DISCUSSION PAPERS IN STATISTICS AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

No. 2/98

Statistical Inference for Inequality Measurement with Dependent Data

by

Christian Schluter and Mark Trede

September 1998

DISKUSSIONSBEITRÄGE ZUR STATISTIK UND ÖKONOMETRIE

SEMINAR FÜR WIRTSCHAFTS- UND SOZIALSTATISTIK
UNIVERSITÄT ZU KÖLN

Albertus-Magnus-Platz, D-50923 Köln, Deutschland

DISCUSSION PAPERS IN STATISTICS AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

No. 2/98

Statistical Inference for Inequality Measurement with Dependent Data[¤]

by

Christian Schluter[y] and Mark Trede[z]

September 1998

Abstract: Standard methods of statistical inference for inequality (or poverty and mobility) measures are based on the assumption of income being an independent and identically distributed random variable. Unfortunately, income data in most empirical problems are neither identically nor independently distributed and the usual methods cannot be applied. We propose estimators consistent with contemporaneous dependence across members of the same household. Monte Carlo experiments con¯rm that the reliability of con¯dence intervals and tests is increased by our methods if the observations are in fact dependent.

Keywords: Inequality measures; statistical inference; asymptotics, dependent observations

JEL classi¯cation: D31, D63, I32

Correspondence to: Mark Trede, Seminar für Wirtschafts- und Sozialstatistik, Universität zu Köln, 50923 Köln, Germany, Tel: +49-221-4702283, Fax: +49-221-4705074, email: trede@wiso.uni-koeln.de

[¤]Partial funding from the British German Academic Research Collaboration (ARC) is gratefully acknowledged.

[y]Department of Economics, University of Bristol, 8 Woodland Road, Bristol, BS8 1TN, UK, Tel.: +44-117-928 8431, Fax: +44-117-928 8577, e-mail: C.Schluter@bristol.ac.uk

[z]Seminar für Wirtschafts- und Sozialstatistik, Universität zu Köln, 50923 Köln, Germany, Tel.: +49-221-470 2283, Fax: +49-221-470 5074, e-mail: trede@wiso.uni-koeln.de

1 Introduction

The recently observed increases in both earnings and income inequality in most developed countries has brought inequality back on the agenda of applied research (Atkinson, Rainwater, and Smeeding 1994, Gottschalk and Smeeding 1998). Whether the observed movements are indeed statistically significant is a somewhat neglected question. Standard methods of statistical inference for inequality (or poverty and mobility) measures are based on the assumption that income is an independent and identically distributed random variable (Beach and Davidson 1983, Cowell 1998a, Cowell 1998b, Hoeffding 1948). Unfortunately, income data in most empirical problems are neither identically nor independently distributed. The i.i.d. assumption often fails in practice because of three types of violations. First, the assumption of identical distributions is violated because inclusion probabilities are not identical for all individuals. Second, temporal dependencies occur in panel data because the same person is observed at different points in time and her incomes are correlated. Third, contemporaneous dependencies arise because income receivers live in households and labour supply decisions, for instance, are taken jointly.

This paper develops distribution-free methods in the presence of such contemporaneous dependencies, and thereby provides a complement to two recent advances in statistical inferences for inequality indices, i.e., the treatment of non-identical sample inclusion probabilities (Sandström 1987) and functional or temporal dependence of income data for Lorenz curves (Davidson and Duclos 1997).

The considerations are important since inferences based on the wrong data generating mechanism are uninformative as can be seen in the following artificial example: assume that we are interested in the mean income of individuals, that all households consist of two persons, and that income is distributed as

$$(X_1, X_2) \text{ » } N(\mu, \Sigma)$$

with unknown μ but known

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

The data are sampled on the household level, i.e., the sample is clustered. From each household $i = 1, \ldots, n$ we have a pair of incomes (X_{i1}, X_{i2}). Obviously, mean

individual income can be estimated by $\bar{X} = \frac{1}{2n} \sum (X_{i1} + X_{i2})$. To obtain a confidence interval for $\bar{X}$ we need its variance $Var\left(\bar{X}\right)$. Wrongly neglecting the possible intra-household dependence would result in $Var\left(\bar{X}\right) = 1/(2n)$, whereas the variance actually is $Var\left(\bar{X}\right) = (1+\rho)/(2n)$. Note that whether the variance is smaller or larger than in the i.i.d. case depends on the kind of correlation. The more unequal the intra-household distribution the smaller the variance.

