P. Conway
20 January 1999

Maximizing Sign Matches as an Estimation Technique

Consider the stochastic process:

$$y_{jc} = \hat{e}(\hat{a}_c x_{jc} - s_c) + \mathring{a}_{jc} \qquad \hat{e}(0) = 0 \qquad (1)$$

The subscript j indicates an observation from the sequence (1, ..., M), while the subscript c indicates a draw from the sequence (1, ..., C). $\hat{e}(.)$ is an unknown sign-preserving function of its argument. The variables y_{jc}, x_{jc} and s_c are all observed. The variable $\mathring{a}_{jc}$ is a random, zero-mean error. The coefficients $\hat{a}_c$ and the functional form of $\hat{e}(.)$ are unknown. The econometrician is interested in deriving unbiased estimates b_c of the underlying coefficients $\hat{a}_c$.

This is a standard estimation question in international trade theory. Consider the subscript c to index countries and the subscript j to index factors of production. All factor-endowment-based theories of trade are based upon the hypothesis that a country's net exports of a product will be a function of the country's comparative advantage as defined by factor abundance. The competing theories within that class will have competing predictions of the volume of trade (here, the size of y_{jc}). In fact, the monotonicity of $\hat{e}(.)$ is not assured in simple parameterizations of the model. They will all predict, however, that the existence of comparative advantage (here, $\hat{a}_c x_{jc} - s_c > 0$) will lead to net exports of the good ($y_{jc} > 0$) while the converse (comparative disadvantage leading to net imports) will also hold. Trefler (1993, 1995) provides two empirical papers based upon one such theory. His estimation results are potentially biased due to a lack of consideration of the implications of this misspecified $\hat{e}(.)$ function.

Estimation.

Estimation of $\hat{a}_c$ is especially simple if the hypothesis is maintained that $\hat{e}(.)$ is a one-for-one mapping. OLS estimation of the equation

$$z_{jc} = y_{jc} - s_c = b_c x_{jc} + e_{jc} \qquad (2)$$

separately for each c will lead to unbiased estimates b_c of $\hat{a}_c$.

While these estimates are unbiased under the maintained hypothesis, they are not robust to alternative specifications of $\hat{e}(.)$. Consider the example that $\hat{e}(.) = k_j(\hat{a}_c x_{jc} - s_c)$. Substitution of this into (2) will lead to

$$z_{jc} = y_{jc} - s_c = -(1+k_j) s_c + k_j \hat{a}_c x_{jc} + \mathring{a}_{jc} \qquad (3)$$

The estimation technique used above will be misspecified (there should be an intercept term) and will lead to biased estimates of $\hat{a}_c$. Clearly, joint estimation of k_j and $\hat{a}_c$ is the preferable estimation strategy in this case, although the results will depend upon the maintained hypothesis of the functional form specified for $\hat{e}(.)$. The degree of robustness of the results to this hypothesis could be determined through use of non-parametric regression techniques: see Yatchew (1998) for an overview.

Maximum-score estimator.

A sign-match exists when y_{jc} and the difference $(\hat{a}_c x_{jc} - s_c)$ take the same sign, or alternatively when the product $y_{jc}^* (\hat{a}_c x_{jc} - s_c) > 0$. In the absence of the random error $\mathring{a}_{jc}$, our knowledge of the function $\hat{e}(.)$ from (1) assures that a sign-match will be observed for each observation. Violation of the sign-match condition will be observed for large absolute values of $\mathring{a}_{jc}$. I derive an estimator for $\hat{a}_c$ based upon this property.

Manski (1975, 1985) first considered the problem of robust estimation of $\hat{a}_c$ when y_{jc} is not observed but the sign of y_{jc} (denoted $z_{jc} = \text{sgn}(y_{jc})$) is observed. He observed that the coefficients derived from a logit estimation are contingent on the underlying errors following a logistic distribution. By contrast, a median regression that maximizes the number of sign matches is distribution-insensitive, and is a consistent estimator of $\hat{a}_c$.[1] In Manski and Thompson (1986), the authors conduct Monte Carlo experiments to compare maximum likelihood logit and maximum score estimators. They conclude that for most cases of non-logistic distributions of errors and for a wide range of sample sizes the maximum score estimator has less bias and greater precision in coefficient estimation.[2]

The econometric problem specified here is amenable to estimation using the maximum score estimation technique. While y_{jc} is observed, it is in practice difficult for the econometrician to specify and estimate a model that nests all alternative hypotheses. More parsimonious models will be characterized by non-monotonic $\hat{e}(.)$ in the data. The maximum score estimation technique is robust to such non-

---

[1] The median regression case is a special case of the á-quantile regression problem investigated by Koenker and Bassett (1978).

[2] The ratios of bias and of standard deviation of coefficient estimates reported by Manski and Thompson (1986) are:

	25	50	100	200	400
Heteroskedastic errors (Table 2)					
Bias	-2.67	-1.40	-2.00	-2.50	-5.00
Standard deviation	1.28	2.00	1.88	1.67	1.40
Homoskedastic errors (Table 5)					
Bias	0.12	0.07	0.00	0.00	0.00
Standard deviation	1.21	1.33	1.40	1.57	1.70

Simulations of other heteroskedastic error structures yield similar results.

monotonicities in the estimation of $\hat{a}_c$.

