Working Papers in Econometrics and Applied Statistics

Estimating Lorenz Curves Using a Dirichlet Distribution

Duangkamon Chotikapanich and William E Griffiths

No. 110 – November 1999

Working Papers in Econometrics and Applied Statistics
Department of Econometrics
University of New England
Armidale, NSW 2351, Australia.

http://www.une.edu.au/febl/EconStud/emetwp.htm

ISSN 0157-0188
ISBN 1 86389 612 0

Estimating Lorenz Curves
Using a Dirichlet Distribution

Duangkamon Chotikapanich
Curtin University of Technology

and

William E Griffiths
University of New England

Abstract
The Lorenz curve relates the cumulative proportion of income to the cumulative proportion of population. When a particular functional form of the Lorenz curve is specified it is typically estimated by linear or nonlinear least squares assuming that the error terms are independently and normally distributed. Observations on cumulative proportions are clearly neither independent nor normally distributed. This paper proposes and applies a new methodology which recognizes the cumulative proportional nature of the Lorenz curve data by assuming that the proportion of income is distributed as a Dirichlet distribution. Five Lorenz-curve specifications were used to demonstrate the technique. Once a likelihood function and the posterior probability density function for each specification are derived we can use maximum likelihood or Bayesian estimation to estimate the parameters. Maximum likelihood estimates and Bayesian posterior probability density functions for the Gini coefficient are also obtained for each Lorenz-curve specification.

Keywords: posterior distribution; Metropolis-Hastings algorithm; Gini coefficient;

1. Introduction

The Lorenz curve is one of the most important tools upon which the measurement of income inequality is based. For a given economy or region, it relates the cumulative proportion of income to the cumulative proportion of population, after ordering the population according to increasing level of income. Two general approaches to Lorenz curve estimation have been adopted. In the first, a particular assumption about the statistical distribution of income is made, the parameters of this income distribution are estimated, and a Lorenz curve consistent with the distributional assumption, and consistent with the parameter estimates for that distribution, is obtained. See, for example, McDonald (1984) and McDonald and Xu (1995). In the second approach, a particular functional form for the Lorenz curve is specified and estimated directly. It is this second approach which is the focus of this paper.

Early breakthroughs on Lorenz curve estimation were those of Gastwirth (1972) and Kakwani and Podder (1973, 1976). Kakwani and Podder recognized the multinomial nature of grouped data and used a Lorenz curve specification that, after transformation, could be placed in an approximate linear model framework. Other specifications have typically been estimated by linear or nonlinear least squares without any regard for the fact that the assumption of independent normally distributed errors is unrealistic (Kakwani 1980, Basmann et al 1990, Chotikapanich 1993). Clearly, observations on cumulative proportions, or even their logarithms if such a transformation is convenient, will be neither independent nor normally distributed. Sarabia et al (1999) overcome this problem by suggesting a distribution-free method of estimation. Suppose that a Lorenz curve has n unknown parameters, and that M observations on the cumulative proportions are available. They find a set of parameter estimates for each of the $K = \binom{M}{n}$ subsets of n observations. Since each of the subsets yields n equations in n unknown parameters, a set of parameter estimates is obtained by solving these equations. The medians of the sets of parameter estimates are recommended as the final set of estimates. No distribution theory is available for this procedure, but the authors do provide some bootstrap standard errors.

An alternative way to proceed, and the approach adopted in this paper, is to choose a distributional assumption that is consistent with the proportional nature of the data and to pursue maximum likelihood or Bayesian estimation. Maximum likelihood estimators have well known statistical properties, and Bayesian estimation provides a framework for finite sample inference with several well recognized advantages. See, for example, Poirier (1995). One multivariate distribution which has shares which sum to one as its vector of random variables is the Dirichlet distribution. By relating the parameters of the Dirichlet distribution to Lorenz curve differences, we can allow for the cumulative proportional nature of the Lorenz curve data, and set up a likelihood function dependent on the unknown parameters of the Lorenz curve. A similar approach was adopted by Woodland (1979) for estimation of share equations that arise in demand and production theory. Although our discussion and examples relate to the use of grouped data, our methodology could also be applied to unit recorded data.

In Section 2, we outline the distributional assumptions and how they relate to Lorenz curve estimation. The likelihood function and a general posterior probability density function (pdf) for a set of unknown Lorenz curve parameters are derived. A Metropolis-Hastings algorithm that can be used to estimate marginal posterior pdfs for the parameters and their moments is described. To illustrate our suggested techniques we use data on Sweden and Brazil considered earlier by Shorrocks (1983) and revisited by Sarabia et al (1999). These data are described in Section 3; five different Lorenz functions that we use in the empirical work are presented. The results are given and discussed in Section 4. Several questions are investigated. To examine whether the results are sensitive to the chosen estimation technique we compare our estimates and their standard errors (and posterior standard deviations) to those obtained by Sarabia et al (1999), and those obtained using least squares (after taking logarithms where relevant). Since Lorenz-curve estimation is usually a first step towards estimating inequality, maximum likelihood (ML) estimates and Bayesian posterior pdfs for the Gini coefficient are obtained for each Lorenz-curve specification. A comparison of the ML and Bayesian results gives an indication of any differences between asymptotic and finite sample inferences. Finally, we examine whether functional form preference is sensitive to the chosen estimation technique and form of inference.

2. Models, Assumptions and Estimation

Suppose we have available observations on cumulative proportions of population $(\pi_1, \pi_2, \ldots, \pi_M$ with $\pi_M = 1)$ and corresponding cumulative proportions of income $(\eta_1, \eta_2, \ldots, \eta_M$ with $\eta_M = 1)$ obtained after ordering population units according to increasing income. We wish to use these observations to estimate a parametric version of a Lorenz curve that we write as $\eta = L(\pi;\beta)$ where β is an $(n\times 1)$ vector of unknown parameters. Clearly, one would not expect all data points to lie exactly on the curve $\eta_i = L(\pi_i;\beta)$. It seems reasonable to assume, however, that conditional on the population proportions π_i, the income shares $q_i = \eta_i - \eta_{i-1}$ are random variables with means

$$E(q_i) \;=\; E(\eta_i) - E(\eta_{i-1}) \;=\; L(\pi_i;\beta) - L(\pi_{i-1};\beta) \tag{1}$$

Our proposal is to also assume $q = (q_1, q_2, \ldots, q_M)'$ follows a Dirichlet distribution which is a distribution consistent with the share nature of the random vector q. The probability density function (pdf) for the Dirichlet distribution is given by

$$f(q\,|\,\alpha) \;=\; \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_M)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_M)}\, q_1^{\alpha_1 - 1} q_2^{\alpha_2 - 1} \ldots q_M^{\alpha_M - 1} \tag{2}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_M)'$ are the parameters of the pdf and $\Gamma(.)$ is the gamma function. By relating the α_i to the Lorenz function, we can find a pdf for q which has the mean given in equation (1) and which is a function of the Lorenz curve parameters. Working in this direction, we set

$$\alpha_i \;=\; \lambda\left[L(\pi_i;\beta) - L(\pi_{i-1};\beta)\right] \tag{3}$$

where λ is an additional unknown parameter. This definition for α_i gives the desired result because the mean of the Dirichlet distribution is given by

$$E(q_i) = \frac{\alpha_i}{\alpha_1 + \alpha_2 + \cdots + \alpha_M}$$

$$= \frac{\lambda[L(\pi_i;\beta) - L(\pi_{i-1};\beta)]}{\lambda\sum_{i=1}^{M}[L(\pi_i;\beta) - L(\pi_{i-1};\beta)]}$$

$$= L(\pi_i;\beta) - L(\pi_{i-1};\beta) \tag{4}$$

since $L(\pi_M;\beta) = 1$ and $L(\pi_0;\beta) = 0$. We can now write the pdf for q as

$$f(q\,|\,\theta) = \Gamma(\lambda)\prod_{i=1}^{M}\frac{q_i^{\lambda[L(\pi_i;\beta)-L(\pi_{i-1};\beta)]-1}}{\Gamma(\lambda[L(\pi_i;\beta) - L(\pi_{i-1};\beta)])} \tag{5}$$

where $\theta = (\beta', \lambda)'$.