This paper is organized in the following way: Section 2 introduces the notation and gives a brief overview over standard statistical inference for inequality measures when the observations are in fact i.i.d. Section 3 is the main contribution of this paper: we develop nonparametric methods of statistical inference when there are contemporaneous dependencies. We consider both scalar measures of inequality (moments based indices and the Gini coefficient) and Lorenz curves. Section 4 compares our estimates with the standard i.i.d. case by means of Monte-Carlo simulation. An empirical illustration using earnings data from the German Socio-economic panel is also given. Finally, section 5 concludes.

2 Independent observations

This section briefly recalls the standard theory of inequality measurement and inferences based on i.i.d. income data. We examine members of two classes of measures, namely moments-based measures such as the Generalised Entropy index and quantile-based measures such as Lorenz curves and the Gini coefficient.

Let X represent the random variable income with distribution $F(x)$, giving the proportion of the population with income less than, or equal to, x. X_i, $i = 1, \ldots, n$ denotes an i.i.d. sample of size n from F. It is this assumption which will be abandoned in subsequent sections. The empirical distribution function of X is defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x)$$

where $1(c)$ is an indicator function equal to 1 if condition c is met and 0 otherwise.

Let I be a moment-based inequality measure such as a Generalized Entropy measure or Atkinson's inequality index. Let $g_i(X)$, $i = 1, \ldots, K$ be transformations of the random variable, and $\mu(g_i(X))$ their expectations. For ease of notation let μ denote

the column vector $(\mu(g_1(X)), \ldots \mu(g_K(X)))'$. All members of the class of moment-based inequality measures can be written as $I = I(\mu)$. For instance, the Generalised Entropy index $GE_\alpha(F)$ with sensitivity parameter $\alpha \neq 0, 1$ is defined by

$$I_{GE_\alpha} = \frac{1}{\alpha^2 - \alpha} \left[\frac{\mu(g_2(X))}{[\mu(g_1(X))]^\alpha} - 1 \right] \tag{1}$$

with the transformation functions $g_1(X) = X$ and $g_2(X) = X^\alpha$, i.e., $K = 2$.

The unbiased sample estimator of $\mu(g_i(X))$, denoted by $\hat{\mu}(g_i(X))$, is

$$\hat{\mu}(g_i(X)) = \int g_i(x) d\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} g_i(X_j), \quad i = 1, \ldots, K. \tag{2}$$

The method-of-moments estimator of the class of moments-based measures is $\hat{I} = I(\hat{\mu})$. Its asymptotic distribution is easily derived by the Cramer-Wold-Device (Serfling 1980) and the delta method (Cramér 1946). For $n \to \infty$ the moments are jointly normally distributed, $\sqrt{n}(\hat{\mu} - \mu) \to N(0, \Sigma)$ with covariance matrix Σ whose elements are

$$\Sigma_{ij} = \mu(g_i(X)g_j(X)) - \mu(g_i(X))\mu(g_j(X)), \quad \text{for } i, j = 1, \ldots, K. \tag{3}$$

This covariance matrix can be estimated consistently by replacing the population moments μ by their empirical counterparts $\hat{\mu}$.

Since the moments are asymptotically normally distributed the inequality estimator itself, being a function of the moments, is normally distributed as well:

$$\sqrt{n}(\hat{I} - I) \to N\left(0, \frac{\partial I}{\partial \mu}' \Sigma \frac{\partial I}{\partial \mu}\right) \tag{4}$$

where $(\partial I / \partial \mu)$ is the gradient of I. Both the covariance matrix and the derivative of I at the true value of μ are obviously unknown. However, according to Slutsky's theorem the asymptotic distribution of $\hat{I}$ is unchanged if we replace Σ and the partial derivatives by consistent estimates.

As Sandström (1987) has demonstrated, taking into account different inclusion probabilities is straightforward. However, in order to concentrate on the issue of contemporaneous dependence, we will assume in subsequent sections that the data are identically distributed. The modifications necessary to take nonidentical inclusion probabilities into account are obvious.

The Lorenz curve depicts the cumulative income share of the least well-off fraction of the population. Let x_p and p denote a quantile of the income variable and its population share, $x_p = F^{-1}(p)$. A coordinate of the Lorenz curve is a pair $(p; \Phi(p; F))$ where

$$\Phi(p; F) = \frac{1}{\mu(X)} \int_0^{x_p} x dF(x) \tag{5}$$

and $\mu(X)$ is the population mean. The consistent estimator is obtained by using the sample analogues $\Phi(p; \hat{F}_n)$, whose asymptotic normality, shown amongst others in Beach and Davidson (1983), follows from the fact that order statistics are asymptotically normally distributed around the respective population quantiles. For the variance estimator see Beach and Davidson (1983).

