응용통계연구 제 12권 1호, 1999년, pp. 153-163

변환된 LORENZ CURVE를 이용한 분포 연구

조영석 [1] 이제영[2] 강석복[3]

요 약

경제학분야에서 소득분배의 불균형 정도에 대한 척도로 널리 이용되는 Lorenz curve를 소개하고, 변환에 의한 Lorenz curve기법을 도입하여 통계학의 주요 주제인 모집단의 분포에 대한 추정에 적용할 수 있는지 조사한다. 여러 특정분포에 대한 변환된 Lorenz curve의 특징을 제시하고 예제로 Hodgkin's disease 데이터를 이용하여 Q-Q 플롯과 새로 제시한 변환된 Lorenz curve를 비교한다.

1. 서론

모집단의 분포에 관한 추정은 통계학에 있어서 매우 중요하게 다루는 주제 가운데 하나이다. 흔히, 통계적 분석에 앞서 실시하는 정규성검정도 모집단의 분포를 정규분포로 추정할 수 있는가 하는 문제이다. 특히 분포의 형태에 관한 추정은 히스토그램이나 Q-Q 플롯과 같은 그래프를 이용하여 접근하기도 하는데, 이들 연구는 컴퓨터의 발달로 최근 들어 더욱 활발해, Wilk와 Gnanadesikan (1968), Jackson *et al.* (1989), Endrenyi와 Patel (1991), Holmgren (1995), Lee와 Rhee (1997) 그리고 Lee *et al.* (1998) 등에 의해 연구되었다.

한편, 경제학분야에서 가장 대표적인 소득분배의 불균형 지표인 Gini index를 구하는 과정에서 Lorenz curve를 이용한다. 몇몇 소득분포들에 대하여 Lorenz curve를 이용한 Gini index의 연구가 계속 진행되어 왔으며, 특히 Moothathu (1985a, b)는 척도모수와 위치모수를 가지는 지수분포와 위치모수와 형상모수를 가지는 파레토분포에서 Lorenz curve와 Gini index의 최우추정량에 대한 확률분포를 구하였고, 또한 Moothathu (1990)는 파레토분포에서 Lorenz curve, Gini index, 그리고 Theil Entropy Index에 대한 균일최소분산비편향추정량과 강일치 점근정규 비편향추정량도 제시하였다.

본 논문에서는 Lorenz curve를 변환하여 여러 특정분포에 대한 추정에 이 변환된 그래프를 이용하려는 시도를 하였다. 먼저, Lorenz curve에 대한 소개와 변환을 통한 새로운 그래프를 제시하고 (2절), 이 변환된 그래프를 이용하여 소득분포로 대표되는 몇 가지 분포들을 포함한 여러 분포에 대해 모의실험을 통한 분포들의 특성을 비교하고, 예제로 Hodgkin's disease 데이터를 이용하여 정규성 검정에서 사용되는 Q-Q 플롯과 새로 제시한 변환된 Lorenz curve를 비교하였다(3절).

1) (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 겸임조교수
2) (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 부교수
3) (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수