The variances and covariances between the shares are given by

$$\text{var}(q_i) = \frac{E(q_i)[1 - E(q_i)]}{\lambda + 1} \tag{6}$$

$$\text{cov}(q_i, q_j) = -\frac{E(q_i)E(q_j)}{\lambda + 1} \tag{7}$$

Thus, the income shares are correlated, with correlations given by

$$r_{ij} = -\left(\frac{E(q_i)E(q_j)}{[1 - E(q_i)][1 - E(q_j)]}\right)^{1/2} \tag{8}$$

Since the variances depend on $E(q_i)$, the shares are also heteroskedastic. The parameter λ acts as an inverse variance parameter. The larger the value of λ, the better the fit of the Lorenz curve to the data.

The maximum likelihood estimate for θ can be found by maximizing the log-likelihood function

$$\log[f(q\,|\,\theta)] = \log\Gamma(\lambda) + \sum_{i=1}^{M}(\lambda[L(\pi_i;\beta) - L(\pi_{i-1};\beta)] - 1)\log q_i$$
$$- \sum_{i=1}^{M}\log\Gamma(\lambda[L(\pi_i;\beta) - L(\pi_{i-1};\beta)]) \tag{9}$$

For Bayesian estimation we use uniform priors on the elements of β, over the feasible ranges for those parameters. Since $(\lambda+1)$ is like an inverse variance parameter, we use a uniform prior for $\log(\lambda+1)$. Also, assuming *a priori* independence of β and λ, yields the prior pdf

$$f(\theta) \;=\; f(\beta,\lambda) \;\propto\; \frac{I(\beta)}{\lambda+1} \qquad\qquad \lambda > 0 \tag{10}$$

where $I(\beta)$ is am indicator function equal to unity for feasible values of β and zero if β falls outside the region that defines $L(\pi;\beta)$ as a Lorenz curve. Application of Bayes theorem involves multiplying together equations (5) and (10) to obtain the kernel of the posterior pdf for θ

$$f(\theta\,|\,q) \;\propto\; f(\theta)\; f(q\,|\,\theta) \tag{11}$$

For all the Lorenz-curve specifications that we estimate, the posterior pdf in (11) is analytically intractable in the sense that we cannot carry out the necessary integration to obtain marginal posterior pdfs for individual parameters and the posterior moments of these parameters. These quantities can be estimated, however, by using a Metropolis-Hastings algorithm to draw observations on θ from the posterior pdf $f(\theta\,|\,q)$. See, for example, Albert and Chib (1996) and Geweke (1999). We used the following random-walk algorithm with the maximum likelihood covariance V_θ used as a covariance matrix for the random-walk generator function. The steps for drawing the $(m+1)th$ observation $\theta_{(m+1)}$ are:

1. Draw a candidate value θ^* from a $N(\theta_{(m)}, cV_\theta)$ distribution where c is a scalar set such that θ^* is accepted approximately 40-50% of the time.
2. Compute

$$r = \frac{f(\theta^* \mid q)}{f(\theta_{(m)} \mid q)}$$

   Note that this ratio can be computed without knowledge of the normalising constant for $f(\theta \mid q)$. Also, if any of the elements of θ^* fall outside the feasible parameter region, then $f(\theta^* \mid q) = 0$.
3. Draw a value u for a uniform random variable on the interval (0,1).
4. If $u \le r$, set $\theta_{(m+1)} = \theta^*$.

   If $u > r$, set $\theta_{(m+1)} = \theta_{(m)}$.
5. Return to step 1, with m set to $m+1$.

Observations generated in this way can be placed in histograms to estimate marginal posterior pdfs, and sample means and standard deviations can be used to estimate posterior means and standard deviations.

3. Data and Lorenz Curves

To illustrate our suggested techniques we use income distribution data on national samples of income recipients for a year close to 1970, for two countries: Sweden and Brazil. These data were used by Sarabia et al (1999). They were derived from Jain (1975) and first published in Shorrocks (1983). The data are in the form of decile cumulative income shares. Shorrocks used the data on these two countries as part of a group of twenty countries to examine the ranking of income distributions given different social states. Sarabia et al (1999) used the data to illustrate their proposed method for the estimation of Lorenz curves. The data on these two countries were chosen because of their differences in the degree of inequality in income distributions.

A large number of functional forms have been suggested in the literature for modelling the Lorenz curve. For details of the various alternatives, see Sarabia et al (1999), and references therein. To keep our study manageable, we chose only 5, ranging from one simple function with only one unknown parameter, to two three-parameter functions which are more flexible, but also harder to estimate precisely. The 5 different Lorenz functions to which we applied the two data sets are:

$$L_1(\pi; k) = \frac{e^{k\pi} - 1}{e^k - 1} \qquad k > 0 \tag{12}$$

$$L_2(\pi; \alpha, \delta) = \pi^{\alpha}[1 - (1 - \pi)^{\delta}] \qquad \alpha \geq 0,\ 0 < \delta \leq 1 \tag{13}$$

$$L_3(\pi; \delta, \gamma) = [1 - (1 - \pi)^{\delta}]^{\gamma} \qquad \gamma \geq 1,\ 0 < \delta \leq 1 \tag{14}$$

$$L_4(\pi; \alpha, \delta, \gamma) = \pi^{\alpha}[1 - (1 - \pi)^{\delta}]^{\gamma} \qquad \alpha \geq 0, \gamma \geq 1,\ 0 < \delta \leq 1 \tag{15}$$

$$L_5(\pi; a, b, d) = \pi - a\pi^{d}(1 - \pi)^{b} \qquad a > 0,\ 0 < d \leq 1, 0 < b \leq 1 \tag{16}$$

The function L_1 is the relatively simple one-parameter function suggested by Chotikapanich (1993); L_2 coincides with the proposal of Ortega et al (1991). L_3 is a well-known form of Lorenz curve suggested by Rasche et al (1980) and L_4 is an extension of L_3 and L_2 introduced by Sarabia et al (1999). Note that L_4 nests both L_2 and L_3, with L_2 being L_4 with $\gamma = 1$ and L_3 being L_4 with $\alpha = 0$. Setting both $\gamma = 1$ and $\alpha = 0$ yields the Lorenz curve $L = 1 - (1 - \pi)^{\delta}$ which originates from the classical Pareto distribution. The function L_5 is the "beta function" proposed by Kakwani (1980). It is considered one of the best performers among a number of different functional forms for Lorenz curves. See, for example, Datt (1998). Note that, when $a = 1$ and $d = 1$, L_5 is the same as L_2 with $\alpha = 1$.

Once a Lorenz curve has been estimated, one is usually interested in various inequality measures that are related to it. As an example, we compute maximum likelihood estimates and posterior pdfs for the Gini coefficients that can be derived from each of the Lorenz functions. In each case the Gini coefficient is defined as

$$G = 1 - 2\int_0^1 L(\pi;\beta)\,d\pi \tag{17}$$

Alternative expressions for G can be found for some of the Lorenz curves. However, with the exception of L_1, they still generally involve a numerical integral. We obtain ML and Bayesian estimates by numerically evaluating (17) in each case. For ML estimation, numerical integration is performed with β replaced by the ML estimate $\hat{\beta}$. For Bayesian estimation, the integral is evaluated for each draw of β from the posterior pdf of β.

