THE EGYPTION POPULATION AND
FAMILY PLANNING REVIEW
ISSR, CAIRO UNV. Vol. 37, No. 1, 2004

Estimating disease risk using Lorenz curve and negative binomial regression

Ibrahim M Abdalla

Department of Statistics
College of Business and Economics
United Arab Emirates University
P. O. Box 17555, Al-Ain
Tel. +9713 7051426
i.abdalla@uaeu.ac.ae

SUMMARY

The paper proposes a parametric approach to estimate the Lorenz curve and the Gini index in the context of describing exposure-disease association. Nonparametric bootstrap statistical inference method is employed for generating estimates of statistical variability for the Gini index. The index describes the overall degree of risk variation in a population; it does not indicate where in the distribution the variation may be occurring. To remedy this limitation, analysis based on the Gini index is interpreted in conjunction with percentile estimates and a measure of skewness of the Lorenz curve. To demonstrate the proposed methodology, international data on AIDS incidence for selected countries is used. Results obtained using the Lorenz-Gini methodology for estimating disease risk are compared with results obtained from an alternative approach utilizing the negative binomial regression.

KEY WORDS: Lorenz curve; Gini index; Bootstrap; Negative binomial; Disease risk.

1. INTRODUCTION

The measurement of variation in the distribution of incidence of a particular disease in a population according to a range of factors or exposures is a major concern of epidemiologists. One such particular epidemic is the occurrence of AIDS (Acquired Immune Deficiency Syndrome) incidence. Measuring variation in the risk of AIDS infection according to, for example, demographic factors, is a fundamental concern of public health practitioners and civil establishments. Quantitative measures which can adequately catch and describe variation of disease risk are of much importance. Based on these measures of exposure-disease association, it is then possible to set priorities for intervention and to implement preventive measures to reduce the risk.

Two commonly used indices to characterize exposure-disease association in populations are the relative risk [1] and the attributable risk [2] which is recently referred to as the generalized impact fraction [3, 4]. The relative risk provides a stable measure of association in a variety of populations. However, it possesses some limitations when used to compare the strength of the association between several exposures and a particular disease or when used to compare the impact of a particular exposure upon different diseases [4]. On the other hand the attributable risk and /or the generalized impact fraction are shown to be inadequate measures to catch and describe variation of disease risk in populations [4].

Recently, the Gini index, a popular measure used by economists to describe variation in the distribution of income, is employed in detecting and testing temporal clustering of disease occurrences and in characterization of exposure-disease association in human

populations [4, 5]. The Gini index discussed by Lee [4] is defined as twice the area between a 45 degree line and a Lorenz curve, where the Lorenz curve is a graph describing the cumulative proportion of disease incidence against the cumulative proportion of population. The index ranges from a minimum value of zero, when the distribution of disease risk is exactly equal in different exposure (risk) levels, i.e. the Lorenz curve would strictly follow the 45 degree line. This indicates that the exposure under study is not a risk factor. The index has a maximum value of one, that is, if all disease incidence could only occur in one particular exposure level. Thus a larger value of the Gini index (close to 1) means that the risk of the disease is more variable in the population, while a smaller value indicates a more uniform distribution of disease risk.

The definition and the estimation procedure of the Lorenz curve adopted in this paper to characterize exposure-disease association is different from Lee [4]. For each number π which denotes the proportion of total individuals infected with the disease defined in the interval [0,1], the Lorenz curve $L(\pi)$ is the proportion of total age associated with the youngest $(100\pi)\%$ of the given diseased population ranked according to age, x (exposure).

Although the Lorenz curve can be calculated directly from empirical data (see for example [4]), parametric estimation of the curve remains useful particularly when dealing with grouped data. From economics literature, two general parametric approaches to Lorenz curve estimation have been used. 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 [6] and McDonald and Xu [7]. In the second approach, a particular functional form for the Lorenz curve is specified and estimated directly, see for example, Chotikapanich [8]. The first approach is the one utilized in this paper and income is replaced by age which is defined as the exposure or risk factor that influences the distribution of disease incidence. Estimation of parametric Lorenz curve and the calculation of the Gini index based on grouped data discussed in this paper, do not require the determination of the size of the population at risk associated with each exposure level. This is another advantage over the use of the traditional relative risk approach, where determination of the size of the population at risk is essential. To demonstrate the methodology developed in this paper, data from an international database maintained by the UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance for the years 1997 to 2000 are used.