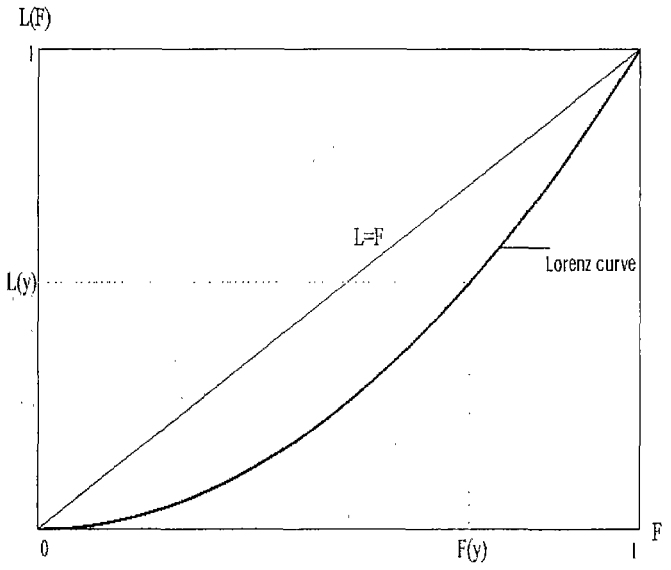

그림 2.1: Lorenz curve

2. Lorenz curve와 변환 그래프

소득분배의 측면에서 경제학자들이 많은 관심을 기울여 왔던 것이 소득의 불균형에 관한 연구이다. 따라서 소득분배의 정도를 측정하기 위해 많은 지표들이 제안되어 왔다. 불균형 정도의 유용한 척도는 소득이 고르게 분배되는 경우와 어느 정도의 차이가 있는지를 측정하는 것이다. 사람들을 소득 순위로 나열할 때 주어진 수준보다 소득이 적은 사람들의 누적비율을 수평축에 나타내고, 이 사람들에 의한 총수입의 누적비율을 수직축에 나타낼 때, 다양한 수준에서 점들을 좌표에서 얻을 수 있다. 이 점들을 연결하여 그린 그래프가 Lorenz curve이다. 이 그래프는 소득분배의 불균형에 관한 척도 중 가장 널리 이용되고 있다.

이 Lorenz curve에 대한 수학적인 형태를 알아보면 다음과 같다.

$$L(y) = \int_0^y x dF(x)/E(Y) \tag{2.1}$$

여기서 Y는 기대값 $E(Y)$가 존재하는 음이 아닌 소득변수이고, $p = F(y)$는 전체 소득수입자의 누적분포함수(cdf)이다.

이와 같이 정의된 변수를 이용하여, $F(y)$를 수평축에 표시하고 $L(y)$를 수직축에 표시하면, 그림 2.1과 같은 Lorenz curve을 그릴 수 있다.

이 소득분포의 cdf는 증가하고 미분가능하기 때문에, $y = F^{-1}(p)$가 정의된다. 이것을 식 (2.1)에 적용하여 Lorenz curve를 바꾸어 나타내면,

$$L(p) = \int_0^p F^{-1}(x)dx/E(Y) \tag{2.2}$$

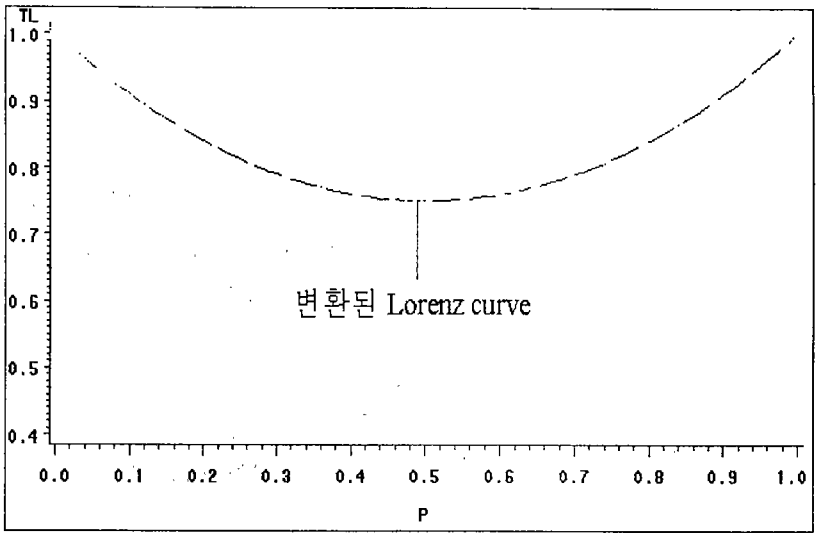

그림 2.2: Lorenz curve의 변환 그래프

가 된다.

이 Lorenz curve의 몇 가지 성질을 살펴보면,

1) Lorenz curve는 직선 $L = F$ 아래의 볼록한 곡선 (convex curve)이 되고,
2) $L(0) = 0$이고,
3) $L(1) = 1$이다.

여기에서 적용되는 Lorenz curve는 소득분배의 불균형 정도의 측도에 널리 이용되고 있지만 그 특징이 대각선 $L = F$ 아래의 볼록한 곡선, 즉 측면에 의존해서 파악해야 하기 때문에 실제 치우친 정도를 파악하는데 약점을 안고 있다. 이 점을 보완하기 위해 우리는 Lorenz curve의 수평적 변환을 통한 새로운 기법을 소개한다. 그 변환 방식을 수평축은 그대로 $F(y)$로 두고 수직축을 $1 + L(F(y)) - F(y)$로 변환한다. 그림 2.2와 같은 형태의 변환된 Lorenz Curve $TL(p) = 1 + L(p) - p$를 얻는다. 이 때 얻어진 변환된 Lorenz curve를 이용하면 좌우대칭을 이루는 표준정규분포 (N(0,1)) (혹은 UNIF(0,1)) 뿐만 아니라 좌우로 치우친 분포의 특징을 더욱 쉽게 알 수 있다.