4. Results

In addition to ML and Bayesian estimation using the assumption of a Dirichlet distribution, we also estimated each function using nonlinear least squares. Nonlinear least squares is "optimal" under the assumption that the η_i are independent normally distributed random variables with mean $L(\pi_i, \beta)$ and constant variance. Although this assumption is not realistic for data which are cumulative proportions, nonlinear least squares is a popular estimation technique, and so the sensitivity of parameter estimates to the choice of technique is useful information.

Point estimates of the Lorenz curve parameters and the corresponding Gini coefficients for Sweden and Brazil are presented in Tables 1 and 2, respectively. The Bayesian point estimates are the posterior means estimated from 75,000 draws using the random-walk Metropolis algorithm, after discarding the first 10,000 draws as a "burn in". The estimates obtained by Sarabia et al (1999), using their proposed technique, are also given for L_2, L_3 and L_4.

[Table 1 near here]

Table 1 provides the estimates for Sweden. For L_1, L_2, L_3 and L_5 the estimates of the Lorenz parameters and the Gini coefficients are not sensitive to the estimation techniques. For L_4 different estimation techniques give very different Lorenz parameter estimates. Despite these differences, the estimates for the Gini coefficient are very similar across all functional forms and estimation techniques. An exception is the one obtained from L_4 using Sarabia's method. Reasons for the atypical outcomes from L_4 are addressed later.

[Table 2 near here]

The remarks made about Sweden also hold for the estimates for Brazil given in Table 2. One difference is the Gini coefficient estimates obtained from ML and Bayes, when using L_1. They are 0.50 and 0.52, when all other estimates are approximately 0.63. When we discuss goodness of fit, we discover that this difference can be attributable to a poor fit. Tables 1 and 2 also reveal the difference in inequality in Sweden and Brazil, with Sweden exhibiting the lower level of inequality.

Standard errors for the ML and nonlinear least squares estimates, and posterior standard deviations for the parameters from Bayesian estimation, are presented in Tables 3 and 4 for Sweden and Brazil, respectively. The posterior standard deviations are estimated from the 75,000 Metropolis draws, and corresponding values of the Gini coefficient. The standard errors for the Gini coefficient for ML and nonlinear least squares were calculated using the asymptotic approximation

$$\text{var}(\hat{G}) = \frac{\partial G}{\partial \beta'} V_\beta \frac{\partial G}{\partial \beta} \tag{18}$$

where V_β is the asymptotic covariance matrix for the ML (or nonlinear least squares) estimator for β. Expressions derived using (18) for each of the Lorenz curves are given in the Appendix.

[Tables 3 and 4 near here]

From Tables 3 and 4, we make the following observations:

1. With the exception of L_4, to which special attention is devoted later, the Bayesian posterior standard deviations are larger than the ML standard errors. Since the ML standard errors are large-sample approximations, whereas the posterior standard deviations reflect finite sample uncertainty, this comparison reveals the extent to which misleading inferences can be made from a large-sample approximation. To illustrate this point further, we plotted the estimated posterior pdfs for (i) α in the function L_2 for Sweden (Figure 1), (ii) the Gini coefficient from L_4 for Sweden (Figure 2), and (iii) the Gini coefficient from L_5 for Brazil (Figure 3). Normal pdfs, centred at the ML estimates, and with standard deviations equal to the ML standard errors, were also drawn on these figures. When viewed through Bayesian eyes, these are the pdfs typically used to make large sample inferences. In all three figures, the Bayesian pdfs have fatter tails, suggesting that ML estimation understates the uncertainty about these quantities.

2. The bootstrap standard errors computed by Sarabia et al (1999) are vastly different from those provided by the other approaches. The difference is sufficiently great to cast doubt on their validity, particularly when the distribution theory for the Sarabia et al technique is not available.
3. The standard errors for nonlinear least squares (which is optimal when the cumulative income proportions are normally distributed) are also quite different. Thus, although the point estimates of the Lorenz parameters and the Gini coefficient are quite insensitive to the chosen estimation technique, interval estimates, and the assessment of estimation precision, depend heavily on the distributional assumption and related method of estimation.
4. Overall, point estimates of the Gini coefficient are insensitive to the Lorenz curve specification. (Those for L_1 from ML and Bayes, using the Brazilian data, are exceptions.) There is, however, considerable variation in the standard errors and posterior standard deviations. Thus, our knowledge or degree of uncertainty about the value of the Gini coefficient does depend on the functional form chosen for the Lorenz curve. This fact is clearly depicted by the posterior pdfs that are graphed in Figures 4 and 5. Figure 4 contains the posterior pdfs for Sweden's Gini coefficient, obtained using L_1, L_4 and L_5. The 3-parameter Lorenz curves L_4 and L_5 suggest relatively precise information about the Gini coefficient. The 1-parameter function L_1 exhibits considerable uncertainty. Figure 5 contains the posterior pdfs for Brazil's Gini coefficient, obtained using L_2, L_4 and L_5. Here, the story is similar, except that the precision in estimation implied by L_5 is much greater than that implied by L_2 and L_4.

We turn now to the question of goodness of fit. Which of the Lorenz functions best fits the data? As we will see, the answer to this question has a bearing on precision of estimation that we discussed under the last point (4). The problem of choosing between the alternative functions can be addressed in a number of ways. For a straight goodness-of-fit comparison, we compare values of information inaccuracy (Theil 1967, 1975). For testing nested functional forms we use likelihood ratio tests for the ML estimates; from a Bayesian perspective, we assess whether various parametric restrictions are true by examining the posterior probability in the region near the restrictions.

Let $\hat{q}_i$ denote the predicted income shares obtained from an estimated model. Theil's (1967) measure of information inaccuracy is defined as

$$I = \sum_{i=1}^{M} q_i \log\left(\frac{q_i}{\hat{q}_i}\right) \tag{19}$$

Functions with smaller values of I are better fits than those with larger values. If the q_i are similar to the $\hat{q}_i$, then knowing their values provides little information relative to knowledge of the predictions. The function is a good fit. On the other hand, q_i quite different from the $\hat{q}_i$ convey considerable information, leading to a large value of I and a poor fit.

The information inaccuracy measure was computed using predictions from the ML estimates, and predictions from the Bayesian posterior means. The outcomes are presented in Table 5. In both countries, L_5 is the best fit, L_4 and L_3 are approximately the same in terms of fit, and are preferred to L_2, which, in turn, is preferred to L_1. There is virtually no difference in the measures obtained from the

ML estimates and those obtained from the Bayesian estimates. There is a difference between Sweden and Brazil, however. For Brazil, the fit of the best function L_5 is much better, and the fit of the worse function L_1, is worse. Also, for Sweden, the function L_2 is only marginally worse than L_3 and L_4. In the case of Brazil it is noticeably inferior.

It is interesting that the precision with which the Gini coefficient is estimated is directly related to how well the function fits the data. The relative magnitudes of the posterior standard deviations for the Gini coefficients (Tables 3 and 4) reflect the relative magnitudes of the information inaccuracy measures. These relativities are also conveyed by the posterior pdfs in Figures 4 and 5.

The second way that we investigated choice of functional form was by examining whether nested versions of L_4 and L_5 would be adequate. Given the results on goodness of fit, one would expect that at least L_3 would be an acceptable restricted version of L_4. Table 6 contains χ^2 values for likelihood ratio tests for various hypotheses. These results confirm our conjecture about the relationship between L_3 and L_4 for both Sweden and Brazil. Also, L_2 is an acceptable restricted version of L_4 for Sweden, but not for Brazil, a conclusion consistent with goodness-of-fit results. Finally, a restricted version of L_2, obtained by setting $\alpha = 1$, is clearly rejected relative to the best-fitting L_5.