The present research effort constitutes an attempt to respond to two questions. First, it demonstrates the utility of the parametric Lorenz curve and the associated Gini index in describing variation of disease risk in human populations. The methodology discussed in this context is particularly useful when the data available for the analysis is presented as a frequency table which details the different exposure (age) levels, and the associated frequencies of disease incidence. A Gini index measured for the entire population does not show the pattern of risk variation - whether more risk is observed, for example, within the young population or the older population. This constitutes the second question. To get further understanding about risk distribution, the results conveyed by the proposed methodology are discussed in conjunction with percentile estimates and a

measure of the skewness of the Lorenz curve. Moreover, in order to assess the performance of the proposed methodology, an alternative statistical model that links disease rates to the exposures under study is discussed. The Poisson model is a suitable choice for this type of data, however, due to extra-Poisson variation often detected in such data, the negative binomial model can be utilized. The Lorenz-Gini approach, however, has an advantage over the Poisson/negative binomial model as it does not require the knowledge of the size of the population at risk (the denominator of rate).

The rest of the paper is organized as follows: Section 2 describes the data employed in the analysis and the research methodology. Section 3 presents analysis results that demonstrate the proposed approach. Concluding remarks are presented in Section 4.

2. METHODOLOGY

2.1. The data

The data utilized to demonstrate the proposed Lorenz curve and the associate Gini index methodology and to fit a statistical model for AIDS incidence are retrieved from a database generated by the UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance, for the period from 1997 to 2000. These data are maintained on-line via http://www.who.int/hiv and http://www.unaids.org, last accessed 4/4/2004. The database contains the WHO records on individual AIDS cases aggregated at the national level by official agencies in different countries. Due to weaknesses/strengths in the health care and epidemiological systems, the quality of the data varies from one country to the other. For each country in the database AIDS cases are reported by gender, age and mode of transmission. People are typically infected years before they are diagnosed

with AIDS. Age at infection is almost never known, however, the time between infection and diagnosis has lengthened by years as a result of new therapies. Consequently, AIDS case reports and AIDS deaths have been dramatically reduced in industrialized countries. Analysis based on this data is intended to estimate AIDS infection rates at the time/age of disease diagnosis. Henceforth, AIDS infection rate is meant AIDS infection rate at the age of diagnosis.

Six countries are included in this study; namely, the United Kingdom, Sweden, Brazil, Thailand, Saudi Arabia and the Republic of Niger. The selection is meant to reflect different geographical, cultural, social and religious backgrounds.

In order to estimate AIDS infection rate based on a Poisson or a negative binomial models, population estimates are obtained for all countries involved in the study for the period from 1997 to 2000. Population estimates and summary demographic data are retrieved from on-line sources; namely, the International Data Base Bank (IDB), maintained by the U.S. Census Bureau under the address http://www.census.gov/ipc/www/idbnew.html, last accessed 4/4/2004.

2.2. Parametric Lorenz curve and the Gini index

Let n_i denote the number of individuals infected with a disease that falls in the i^{th} age group, exposure level, $i = 1,\ldots,k$. Let p_i denote the probability that a randomly selected infected individual falls in age group i. Let $\mathbf{n} = (n_1, n_2, \ldots, n_k)$ and $\mathbf{p} = (p_1, p_2, \ldots, p_k)$. The joint probability density function $f(\mathbf{n}|\mathbf{p})$, which gives the

probability that the number of individuals infected with the disease in age group i is $n_i, i = 1,\ldots,k$, follows the multinomial distribution. This is expressed as

$$f(\mathbf{n}|\mathbf{p}) \propto p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k},$$

where $\sum_{i=1}^{k} p_i = 1$ and $\sum_{i=1}^{k} n_i = N$, where N is the total number of individuals infected with the disease.