3. 모의실험을 통한 분포의 추정

모의실험(simulation)을 통하여 위에서 제안한 변환된 Lorenz curve의 그래프를 이용하여 분포의 특성을 고찰함으로써 여러 가지 특정분포들(좌측으로 치우친 분포, 우측으로 치우친 분포, 꼬리부분이 두툼한 분포 등)에 대하여 곡선의 특성을 살펴 보기로 한다. 먼저, 소득분포로 대표되는 지수분포에 대해 알아보자.

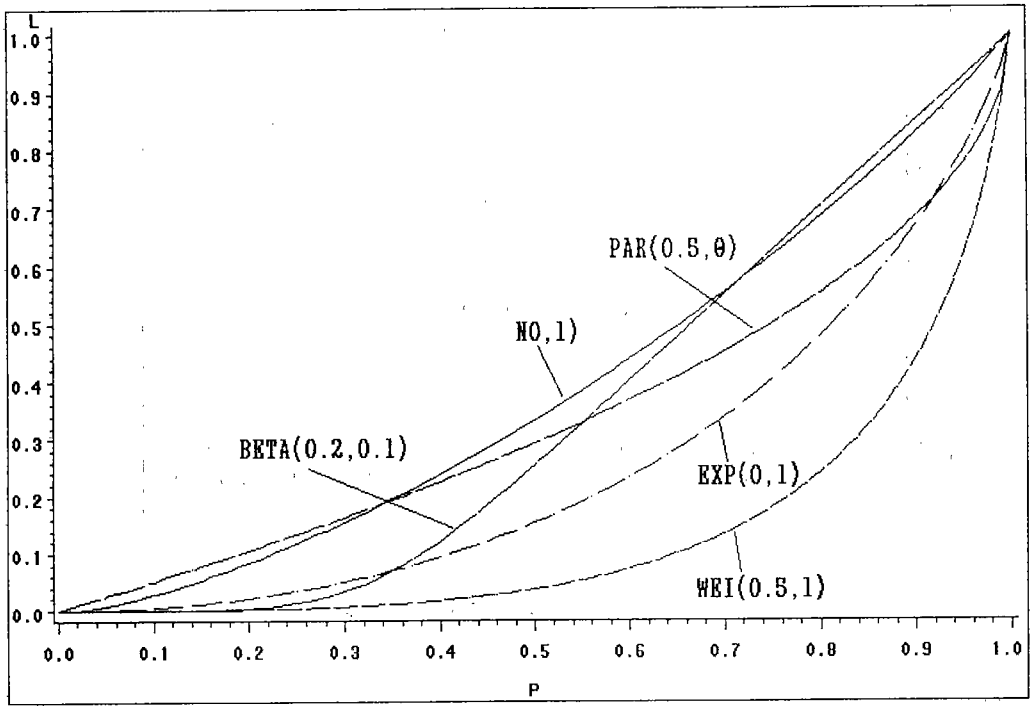

그림 3.1: 특정 분포에 대한 Lorenz curve

연속확률변수 X 의 확률밀도함수가

$$f(x;\mu,\ \sigma) = \frac{1}{\sigma} e^{-\frac{x-\mu}{\sigma}},\ \mu < x,\ 0 < \sigma \tag{3.1}$$

인 척도모수 σ와 위치모수 μ를 갖는 지수분포 (EXP(μ,σ))에 대해 생각해 보면, 이 분포에서 Lorenz curve는

$$\begin{aligned} L(p) &= \int_0^p F^{-1}(t)dt/E(x) \\ &= p + (1-p)\log(1-p)\frac{\sigma}{\mu+\sigma},\ 0 \le p \le 1 \end{aligned} \tag{3.2}$$

이다.

또한 이 Lorenz curve을 이용하여 변환된 Lorenz curve을 구하면

$$TL(p) = 1 + (1-p)\log(1-p)\frac{\sigma}{\mu+\sigma},\ 0 \le p \le 1 \tag{3.3}$$

이다.

