How Much Do Starting Values Really Matter? An Empirical Comparison of Genetic Algorithm and Traditional Approaches

Glynn T. Tonsor
Assistant Professor, Dept. of Agricultural Economics
Michigan State University
213D Agriculture Hall
East Lansing, MI 48824-1039
Phone: (517) 353-9848
gtonsor@msu.edu

and

Terry Kastens
Professor, Dept. of Agricultural Economics
Kansas State University
Manhattan, KS

Working paper being prepared for presentation at the
American Agricultural Economics Association Annual Meetings
Long Beach, CA, July 23-26, 2006

How Much Do Starting Values Really Matter? An Empirical Comparison of Genetic Algorithm and Traditional Approaches

Abstract

This research evaluates the impact of using different starting conditions in estimating meat demand systems. Results suggest that as the econometric task becomes increasingly nonlinear, specification of starting conditions becomes increasingly important. This work demonstrates implications of failing to use the best available starting value conditions and how these implications vary with the complexity of the underlying econometric model of interest. Furthermore, this piece proposes a universal approach to be used by all applied econometric practioners to developing appropriate starting values for use in subsequent model estimation.

Key words: applied econometrics, genetic algorithms, starting values

Introduction

It is often said that applied empirical research is "as much an art as a science." Such statements arrive due to the fact that empirical results are frequently very sensitive to the selection of data sets, model specification, and econometric techniques used in the empirical exercise. While a whole career could be made in analyzing each of these factors (and the numerous sub-factors impacting them), this research seeks to shed light on the extent to which different starting values of parameters to be estimated impact empirical findings.

Most applied empirical work fails to disclose the procedures used in establishing starting conditions for their estimated models. In fact, the vast majority of authors don't even bother to acknowledge starting conditions in their discussion. This raises two questions: 1) "Do authors even consider the impact of start values on their models?" and 2) "Just how sensitive are subsequent model results and implications to the actual starting conditions employed in the model estimation process?" Furthermore, this issue is becoming of increasing importance as econometric models continue to become more sophisticated and highly non-linear as allowed by constantly increasing computer power. These increasingly complex models are inherently more likely to be estimated with algorithms converging to local rather than global solutions as the number of local solutions tends to increase quickly with the level of complexity inherent in the underlying model.

There are two primary objectives of this research: 1) to examine the magnitude of differences and economic implications of these differences in applying various starting condition rules to recently published applied empirical exercises; 2) to develop a practical approach to recommend for use in deriving appropriate starting conditions that can be used by future applied econometricians.

Genetic Algorithm Introduction

Economic modeling has benefited significantly from relatively recent developments in empirical research methodologies and enhancements in computing power enabling more sophisticated modeling techniques to be evaluated. These advancements have, at least in theory, led to overall improvements in the quality and reliability of model results; therefore improving the ability of economists to provide decision makers with appropriate advice and valuable information.

An enhancement of high interest in this research is the increasing use of genetic algorithm techniques. While not originating from economists, economists are increasingly using these tools to improve their empirical modeling methodologies. Genetic algorithms are utilized by econometricians to increase confidence in finding globally optimum solutions rather than local optima. Researchers are never fully assured that their empirical search techniques have revealed global optimums. The more nonlinear the optimization functions are, the more likely traditional algorithms are to stop iterating and propose results that in actuality are local optimums. By design, all gradient-type algorithms take a starting point (or vector of starting values) and search from that point to another, gravitating towards the local optimum nearest to the starting point. This is why a truly exhaustive search, including multiple starting points, is needed to gain confidence in finding global optima solutions.

Dorsey and Mayer often are credited with being the first to analyze the ability of genetic algorithms to solve optimization problems plagued by the problems just discussed. They provide a nice introductory and application discussion of how genetic

algorithms work.[1] Genetic algorithms (GA) iterate towards a solution through a process very similar to that of natural evolution (Goldberg). The GA takes an initial population of values (similar to the starting values used by traditional algorithm approaches) and randomly selects a subset of this initial population to utilize in generating "offspring" which are the next set of candidate values. The success of such an approach hinges on the "proper selection" to use in generating the next set of candidate values. As iterations continue, the traits yielding the most preferred objective value continue to persist while less desirable traits die out. This part of the GA is what is similar to evolutionary process and the theory of "survival of the fittest."

This GA process is different from traditional algorithms in the sense that it does not move "from point to point" along a function being evaluated, but rather it randomly (where this random process incorporates an evaluation of the desirability of each candidate value) chooses a set of values to evaluate in the next iteration. This randomness is what characterizes GA processes to be less susceptible to "getting stuck" on local solutions or excessively struggling with non-differentiable issues. The "random selection" process is similar to having the search algorithm "jump" along the objective function as opposed to the traditional gradient-based approach of "moving along" or "walking along" the objective function.