Assume that age x can be represented by a particular density function, then the parameters p_i can be expressed as functions of a number of parameters, $p_i = p_i(\lambda)$. That is, if $F(x|\lambda)$ is the cumulative distribution function (cdf) for age distribution, where λ is the vector of parameters upon which $F(x|\lambda)$ depends, then

$$f(\mathbf{n}|\mathbf{p}) \propto \prod_{i=1}^{k} [F(x_{i+1}|\lambda) - F(x_i|\lambda)]^{n_i}.$$

Assuming that the distribution of age is lognomal with parameters $\lambda = (\mu, \sigma)$, then $F(x_i|\lambda)$ is given by

$$F(x_i|\lambda) = \Phi\left(\frac{\log(x_i) - \mu}{\sigma}\right),$$

where Φ denotes the standard normal cdf. Maximum likelihood estimates of μ and σ can be obtained by maximizing the likelihood function $f(\mathbf{n}|\mathbf{p})$.

The functional form of the Lorenz curve which is derived from the lognormal model (see [9]) is given as

$$L(\pi) = \Phi(\Phi^{-1}(\pi) - \sigma),$$

where π denotes the cumulative proportion of individuals infected with the disease, $\pi = \sum_{j \le i} n_i / N$. The Gini index associated with this Lorenz curve is given by [9] as

$$\text{Gini} = 2\Phi(\frac{\sigma}{\sqrt{2}}) - 1.$$

2.3. Negative binomial regression

Poisson regression is often used to analyze count or rate data such as number/rate of AIDS incidence in a particular region or period of time. A key assumption of the Poisson model is that the variance equals the mean. However, count data, y, often exhibit over-dispersion, with a variance larger than the mean. Evidence of over-dispersion indicates inadequate fit of the Poisson model. Corrective measures include using a simple scale-factor adjustment by setting the deviance or Pearson Chi-square divided by degrees of freedom as an estimate of the dispersion parameter. However, this might not be appropriate if the variance of y is increasing faster than the Poisson model allows. One way to handle this situation is to fit a parametric model that is more dispersed than the Poisson. A natural choice is the negative binomial. Suppose that $y \sim \text{Poisson}(\lambda)$, but λ itself is a random variable with a gamma distribution. That is,

$$y \mid \lambda \sim \text{Poisson}(\lambda),$$
$$\lambda \sim \text{Gamma}(\alpha, \beta),$$

where $\text{Gamma}(\alpha,\beta)$ is the gamma distribution with mean $\alpha\beta$ and variance $\alpha\beta^2$, whose density is

$$f(\lambda) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\lambda/\beta),$$

for $\lambda > 0$ and zero otherwise. Based on this, the unconditional distribution of y is negative binomial,

$$f(y) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)(y!)}(\frac{\beta}{1+\beta})^{y}(\frac{1}{1+\beta})^{\alpha},$$

$y = 0,2,\cdots$. The mean and variance for the negative binomial are given as

$$E(y) = E[E(y \mid \lambda)] = E(y) = \alpha\beta$$

$$\begin{aligned} Var(y) &= E[Var(y \mid \lambda)] + Var[E(y \mid \lambda)] \\ &= Var(\lambda) + E(\lambda) \\ &= \alpha\beta + \alpha\beta^2. \end{aligned}$$

To build a regression model, the negative binomial is expressed in terms of the parameters $\mu = \alpha\beta$ and $k = \alpha$, so that $E(y) = \mu$ and $Var(y) = \mu + k\mu^2$. The negative binomial distribution approaches Poisson(μ) as $k \to 0$. Thus, it is typically assumed that

$$y_i \sim Negbin(\mu_i, k)$$

and apply a log link, so that

$$\log(\mu_i) = \eta_i = offset + x_i^T\beta.$$

The offset is a regression variable with constant coefficient of 1 for each observation. It is usually set as the logarithm of the population at risk. x_i are the predictor variables and β are parameters to be estimated.

