THE UNIVERSITY of York

Discussion Papers in Economics

No. 2001/17

Measuring Income Related Inequality in Health and Health Care: the Partial Concentration Index with Direct and Indirect Standardisation

by

Hugh Gravelle

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Measuring income related inequality in health and health care: the partial concentration index with direct and indirect standardisation

Hugh Gravelle*

October 2001

Abstract. The partial concentration index measures income related inequality in health (or health care) after removing the effects of standardising variables which affect health (or health care), are correlated with income but not amenable to policy. When the marginal effects of income are independent of the standardising variables, direct standardisation yields consistent estimates of the partial concentration index. Indirect standardisation underestimates the partial concentration index whenever the standardising variables are correlated with income, irrespective of the signs of the correlation of standardising variables and income with each other and with health. A generalised version of the partial concentration index is proposed for cases where the marginal effect of income depends on the standardising variables. Direct standardisation again yields a consistent estimate but indirect standardisation does not. It is also shown that the direct standardisation procedure can be applied to individual or grouped data and that the conclusions about the merits of direct and indirect standardisation hold for grouped data.

Keywords: concentration index; inequality; direct standardisation

JEL codes: I18, I32

---

* National Primary Care Research and Development Centre, Centre for Health Economics, University of York, York YO10 5DD; email hg8@york.ac.uk; fax 01904 432700; phone 01904 433663.
NPCRDC receives funding from the Department of Health. The views expressed are not necessarily those of the DH. I am grateful for comments from Mark Dusheiko, Andrew Jones, Matt Sutton, and participants in a seminar at the University of Bergen.

1. Introduction

The concentration index is a commonly used measure of income related inequality in health and health care (van Doorslaer, et al, 2000; Wagstaff and van Doorslaer, 2000a). It is a generalisation of the Gini coefficient (Lambert, 1993). In the case of income related inequality in health, the index is derived from the concentration curve $L(s)$ which graphs the cumulative proportion of health against the cumulative proportion of the population ranked by income (see Figure 1). If there is no income related inequality in health the poor will be, other things equal, as healthy as the rich and the poorest k% of the population will have k% of total population health. The concentration curve will then coincide with the 45^o line. If poor people are less healthy than the rich the poorest k% will have less than k% of total health and the concentration curve will lie below the 45^o line, as in Figure 1.

The concentration index C_{hy} summarises the total amount of income related inequality in health. It is defined as twice the area between the health concentration curve $L(s)$ and the 45^o line:

$$C_{hy} = 1 - 2\int_0^1 L(s)ds \tag{1}$$

When the poor have a disproportionately small share of health the concentration curve $L(s)$ lies below the diagonal and C_{hy} is positive.

The concentration index for health care is derived in an exactly analogous manner, with health care replacing health in the above account. We present the analysis in terms of the concentration index for health but the results derived hold also for concentration indices for health care. Whenever it is felt that some characteristic of individuals (health, health care, payments for health care, consumption of other specific goods …) should not vary with their income, the concentration index of that characteristic against income is an appealing measure of income related inequality. The arguments in the paper on the appropriate way to define and estimate income related inequality therefore have wide applicability to discussions of horizontal equity.

An obvious potential difficulty with C_{hy} as a measure of income related inequality in health is that other factors z (age, sex, education…) are likely to affect health and to be correlated with income (van Doorslaer and Koolman, 2000).[1] For example if health increases with income and decreases with age, C_{hy} will be smaller the greater the extent to which age is positively correlated with income. The measured health of individuals will reflect age and income and the positive effect of income will on average be reduced by the effect of age. Since age is not a factor susceptible to policy the true extent of policy relevant income related inequality is understated by C_{hy}. The effect of *standardising* or policy irrelevant variables like age should be removed from a measure of income related inequality.

The *partial concentration index* I_{hy} is a measure of policy relevant income related inequality. It is derived by removing the effect of the policy irrelevant variables from C_{hy}. It is an intuitively appealing measure of income related inequality in that it focuses attention on the sources of income related inequality which are amenable to policy. It has been used extensively in the inequality literature (Wagstaff and van Doorslaer, 2000). Section 2 discusses its properties and its interpretation when there are policy relevant variables affecting health and correlated with income.

Direct and indirect standardisation are two methods of estimating the effect of the policy irrelevant variables and removing them from C_{hy} to yield partial concentration indices. With direct standardisation the effect of the standardising variables on health is estimated via a health function which includes the standardising variables and income. For indirect standardisation the effect of the standardising variables is estimated from a health function which includes only the standardising variables. Both approaches can be found in the literature. For example, Propper and Upward (1992), Sutton (2001), and van Doorslaer, E., et al (1997) use direct standardisation; Kakwani et al (1997), Urbanos-Garrido (2001), Wagstaff and van Doorslaer, (2000) and van Doorslaer and Koolman (2000) use indirect standardisation.

