Comparing Cross-Country Estimates of Lorenz Curves Using a Dirichlet Distribution Across Estimators and Datasets

Andrew C. Chang* Phillip Li† Shawn M. Martin‡

May 1, 2017

Abstract

Chotikapanich and Griffiths (2002) introduced the Dirichlet distribution to the estimation of Lorenz curves. This distribution naturally accommodates the proportional nature of income share data and the dependence structure between the shares. Chotikapanich and Griffiths (2002) fit a family of five Lorenz curves to one year of Swedish and Brazilian income share data using unconstrained maximum likelihood and unconstrained non-linear least squares. We attempt to replicate the authors' results and extend their analyses using both constrained estimation techniques and five additional years of data. We successfully replicate a majority of the authors' results and find that some of their main qualitative conclusions also hold using our constrained estimators and additional data.

JEL Codes: C24; C51; C87; D31

Keywords: Dirichlet; Gini Coefficient; Income Distribution; Lorenz Curve; Replication

*Chang: Board of Governors of the Federal Reserve System. 20th St. NW and Constitution Ave., Washington DC 20551 USA. +1 (657) 464-3286. a.christopher.chang@gmail.com. https://sites.google.com/site/andrewchristopherchang/.

†Li: Office of Financial Research. phillip.li@ofr.treasury.gov.

‡Martin: University of Michigan, Ann Arbor. shawnmm@umich.edu.

§The views and opinions expressed here are those of the authors and are not necessarily those of the Board of Governors of the Federal Reserve System, Department of the Treasury, or the Office of Financial Research. This paper was started before Phillip Li joined the Office of Financial Research. We thank conference participants at the 2015 International Atlantic Economic Conference, especially Tiago Pires and Keshab Bhattarai, for helpful comments. We are responsible for any errors.

Introduction

The Lorenz curve is a commonly used tool to illustrate income distributions and income inequality. It is constructed by relating ordered cumulative proportions of income to ordered cumulative population shares. The curve is then used to estimate income inequality measures, such as the Gini coefficient or Atkinson's inequality measure.

Unfortunately, estimates of inequality from Lorenz curves can depend crucially on distributional assumptions, functional form assumptions, and estimation methodologies (Cheong, 2002; Chotikapanich and Griffiths, 2002, 2005; and Abdalla and Hassan, 2004). Therefore, the literature proposes different functional forms and re-parameterizations for both the Lorenz curve and income distributions.[1] Estimation is commonly based on least squares techniques, with more recent studies using Bayesian and maximum likelihood estimation.[2]

We have three main objectives in this paper. We first attempt a narrow replication of Chotikapanich and Griffiths (2002), hereafter CG, who propose using a Dirichlet distribution to model the proportional nature and dependence structure of cumulative income share data. CG estimate five Lorenz curves using both maximum likelihood (ML) and non-linear least squares (NL) on one year of Brazilian and Swedish data, obtaining implied Gini coefficients. CG have three main findings: (1) the point estimates of the parameters and of the Gini coefficients are generally insensitive to the choice of Lorenz curve specification and estimator, (2) the standard errors are sensitive to the specification and estimator, and (3) ML performs better than NL under the Dirichlet distributional assumption.

We replicate a majority of CG's three main findings. For less parameterized Lorenz curves, our point estimates and standard errors match CG. We experience considerable instability in estimating the more parameterized Lorenz curves, consistent with CG. Our successful narrow replication contributes to the current push for replication and robustness in economics research (Chang and Li, 2015; Welch, 2015; Zimmermann, 2015).

Our second objective is to extend CG in a scientific replication by using constrained esti-

[1]For example, Kakwani (1980), Rasche et al. (1980), Ortega et al. (1991), Chotikapanich (1993), Sarabia et al. (1999, 2001, 2005), Rohde (2009), Helene (2010), and Wang and Smyth (2015).

[2]See Chotikapanich and Griffiths (2002, 2008), Hasegawa and Kozumi (2003).

mators. We apply constrained maximum likelihood (CML) and constrained non-linear least squares (CNL) to the same functional forms and data as CG. We use constrained estimators because the parameters from the Lorenz curve specifications in CG should be constrained to ensure that the curves are invariant to increasing convex exponential and power transformations (Sarabia et al. (1999)). Although these restrictions are mentioned in CG, some of CG's estimates violate the constraints. We find that some parameter estimates differ between constrained and unconstrained estimators, but the implied Gini coefficients are similar between constrained and unconstrained estimators.