Test for over-dispersion can be conducted using a *likelihood ratio test* based on Poisson and negative binomial distributions. This tests equality of the mean and the variance imposed by the Poisson distribution (i.e. $k = 0$) against the alternative that the variance exceeds the mean (i.e. $k > 0$). The test statistic (the *likelihood ratio, LR*) is given by

$$LR = -2[\log likelihood(Poisson) - \log likelihood(negative\ binomial)]$$

The asymptotic distribution of the likelihood ratio is Chi-square distribution with 1 degree of freedom [10]. Over-dispersion is rejected if the likelihood ratio is greater than $\chi^2_{(1-2\alpha,1\,d.f.)}$, where α is the level of significance.

3. ANALYSIS RESULTS

3.1. Fitting a Lorenz curve and Gini index For AIDS data

The data comprising the combined AIDS incidence covering the period from 1997 to 2000 for all countries described in Section 2.1 are used to fit the Lorenz curve and the associated Gini index. Based on assuming a lognormal distribution for age, the parameters μ and σ are estimated via maximum likelihood estimation technique employing the likelihood function $f(\mathbf{n}|\mathbf{p})$ outlined in Section 2.2 above. Consequently, Lorenz curves, $L(\pi)$s, and the corresponding Gini indices are estimated by country. Generated 250 bootstrap samples are used to estimate the standard error and the 95% confidence interval associated with each Gini index. The confidence intervals are based on bias-corrected and adjusted percentiles (2.5% and 97.5%) obtained using the bootstrap function in the S-plus package, see Table 1.

Figure 1 demonstrates the size of variation in AIDS incidence due to age revealed by the fitted Lorenz curves. Results suggest that the risk of AIDS is more variable in the UK than in the other countries in the group (Gini index = 0.27, s.e.=0.11). Much of the risk is amongst the youngest and the old population, see discussion of percentile estimates below. The Republic of Niger is linked with the lowest variation in AIDS incidence due to age factor (Gini index = 0.21, s.e.=0.07). When these results are discussed in conjunction with percentile estimates and the skewness of the Lorenz curve, a clear image regarding the distribution of AIDS risk emerges. The mean, $e^{(\mu+\frac{1}{2}\sigma^2)}$, and median, e^{μ}, age of infection with AIDS for each country together with the estimated skewness coefficient, $(e^{\sigma^2}+2)\sqrt{e^{\sigma^2}-1}$, for a lognormal model are given

in Table 2. All curves indicate moderate skewness of AIDS infection towards old age. Using the constructed Lorenz curves, percentile estimates indicate that the share of AIDS infections amongst the youngest 25% of the population is lowest in Sweden (1.9%) and highest in Thailand and the UK (4.6%). Overall, infections amongst the young population are low compared to the old population. The bulk of AIDS infections are reported in middle age range (between the 25th and the 75th percentiles). Thailand stands odd with 85.6% of the AIDS infected population contained in this group, followed by the Republic of Niger (82.3%), see Table 2. Based on the Gini index alone, results indicate that the UK has a higher variation in AIDS risk compared to Thailand. It is clear now that the middle age group in Thailand explains much of the variation in the distribution of AIDS risk compared to the same group in the UK. Thus a Gini index complemented with another measure of variation (percentile estimates) would result in a better understanding of the distribution of disease risk.

Table 1: Estimated Gini index, Standard errors (s.e.), and 95% Confidence interval (C.I.).

Country	Gini index	s.e.	95% C.I.
UK	0.27	0.11	0.13 – 0.61
Sweden	0.22	0.06	0.14 – 0.47
Brazil	0.24	0.09	0.13 – 0.57
Thailand	0.24	0.10	0.12 – 0.60
Saudi Arabia	0.23	0.08	0.14 – 0.59
Republic of Niger	0.21	0.07	0.12 – 0.43