[1] Our discussion on genetic algorithms is not intended to be exhaustive, but rather a "sufficient introduction." Those interested in a more in-depth discussion are advised to consult the Dorsey and Mayer article.

Methods

The methodology employed in this paper includes estimating a series of recently used applied meat demand models under different sets of starting conditions. Meat demand models are considered as they are frequently used in the development and presentation of new, usually more sophisticated demand models as well as the fact that a whole wealth of meat demand models exist with little explicit consideration of starting value condition impacts on subsequent model results. The demand models considered are of varying degrees of nonlinearity further allowing us to gauge the relative importance of employing different starting conditions for various levels of model complexity. These models include variations of recently used AIDS (Almost Ideal Demand System) models. The starting conditions considered can be broken down into two sets: 1) default starting values of the statistical package (e.g., 0.01 in SAS) and 2) starting values implied by using a genetic algorithm search technique.

To evaluate the economic impact of starting conditions we estimate each of the considered demand models (e.g., variations of the AIDS models) using each of these starting conditions. The resulting parameters, elasticities, etc. are then statistically compared in both in-sample and out-of-sample exercises to assess the implications of employing the different starting conditions. From these exercises, a generalized approach is developed with the intention of being used by future researchers as an accepted standardized approach to generating appropriate starting conditions prior to actual estimation of the final model used in drawing the economic implications of actual focus in the research at hand. This generalized approach includes the employment of a genetic algorithm in improving starting conditions.

Data

Data used in this analysis consists of quarterly per capita disappearance and price series for beef, pork, poultry, and fish for the US domestic market. This data was collected over the 1976(1) -2001(4) period yielding 104 total observations. Quarterly price and disappearance data ranging from 1976(1) through the 1993(4) were obtained from Dr. Henry Kinnucan and are identical to that used by Kinnucan et al. Subsequent beef, pork, and poultry per capita disappearance data from 1994(1) to 2001(4) were obtained from the United States Department of Agriculture (USDA), Economic Research Service (ERS) supply and utilization tables published in the *Red Meat Yearbook*. Beef, pork, and poultry price data are average retail prices obtained from ERS.[2] Corresponding fish per capita disappearance data were obtained following the same procedure used by Kinnucan et al. and discussed in more detail by Schmitz and Capps. Using a fish consumer price index obtained from ERS and a base price from 1983(1), quarterly fish price data spanning from 1994(1) to 2001(4) were derived for this analysis.[3]

Table 1 provides summary statistics of the entire dataset and the estimated expenditure share allocated to beef, pork, poultry, and fish consumption for US consumers. Upon inspection of the budget share estimates, it is apparent that the representative US household allocates a high percentage of its animal protein

[2] More specifically, the beef and pork prices used have variable names BFVRCCUS and PKVRCCUS, respectively. Furthermore, the poultry price is calculated as the sum of expenditures on whole fryers and turkey divided by the sum of per capita disappearance of chicken and turkey. For additional details on these prices, readers are referred to USDA, 2006.

[3] More specifically, observed fish consumption (obtained from Kinnucan et al.) was regressed against quarterly dummy and annual trend variables. Corresponding regression coefficients (which were all significant in a model with an R-squared of 0.82) were then used to quarterize annual consumption over the 1994(1) to 2001(4) period. Any error associated with this allocation was then evenly distributed across all four quarters. As noted by previous authors (Kinnucan et. al.; Dameus et. al.), US fish data is poor and procedures undertaken in this study are necessary to analyze quarterly US fish demand. This data and additional details are available upon request.

expenditures (with nearly 50% being distributed to beef) to beef, pork, and poultry and a lower percentage to fish.

Initial Results

In estimating each combination of starting value approach and the different AIDS demand systems; beef, pork, poultry, and fish are treated as a weakly separable group. With homogeneity, Engle aggregation, and symmetry imposed, iterated seemingly unrelated regression estimates were calculated while dropping one equation to avoid singularity of the error covariance matrix. The parameters of this omitted equation are obtained by utilizing the imposed theoretical restrictions noted above. More details on each of the three AIDS model specifications can be found in Appendix I.

As shown in Tables 2-4, estimation of the traditional AIDS model is not impacted by the method used in determining starting values. This is observed by the fact that all three starting value specifications result in models converging to a solution, yielding the same coefficient estimates, and hence the same model fit statistics.

Conversely, the second demand model specification considered (*Basic GAIDS* in the tables), converged to a solution using either SAS default starting values or starting values implied by use of a Genetic Algorithm. However, Tables 3 & 4 reveal that using the Genetic Algorithm to develop starting conditions yields different coefficient estimates that in turn describe a model with a better in-sample fit than using the default SAS starting values.