Our third objective is to fit the various Lorenz curve specifications with both constrained and unconstrained estimators on five additional years of Swedish and Brazilian income distribution data from the World Bank: data not used by CG. In this scientific replication, we find that a few of the main conclusions from CG also hold using the constrained estimators and these additional data. Similar to Abdalla and Hassan (2004), who apply the methodologies from CG to data from the Abu Dhabi Emirate and their own Lorenz curve form, we find that Gini coefficient point estimates are robust to different functional forms and estimation methods when applied to additional data.

Narrow Replication

The data are the cumulative proportions of income ($\eta_1, \eta_2, ..., \eta_M$ with $\eta_M = 1$) and corresponding cumulative population shares ($\pi_1, \pi_2, ..., \pi_M$ with $\pi_M = 1$). Let $q_i = \eta_i - \eta_{i-1}$ be the income shares. CG assume that $(q_1, \ldots, q_M)$ has a Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_M)$, where $\alpha_i = \lambda[L(\pi_i; \beta) - L(\pi_{i-1}; \beta)]$. $L(\cdot)$ is the Lorenz curve specification with an associated vector of unknown parameters β, and $\lambda > 0$ is an unknown scalar parameter from the Dirichlet distribution.

CG apply five Lorenz curve specifications to one year of Brazilian and Swedish data:

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

$$L_2(\pi_i; \alpha, \delta) = \pi^\alpha[1 - (1 - \pi)^\delta], \qquad \alpha \geq 0, 0 < \delta \leq 1 \qquad (2)$$

$$L_3(\pi_i;\delta,\gamma) = [1-(1-\pi)^{\delta}]^{\gamma}, \qquad \gamma \geq 1, 0 < \delta \leq 1 \tag{3}$$

$$L_4(\pi_i;\alpha,\delta,\gamma) = \pi^{\alpha}[1-(1-\pi)^{\delta}]^{\gamma}, \qquad \alpha \geq 0, \gamma \geq 1, 0 < \delta \leq 1 \tag{4}$$

$$L_5(\pi_i;a,d,b) = \pi - a\pi^{d}(1-\pi)^{b}. \qquad a > 0, 0 < d \leq 1, 0 < b \leq 1 \tag{5}$$

Each specification is then estimated with ML based on the Dirichlet distributional assumption or with NL without the distributional assumption.[3] We use the Matlab function *fminunc* to maximize the log likelihood functions. The standard errors are from the negative inverse of the numeric Hessian matrix evaluated at the maximum. We use the Matlab function *lsqcurvefit* and the Stata command *nl* for the NL optimizations. For NL, CG suggest using Newey and West (1987) standard errors.[4]

Table 1 shows our narrow replication results. For Lorenz curves L_1 to L_3 and for both countries, our ML point estimates and standard errors more or less match those from CG. Our ML estimation for L_4 is unstable, with more stable estimation using Brazilian data than Swedish data, consistent with CG. However, the Swedish ML point estimates for α fluctuate around values that are often greater than CG's estimates. When we perform ML with random starting values on Swedish data, the point estimates are similar to CG's but the standard errors are unstable.[5] This instability may indicate that the area around the maximum is flat, yielding point estimates and variances that are not unique (Gill and King, 2003). In addition, the numeric variance-covariance matrix evaluated at the converged values is not positive definite for over 50% of the random starting values. As a result, we do not report ML standard errors for L_4 with Swedish data. For L_4 with Brazilian data, our point estimates and standard errors more or less match those from CG.

We are unable to replicate CG's ML results for L_5 for both countries, despite attempting estimation using a grid of starting values. As noted in Ortega et al. (1991) and Sarabia et al.

[3]We conduct the replications without assistance from the authors and without their code, using data from the original source (Jain, 1975). We use Matlab R2013a and Stata 13MP on the Windows 7 Enterprise (64-bit) and OS X Version 10.9.5 operating systems respectively.