The likelihood ratio test is a large-sample approximate test whose properties can be questionable in small samples, particularly in our case, where there are only 10 observations. An alternative procedure, valid in finite samples, is to examine the posterior probability mass in the region where the restrictions hold. Proceeding in this direction, we obtained scatter plots of the Markov-Chain Monte-Carlo observations for *a* and *d* in L_5. These scatter plots appear in Figures 6 and 7, for Sweden and Brazil, respectively. Setting $a = 1$ and $d = 1$ in L_5, and $\alpha = 1$ in L_2, gives the same restricted version of a Lorenz function. Both plots show no probability in the vicinity of $a = 1$ and $d = 1$. For Brazil there is a concentration of probability around $d = 1$, but this concentration does not extend beyond a= 0.92, indicating no support for both restrictions.

The posterior pdfs for α from L_4 were plotted (Figures 8 and 9) to see if L_3 is an acceptable restricted version of L_4 from a Bayesian perspective. For both Brazil and Sweden, these pdfs have modes near zero. The Swedish one declines very slowly – it is almost uniform – from zero to 0.5, then sharply to 0.7. That for Brazil declines almost linearly from zero to 0.6. Both suggest $\alpha = 0$ is an acceptable value and hence there is nothing to gain by moving from the 2-parameter function L_3 to the 3-parameter function L_4. Figures 8 and 9 also explain why, for L_4, the estimates of α were very sensitive to estimation technique (Tables 1 and 2). The ML estimate is approximately equal to the mode of the pdf which is near zero. The Bayesian estimate is the posterior mean which is near the centre of the distribution in each case.

The above exercise was repeated for the parameter γ from L_4. See Figures 10 and 11. Interestingly, there was a symmetry between the pdfs for α and γ. For

Sweden, the pdf for γ was gradually increasing, but almost uniform, from 1 to 1.55. For Brazil it increased linearly from 1 to 1.35. After the increasing part of the functions, there was a sharp decline at the right side of the distributions. The reason that a hypothesis test suggested L_2 was an acceptable restricted version of L_4 for Sweden, but not for Brazil, is clear. There is substantial probability mass at 1 for the former, but not for the latter.

A remaining puzzle is: Why is the Gini coefficient from L_4 estimated relatively accurately, as reflected by the standard errors and standard deviations in Tables 3 and 4, and posterior pdfs in Figures 4 and 5, when the parameters α and γ from L_4 are estimated with little precision? We shed light on this question by examining scatter plots of the Markov Chain Monte Carlo observations on α and γ. See Figures 12 and 13. The cigar-shaped nature of these plots indicates a very high correlation between the parameters. Thus, although we cannot estimate the parameters accurately individually, we can estimate combinations of the parameters very accurately. It appears that the data does not discriminate between large γ with small α and small γ with large α, and that these combinations have similar implications for the value of the Gini coefficient. Also, we observe in the Swedish case that, although the hypotheses $\alpha = 0$ and $\gamma = 1$ are reasonable when considered separately, the joint hypothesis ($\alpha = 0, \gamma = 1$) is clearly rejected.

Conclusions and Summary

One way of estimating a Lorenz curve is to assume a particular distribution for income, estimate the parameters of that distribution, and derive the corresponding Lorenz curve. Another way is to assume a particular Lorenz curve, and estimate its parameters. For this second approach we have suggested a distributional assumption and corresponding estimation techniques which are consistent with the proportional nature of Lorenz-curve data, can be employed with any Lorenz-curve specification and can be used with grouped data or unit-record data.

Our model and estimation techniques were applied to two data sets that have been the subject of past analyses, one for Sweden, a country with relatively low inequality, and one for Brazil, a country with relatively high inequality. Results were obtained for 5 different Lorenz-curve specifications. Our findings suggest that point estimation of the Gini coefficient is generally insensitive to choice of distributional assumption, estimation technique and Lorenz-curve specification. There were two exceptions to this conclusion. One was for the function L_1 applied to the Brazilian data, using the Dirichlet distribution. In this case, the different estimates were attributable to a poor fit. The second exception was the estimate from L_4 with the Swedish data and the estimation technique of Sarabia et al. This discrepancy is likely to be a consequence of estimation instability associated with the overparameterized function L_4.

Although point estimation of the Gini coefficient was robust, assessment of the precision of estimation was not. It depended heavily on choice of functional form and the distributional assumption, and, to a lesser extent, on whether ML or

Bayesian inference was adopted. With respect to choice of functional form, we found that L_5 provided the best fit, L_4 tends to be an unnecessary overparameterisation, and L_1 can fit poorly. With respect to tools of analysis, we showed how Bayesian posterior pdfs can be an effective means for conveying knowledge about unknown parameters and inequality measures, and how they can be used to assess the validity of parametric restrictions on Lorenz functions.

Reference

Albert, J. H. and Chib, S. (1996), "Computation in Bayesian Econometrics: An Introduction to Markov Chain Monte Carlo", in R. C. Hill (ed.), *Advances in Econometrics Volumn 11A: Bayesian Computational Methods and Applications*, JAI Press, Greenwich.

Basmann, R.L., Hayes, K.J., Slottje, D.J. and Johnson J.D. (1990), "A General Functional Form for Approximating the Lorenz Curve", *Journal of Econometrics*, 43, 77-90.

Chotikapanich, D. (1993), "A Comparison of Alternative Functional Forms for the Lorenz Curve", *Economics Letters*, 41, 129-138.

Datt, G. (1998), "Computational Tools for Poverty Measurement and Analysis", *FCND Discussion Paper No. 50*, International Food Policy Research Institute, World Bank.

Gastwirth, J.L. (1972), "The Estimation of the Lorenz Curve and Gini Index", *Review of Economics and Statistics*, 54, 306-316.

Geweke, J. (1999), "Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication", *Econometric Reviews*, 18, 278-292.

Jain, S. (1975), *Size Distribution of Income*, World Bank, Washington.

Kakwani, N.C. (1980), "On a Class of Poverty Measures", *Econometrica*, 48, 437-446.

Kakwani, N.C. and N. Podder (1973), On Estimation of Lorenz Curves from Grouped Observations" *International Economic Review*, 14, 278-292.

Kakwani, N.C. and N. Podder (1976), "Efficiency Estimation of the Lorenz Curve and Associated Inequality Measures from Grouped Observations", *Econometrica*, 44, 137-148.

McDonald, J. B. (1984), "Some Generalized Functions for the Size Distribution of Income", *Econometrica*, 52, 647-663.

McDonald J.B. and Y.J. Xu (1995), "A Generalization of the Beta Distribution with Applications", *Journal of Econometrics*, 66, 133-152.

Ortega, P., Fernandez, M.A., Lodoux, M. and A. Garcia (1991), "A New Funational Form for Estimating Lorenz Curves", *Review of Income and Wealth*, 37, 447-452.

Poirier, R.A. (1995), *Intermediate Statistics and Econometrics: A Comparative Approach*, Cambridge: MIT Press.

Rasche, R.H.,Gaffney, J., Koo, A. and N. Obst (1980), "Functional Forms for Estimating the Lorenz Curve", *Econometrica*, 48, 1061-1062.

Sarabia, J-M, Castillo, E. and D.J. Slottje (1999), “An Ordered Family of Lorenz Curves”, *Journal of Econometrics*, 91, 43-60.

Shorrocks, A.F. (1983), “Ranking Income Distributions”, *Economica*, 50,3-17.

Theil, H. (1967), *Economics and Information Theory*, Amsterdam, North Holland.

Theil, H. (1975), *Theory and Measurement of Consumer Demand*, Amsterdam, North Holland.