다음은 연속확률변수 X의 확률밀도함수가

$$f(x;\theta,\ \xi) = \frac{\xi\theta}{x^{\xi+1}},\ x \ge \theta \tag{3.4}$$

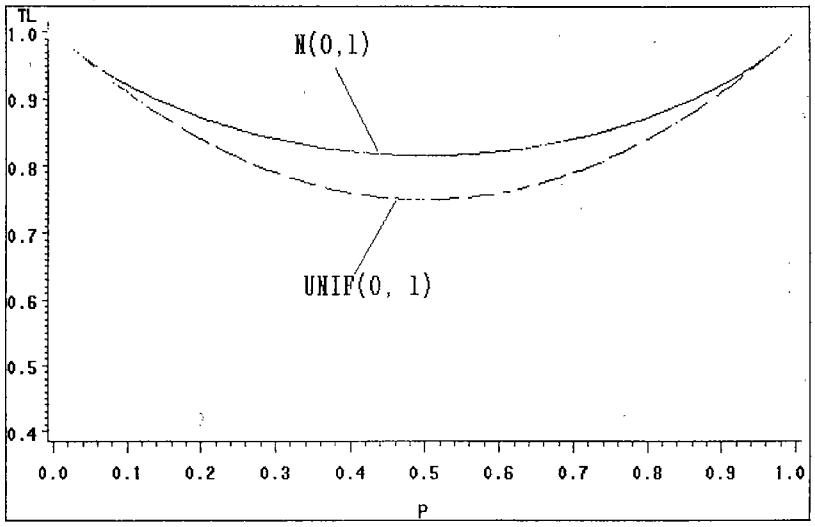

그림 3.2: 좌우대칭인 분포에 대한 변환된 Lorenz curve

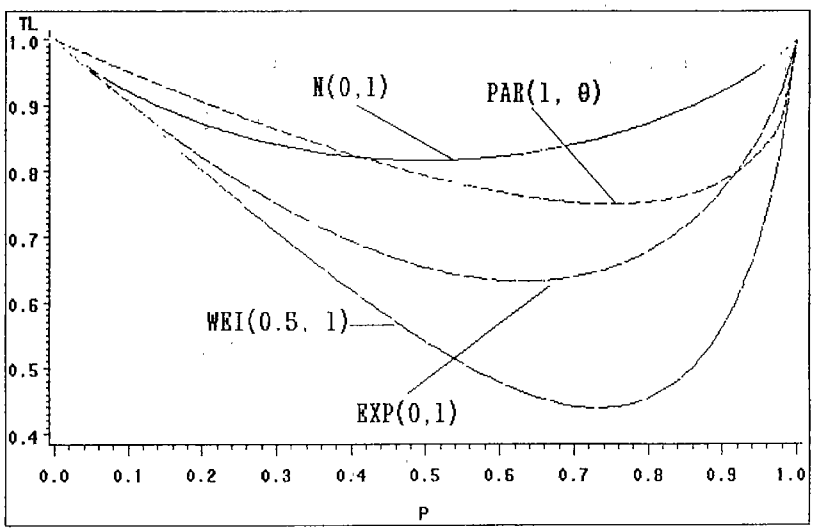

그림 3.3: 우측으로 치우친 분포에 대한 변환된 Lorenz curve

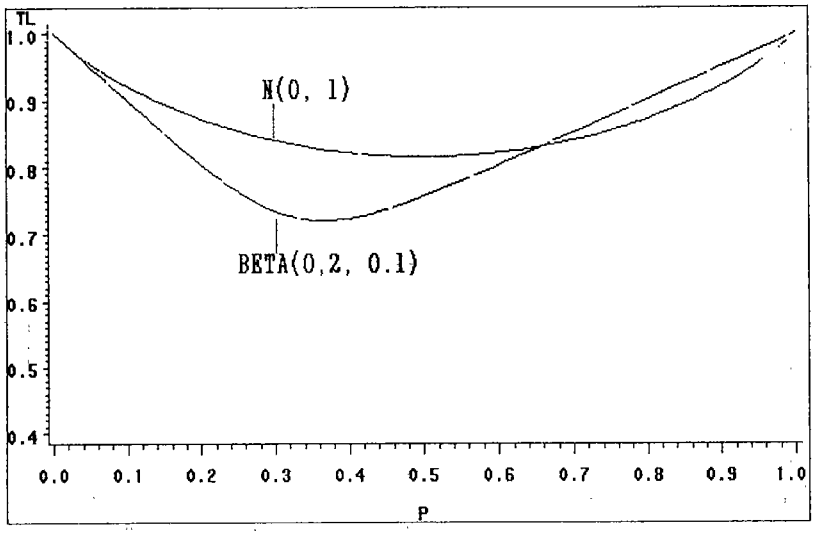

그림 3.4: 좌측으로 치우친 분포에 대한 변환된 Lorenz curve

인 형상모수 ξ와 위치모수 θ를 갖는 파레토분포 (PAR(ξ, θ))에 대해 생각해 보면, 이 분포에서 Lorenz curve는

$$L(p,\xi) = 1-(1-p)^{1-\xi^{-1}},\ 0 \le p \le 1 \tag{3.5}$$

이다. 그러므로 이 Lorenz curve를 이용하여 변환된 Lorenz curve는

$$TL(p,\xi) = 2-p-(1-p)^{1-\xi^{-1}},\ 0 \le p \le 1 \tag{3.6}$$

이다.

다음은 확률밀도함수가

$$f(x;\theta) = \frac{1}{\theta},\ 0 < x < \theta \tag{3.7}$$

인 좌우대칭인 균일분포 (UNIF$(0,\theta)$)에 대해 생각해 보면, 이 분포에서 Lorenz curve는

$$L(p,\theta) = p^2,\ 0 \le p \le 1 \tag{3.8}$$

이다. 그러므로 이 Lorenz curve를 이용하여 변환된 Lorenz curve는