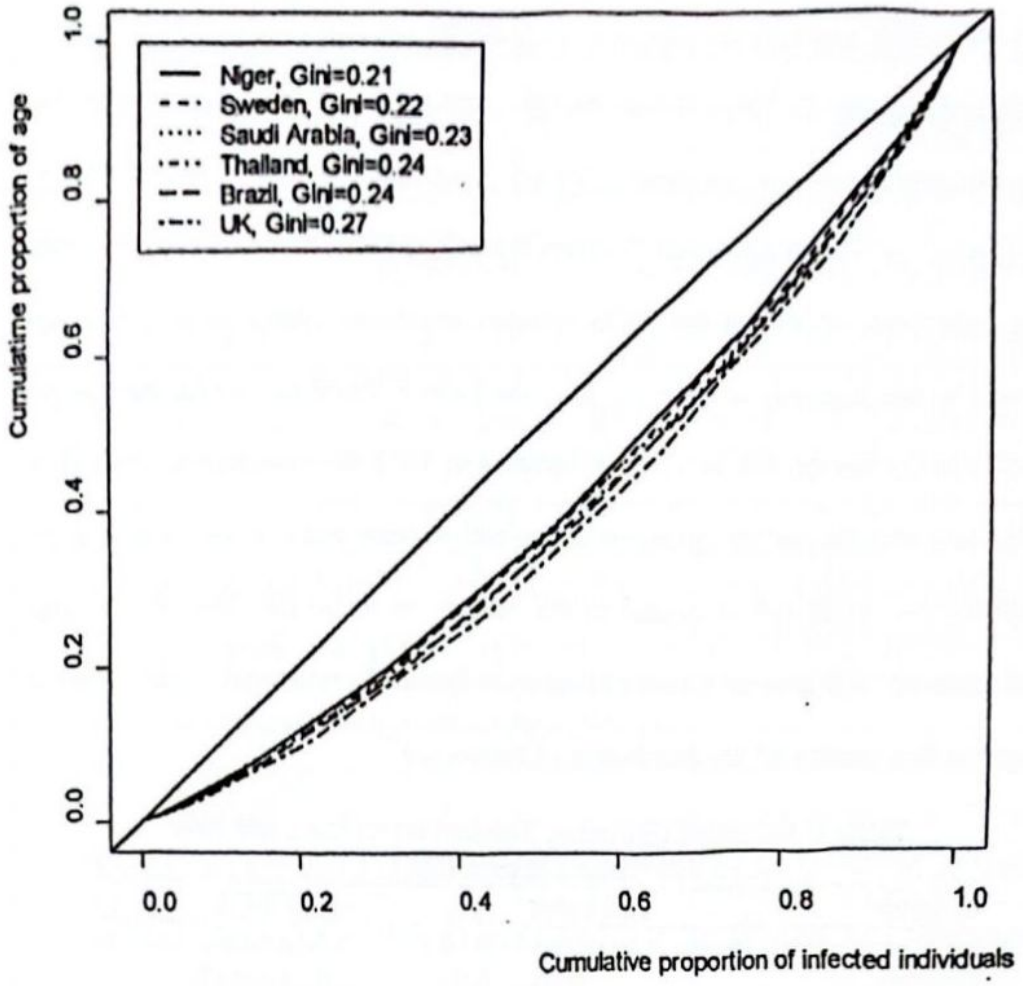

Figure 1: Lorenz curves, distribution of AIDS incidence in selected countries.

Table 2: Estimated mean and median age of infected population, together with estimated skewness and selected age percentiles of the Lorenz curve by country

				% of AIDS cases in the population		
Country	Mean	Median	Skewness	Amongst the youngest 25%	Amongst the middle 25%-75%	Amongst the oldest 25%
UK	38.9	34.6	1.68	4.6	72.6	22.8
Sweden	38.2	35.4	1.27	1.9	71.1	27.0
Brazil	35.5	32.3	1.48	4.3	79.2	16.5
Thailand	33.1	30.0	1.48	4.6	85.6	9.8
Saudi Arabia	34.7	31.8	1.41	4.5	74.1	21.4
Republic of Niger	25.1	32.8	1.22	3.4	82.3	14.3

To describe the pattern of association between exposure (age) and the risk of AIDS disease over time, the Gini indices associated with all Lorenz curves fitted for each country using annual data are shown in Figure 2. The figure suggests that the effect of age as a factor that determines AIDS risk in Thailand is declining over the last two years, 1999 - 2000. To some extent, the same pattern can be observed in Sweden data. This contrasts with the pattern depicted by British and Saudi data. In Brazil, after a drop in the risk of AIDS due to age in 1997, an increasing pattern is observed over 1998 to 2000. The Republic of Niger has the lowest association between AIDS and age exposure in 1997 compared to the other countries in the group. However, this has increased significantly in 1998 and started to level down in the following two years 1999 and 2000.