The last inequality measure to be examined is the Gini coefficient, perhaps the most well-known quantile-based measure. The Gini coefficient is defined as

$$Gini(F) = \frac{\delta(F)}{2\mu(X)} \tag{6}$$

where $\delta(F) := \int\int |x - y| dF(x) dF(y)$. Hoeffding (1948) shows that an unbiased estimator of $\delta(F)$ is $[n/(n-1)]\delta(\hat{F}_n)$ which is a member of the class of U-statistics. A limit theorem (the delta-method for U-statistics) implies that the sample estimator of the Gini coefficient is asymptotically normal. For the cumbersome variance expression see Hoeffding (1948).

3 Contemporaneous Dependencies

The usual variance estimators are based on the assumption that the data are independent and identically distributed. Unfortunately, this assumption often fails in practice, since, for instance, labour supply decisions are taken jointly within households. Ignoring such contemporaneous dependencies will produce wrong inferences. In this section we propose variance estimators which allow for such dependencies.

3.1 Moment-based inequality measures

Let X_i be the income of individual i. Each individual belongs to a (unique) household, the households may have different sizes. Let the index sets H_h, $h = 1, \ldots, H$ denote

the households, H being the number of households. $|H_h|$ is the size of household h, the total number of persons in the sample is $n = \sum_h |H_h|$. With such contemporaneous dependencies X_i, $i = 1, \ldots, n$ is – if ordered by households – a sequence of m-dependent random variables with $m = \max_h |H_h|$. Using a limit theorem for m-dependent processes (Spanos 1986, p. 179), the Cramer-Wold-Device and the delta-method give that

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \Sigma) \quad \text{and} \quad Var(\hat{I}) = \frac{1}{n} \left(\frac{\partial I}{\partial \mu} \right)' \Sigma \left(\frac{\partial I}{\partial \mu} \right) \tag{7}$$

as in (4) where I is the inequality measure.

In order to derive the asymptotic variance of the inequality index we only need to know the covariance matrix Σ of the empirical moments $\hat{\mu} = (\hat{\mu}(g_1(X)), \ldots, \hat{\mu}(g_K(X)))'$ which is

$$\Sigma = Cov(\hat{\mu}) = E(\hat{\mu}\hat{\mu}') - \mu\mu', \tag{8}$$

since $\hat{\mu}$ is unbiased. For ease of notation, abbreviate $g_p(X)$ by g_p and $g_p(X_i)$ by g_{pi}. Consider a typical element (p, q) of the $(K \times K)$-matrix $E(\hat{\mu}\hat{\mu}')$:

$$\begin{aligned} E(\hat{\mu}(g_p)\hat{\mu}(g_q)) &= E\left[\frac{1}{n} \sum_i g_{pi} \cdot \frac{1}{n} \sum_j g_{qj} \right] \\ &= \frac{1}{n^2} \sum_h \sum_{i \in H_h} E\left(g_{pi} g_{qi}\right) + \frac{1}{n^2} \sum_h \sum_{i \in H_h} \sum_{j \in H_h, j \neq i} E\left(g_{pi} g_{qj}\right) \\ &\quad + \frac{(n^2 - \sum_h |H_h|^2)}{n^2} \mu(g_p)\mu(g_q) \end{aligned} \tag{9}$$

The first term in (9) refers to all individuals, neglecting any dependencies. The second term in (9) may be written as

$$\sum_h \sum_{i \in H_h} \sum_{j \in H_h, j \neq i} E\left(g_{pi} g_{qj}\right) = \sum_{s=1}^{\max_h |H_h|} \sum_{h:|H_h|=s} \sum_{i \in H_h} \sum_{j \in H_h, j \neq i} E\left(g_{pi} g_{qj}\right),$$

where s is an index for household size. Assuming that the (theoretical) expectation $E\left(g_{pi} g_{qj}\right) =: \mu^{(s)}_{p,q,i,j}$ is constant across households of equal size, the last equation may be written as

$$\sum_s \sum_{i \neq j} \mu^{(s)}_{p,q,i,j} n^{(s)}, \tag{10}$$

where $n^{(s)}$ is the number of households of size s. Constructing an estimator for $\mu^{(s)}_{p,q,i,j}$ is straightforward: compute the sample cross products $g_p g_q$ for all households of size

s. A natural estimator for $\mu^{(s)}_{p,q,i,j}$ thus is

$$\hat{\mu}^{(s)}_{p,q,i,j} = \frac{1}{n^{(s)}} \sum_{h:|H_h|=s} \frac{1}{s(s-1)} \sum_{i \in H_h} \sum_{j \in H_h, j \neq i} g_p(X_i) g_q(X_j). \tag{11}$$

The first and third term in (9) are estimated in the obvious way. Thus, all the components necessary for calculating the asymptotic variance of the inequality index are available: Having estimated $E(\hat{\mu}\hat{\mu}')$ in (8) by (9) we immediately arrive at the variance estimator (7). If there are no intra-household dependencies only the first term in (9) is non-zero and we arrive at the usual (i.i.d.) variance estimator.