[4]We implement *nl* in Stata with different lag values for the Newey-West standard errors and find that a lag of 2 matches the standard errors reported by CG. These are the standard errors we report. We use the Stata option *vce(hac nwest 2)* in the *nl* command.

[5]We use 2000 sets of random starting values from a standard normal distribution.

(1999), L_5 can result in a negative income share η_i for a population share π_i, leading to the difference $L_5(\pi_i;\beta) - L_5(\pi_{i-1};\beta)$ being negative and the term $\log\Gamma(\lambda[L_5(\pi_i;\beta) - L_5(\pi_{i-1};\beta)])$ from the log likelihood function being computationally infeasible.

We initialize NL for each Lorenz curve over a grid of starting values that spans the parameters' support. We find that all Lorenz curve specifications except L_1 display some instability.[6] Instability is most frequent for L_4 and L_5. However, the parameter estimates that minimize the NL objective function and the corresponding standard errors are equivalent to CG.

For both ML and NL we also attempt to replicate the Gini coefficient $G = 1-2\int_0^1 L_j(\pi;\beta)\, d\pi$. Following CG, we obtain point estimates of G by replacing β with the ML or NL $\hat{\beta}$s for each Lorenz curve specification. With the exception of L_1, we successfully replicate the Gini point estimates and standard errors for all estimators and Lorenz curve specifications.[7,8]

Similar to CG, we find that the Gini point estimates are insensitive to the choice of Lorenz curve specification and estimator, although L_1 fitted with Brazilian data is an exception. Given our inability to estimate L_5 using ML and the non-positive definite numeric Hessian for L_4 using Swedish data, we do not report an ML Gini coefficient for L_5 or standard errors of the Gini coefficient for both L_4 and L_5.

We also successfully replicate the information inaccuracy measures suggested by Theil (1967)[9] and the likelihood ratio test (LRT) results except for L_5 vs. L_2 (with α=1) for Brazil. We obtained 51.355 as the test statistic compared to 31.355 from CG. Both likelihood ratio statistics, however, lead to the same conclusion that the functional form L_2, with $\alpha = 1$, is

[6] A majority of the parameter estimates are similar. However, some initial values lead to NL point estimates with larger residual sum of squares and, in some cases, infinite Gini coefficients.

[7] Our initial attempt to replicate CG's ML standard errors for L_1's Gini coefficients led us to analytically verify CG's formula for the variance of the Gini coefficient, $var(\hat{G})$. We find a typo in CG's L_1 formula for $var(\hat{G})$ but are able to replicate the ML standard errors for the Gini coefficient with our corrected formula. CG report $\text{var}(\hat{G}) = \left[\frac{2(e^{\hat{k}}(e^2-\hat{k}^2-2)+1)}{(\hat{k}(e^{\hat{k}}-1))^2}\right]^2 \text{var}(\hat{k})$ but we analytically find $\text{var}(\hat{G}) = \left[\frac{2(e^{\hat{k}}(e^{\hat{k}}-\hat{k}^2-2)+1)}{(\hat{k}(e^{\hat{k}}-1))^2}\right]^2 \text{var}(\hat{k})$.

[8] We discover a minor computational issue in the calculation of the NL standard errors for L_1 by CG. We find that the CG standard errors for the L_1 NL Gini coefficient are calculated as $\text{var}(\hat{G}) = \frac{\partial G}{\partial \beta}\text{var}(\hat{k})$ when the correct formula is $\text{var}(\hat{G}) = \frac{\partial G}{\partial \beta'}\text{var}(\hat{k})\frac{\partial G}{\partial \beta}$. We verify this using CG's reported Brazilian values for $\text{SE}(\hat{G})$ and $\text{SE}(\hat{k})$, .1647 and .6726, in the formula of $\text{var}(\hat{G})$ corrected for the typo detailed in footnote 7: $.1647^2 = \frac{\partial G}{\partial \beta} \times .6726^2 \times \frac{\partial G}{\partial \beta}$, which implies $[\frac{\partial G}{\partial \beta}]^2 = .0600$ and $\frac{\partial G}{\partial \beta} = .2449$, however $\frac{\partial G}{\partial \beta}$ evaluated at $\hat{k} = .0600$. Therefore the variance of $\hat{G}$ should be $\text{var}(\hat{G}) = .0600 \times .6742^2 \times .0600 = .0016$ and $\text{SE}(\hat{G}) = .0403$. A similar computational error occurs for Swedish data.