$$TL(p,\theta) = p^2 - p + 1,\ 0 \le p \le 1 \tag{3.9}$$

이다.

일반적으로 확률변수의 누적분포함수가 특성함수로 표시되는 경우에 Lorenz curve를 정확히 계산 할 수 없으므로 다른 방법으로 추정해야 한다. 양의 값을 갖는 확률분포이거나

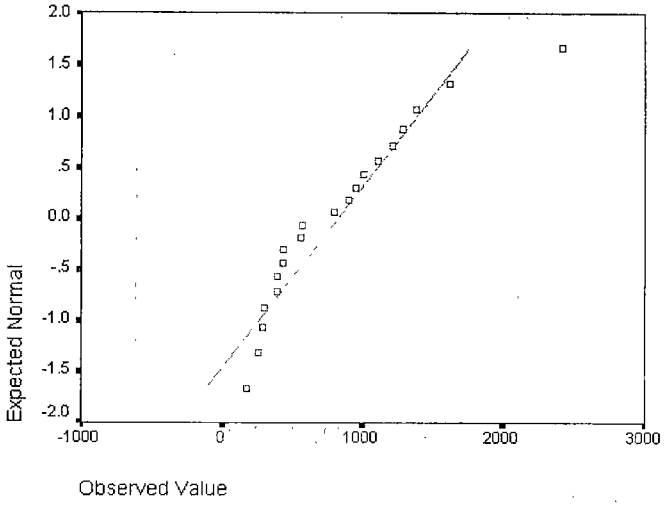

그림 3.5: Hodgkin's disease의 Normal Q-Q 플롯

분포를 알 수 없는 경우에서의 확률표본 $X_1, X_2, \ldots, X_n$ 에 대해 $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$를 순서통계량이라 하면, Lorenz curve는

$$\widehat{L}(p) = \frac{\sum_{j=1}^{i} X_{(j)}}{\sum_{j=1}^{n} X_j}, \; p = i/n, i = 1, 2, \ldots, n \tag{3.10}$$

으로 추정하고 만일 정규분포와 같이 음수를 갖는 경우에는 표본을 양수로 변수 변환하여 Lorenz curve를 다음과 같이

$$\widehat{L}(p) = \frac{\sum_{j=1}^{i} (X_{(j)} - X_{(1)})}{\sum_{j=1}^{n} X_j - nX_{(1)}}, i = 1, 2, \ldots, n \tag{3.11}$$

으로 추정한다. 이 추정된 Lorenz curve를 이용하여 다음과 같은 변환된 Lorenz curve를 추정한다.

$$\widehat{T}L(p) = \frac{\sum_{j=1}^{i} (X_{(j)} - X_{(1)})}{\sum_{j=1}^{n} X_j - nX_{(1)}} - p + 1, \; 0 \le p \le 1. \tag{3.12}$$