Note that we do not need any information regarding the position of individual i in the household. We simply impose symmetry on all household members and assume that the stochastic nature of the dependencies between any two members are symmetric as well. This assumption is rather strong: it presumes that the correlation of earnings between husband and wife is the same as the correlation between, say, son and father. However, using the suggested approach immediately permits further refinement of such assumptions.

3.2 Lorenz curves

The Lorenz curve ordinates as defined in (5) may be written as

$$\Phi(p, F) = \frac{p\gamma_p}{\mu(X)}$$

with

$$p\gamma_p = \int_0^{x_p} x dF(x) \tag{12}$$

being the conditional mean income of persons with income less than the p-quantile of the distribution, $x_p = F^{-1}(p)$. The unconditional mean $\mu(X)$ can be interpreted as a special conditional mean as well ($\mu(X) = \gamma_1$). In order to obtain the joint asymptotic distribution of the Lorenz curve ordinates we first derive the covariance matrix of the estimators of (12) and then apply the delta method. For notational convenience and without loss of generality, we restrict attention to households of identical size m and we make a slight change of notation. Income of individual $j = 1, \ldots, m$ in household $i = 1, \ldots, n$ will be denoted by X_{ij}. The marginal income distribution (of individuals) has distribution function F, the joint distribution of any two household members is written as $G($ ¢ ¢, see below.

As in Beach, Davidson, and Slotsve (1995) we will work with more general conditional moments than (12). Let

$$p\gamma_p = \int_0^{x_p} h(x)dF(x) \tag{13}$$

$$q\delta_q = \int_0^{x_q} g(x)dF(x) \tag{14}$$

be conditional moments of some functions $h()$ and $g()$ of income. Obviously, (12) is just a special case of (13), $h(x) = x$. The quantities (13) and (14) can be estimated nonparametrically by

$$\hat{\gamma}_p = \frac{1}{nm} \sum_{i,j} h(X_{ij}) 1(X_{ij} \cdot x_p)$$

and similarly for the other estimator. The estimators can (with an error of order at most $o(n^{-1})$) also be written as (see Beach and Davidson 1983)

$$\begin{aligned} p\hat{\gamma}_p &= p(h(x_p)) + \frac{1}{nm} \sum_{i,j} (h(X_{ij}) - h(x_p)) 1(X_{ij} \cdot x_p) \\ q\hat{\delta}_q &= q(g(x_q)) + \frac{1}{nm} \sum_{i,j} (g(X_{ij}) - g(x_q)) 1(X_{ij} \cdot x_q). \end{aligned}$$

By definition (and again with an error of order at most $o(n^{-1})$)

$$\begin{aligned} & Cov\left[p\hat{\gamma}_p, q\hat{\delta}_q\right] \\ = & \frac{1}{(nm)^2} \sum_{i,j} \sum_{k,l} E\left[(h(X_{ij}) - h(x_p)) 1(X_{ij} \cdot x_p)(g(X_{kl}) - g(x_q)) 1(X_{kl} \cdot x_q)\right] \\ & - E\left[(h(X_{ij}) - h(x_p)) 1(X_{ij} \cdot x_p)\right] E\left[(g(X_{kl}) - g(x_q)) 1(X_{kl} \cdot x_q)\right]. \end{aligned} \tag{15}$$