[9] The measure, $I = \sum_{i=1}^{M} q_i log(\frac{q_i}{\hat{q}_i})$, compares actual income shares, q_i, to predicted income shares, $\hat{q}_i$. Smaller values of I indicate a better fit.

rejected relative to L_5. The L_5 LRT and information inaccuracy measure for L_5 are calculated using CG's reported point estimates.

Scientific Replication: Constrained Optimization

Table 1 shows our results.[10] Point estimates for L_1 to L_3 are identical for constrained and unconstrained estimation. For L_4 with Swedish data, the CNL estimates deviate the most from the NL estimates. For example, the CNL estimate for α is close to 0 while the NL estimate is -0.7549. For L_4 with Brazilian data, the CNL estimates are close to the NL estimates. In terms of CML results for L_4, the constrained estimates and standard errors match the unconstrained quantities. Though we were unable to replicate unconstrained ML point estimates for L_5, with parameter constraints imposed the CML point estimates are close to the unconstrained estimates from CG; we were unable to generate standard error estimates as the numeric hessians were quite unstable across different sets of starting values. The CNL point estimates for L_5 are either identical to or very close to the NL quantities.

Overall, we find that the CML and CNL estimates of the model parameters can differ from their ML and NL counterparts, but the implied Gini coefficients are similar across estimators.

Scientific Replication: Extension to World Bank Data

We further extend CG using data from the World Bank Poverty and Equity Database (World Bank, 2015a,b). We construct a dataset of seven quantiles of cumulative income shares for Brazil in 1987, 1992, 1995, 2001 and 2005 and for the equivalent years for Sweden, with 2001 replaced by 2000.

Our results are in Tables A1 and A2 in the online appendix. We find that unconstrained and constrained estimation applied to World Bank data yield qualitative conclusions similar to those reported by CG, who use data from Jain (1975). With the exception of L_4, the point

[10] Table 1 uses the Matlab functions *fmincon* and *lsqcurvefit*. In unreported results we also attempt to use the Matlab function *patternsearch* to apply ML and CML. *Patternsearch* yields parameter estimates that are either identical to *fminunc* and *fmincon* or imply a smaller log-likelihood; it also tends to be less stable than *fminunc* and *fmincon*.

estimates of the parameters for all Lorenz curve specifications are similar across estimation techniques, but there are differences in the standard errors. ML and NL point estimates for L_4 differ for all years of Brazilian and Swedish World Bank data. Similar to our narrow replication, we experience the same computational instability with unconstrained ML for L_4 and computational infeasibility for L_5 with World Bank data.

We also find that, for a given year, Gini coefficients are similar across Lorenz curve specifications and estimators (with the exception of L_1 with Brazilian data), even though some unconstrained parameter estimates violate the restricted ranges. For Brazil, ML estimation of L_1 results in Gini coefficients that are lower than other functional forms, and NL estimation results in higher Gini coefficients. In addition, our point estimates of the Gini coefficients are similar to those officially reported by the World Bank (see Table A3 in the online appendix).

Table A4 in the online appendix compares the fit using the Theil (1967) information inaccuracy measure. Our results are largely consistent with CG.

Conclusion and Recommendation

Our narrow replication of CG verifies a majority of their results. Our scientific replication extends the analysis from CG to constrained estimators and additional data, where we find that some of their qualitative results still carry through. However, in both our narrow and scientific replications we find instabilities in the estimation across different datasets, estimators, and optimization algorithms (see footnote 10). These instabilities are possibly due to the Dirichlet and Lorenz curve functional form specifications.

Although we have explored different functional forms and estimators for modeling Lorenz curves, it is difficult for us to make a sweeping recommendation as to which estimator and functional form that researchers should use. However, assuming you only care about the Gini coefficient, and not the fit of actual income shares, then we feel the parsimonious L_1 is the best option. L_1's implied Gini coefficient is relatively, though not completely, invariant to estimator choice and also is stable across initialized starting values.