Woodland, A.D. (1979), “Stochastic Specification and the Estimation of Share Equations”, *Journal of Econometrics*, 10, 361-383.

Appendix: Expressions for variances of the Gini coefficient.

For L_1: $$\mathrm{var}(\hat{G}) = \left(\frac{2(e^{\hat{k}}(e^2 - \hat{k}^2 - 2) + 1)}{(\hat{k}(e^{\hat{k}} - 1))^2} \right)^2 \mathrm{var}(\hat{k})$$

For L_2 : $$G = 1 - 2\int_0^1 \pi^\alpha [1 - (1-\pi)^\delta]\, d\pi$$

$$\mathrm{var}(\hat{G}) = \begin{bmatrix} \frac{\partial G}{\partial \alpha} & \frac{\partial G}{\partial \delta} \end{bmatrix} \begin{bmatrix} \mathrm{var}(\hat{\alpha}) & \mathrm{cov}(\hat{\alpha}, \hat{\delta}) \\ \mathrm{cov}(\hat{\alpha}, \hat{\delta}) & \mathrm{var}(\hat{\delta}) \end{bmatrix} \begin{bmatrix} \frac{\partial G}{\partial \alpha} \\ \frac{\partial G}{\partial \delta} \end{bmatrix}$$

where $$\frac{\partial G}{\partial \alpha} = -2\int_0^1 \pi^\alpha \log(\pi)[1 - (1-\pi)^\delta]\, d\pi$$

and $$\frac{\partial G}{\partial \delta} = 2\int_0^1 \pi^\alpha (1-\pi)^\delta \log(1-\pi)\, d\pi$$

For L_3 : $$G = 1 - 2\int_0^1 [1 - (1-\pi)^\delta]^\gamma\, d\pi$$

$$\mathrm{var}(\hat{G}) = \begin{bmatrix} \frac{\partial G}{\partial \delta} & \frac{\partial G}{\partial \gamma} \end{bmatrix} \begin{bmatrix} \mathrm{var}(\hat{\delta}) & \mathrm{cov}(\hat{\delta}, \hat{\gamma}) \\ \mathrm{cov}(\hat{\delta}, \hat{\gamma}) & \mathrm{var}(\hat{\gamma}) \end{bmatrix} \begin{bmatrix} \frac{\partial G}{\partial \delta} \\ \frac{\partial G}{\partial \gamma} \end{bmatrix}$$

where $$\frac{\partial G}{\partial \delta} = 2\int_0^1 \gamma [1 - (1-\pi)^\delta]^{\gamma-1} (1-\pi)^\delta \log(1-\pi)\, d\pi$$

and $$\frac{\partial G}{\partial \gamma} = -2\int_0^1 [1 - (1-\pi)^\delta]^\gamma \log[1 - (1-\pi)^\delta]\, d\pi$$

For L_4 : $$G = 1 - 2\int_0^1 \pi^{\alpha}[1-(1-\pi)^{\delta}]^{\gamma}\, d\pi$$

$$\operatorname{var}(\hat{G}) = \begin{bmatrix} \frac{\partial G}{\partial \alpha} & \frac{\partial G}{\partial \delta} & \frac{\partial G}{\partial \gamma} \end{bmatrix} \begin{bmatrix} \operatorname{var}(\hat{\alpha}) & \operatorname{cov}(\hat{\alpha},\hat{\delta}) & \operatorname{cov}(\hat{\alpha},\hat{\gamma}) \\ \operatorname{cov}(\hat{\alpha},\hat{\delta}) & \operatorname{var}(\hat{\delta}) & \operatorname{cov}(\hat{\delta},\hat{\gamma}) \\ \operatorname{cov}(\hat{\alpha},\hat{\gamma}) & \operatorname{cov}(\hat{\delta},\hat{\gamma}) & \operatorname{var}(\hat{\gamma}) \end{bmatrix} \begin{bmatrix} \frac{\partial G}{\partial \alpha} \\ \frac{\partial G}{\partial \delta} \\ \frac{\partial G}{\partial \gamma} \end{bmatrix}$$

where $$\frac{\partial G}{\partial \alpha} = -2\int_0^1 \pi^{\alpha} \log(\pi)[1-(1-\pi)^{\delta}]^{\gamma}\, d\pi$$

$$\frac{\partial G}{\partial \gamma} = -2\int_0^1 \pi^{\alpha} [1-(1-\pi)^{\delta}]^{\gamma} \log[1-(1-\pi)^{\delta}]\, d\pi$$

$$\frac{\partial G}{\partial \delta} = 2\int_0^1 \pi^{\alpha}\gamma[1-(1-\pi)^{\delta}]^{\gamma-1} (1-\pi)^{\delta} \log(1-\pi)]\, d\pi$$

For L_5 : $$G = 1 - 2\int_0^1 [\pi - a\pi^{d}(1-\pi)^{b}]\, d\pi$$

$$\operatorname{var}(\hat{G}) = \begin{bmatrix} \frac{\partial G}{\partial a} & \frac{\partial G}{\partial d} & \frac{\partial G}{\partial b} \end{bmatrix} \begin{bmatrix} \operatorname{var}(\hat{a}) & \operatorname{cov}(\hat{a},\hat{d}) & \operatorname{cov}(\hat{a},\hat{b}) \\ \operatorname{cov}(\hat{a},\hat{d}) & \operatorname{var}(\hat{d}) & \operatorname{cov}(\hat{d},\hat{b}) \\ \operatorname{cov}(\hat{a},\hat{b}) & \operatorname{cov}(\hat{d},\hat{b}) & \operatorname{var}(\hat{b}) \end{bmatrix} \begin{bmatrix} \frac{\partial G}{\partial a} \\ \frac{\partial G}{\partial d} \\ \frac{\partial G}{\partial b} \end{bmatrix}$$

where $$\frac{\partial G}{\partial a} = 2\int_0^1 \pi^{d}(1-\pi)^{b}\, d\pi$$

$$\frac{\partial G}{\partial d} = 2\int_0^1 a\pi^{d}(1-\pi)^{b} \log(\pi)\, d\pi$$

$$\frac{\partial G}{\partial b} = 2\int_0^1 a\pi^{d}(1-\pi)^{b} \log(1-\pi)\, d\pi$$

Table 1
Estimates for Lorenz Parameters and Gini Coefficients

Sweden

		α	δ	γ	Gini
L_2					
	NL	0.5954	0.6352		0.3880
	ML	0.6068	0.6412		0.3872
	Bayes	0.6073	0.6418		0.3870
	Sarabia	0.5960	0.6400		0.3850
L_3					
	NL		0.7269	1.5602	0.3871
	ML		0.7335	1.5767	0.3877
	Bayes		0.7337	1.5766	0.3875
	Sarabia		0.7300	1.5620	0.3860
L_4					
	NL	-0.7550	0.7931	2.2891	0.3864
	ML	0.0048	0.7330	1.5721	0.3876
	Bayes	0.2753	0.6970	1.3141	0.3872
	Sarabia	0.0769	0.6490	1.1740	0.3210

L_1		k			Gini
	NL	2.5029			0.3792
	ML	2.5313			0.3828
	Bayes	2.5256			0.3814

L_5		a	d	b	Gini
	NL	0.7664	0.9397	0.5929	0.3876
	ML	0.7492	0.9199	0.5862	0.3870
	Bayes	0.7490	0.9201	0.5865	0.3866