그림 3.2에서 좌우대칭인 정규분포와 균일분포에서 변환된 Lorenz curve는 $p = 0.5$를 중심으로 좌우대칭의 그래프이며 꼬리부분이 정규분포보다 두툼한 균일분포의 변환된 Lorenz curve가 정규분포의 변환된 Lorenz curve보다 아래 위치하고 있다. 그림 3.1에 제시한 특정분포의 Lorenz curve을 보면 서로 교차하는 경우에는 치우침의 정도를 파악하기 용이하나 서로 교차하지 않는 정규분포와 지수분포 그리고 와이블분포에서 치우침의 차이를 비교하기가 쉽지 않다 그러나 그림 3.3과 3.4에 제시한 변환된 Lorenz curve를 비교하면 우측으로

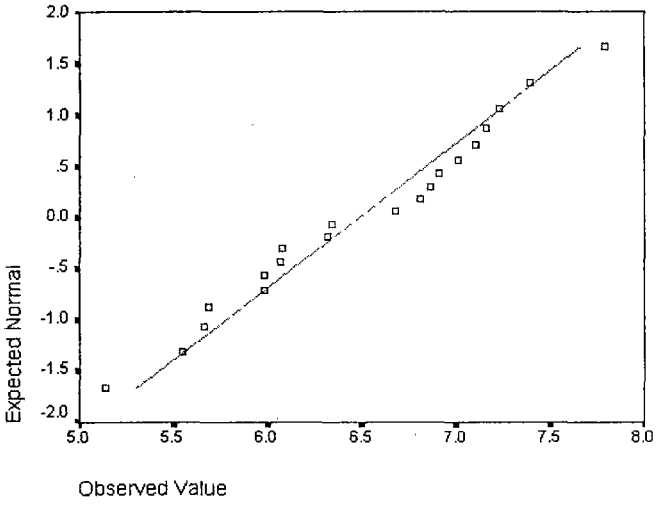

그림 3.6: Log Hodgkin's disease의 Normal Q-Q 플롯

긴 꼬리를 갖는 분포의 변환된 Lorenz curve는 우측으로 치우친 형태를 뚜렷하게 띠고 있고 좌측으로 긴 꼬리를 갖는 분포의 변환된 Lorenz curve는 좌측으로 뚜렷하게 치우친 형태를 띠고 있음을 알 수 있었다.

그리고 간단한 예로서 Hodgkin's disease 데이터 (Alterman(1992))에서 회복된 20명의 환자 혈액샘플에서 mm^3당 세포의 개수를 조사한 자료를 이용하여 정규성을 검정한 결과 정규 Q-Q 플롯은 그림 3.5와 같이 정규성에 벗어난 것처럼 보인다. 이 데이터의 분포가 정규성을 따른다는 가설을 기각하므로 (실제 Shapiro-Wilk 검정통계량의 $P-$값은 0.031) 데이터를 Log변환하여 정규 Q-Q 플롯를 나타낸 결과 그림 3.6과 같이 직선형태의 정규성을 따랐다 (Shapiro-Wilk 검정통계량의 $P-$값은 0.772). 한편, 우리는 Hodgkin's disease 데이터와 Log 변환된 Hodgkin's disease 데이터를 먼저 평균과 표준편차로 표준화하고 식 (3.12)에 의하여 변환된 Lorenz curve을 구하여 그림 3.7에 제시했다. 실제로 Hodgkin's disease 데이터의 정규 Q-Q 플롯인 그림 3.5와 Log Hodgkin's disease 데이터의 정규 Q-Q 플롯인 그림 3.6의 변화를 비교하는 것 보다 새로 제시한 변환된 Lorenz curve의 그림 3.7에서 비교하는 것이 변화를 잘 감지할 수 있다. 물론 이 예제는 단편적인 예제에 불과 하지만 변환된 Lorenz curve를 이용한 정규성검정에도 적용될 수 있다는 확신을 가진다.

4. 결론

경제학분야에서 소득분배의 불균형 정도에 대한 척도로 널리 이용되는 Lorenz curve의 특징은 $L = F$ 대각선을 기준으로 그래프가 기울어져 있어 그 분포별 특징을 파악하는데