The last two expectations in (15) are taken with respect to $F(x)$:

$$\int [h(x) - h(x_p)] 1(x \cdot x_p) dF(x) = p(\gamma_p - h(x_p))$$

and analogously for the other term. The first expectation in (15) is investigated in some detail in the appendix. Assuming without loss of generality that $p \cdot q$ we finally arrive at

$$\begin{aligned} Cov(p\hat{\gamma}_p, q\hat{\delta}_q) = & \frac{m-1}{nm} C + \frac{p}{nm} \phi_p - \gamma_p \delta_p \\ & + [1 - mq][\gamma_p - h(x_p)][\delta_q - g(x_q)] \\ & + [h(x_p) - \gamma_p][\delta_q - \delta_p] \end{aligned} \tag{16}$$

where

$$p\phi_p = E\left[h\left(X_{ij}\right) g\left(X_{ij}\right) 1\left(X_{ij} \cdot x_p\right)\right]$$

and

$$\begin{aligned} C &= G\left(x_p, x_q\right) \eta_{p,q} + h\left(x_p\right) g\left(x_q\right) G\left(x_p, x_q\right) \\ &\quad - g\left(x_q\right) \int_0^{x_p} h\left(x\right) dG\left(x, x_q\right) - h\left(x_p\right) \int_0^{x_q} g\left(x\right) dG\left(x_p, x\right) \end{aligned}$$

with

$$G\left(x_p, x_q\right) \eta_{p,q} = E\left[h\left(X_{ij}\right) g\left(X_{il}\right) 1\left(X_{ij} \cdot x_p\right) 1\left(X_{il} \cdot x_q\right)\right].$$

In the special case where all observations are independent, i.e., $G(x, y) = F(x)F(y)$, equation (16) reduces to the result derived in Beach and Davidson (1983).

It is possible to estimate the unknown quantities in (16) consistently from the sample without making any distributional assumptions.

3.3 The Gini coefficient

The Gini coefficient can be written in a form equivalent to (6) but more convenient as

$$Gini(\mu, A) = 1 - \frac{2A}{\mu}$$

where $\mu = \mu\left(X\right)$ is the population mean and $A = \int_0^1 p\gamma_p dp$ is the integral of the conditional moments defined in (12). Its estimator uses the sample counterparts $\hat{\mu}\left(X\right)$ and $\hat{A} = \int_0^1 p\hat{\gamma}_p dp$, i.e. $\widehat{Gini} = Gini(\hat{\mu}\left(X\right), \hat{A})$. Since both $\hat{\mu}$ and $\hat{A}$ are normally distributed, application of the delta method implies that the Gini coefficient is normally distributed as well. Let Ψ denote the covariance matrix of the vector $(\hat{\mu}, \hat{A})$. The variance of $\hat{\mu}$ has already been derived in section 3.1. Consider next the second moment and variance of $\hat{A}$. The notation $\hat{\gamma}(p, x)$ is used to emphasise that the estimated Lorenz ordinate depends on the realisation of the random variable and that integration is over the population share p. Expectations are taken with respect to the distribution function $F(x)$.

$$E(\hat{A}^2) = E\left[\int_0^1 p\hat{\gamma}(p, X)dp \int_0^1 q\hat{\gamma}(q, X)dq\right]$$

$$= \int_0^1 \int_0^1 \int_{-\infty}^{\infty} p\psi(p,x)q\psi(q,x)dF(x)dpdq$$
$$= \int_0^1 \int_0^1 Cov(p\psi_p, q\psi_q)dpdq + A^2$$

using (16). Thus $Var(\hat{A}) = \int_0^1 \int_0^1 Cov(p\psi_p, q\psi_q)dpdq$.

The off-diagonal element of - is $Cov(\hat{\mu}, \hat{A})$ but noting that $\hat{\mu} = \psi_1$ we have immediately

$$Cov(\hat{\mu}, \hat{A}) = \int_0^1 Cov(p\psi_p, \psi_1)dp$$

The variance estimator of $\hat{Gini}$ is thus $(\partial Gini/\partial(\mu, A))'$- $(\partial Gini/\partial(\mu, A))$ evaluated at $(\hat{\mu}, \hat{A})$.

4 Simulation and empirical illustration

We turn to examining the precision of the variance estimator proposed in the previous section and compare its performance to that of the usual (i.i.d.) estimator given in (2). Using artificially generated data the performance is assessed in terms of the coverage success of confidence intervals. In each iteration of the experiment, we check whether or not the $(1 - \alpha)$-confidence interval contains the true value of the inequality index. If the number of repetitions is large, a good variance estimator would generate confidence intervals which fail to capture the true population value $\alpha \cdot 100\%$ of the time. We have used 1000 repetitions.

Two data models are examined in the simulation study. Model I is a traditional male breadwinner model. The population consists of two-person households composed of a spouse who does not earn and a breadwinner, whose earnings are drawn from a lognormal distribution $LN(\mu = 1, \sigma = 0.5)$ with density

$$f(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right).$$

Model II is an assortative mating model. The population are two-person households whose members receive earnings from a bivariate lognormal distribution $LN(1, 0.5)$ with the same marginal densities as above and correlation coefficient ρ where either $\rho = 0.75$ or $\rho = 0.3$.