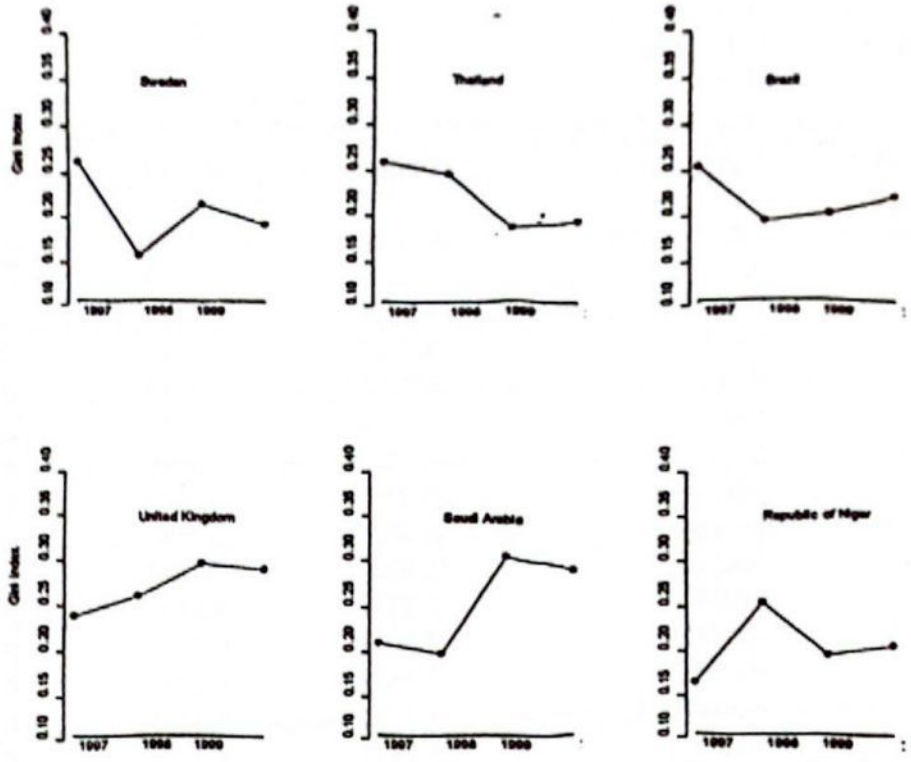

Figure 2: Distribution of the Gini index by country and time (years).

3.2. Negative binomial model application

Before using the negative binomial model as an alterative for a Poisson model, the significance of over-dispersion of the data under the Poisson model is tested by carrying out the likelihood ratio test (*LR*). Using the GENMOD procedure in SAS statistical package, Poisson and negative binomial models are fitted separately. Predictors x_i employed in the fitting are two class variables; namely the country where AIDS incidence are reported (UK, Sweden, Brazil, Saudi Arabia, Republic of Niger and Thailand. Thailand is set as a reference) and age (eleven 5 years interval levels, 0-4, 15-19,...,60+ , with the 60+ years set as a reference). *LR* is computed using the formula outlined in Section 2.3 and is given as $LR = 6856.79$. This value corresponds to a p-value=0.0. Hence the hypothesis that $k = 0$ is rejected and it is concluded that the mean and the variance are not equal and that the Poisson distribution assumption is not suitable, so it has to be abandoned.

Table 3: Negative binomial regression - maximum likelihood parameter estimates.

