sum_{i=1}^{n} [(2Y_i - 1) - (1 - 2\alpha)]\{X_i'b > 0\}\,. \tag{5}$$

Clearly, $S_n^{0.5}(b) = S_n(b)$ in (4). Assume that the linear α-quantile regression specification holds for some $\alpha \in (0,1)$. To see that it makes sense, under this assumption, to estimate β_α with the maximizer of $S_n^\alpha(b)$, consider $S^\alpha(b) = ES_n^\alpha(b)$. We see that

$$\begin{aligned} S^\alpha(b) &= E^X\left[E\left[(2Y-1) - (1-2\alpha)\right]\{X'b > 0\} \mid X\right] \\ &= E^X\left[\left[(2P\{-\epsilon < X'\beta_\alpha \mid X\} - 1) - (1-2\alpha)\right]\{X'b > 0\}\right]\,. \end{aligned}$$

The linear α-quantile regression specification implies a zero conditional α-quantile restriction on ϵ: for all x, $P\{\epsilon \leq 0 \mid x\} \geq \alpha$ and $P\{\epsilon \geq 0 \mid x\} \geq 1 - \alpha$. Thus, $x'\beta_\alpha > 0$ if and only if $P\{-\epsilon \leq x'\beta_\alpha \mid x\} \geq P\{-\epsilon \leq 0 \mid x\} \geq 1 - \alpha$. Deduce that for each possible value of X, the term in outer brackets in the last expression is maximized at $b = \beta_\alpha$. It follows that $S^\alpha(b)$ is maximized at $b = \beta_\alpha$. The analogy principle (e.g., Manski, 1988) prescribes using a maximizer of $S_n^\alpha(b)$ to estimate β_α.

The summands of the criterion function in (5) depend on b only through indicator functions of sets. As such, each summand has a "sharp edge", to use the terminology of Kim and Pollard (1990). These authors provide a beautiful heuristic for why estimators that optimize empirical processes with "sharp-edge" summands converge at rate $n^{-1/3}$, rather than the usual $n^{-1/2}$ rate. They decompose the sample criterion function into a deterministic trend plus noise. Then, for each possible parameter value, they consider how the trend and the noise compete for dominance. The key insight is that only a parameter value for which the trend does not overwhelm the standard deviation of the noise has a fighting chance of being an optimizer. Sharp edges produce standard errors with nonstandard sizes leading to the nonstandard $n^{-1/3}$ rate. We now examine how their argument works for the MSE for a very simple model.

Assume the median regression specification for the model $Y = \{\beta - X - \epsilon > 0\}$. Thus, $\beta_{0.5} = (\beta, -1)$ where the slope coefficient is known to equal -1 and and the intercept β is the unknown parameter of interest. Assume that ϵ has median zero and is independent of X, so that the conditional median zero restriction is trivially satisfied. Also, assume that the distributions of X and ϵ have everywhere positive Lebesgue densities.

Refer to (4). Define $S(b) = ES_n(b) = E(2Y - 1)\{X'b > 0\}$. In the intercept example, $S(b) = E(2\{\epsilon < \beta - X\} - 1)\{X < b\}$. Simple calculations show that

$$S(b) = 2\int_{-\infty}^{b} F_\epsilon(\beta - t) f_x(t) dt - F_x(b)$$

where $F_\epsilon(\cdot)$ is the cdf of ϵ, $f_x(\cdot)$ is the pdf of X, and $F_x(\cdot)$ is the cdf of X. Write $f_\epsilon(\cdot)$ for the pdf of ϵ. Again, simple calculations show that

$$S'(b) \quad = \quad 2F_\epsilon(\beta - b) f_x(b) - f_x(b)$$

$$S''(b) \quad = \quad 2F_\epsilon(\beta - b)f'_x(b) - f_x(b)2f_\epsilon(\beta - b) - f'_x(b)\,.$$

By the median restriction, we see that $S'(\beta) = 0$ and $S''(\beta) = -2f_x(\beta)f_\epsilon(0) < 0$. Thus, $S(b)$ is locally maximized at $b = \beta$. In fact, the given assumptions imply that $S(b)$ is globally and uniquely maximized at $b = \beta$. The MSE maximizes $S_n(b) - S_n(\beta)$. For each b, decompose $S_n(b) - S_n(\beta)$ into a sum of a deterministic trend and a random perturbation:

$$S_n(b) - S_n(\beta) = S(b) - S(\beta) + [S_n(b) - S_n(\beta) - [S(b) - S(\beta)]]\ .$$

A Taylor expansion about β shows that for b near β, the trend $S(b) - S(\beta)$ is approximately quadratic with maximum value zero at $b = \beta$:

$$S(b) - S(\beta) \approx S''(\beta)(b - \beta)^2\,.$$

By a central limit theorem, for large n, the random contribution $S_n(b) - S_n(\beta) - [S(b) - S(\beta)]$ is approximately normally distributed with mean zero and variance σ_b^2/n where

$$\sigma_b^2 = E\left[(2Y - 1)[\{X < b\} - \{X < \beta\}]\right]^2 - \left[E(2Y - 1)[\{X < b\} - \{X < \beta\}]\right]^2\ .$$