Since model I generates observations where half of the population have zero earnings we need an inequality index capable of handling zeros. We opted for the Generalized Entropy index with parameter $\alpha = 2$ and parameter $\alpha = 1$.

Using the decomposability of the Generalized Entropy index the population inequality values are easily computed from the parameters of the lognormal distributions. We will derive the true population values for $\alpha = 2$, the same approach can be used for $\alpha = 1$. There are two groups, earners (denoted by subscript 1) and non-earners (subscript 2). Mean earnings are zero for the latter group and $\exp(\mu + \sigma^2/2) = 3.0802$ for the former; $E(X_1) = 3.0802$ and $E(X_2) = 0$. The overall mean is therefore $E(X) = 1.5401$. Although the Generalized Entropy index is not properly defined for distributions with zero incomes only, it obviously makes sense to set its value to zero. Inequality in the earners group is

$$GE_\alpha = \frac{\exp\left[\frac{1}{2}(\alpha^2 - \alpha)\sigma^2\right] - 1}{\alpha^2 - \alpha}, \tag{17}$$

which is $GE_{2,1} = 0.1420$ for the earners. The Generalized Entropy can be decomposed as

$$\begin{aligned} GE_\alpha &= GE_{\alpha,between} + GE_{\alpha,within} \\ &= \frac{1}{\alpha^2 - \alpha}\left[\sum_{j=1}^{2} f_j \left(\frac{E(X_j)}{E(X)}\right)^\alpha - 1\right] + \sum_{j=1}^{2} w_j GE_{\alpha,j} \end{aligned}$$

where f_j is the relative size of group j, hence $f_1 = f_2 = 1/2$, and the weights $w_j = f_j\left(E(X_j)/E(X)\right)^\alpha$, hence $w_1 = 0$ and $w_2 = 2^{\alpha-1}$. Therefore, under model I the true population value is $GE_2 = 0.7840$. For $\alpha = 1$ the population value can be computed along the same lines, it is $GE_1 = 0.8178$.

Under model II there are no between inequalities since both groups have the same distribution. Applying (17) yields $GE_1 = 0.1250$ and $GE_2 = 0.1420$.

For the simulation study we varied the coverage probability of the confidence intervals, the sensitivity parameter of the inequality index, and the correlation coefficient ρ in Model II. The number of observations is $n = 2500$ households (i.e. 5000 individuals). Table 1 reports the results of the simulations. The most noteworthy result is the good coverage performance of confidence intervals constructed from our variance estimator. In contrast, the confidence intervals build on the iid variance estimator

perform poorly, particularly so in model I where the coverage failure is much too small (i.e. the variances are grossly over-estimated). In model II we observe the opposite: the iid variances are under-estimated and, as a result, the confidence intervals are too narrow. This effect vanishes if the correlation is small – for $\rho = 0.3$ there is hardly any difference in the coverage performances – as it should be since as the correlation falls, the data generating mechanism approaches the iid case.

Table 1: Proportion of coverage failure

Index		Model I		Model II, $\rho = 0.75$		Model II, $\rho = 0.3$	
	α	50%	5%	50%	5%	50%	5%
GE(2)	iid	0.250	0.001	0.590	0.098	0.506	0.052
	dep	0.524	0.064	0.524	0.052	0.493	0.047
GE(1)	iid	0.011	0.000	0.597	0.134	0.531	0.050
	dep	0.537	0.052	0.515	0.062	0.511	0.044

To apply the proposed variance estimator on real data, we estimate the inequality of annual earnings for the 1983 cross-section of the German Socio-Economic Panel (GSOEP). The cross-section consists of 4253 households, whose size varies from 1 to 10, the average size being 3.06. Although the sampling probabilities of the households are in fact not exactly identical we assume for simplicity that the sample is identically distributed. Roughly 47% of the individuals in the sample have zero earnings. We compare the variance estimates from section 3.1 to the iid variance estimates. Because of the zero observations we use the same inequality indices as above, i.e., the Generalized Entropy index with $\alpha > 0$.

Table 2 lists the results. Obviously, the variance estimates differ a lot for smaller values of the parameter α (high sensitivity at low incomes, in particular at zero) while the differences are rather small for large values of α. This findings are consistent with the simulation results above.