References

Abdalla, I. M., Hassan, M. Y., 2004. Maximum likelihood estimation of lorenz curves using alternative parametric model. Metodoloski Zvezki 1 (1), 109–118.

Chang, A. C., Li, P., 2015. Is economics research replicable? sixty published papers from thirteen journals say "usually not". Finance and Economics Discussion Series 2015-083. Washington: Board of Governors of the Federal Reserve System.

Cheong, K. S., 2002. An empirical comparison of alternative functional forms for the lorenz curve. Applied Economics Letters 9 (3), 171–176.

Chotikapanich, D., 1993. A comparison of alternative functional forms for the lorenz curve. Economics Letters 41 (2), 129–138.

Chotikapanich, D., Griffiths, W. E., 2002. Estimating lorenz curves using a dirichlet distribution. Journal of Business & Economic Statistics 20 (2), 290–295.

Chotikapanich, D., Griffiths, W. E., 2005. Averaging lorenz curves. The Journal of Economic Inequality 3 (1), 1–19.

Chotikapanich, D., Griffiths, W. E., 2008. Estimating Income Distributions Using a Mixture of Gamma Densities. Springer.

Gill, J., King, G., 2003. Numerical issues involved in inverting hessian matrices. Numerical Issues in Statistical Computing for the Social Scientist.

Hasegawa, H., Kozumi, H., 2003. Estimation of lorenz curves: a bayesian nonparametric approach. Journal of Econometrics 115 (2), 277–291.

Helene, O., 2010. Fitting lorenz curves. Economics Letters 108 (2), 153–155.

Jain, S., 1975. Size distribution of income: A compilation of data.

Kakwani, N., 1980. On a class of poverty measures. Econometrica 48 (2), 437–446.

Newey, W. K., West, K. D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Ortega, P., Martin, G., Fernandez, A., Ladoux, M., Garcia, A., 1991. A new functional form for estimating lorenz curves. Review of Income and Wealth 37 (4), 447–452.

Rasche, R., et al., 1980. Functional forms for estimating the lorenz curve: Comment. Econometrica 48 (4), 1061–62.

Rohde, N., 2009. An alternative functional form for estimating the lorenz curve. Economics Letters 105 (1), 61–63.

Sarabia, J. M., Castillo, E., Pascual, M., Sarabia, M., 2005. Mixture lorenz curves. Economics Letters 89 (1), 89–94.

Sarabia, J.-M., Castillo, E., Slottje, D. J., 1999. An ordered family of lorenz curves. Journal of Econometrics 91 (1), 43–60.

Sarabia, J.-M., Castillo, E., Slottje, D. J., 2001. An exponential family of lorenz curves. Southern Economic Journal 67 (3), 748–756.

Theil, H., 1967. Economics and Information Theory. North-Holland Amsterdam.

Wang, Z., Smyth, R., 2015. A hybrid method for creating lorenz curves. Economics Letters 133, 59–63.

Welch, I., 2015. Plausibility. Available at SSRN 2570577.

World Bank, 2015a. Povcalnet: an online analysis tool for global poverty monitoring. http:\iresearch.worldbank.org/PovcalNet, accessed 2015-08-05.

World Bank, 2015b. Poverty & equity data. http://data.worldbank.org/data-catalog/poverty-and-equity-database, accessed 2015-06-04.

Zimmermann, C., 2015. On the need for a replication journal. FRB St Louis Paper No. FEDLWP2015-016.

Table 1: Sweden and Brazil estimates using data from Jain (1975)[a]