For b near β, the second term is much smaller than the first. It is the first term that accounts for the sharp-edge effect. It equals

$$F_x(\beta) + F_x(b) - 2\left[F_x(\beta)\{b > \beta\} + F_x(b)\{b < \beta\}\right]\ .$$

A Taylor expansion of both $F_x(b)$ terms about β shows that this term is approximately equal to $|b - \beta|f_x(\beta)$ for b near β. Thus, near β, the criterion function $S_n(b) - S_n(\beta)$ is approximately equal to a quadratic maximized at β, namely, $-c_1(b - \beta)^2$ for $c_1 > 0$, plus a zero-mean random variable with standard deviation equal to $c_2 n^{-1/2}|b - \beta|^{1/2}$ for $c_2 > 0$. Values of b for which $-c_1(b - \beta)^2$ is much bigger in absolute value than $c_2 n^{-1/2}|b - \beta|^{-1/2}$ have little chance of

maximizing $S_n(b) - S_n(\beta)$. Rather, the maximizer is likely to be among those b values for which, for some $c > 0$,

$$(b-\beta)^2 \leq cn^{-1/2}|b-\beta|^{1/2}\,.$$

Rearranging, we see that the maximizer is likely to be among the b values for which

$$|b-\beta| \leq cn^{-1/3}\,.$$

This is the essence of the heuristic presented by Kim and Pollard (1990) for $n^{-1/3}$ convergence rates. These authors also note that when criterion functions are smooth, the variance of the random perturbation usually has order $|b-\beta|^2$ (instead of $|b-\beta|$) which, by the same heuristic, leads to the faster $n^{-1/2}$ convergence rate.

Smoothing the MSE

In order to remedy some of the shortcomings of the MSE, Horowitz (1992) develops a smoothed maximum score estimator (SMSE) under a linear median regression specification for the latent variable in the binary model. He replaces the indicator function in (4) with a smooth approximation. His SMSE maximizes a criterion function of the form

$$n^{-1}\sum_{i=1}^{n}(2Y_i - 1)K(X_i'b/\sigma_n)$$

where K is essentially a smooth cdf and σ_n approaches zero as the sample size increases. Thus, $K(X_i'b/\sigma_n)$ approaches the indicator function $\{X_i'b > 0\}$ as $n \to \infty$. By smoothing out the "sharp-edge" of the indicator function in (4), Horowitz is able to use Taylor expansion arguments to show that the SMSE, under slightly stronger conditions than those required for consistency of the MSE, converges at rate n^{δ} for $2/5 \leq \delta < 1/2$ and has a normal limit. The

exact rate of convergence depends on certain smoothness assumptions and satisfies an optimality property (see Horowitz, 1993). The normality result makes it possible to do standard asymptotic inference with the SMSE.

Kordas (2005) applies the smoothing technique of Horowitz (1992) to the criterion function in (5) and obtains asymptotic results similar to those of Horowitz (1992) for any $\alpha \in (0, 1)$.

References

Abrevaya, J. and J. Huang (2005): "On the Bootstrap of the Maximum Score Estimator," *Econometrica*, to appear.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60(3), 505–531.

Horowitz, J. L. (1993): "Optimal Rates of Convergence of Parameter Estimators in the Binary Response Model with Weak Distributional Assumptions," *Econometric Theory*, 9, 1–18.

Horowitz, J. L. (2002): "Bootstrap Critical Values for Tests based on the Smoothed Maximum Score Estimator," *Econometrica*, 111, 141–167.

Kim, J. and D. Pollard (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18(1), 191–219.

Koenker, R. and G. Bassett, Jr. (1978): "Regression Quantiles," *Econometrica*, 46(1), 33–50.

Kordas, G. (2005): "Smoothed Binary Regression Quantiles," *Journal of Applied Econometrics*, to appear.

Manski, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice,"

Journal of Econometrics], 3, 205–228.

MANSKI, C. F. *(1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator,"* Journal of Econometrics], 27, 313–333.

MANSKI, C. F. (1988): *Analog Estimation Methods in Econometrics.* Chapman and Hall, New York.

MANSKI, C. F. (1995): *Identification Problems in the Social Sciences.* Harvard University Press, Cambridge, Mass.

MCFADDEN, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," Chapter 4 in *Frontiers in Econometrics*, Editor Zarembka, Academic Press, 105–142.

POWELL, J. L. (1994): "Estimation of Semiparametric Models," Chapter 41 in *Handbook of Econometrics, Volume 4*, Editors Engle and McFadden, 2443–2521.