5 Conclusion

In this paper we consider the effect of intra-household dependencies of earnings or income on the statistical inference of inequality indices and Lorenz curves. The stan-

Table 2: Variance estimates under i.i.d. and dependence

Parameter α	Variance estimate under i.i.d. assumption	Variance estimate under dependence
0.01	1.927e-01	6.351e-03
0.05	7.952e-03	2.024e-04
0.10	2.085e-03	4.632e-05
0.50	1.768e-04	4.196e-05
1.00	2.985e-04	2.210e-04
1.50	1.681e-03	1.578e-03
2.00	1.733e-02	1.706e-02

dard methods of statistical inference are usually based on the assumption of income being an independent and identically distributed random variable. Most micro data sets have information about individuals but are sampled on the household level. It is unlikely that earnings or incomes of individuals living in the same household are independent since, for instance, labour supply decisions are taken either simultaneously or stepwise.

We show that the variance estimates based on the i.i.d. assumption may be misleading if there are in fact intra-household dependences. Theoretical considerations as well as Monte Carlo simulations show that whether the i.i.d. estimates are too wide or too narrow depends on the kind of dependence.

Appendix

In order to derive the covariance structure of the Lorenz curve ordinates, we adapt an argument suggested by Beach, Davidson, and Slotsve (1995) to the case with contemporaneous dependencies. For the first expectation in (15) we need to distinguish three cases; we assume without loss of generality that $p \cdot q$.

1. $i = k$ and $j = l$, there are nm such cases.

$$\begin{aligned}
& E\left[\left(h\left(X_{ij}\right) - h\left(x_p\right)\right) 1\left(X_{ij} \cdot x_p\right)\left(g\left(X_{ij}\right) - g\left(x_q\right)\right) 1\left(X_{ij} \cdot x_q\right)\right] \\
= \; & E\left[1\left(X_{ij} \cdot x_p\right)\left(h\left(X_{ij}\right) g\left(X_{ij}\right) - h\left(X_{ij}\right) g\left(x_q\right) - g\left(X_{ij}\right) h\left(x_p\right) + h\left(x_p\right) g\left(x_q\right)\right)\right] \\
= \; & \underbrace{E\left[1\left(X_{ij} \cdot x_p\right) h\left(X_{ij}\right) g\left(X_{ij}\right)\right]}_{=:p\phi_p} - pg\left(x_q\right)\gamma_p - ph\left(x_p\right)\delta_p + ph\left(x_p\right) g\left(x_q\right) \\
= \; & p\left[\phi_p - g\left(x_q\right)\gamma_p - h\left(x_q\right)\delta_p + h\left(x_p\right) g\left(x_q\right)\right]
\end{aligned}$$

2. $i = k$ but $j \neq l$, there are $nm\left(m - 1\right)$ such cases. Let the joint income distribution of two members of the household be denoted by $G(x, y)$.

$$\begin{aligned}
& E\left[\left(h\left(X_{ij}\right) - h\left(x_p\right)\right)\left(g\left(X_{il}\right) - g\left(x_q\right)\right) \underbrace{1\left(X_{ij} \cdot x_p\right) 1\left(X_{il} \cdot x_q\right)}_{=:i\left(X_{ij}, X_{il}\right)}\right] \\
= \; & E\left[i\left(X_{ij}, X_{il}\right)\left(h\left(X_{ij}\right) g\left(X_{il}\right) - h\left(X_{ij}\right) g\left(x_q\right) - h\left(x_p\right) g\left(X_{il}\right) + h\left(x_p\right) g\left(x_q\right)\right)\right].
\end{aligned}$$

Examine this expression term by term:

$$\begin{aligned}
E\left[i\left(X_{ij}, X_{il}\right) h\left(X_{ij}\right) g\left(X_{il}\right)\right] &= \int h\left(x\right) g\left(y\right) i\left(x, y\right) dG\left(x, y\right) \\
&= : G\left(x_p, x_q\right)\eta_{p,q} \\
E\left[i\left(X_{ij}, X_{il}\right) h\left(x_p\right) g\left(x_q\right)\right] &= h\left(x_p\right) g\left(x_q\right) G\left(x_p, x_q\right) \\
E\left[i\left(X_{ij}, X_{il}\right) h\left(X_{ij}\right) g\left(x_q\right)\right] &= \int h\left(x\right) i\left(x, y\right) g\left(x_q\right) dG\left(x, y\right) \\
&= g\left(x_q\right) \int_0^{x_p} \int_0^{x_q} h\left(x\right) G'\left(x, y\right) dydx \\
&= g\left(x_q\right) \int_0^{x_p} h\left(x\right) dG\left(x, x_q\right) \\
E\left[i\left(X_{ij}, X_{il}\right) g\left(X_{il}\right) h\left(x_p\right)\right] &= h\left(x_p\right) \int_0^{x_q} g\left(x\right) dG\left(x_p, x\right).
\end{aligned}$$