Table 2
Estimates for Lorenz Parameters and Gini Coefficients

Brazil

		α	δ	γ	Gini
L_2					
	NL	0.5727	0.2876		0.6361
	ML	0.5270	0.2857		0.6326
	Bayes	0.5284	0.2861		0.6324
	Sarabia	0.4900	0.2780		0.6350
L_3					
	NL		0.3782	1.4357	0.6328
	ML		0.3721	1.4160	0.6325
	Bayes		0.3721	1.4153	0.6322
	Sarabia		0.3640	1.3960	0.6340
L_4					
	NL	0.2169	0.3467	1.2674	0.6339
	ML	0.0262	0.3683	1.3950	0.6325
	Bayes	0.1850	0.3446	1.2717	0.6327
	Sarabia	0.0770	0.6170	1.1740	0.6440
L_1		k			Gini
	NL	5.3685			0.6368
	ML	3.8438			0.5234
	Bayes	3.7277			0.5063
L_5		a	d	b	Gini
	NL	0.9151	1.0001	0.2698	0.6349
	ML	0.9131	0.9990	0.2685	0.6349
	Bayes	0.9102	0.9970	0.2671	0.6348

Table 3

Standard Errors (Deviations) for Lorenz Parameters and Gini Coefficients

Sweden

		α	δ	γ	Gini
L_2					
	NL	0.0100	0.0037		0.0010
	ML	0.0206	0.0085		0.0041
	Bayes	0.0279	0.0112		0.0054
	Sarabia	0.0018	0.0303		
L_3					
	NL		0.0028	0.0066	0.0007
	ML		0.0072	0.0176	0.0038
	Bayes		0.0107	0.0251	0.0050
	Sarabia		0.0263	0.0022	
L_4					
	NL	0.4822	0.0322	0.4696	0.0000
	ML	0.6612	0.0756	0.6369	0.0036
	Bayes	0.1700	0.0267	0.1601	0.0053
	Sarabia	0.0003	0.0977	0.0002	

L_1		*k*	Gini
	NL	0.0621	0.0219
	ML	0.1831	0.0228
	Bayes	0.2284	0.0286

L_5		*a*	*d*	*b*	Gini
	NL	0.0101	0.0096	0.0075	0.0009
	ML	0.0143	0.0093	0.0109	0.0031
	Bayes	0.0216	0.0137	0.0164	0.0046

Table 4
Standard Errors (Deviations) for Lorenz Parameters and Gini Coefficients

Brazil

		α	δ	γ	Gini
L_2					
	NL	0.0163	0.0019		0.0011
	ML	0.0383	0.0053		0.0052
	Bayes	0.0515	0.0072		0.0072
	Sarabia	0.0038	0.0662		
L_3					
	NL		0.0033	0.0107	0.0009
	ML		0.0068	0.0225	0.0040
	Bayes		0.0093	0.0304	0.0050
	Sarabia		0.0713	0.0004	
L_4					
	NL	0.1322	0.0203	0.1015	0.0019
	ML	0.2148	0.0318	0.1734	0.0039
	Bayes	0.1307	0.0221	0.1041	0.0054
	Sarabia	0.0001	0.1041	0.0091	

L_1		*k*	Gini
	NL	0.4865	0.1192
	ML	0.8237	0.0747
	Bayes	0.8702	0.0883

L_5		*a*	*d*	*b*	Gini
	NL	0.0025	0.0023	0.0014	0.0003
	ML	0.0038	0.0024	0.0021	0.0013
	Bayes	0.0044	0.0023	0.0027	0.0018

Table 5
Information Inaccuracy Measure

	Sweden		Brazil	
	ML	Bayes	ML	Bayes
L_1	0.00888	0.00888	0.10851	0.11382
L_2	0.00029	0.00029	0.00056	0.00056
L_3	0.00025	0.00025	0.00031	0.00031
L_4	0.00025	0.00026	0.00031	0.00033
L_5	0.00017	0.00017	0.00003	0.00003

Table 6
The Likelihood Ratio Test

	Sweden	Brazil	Critical Value
L_4 VS L_2	1.351	5.333	3.841
L_4 VS L_3	0.000	0.015	3.841
L_5 VS L_2	36.907	31.355	5.991

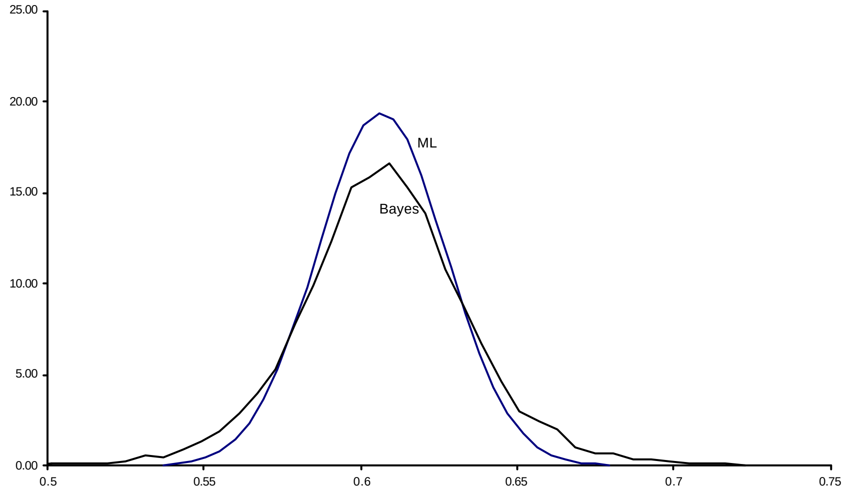

Figure 1: Pdfs for α for L_2 and Sweden

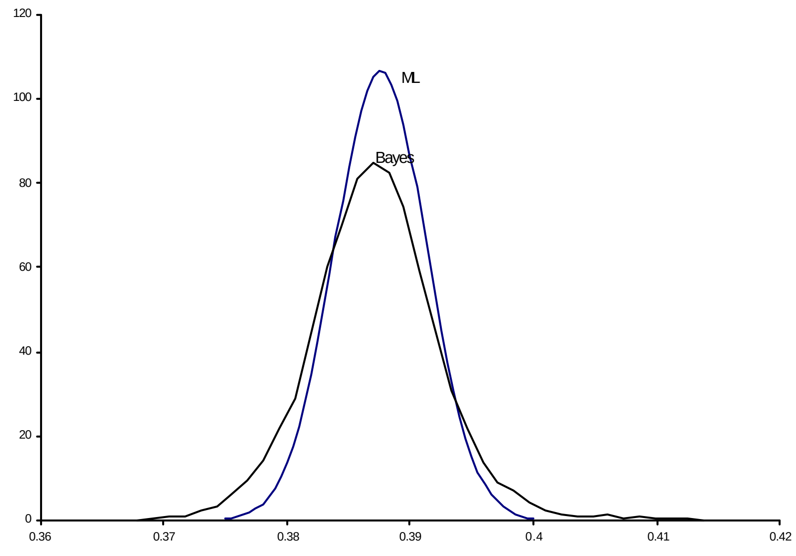

Figure 2: Pdfs for Gini coefficient for L_4 and Sweden.