		Sweden. Chotikapanich & Griffiths				Sweden. Our Unconstrained & Constrained Results			
L_2	NL	α: .60 (.01)	δ: .64 (.01)		G: .388 (.001)	α: .60 (.01)	δ: .64 (.01)		G: .388 (.001)
	ML	α: .61 (.02)	δ: .64 (.01)		G: .387 (.004)	α: .61 (.02)	δ: .64 (.01)		G: .387 (.004)
L_3	NL		δ: .73 (.00)	γ: 1.56 (.01)	G: .387 (.001)		δ: .73 (.01)	γ: 1.56 (.01)	G: .387 (.001)
	ML		δ: .73 (.01)	γ: 1.58 (.02)	G: .388 (.004)		δ: .73 (.01)	γ: 1.58 (.02)	G: .388 (.004)
L_4	CNL	n/a	n/a	n/a	n/a	α: .00	δ: .73	γ: 1.56	G: .387
	NL	α: -.76 (.56)	δ: .79 (.04)	γ: 2.29 (.55)	G: .386 (.001)	α: -.75 (.56)	δ: .79 (.04)	γ: 2.29 (.55)	G: .386 (.001)
	CML	n/a	n/a	n/a	n/a	α: .00	δ: .73	γ: 1.57	G: .388
	ML	α: .00 (.66)	δ: .73 (.08)	γ: 1.57 (.64)	G: .388 (.004)	α: .00 (-)	δ: .73 (-)	γ: 1.57 (-)	G: .388 (-)
L_1	NL	k: 2.50 (.08)			G: .379 (.029)	k: 2.50 (.08)			G: .379 (.010)
	ML	k: 2.53 (.18)			G: .383 (.023)	k: 2.53 (.18)			G: .383 (.023)
L_5	NL	a: .77 (.01)	d: .94 (.01)	b: .59 (.01)	G: .388 (.001)	a: .77 (.01)	d: .94 (.01)	b: .59 (.01)	G: .388 (.001)
	CML	n/a	n/a	n/a	n/a	a: .75 (-)	d: .92 (-)	b: .59 (-)	G: .387 (-)
	ML	a: .75 (.01)	d: .92 (.01)	b: .59 (.01)	G: .387 (.003)	a: - (-)	d: - (-)	b: - (-)	G: - (-)
		Brazil. Chotikapanich & Griffiths				Brazil. Our Unconstrained & Constrained Results			
L_2	NL	α: .57 (.02)	δ: .29 (.01)		G: .636 (.001)	α: .57 (.02)	δ: .29 (.01)		G: .636 (.001)
	ML	α: .53 (.04)	δ: .29 (.01)		G: .633 (.005)	α: .53 (.04)	δ: .29 (.01)		G: .633 (.005)
L_3	NL		δ: .38 (.01)	γ: 1.44 (.01)	G: .633 (.001)		δ: .38 (.01)	γ: 1.44 (.01)	G: .633 (.001)
	ML		δ: .37 (.01)	γ: 1.42 (.02)	G: .633 (.004)		δ: .37 (.01)	γ: 1.42 (.02)	G: .633 (.004)
L_4	CNL	n/a	n/a	n/a	n/a	α: .22	δ: .35	γ: 1.27	G: .634
	NL	α: .22 (.20)	δ: .35 (.03)	γ: 1.27 (.15)	G: .634 (.001)	α: .22 (.20)	δ: .35 (.03)	γ: 1.27 (.15)	G: .634 (.001)
	ML	α: .03 (.21)	δ: .37 (.03)	γ: 1.40 (.17)	G: .633 (.004)	α: .03 (.22)	δ: .37 (.03)	γ: 1.39 (.18)	G: .633 (.004)
L_1	NL	k: 5.37 (.67)			G: .637 (.165)	k: 5.37 (.67)			G: .637 (.040)
	ML	k: 3.84 (.82)			G: .523 (.075)	k: 3.84 (.82)			G: .523 (.075)
L_5	CNL	n/a	n/a	n/a	n/a	a: .91	d: 1.00	b: .27	G: .635
	NL	a: .92 (.01)	d: 1.00 (.01)	b: .27 (.01)	G: .635 (.001)	a: .92 (.01)	d: 1.00 (.01)	b: .27 (.01)	G: .635 (.001)
	CML	n/a	n/a	n/a	n/a	a: .91 (-)	d: 1.00 (-)	b: .27 (-)	G: .635 (-)
	ML	a: .91 (.01)	d: 1.00 (.01)	b: .27 (.01)	G: .635 (.001)	a: - (-)	d: - (-)	b: - (-)	G: - (-)

[a] 'ML': maximum likelihood. 'NL': non-linear least squares. 'CML': constrained maximum likelihood. 'CNL': constrained non-linear least squares. 'G': Gini coefficient. We report constrained estimates only when they differ from the unconstrained estimates. '–' represents estimates we are unable to obtain. 'n/a' represents estimates that CG did not originally obtain.