If $G\left(x, y\right)$ is symmetric and $g\left(x\right) = h\left(x\right) = x$, the last two expressions become

$$\begin{aligned}
& x_q \int_0^{x_p} x dG\left(x, x_q\right) \\
& x_p \int_0^{x_q} x dG\left(x, x_p\right).
\end{aligned}$$

3. $i \neq k$, there are $nm(n-1)m$ such cases.

$$\begin{aligned} & E\left[1\left(X_{ij} \leq x_p\right)\left(h\left(X_{ij}\right)-h\left(x_p\right)\right)1\left(X_{kl} \leq x_q\right)\left(g\left(X_{kl}\right)-g\left(x_q\right)\right)\right] \\ = & \; p\left(\gamma_p-h\left(x_p\right)\right)q\left(\delta_q-g\left(x_q\right)\right). \end{aligned}$$

Putting all terms together we arrive at

$$\begin{aligned} Cov(p\hat{\gamma}_p, q\hat{\delta}_q) = & -p\left(\gamma_p-h\left(x_p\right)\right)q\left(\delta_q-g\left(x_q\right)\right) \\ & + \left(\frac{1}{nm}\right)^2 nmp\left(\phi_p-g\left(x_q\right)\gamma_p-h\left(x_p\right)\delta_p+h\left(x_p\right)g\left(x_q\right)\right) \\ & + \left(\frac{1}{nm}\right)^2 nm(n-1)mp\left(\gamma_p-h\left(x_p\right)\right)q\left(\delta_q-g\left(x_q\right)\right) \\ & + \left(\frac{1}{nm}\right)^2 nm(m-1)C \end{aligned}$$

where

$$\begin{aligned} C = & \; G\left(x_p, x_q\right)\eta_{p,q}+h\left(x_p\right)g\left(x_q\right)G\left(x_p, x_q\right) \\ & -g\left(x_q\right)\int_0^{x_p} h(x)\,dG\left(x, x_q\right)-h\left(x_p\right)\int_0^{x_q} g(x)\,dG\left(x_p, x\right). \end{aligned}$$

and

$$\begin{aligned} p\phi_p = & \; E\left[h\left(X_{ij}\right)g\left(X_{ij}\right)1\left(X_{ij} \leq x_p\right)\right] \\ G\left(x_p, x_q\right)\eta_{p,q} = & \; E\left[h\left(X_{ij}\right)g\left(X_{il}\right)1\left(X_{ij} \leq x_p\right)1\left(X_{il} \leq x_q\right)\right]. \end{aligned}$$

The expression can be simplified to

$$\begin{aligned} Cov(p\hat{\gamma}_p, q\hat{\delta}_q) = & \frac{m-1}{nm}C \\ & +\frac{p}{nm}\left[\phi_p-\delta_p\gamma_p+[1-mq]\left[\gamma_p-x_p\right]\left[\delta_q-x_q^2\right]+\left[x_p-\gamma_p\right]\left[\delta_q-\delta_p\right]\right]. \end{aligned}$$

References

Atkinson, A., L. Rainwater, and T. Smeeding (1994). *Income Distribution in OECD countries: The Evidence from the Luxemburg Income Study*. OECD, Paris.

Beach, C., R. Davidson, and G. Slotsve (1995). Distribution free statistical inference for lorenz dominance with crossing lorenz curves. *Greqam DP 95A03*.

Beach, C. M. and R. Davidson (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies 50*, 723–725.

Cowell, F. A. (1998a). Estimation of inequality indices. In J. Silber (Ed.), *Income Inequality Measurement: From Theory to Practice*. Kluwer, Dewenter.

Cowell, F. A. (1998b). Measurement of inequality. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution*, Chapter 2. Amsterdam: North Holland.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Davidson, R. and J.-Y. Duclos (1997). Statistical inference for the measurement of the incidence of taxes and transfers. *Econometrica 65*, 1453–1465.

Gottschalk, P. and T. Smeeding (1998). Empirical evidence on income inequality in industrialised countries. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution*, Chapter 10. Amsterdam: North Holland.

Hoeffding, W. (1948). A class of statistics with asymptotic normal distribution. *Annals of Mathematical Statistics*, 293–325.

Sandström (1987). Asymptotic normality of linear functions of concomitants of order statistics. *Metrika 34*, 129–142.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. New York: New York.

Spanos, A. (1986). *Statistical Foundations of Econometric Modelling*. Cambridge: Cambridge University Press.