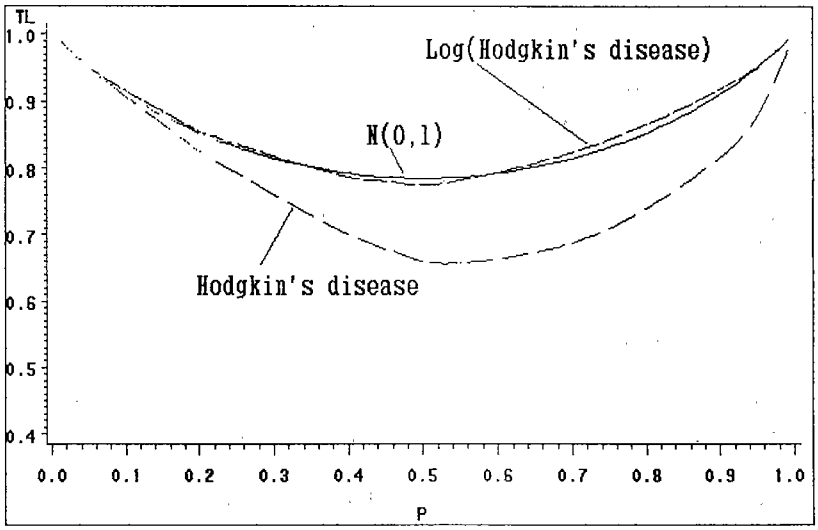

그림 3.7: 변환된 Lorenz curve

어려움이 있었지만 변환된 Lorenz curve의 그래프는 수평축과 평행이므로 그 분포별 특징을 파악하기가 용이하다.

특히 표준정규분포를 기준으로 분포의 좌우대칭과 좌우 치우침을 분석하는데 용이하다고 생각된다. 따라서 변환된 Lorenz curve의 그래프를 관찰함으로서 개략적인 분포를 추정할 수 있다고 생각한다.

참고문헌

[1] Alterman, D. G. (1992). *Practical Statistics for Medical Research*, Chapman and Hall. London.

[2] Endrenyi, L. and Patel, M. (1991). A new, sensitive graphical method for detecting deviations from the normal distribution of drug responses: the NTV plot, *British Journal Clinical Pharmacology*, Vol. 32, 159-166.

[3] Holmgren, E. B. (1995). The P-P plot as a method for comparing treatment effects, *Journal of American Statistical Association*, Vol. 90, 360-365.

[4] Jackson, P. R., Tucker, G. T., and Woods, H. F. (1989). Testing for bimodality in frequency distributions of data suggesting polymorphisms of drug metabolism histograms and probit plots, *British Journal Clinical Pharmacology*, Vol. 28, 647-653.

[5] Lee, J.-Y. and Rhee S.-W. (1997). 특정분포에 따른 확률 Plot들의 정규성과 Bimodality

비교, 〈한국통계학회논문집〉, 제4권 1호, 243-254.

[6] Lee, J.-Y., Woo, J. S., and Chio, D. W. (1998). Using a normal test variable (NTV) for clinical reseach, 〈응용통계연구〉, 제11권 1호, 1-12.

[7] Moothathu, T. S. K. (1985a). Distribution of maximum likelihood estimators of Lorenz curve and Gini index of the exponential distribution, *Annals of Institute of Statistical Mathematics*, Vol. 37, 437-497.

[8] Moothathu, T. S. K. (1985b). Sampling distribution of Lorenz curve and Gini index of the Pareto distribution, *Sankhya*, Series B, Vol. 47, 247-278.

[9] Moothathu, T. S. K. (1990). The best estimator of Lorenz curve, Gini index and Theil entropy index of the Pareto distribution, *Sankhya*, Series B, Vol. 52, 115-127.

[10] Wilk, M. B. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data, *Biometrika*, Vol. 55, 1-17.

[1998년 5월 접수, 1998년 12월 최종수정]

A study on distribution based on the Transformed Lorenz Curve

Young-Suk Cho [1] Jea-Young Lee[2] Suk-Bok Kang [3]

ABSTRACT

In this paper, we introduce the Lorenz curve that is proved to be a powerful tool to measure the income inequality within a population of income receivers. Using the Lorenz curve, we propose the transformed Lorenz curve and compare the transformed Lorenz curves for the special statistical distributions. For two Hodgkin's disease data sets, we compare the Q-Q plots with the transformed Lorenz curves.

1) Adjunct Assistant Professor, Department of Statistics Yeungnam University, 214-1 Daedong, Kyongsan, Kyongbuk 712-749, Korea

2) Associate Professor, Department of Statistics, Yeungnam University, 214-1 Daedong, Kyongsan, Kyongbuk 712-749, Korea

3) Professor, Department of Statistics, Yeungnam University, 214-1 Daedong, Kyongsan, Kyongbuk 712-749, Korea