The third, and most complex model specification, never converged when using the default SAS starting values but did converge using the Genetic Algorithm approach.

By extension, the coefficient estimates and model fit statistics are “different” across the approaches.

In summary across the three model specifications, we have initially found that as the underlying model becoming increasingly complex and more nonlinear, the specification of starting values becomes increasingly important. Furthermore, in some cases, using a Genetic Algorithm to develop starting conditions can alleviate issues of non-convergence. Applying this information; our current proposal to applied econometricians is to utilize a genetic algorithm approach to first develop a set of starting conditions to subsequently be used in solving the actual econometric model of interest.

Conclusions

There are a number of contributions and implications of this research. First, actual acknowledgement is provided of the fact that starting values are not given proper discussion and consideration in applied econometric exercises. Secondly, we demonstrate the implications of failing to use proper starting value conditions and how these implications vary with the complexity of the underlying econometric model of interest. Furthermore, this work suggests a universal approach to be used by all applied econometric practioners to developing appropriate starting values for use in subsequent model estimation. Current extensions being implemented include evaluations of the effect of these different starting value approaches on out-of-sample predictive accuracy and economic implications stemming from different elasticity estimates.

References

Alston J. M., J. A. Chalfant, and N. E. Piggott. "Incorporating Demand Shifters in the Almost Ideal Demand System." *Economic Letters* 70(2001):73-78.

Berndt, E. and N. Savin. "Evaluation and Hypothesis Testing in Singular Equation Systems with Autoregressive Disturbances." *Econometrica* 32(1975):937-957.

Bewley, R. *Allocation Models: Specification, Estimation, and Applications*. Cambridge, MA: Ballinger, 1986.

Bollino, C.A. "GAIDS: A Generalized Version of the Almost Ideal Demand System." *Economic Letters*. 33(1990):127-129.

Bridge Financial Data Center. Futures settlement price and implied volatility data (CD-ROM), 1989-2004. Chicago, IL, 2005.

Capps, Jr. O. "Alternative Estimation Methods of Nonlinear Demand Systems." *Western Journal of Agricultural Economics*. 8:1(1983):50-63.

Dameus, A., F. Richter, B. Brorsen, and K. Sukhdial. "AIDS versus the Rotterdam Demand System: A Cox Test with Parametric Boostrap." *Journal of Agricultural and Resource Economics.* 27:2(2002):335-347.

Deaton, A. and J. Muellbauer. "An Almost Ideal Demand System." *American Economic Review*. 70(1980):312-326.

Dorsey, R.E. and W.J. Mayer. "Genetic Algorithms for Estimation Problems with Multiple Optima, Nondifferentiability, and Other Irregular Features." *Journal of Business and Economics Statistics*. 13:1(1995):53-66.

Goldberg, D.E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

Kinnucan, H.W., H. Xiao, C. Hsia, and J.D. Jackson. "Effects of Health Information and Generic Advertising on US Meat Demand." *American Journal of Agricultural Economics.* 79(1997):13-23.

LMIC. Livestock Marketing Information Center, Lakewood, Colorado. Internet address: http://www.lmic.info/.

MathWorks, MATLAB 7.0.4. Natick, Massachusetts: The MathWorks Inc., 1994-2006.

Park, J.L., R.B. Holcomb, K.C. Raper, and O. Capps, Jr. "A Demand Systems Analysis of Food Commodities by US Households Segmented by Income." *American Journal of Agricultural Economics*. 78:2(1996):290-300.

Piggott, N. “The Nested PIGLOG Model: An Application to U.S. Food Demand.” *American Journal of Agricultural Economics* 85(2003):1-15.

Piggott, N., J. Chalfant, J. Alston, and G. Griffith. “Demand Response to Advertising in the Australian Meat Industry.” *American Journal of Agricultural Economics*. 78:2(1996):226-279.

Piggott, N.E. and T.L. Marsh. “Does Food Safety Information Impact US Meat Demand?” *American Journal of Agricultural Economics*. 86:1(2004):154-174.

Raper, K.C., M.N. Wanzala, and R.M. Nayga Jr. “Food Expenditure and Household Demographic Composition in the US: A Demand Systems Approach.” *Applied Economics*. 34(2002):981-992.

Robert, C.P. and G. Casella. "*Monte Carlo Statistical Methods.*" (second edition). New York: Springer-Verlag, 2004.

SAS Institute Inc., SAS 9.1.3 Help and Doc., Cary, NC: SAS Institute Inc., 2000-2004.