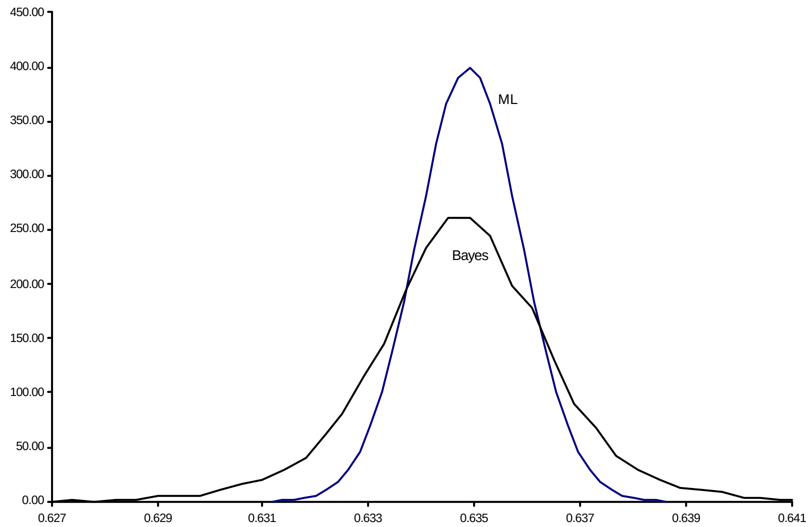

Figure 3: Pdfs for Gini coefficient for L_5 and Brazil

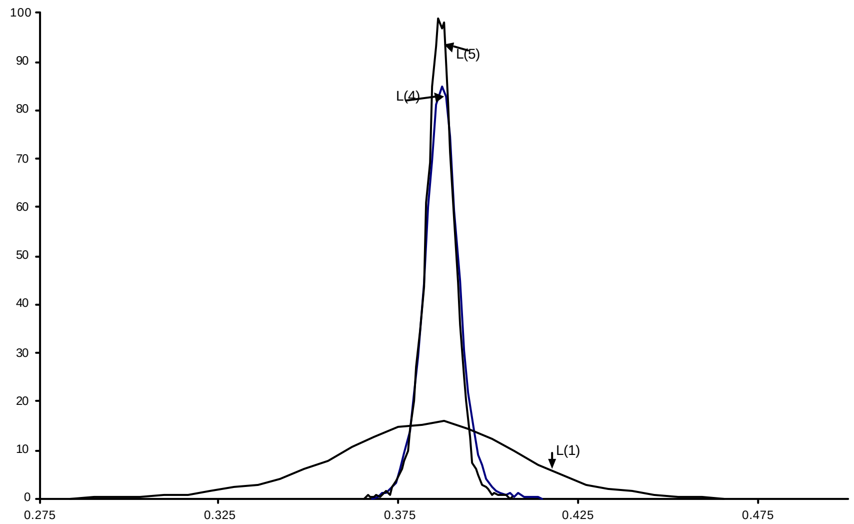

Figure 4: Posterior pdfs for the Gini coefficient for Sweden.