Implementing the maximum score estimator.

The general numerical technique is a simple one. The index c runs from 1 to C, and there are M observations for each value of c. For the M country-specific observations indexed by c, I undertake a grid search of estimators b_c. Those b_c that generate the highest number of sign matches ($y_{jc}^* (b_c x_{jc} - s_c) > 0$) form the basis for the estimate of $\hat{a}_c$.

In contrast to the typical continuous likelihood function maximization, it is common for the grid search, no matter how finely defined, to yield a large number of estimates b_c with equally large sign-match totals. I consider two methods for choosing among these b_c.

- First, I choose the average of the b_c generating the largest number of sign matches (b_c^{avg}). These need not be contiguous points in the grid search, so that the value b_c^{avg} may not itself generate the maximum sign matches.

- Second, I define an initial starting value for b_c (e.g., $b_c^o = 1$). The estimate is then that value of b_c from among those with maximum sign matches that minimizes deviation from the initial starting value (b_c^{md})

Simulation 1: Country-specific deviation in vector s_c.

First, I investigate the comparative properties of the maximum-score and OLS estimators in a model for which OLS is appropriate. In this trial I undertake 30 independent simulations of 5000 observations each. Each simulation included 50 trials of 100 observations each, and the percentage of sign-matches reported is the average over those 50 trials. Each of the 30 simulations was undertaken with the same structure:

$$y_{jc} = .5 (x_{jc} - (.75 + m_c)) + e_{jc} \qquad (4)$$

Each of the 30 simulations (subscript c) had a common but unknown (to the econometrician) vector (.75+ m_c). The maximum score estimator derived an estimate b_c of (.75+m_c). The OLS estimator provided coefficients on x_{jc} and an intercept, using the observed values of y_{jc} as dependent variable. The reported coefficient is derived from these. The across-simulation means and standard deviations are as follows:

	Signs matched (percent)	Mean b_c	Std. Dev. b_c
MD sign-match	88.6	0.77	0.18
AVG sign-match	88.6	0.78	0.21
OLS		0.77	0.10

Mean value of m_c over all simulations: 0.02.

The maximum score estimators yield nearly identical results to those of OLS. However, as expected, the mean standard error for maximum-score estimation was greater than that of OLS. The ratio (.18/.10) of standard deviations is similar to that found by Manski and Thompson (1986) for its largest samples in homoskedastic logistic samples.

Simulation 2: Variations in the ê(.) function.

This simulation has 100 replications of an economy with C=40 and M=9. The dependent variable y_{jc} is derived from the following equation.

$$y_{jc} = k_j (x_{jc} - s_c) + å_{jc} \qquad (5)$$

I specify ê(.) as a function of the index j, taking the values $k_j = .1 * j$ for jå(1,..,M). x_{jc} is created from a unit normal distribution centered at 1, while $s_c = .05 * c$ for cå(1,..,C). $å_{jc}$ is drawn from a unit normal distribution centered on zero and multiplied by .2.

As is evident from the estimation results, the maximum score estimates are robust to the introduction of this varying k_j term. However, as the OLS results derived from estimation of (2) on the same data indicate, the simple OLS estimator will be biased.[3]

	Signs matched (percent)	Mean b_c	Std. Dev. b_c
MD sign-match	86.7	1.003	0.080
AVG sign-match	86.7	1.007	0.091
OLS (no intercept)		0.757	0.022

Conclusion.

This note has indicated the value of use of the maximum-score estimator in a class of problems common to empirical international trade theory. In the absence of knowledge about the specific functional form, the sign-matching technique of the maximum-score estimator will be a robust and consistent estimation procedure.

[3] It will be the case that with proper specification of the underlying model non-linear least squares will return consistent and more precise estimates of $â_c$. When the underlying model is misspecified, however, the bias in results can be extreme (as in this example).

Bibliography

Koenker, R. and G. Bassett: “Regression Quantiles”, Econometrica 46, 1978, pp. 33-50.

Manski, Charles: “Maximum Score Estimation of the Stochastic Utility Model of Choice”, Journal of Econometrics 3, 1975, pp. 205-228.

Manski, Charles: “Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator”, Journal of Econometrics 27, 1985, pp. 313-333.

Manski, Charles and T. Scott Thompson: “Operational Characteristics of Maximum Score Estimation”, Journal of Econometrics 32, 1986, pp. 85-108.

Trefler, Daniel: "International Factor Price Differences: Leontief was Right!", Journal of Political Economy 101/6, 1993, pp. 961-987.

Trefler, Daniel: "The Case of the Missing Trade and Other Mysteries", American Economic Review 85/5, 1995, pp. 1029-1046.

Yatchew, Adonis: “Nonparametric Regression Techniques in Economics”, Journal of Economic Literature 36, 1998, pp. 669-721.