Tonsor, G.T. “Feedlot Cattle Crush: Joint Distribution Model Development, Evaluation, and Applications.” Agricultural Economics Ph.D. Dissertation, 2006. Kansas State University.

United States Department of Agriculture. Economic Research Service. Washington DC.

United States Department of Agriculture. *Beef and Pork Values and Price Spreads Explained*. Economic Research Service Publication LDP-M-118-01. Accessed on May 1, 2006at: http://www.ers.usda.gov/publications/ldp/APR04/ldpm11801/ldpm11801r.pdf.

United States Department of Labor *Consumer Expenditure Survey*. Bureau of Labor Statistics.

United States Department of Labor *Detailed Annual Consumer Price Indices*. Bureau of Labor Statistics.

Table 1. Summary Statistics of Quarterly US Data (1976-2001)

	Mean	Std. Dev.
Beef Consumption (lbs per capita)	18.40	2.09
Pork Consumption (lbs per capita)	12.72	0.91
Poultry Consumption (lbs per capita)	19.87	4.83
Fish Consumption (lbs per capita)	3.62	0.66
Beef Retail Price ($/lb)	2.52	0.45
Pork Retail Price ($/lb)	1.91	0.40
Poultry Retail Price ($/lb)	0.87	0.13
Fish Retail Price ($/lb)	2.51	0.77
Meat and Fish Expenditure ($/capita)	96.93	18.80
Beef Expenditure Share	0.48	0.06
Pork Expenditure Share	0.25	0.01
Poultry Expenditure Share	0.18	0.04
Fish Expenditure Share	0.09	0.02

Table 2. Do each the model and starting value combinations result in model convergence?

	Traditional AIDS	*Basic GAIDS*	*Enhanced GAIDS*
SAS Default 0.01	*YES*	*YES*	*NO*
OLS Implied	*YES*	*NA*	*NA*
Genetic Algorithm	*YES*	*YES*	*YES*

Table 3. Do coefficients differ from using SAS default starting values?

	Traditional AIDS	*Basic GAIDS*	*Enhanced GAIDS*
SAS Default 0.01	-	-	-
OLS Implied	*NO*	*NA*	*NA*
Genetic Algorithm	*NO*	*YES*	*YES*

Table 4. Are the in-sample model fit statistics better than using SAS default starting values?

	Traditional AIDS	*Basic GAIDS*	*Enhanced GAIDS*
SAS Default 0.01	-	-	-
OLS Implied	*NO*	*NA*	*NA*
Genetic Algorithm	*NO*	*YES*	-

Appendix I.

This analysis estimates three different demand models. The models, as listed here and discussed in the text, are noted in order of increasing complexity.

Model Specification #1: The first estimated model is the linear approximation to the traditional AIDS (Almost Ideal Demand System) Model (see Deaton and Muellbauer for details). The employed specification contains 12 parameters to be estimated. This model is estimated under three different starting conditions: 1) using the default starting values implied by SAS of 0.01 for all parameters, 2) using the estimated coefficients found by OLS estimation of each individual equation, and 3) using the coefficients suggested by a Genetic Algorithm technique the iterates 50,000 times attempting to minimize the sum of squared errors in the system.

Model Specification #2: The second estimated model is the most basic specification of a Generalized Almost Ideal Demand System (GAIDS) (see Piggott and Marsh or Bollino for details). As noted by Piggott and Marsh, this demand system specification allows for pre-committed quantities, time effects, food safety issues, etc. to be evaluated in a manner that is consistent with derived elasticities being invariant to the units of measurement employed in the data used. However, the "cost" of this improvement is added nonlinearity to the AIDS model. As such, this makes for a nice transitional model to compare with the linear approximate AIDS specification. The employed specification contains 17 parameters to be estimated.

This model is estimated under two different starting conditions: 1) using the default starting values implied by SAS of 0.01 for all parameters and 2) using the

coefficients suggested by a Genetic Algorithm technique the iterates 50,000 times attempting to minimize the sum of squared errors in the system. It is not feasible to OLS techniques to derive starting values due to the nonlinear price index implicit in estimation of the GAIDS system.

Model Specification #3: The final estimated model is a more complex specification of a Generalized Almost Ideal Demand System (GAIDS) containing 36 parameters to be estimated. This specification adds additional parameters, implicitely assumed to be zero in *Model Specification #*2. Again, this adds nonlinearity to the model making it a nice transitional model to compare.

This model is estimated under two different starting conditions: 1) using the default starting values implied by SAS of 0.01 for all parameters and 2) using the coefficients suggested by a Genetic Algorithm technique the iterates 50,000 times attempting to minimize the sum of squared errors in the system. It is not feasible to OLS techniques to derive starting values due to the highly nonlinear price index implicit in estimation of the GAIDS system.