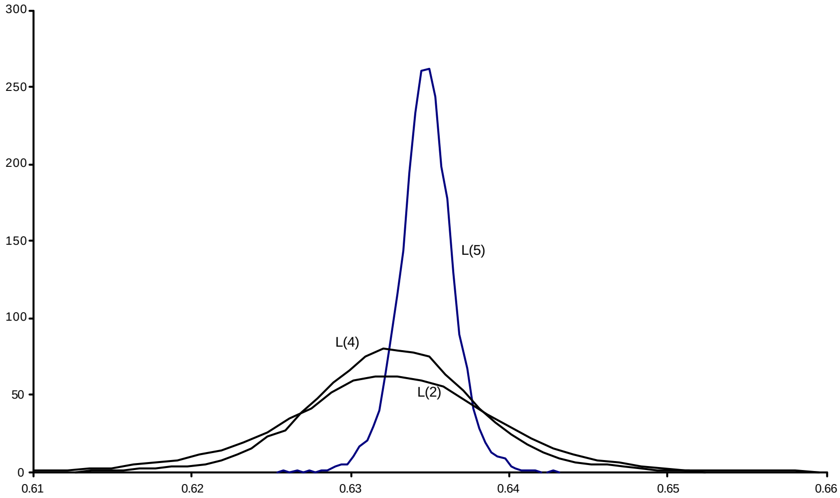

Figure 5: Posterior pdfs for the Gini coefficient for Brazil

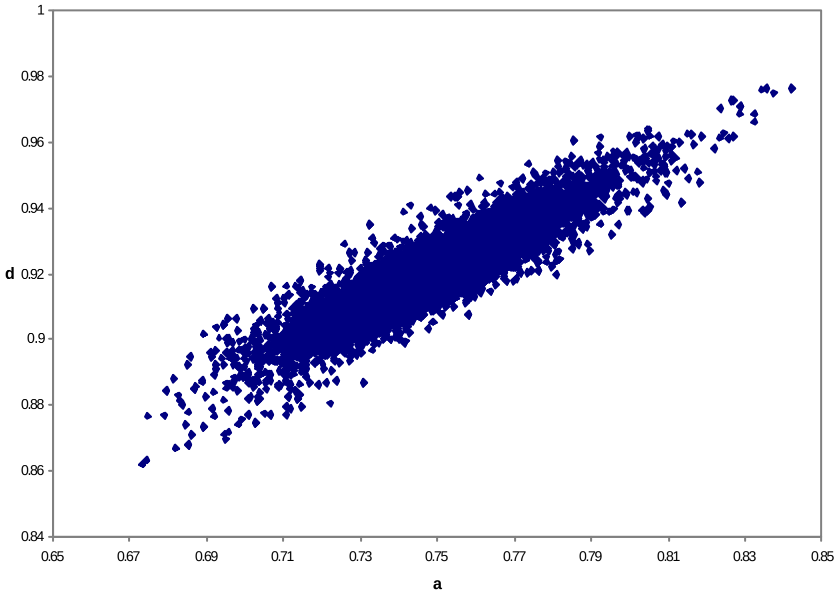

Figure 6: Joint scatter plot (a, d) for L_5, Sweden

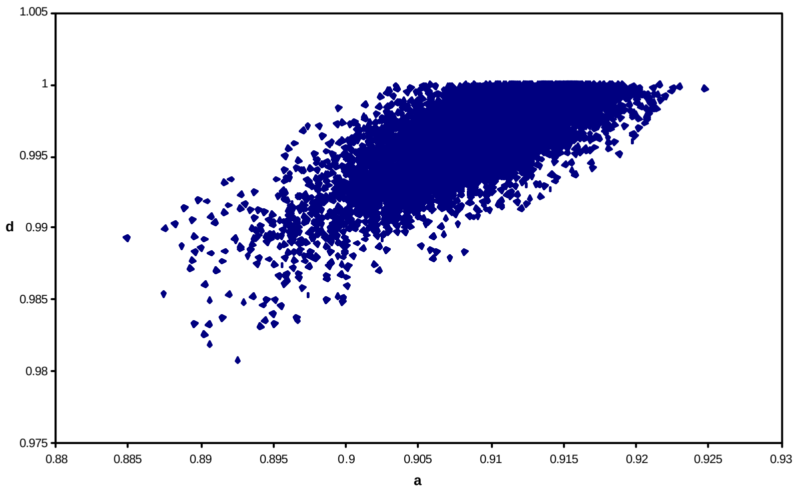

Figure 7: Joint scatter plot (a, d) for L_5, Brazil

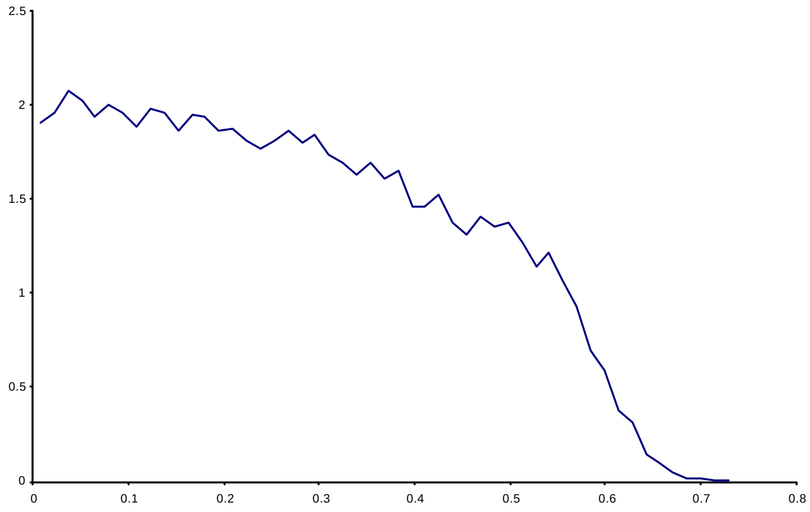

Figure8: Posterior pdf for α for L_4, Sweden

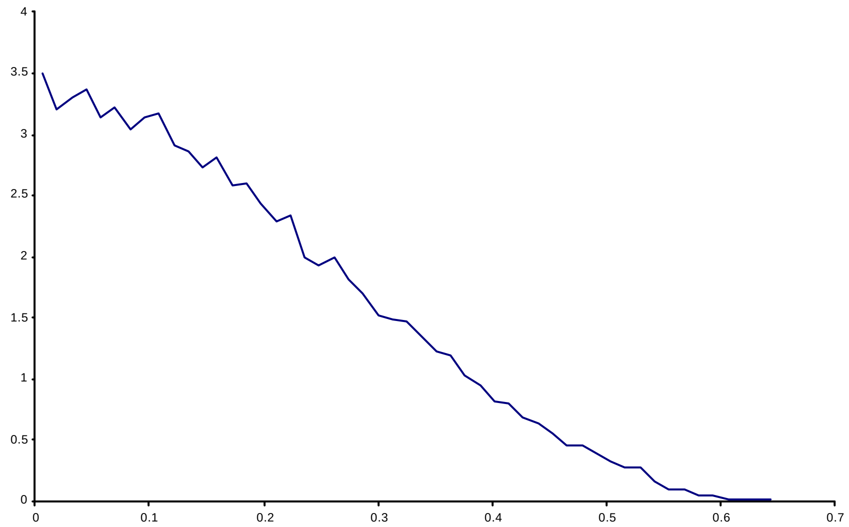

Figure 9: Posterior pdf for α for L_4, Brazil

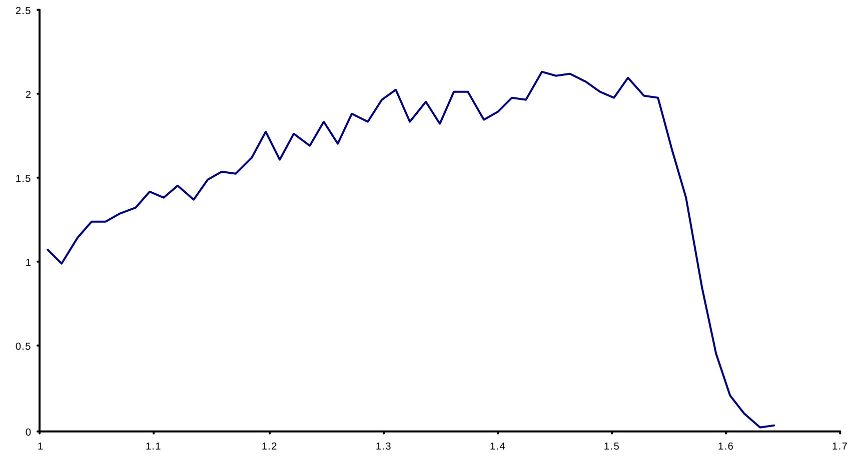

Figure 10 : Posterior pdf for γ for L_4, Sweden

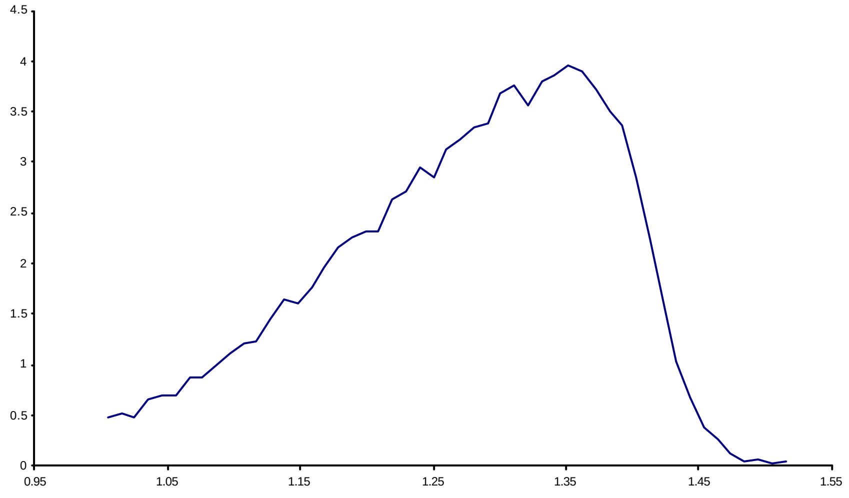

Figure11: Posterior pdf for γ for L_4, Brazil

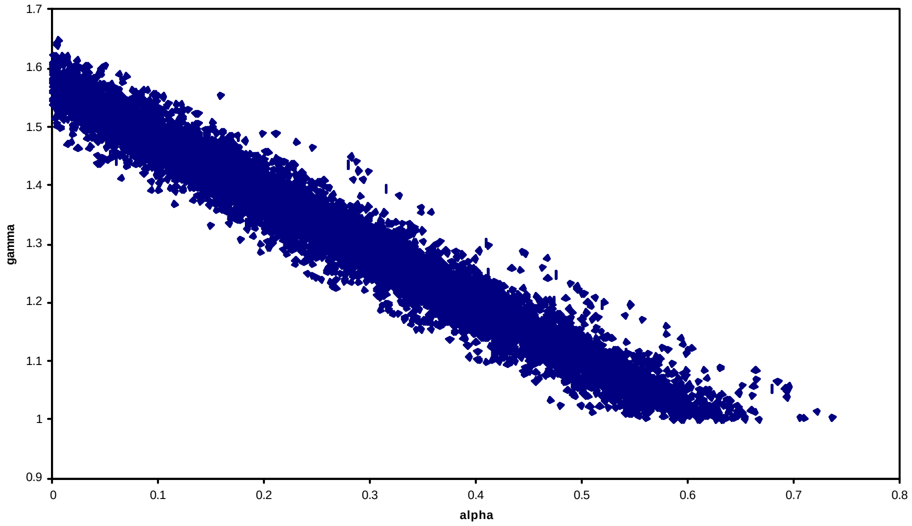

Figure 12: Joint scatter plot (α, γ) for L_4, Sweden

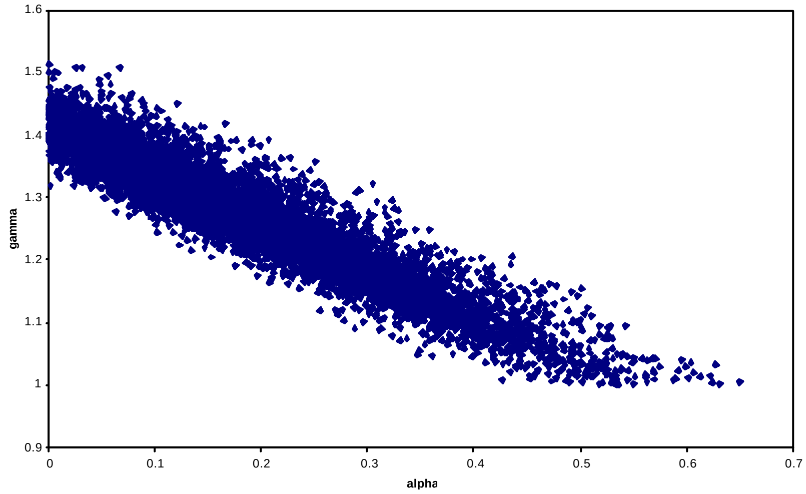

Figure 13: Joint scatter plot (α, γ) for L_4, Brazil