Predictor	Name	Coefficient	Std error	p-value
Intercept	-	-9.5972	0.2765	0.0001
Country	UK	-3.0422	0.2291	0.0001
	Sweden	-3.8671	0.2505	0.0001
	Brazil	-1.1531	0.2252	0.0001
	Saudi	-4.8822	0.2617	0.0001
	Niger	-1.7425	0.2324	0.0001
Age	0-4	0.3854	0.3245	0.2350
	5-9	-0.8887	0.3386	0.0087
	10-14	-2.0259	0.3860	0.0001
	15-19	-0.3937	0.3380	0.2440
	20-24	1.9227	0.3272	0.0001
	25-29	2.4427	0.3167	0.0001
	30-34	2.9031	0.3197	0.0001
	35-39	2.5683	0.3166	0.0001
	40-49	2.1313	0.3168	0.0001
	50-59	1.6192	0.3244	0.0001

Table 3 summarizes the negative binomial model regression results including coefficients and their statistical significance terms. Positive coefficients suggest that as

each input term increases, there is a corresponding increase in the rate of AIDS disease. Similarly, a negative coefficient reflects an inverse relationship. Based on this, it is clear that all countries in the group have lower rates of AIDS infection compared to Thailand, (p-value=0.0001), this confirms Lorenz-Gini results obtained in Table 2. The lowest infection rate compared to Thailand, is reported in Saudi Arabia (see Table 3). Moreover, the population aged 5-14 years have a lower infection rate compared to those aged 60+, the p-value < 0.05. Where those aged 20 - 59 have higher infection rate compared to the 60+, p-value=0.0001. These results are also similar to those based on the Lorenz curve and its percentile estimates, see Table 2. The model scaled deviance and Pearson Chi-square suggest a fair degree of over dispersion in AIDS incidence, 1.60 and 1.36 respectively.

4. CONCLUDING REMARKS

The paper proposed a parametric approach to estimate the Lorenz curve and the Gini index in the context of describing exposure-disease association. As a measure that characterizes exposure-disease association in populations, the Gini index has some limitations. One particular limitation, is that it measures the overall risk variation in a population and does not show the pattern of risk within different exposure levels. A possible remedy for this limitation is by discussing results and implications driven using the Lorenz and the Gini index in conjunction with results obtainable using other measures of variation such as the skewness of the Lorenz curve and percentile estimates.

The methodology is demonstrated using international AIDS data covering a number of countries. Results based on the Lorenz-Gini methodology are similar to those obtained using an alternative technique employing the negative binomial regression. This

suggests that the Lorenz-Gini methodology works well in providing alternative estimates of disease risk. Moreover, the parametric Lorenz-Gini methodology employed in this paper has an advantage over other measures of risk, including the negative binomial regression and the relative risk approaches, in that the estimation of the curve and the Gini index is independent of the size of the population at risk. This is important, as the bulk of published data in the internet and other sources is available in frequency distribution format.

REFERENCES

1. Breslow N E, Day N E. Statistical Methods in Cancer Research, IARC Scientific Publication 1980; I (32), Lyon.

2. Lvein M L. The Occurance of Lung Cancer in Man, Acta Union Internationalis Contra Cancrum 1953; 9, 531-541.

3. Morgenstern H, Bursic E S. A Method for Using Epidemiologic Data to Estimate the Potential Impact of an Intervention on the Health Status of a Target Population, Journal of Community Health 1982; 7, 292-309.

4. Lee W C. Characterizing Exposure-Disease Association in Human Populations Using the Lorenz Curev and Gini Index, Statistics in Medicine 1997; 16, 729-739.

5. Lee W C. Analysis of Seasonal Data Using the Lorenz Curve and the Associated Gini Index, International Journal of Epidemiology 1996; 25, 426-434.

6. McDonald J B. Some Generalized Functions for the Size Distribution of Income, Economitrica 1984; 52, 647-663.

7. McDonald J B, Xu Y J. A Generalization of the Beta Distribution with Applications, Journal of Econometrics 1995; 66, 133-152.

8. Chotikapanich D. A Comparison of Alternative Functional Forms for the Lorenz Curve. Economic Letters 1993; 41: 129-138.

9. Schader M, Schmid F. Fitting Parametric Lorenz Curves to Grouped Income Distributions - A Critical Note, Empirical Economics 1994; 19, 361-370.

10. Cameron A C, Trivedi P K. Regression Analysis of Count Data, Cambridge University Press, 1998.