With full information on the relevant and irrelevant variables affecting health, direct standardisation yields consistent estimates of the partial concentration index. By

[1] Similarly, if the issue is income related inequality in health care, the fact that health status is

contrast, even under full information, estimates of the partial concentration index based on indirect standardisation are inconsistent (section 3). Indirect standardisation underestimates inequality because the estimates of the effect of the standardising variables on health also partially capture some of the effect of income on health which is left out of the estimated health equation for indirect standardisation. Hence deducting the estimated effect of the standardising variables also removes some of the income related inequality. The underestimation is worse the greater the extent to which the irrelevant or standardising variables are correlated with income. Indeed, indirect standardisation will tend to reduce the measured level of inequality even if the "standardizing" variables have no effect on health.

When there is incomplete information on the variables affecting health neither direct or indirect standardisation give consistent estimates of the partial concentration index. Direct standardisation seems preferable in such cases since indirect standardisation suffers from the deliberate omission of the income variable from the health equation as well from the omission of variables on which there is no information.

When income and standardising variables have a linear effect on health it is straightforward to remove the effect of standardising variables correlated with income to calculate the partial concentration index. But it will often be the case that income and the standardising variables interact in determining health. For example the protective effect of income on health is likely to depend on age and gender. Calculating health with the standardising variables fixed at their mean in some reference population is an intuitively appealing method of dealing with the interdependence of income and the standardising variables. The resulting inequality measure depends on the fixed values of the standardising variables but it does so in a transparent way and it is simple to investigate its sensitivity to different assumptions. The procedure also solves the problem caused by the correlation of income and the standardising variables.

What we have just described is direct standardisation as it is understood in the epidemiological literature where it is applied to grouped data. We also investigate the

---

correlated with income and affects consumption should be allowed for.

calculation of the partial concentration index for grouped data using both direct and indirect standardisation. Our conclusions about the relative merits of direct and indirect standardisation are shown to be valid for grouped as well as for individual level data.

2. Partial concentration indices as inequality measures

Consider the individual level[2] health production function

$$h = \beta_0 + \beta_y y + \beta_z z + \beta_x x + \varepsilon \tag{2}$$

where h is a measure of health, y is income, x and z are other variables affecting health. We assume that there are no other factors affecting health which are correlated with income, z or x. ε is therefore an error uncorrelated with any of the factors affecting health. To simplify notation, z and x are interpreted as single variables and proofs of the more general case are relegated to the Appendix.

The concentration index can be written as (Lambert, 1993)

$$C_{hy} = \frac{2}{\mu_h}\mathrm{Cov}(h, F(y)) = \frac{2}{\mu_h}\mathrm{Cov}(\beta_0 + \beta_y y + \beta_z z + \beta_x x + \varepsilon, F(y)) \tag{3}$$

where μ_h is mean population health, and $F(y)$ is the distribution function for income. Since the covariance is additive and the error ε is uncorrelated with y, the concentration index for health against income can be decomposed (Rao, 1969):[3]

$$\begin{aligned} C_{hy} &= \frac{2}{\mu_h}\left[\beta_y \mathrm{Cov}(y, F(y)) + \beta_z \mathrm{Cov}(z, F(y)) + \beta_x \mathrm{Cov}(x, F(y))\right] \\ &= \frac{\beta_y \mu_y}{\mu_h} C_{yy} + \frac{\beta_z \mu_z}{\mu_h} C_{zy} + \frac{\beta_x \mu_x}{\mu_h} C_{xy} \end{aligned} \tag{4}$$

[2] We discuss non linear health production functions in section 4 and grouped data in section 5.

[3] We do not include a term arising from the covariance of the error term with $F(y)$ in the decompositions of the concentration index since

$$\mathrm{Cov}\,(\varepsilon, F(y)) = \mathrm{Cov}\,(\mathrm{E}\varepsilon | F(y), F(y)) = \mathrm{Cov}\,(\mathrm{E}\varepsilon | y, F(y)) = \mathrm{Cov}\,(0, F(y) = 0\,.$$

We can always specify the health equation so that all variable affecting health and correlated with income are included specifically on the right hand side, rather than being wrapped up in the error term. The concentration indices of the residuals from estimates of the health or standardising equations are not zero and are discussed in section 4.

where μ_y, μ_z and μ_x are the population means of y, z, and x. C_{yy} is the concentration index of income against income C_{yy}, otherwise known as the Gini coefficient. C_{zy}, C_{xy} are the concentration indices of z, x against income.

The decomposition of the concentration index reveals the potential problem in using C_{hy} as a measure of income related inequality when other variables affect health ($\beta_z, \beta_x \neq 0$). If z, x are also correlated with income then the concentration indices of z, x against income (C_{zy}, C_{xy}) will be non zero and C_{hy} will reflect non income factors.

The way in which the non-income variables should be dealt with in an index designed to measure income related inequality depends on the view one has of the policy relevance of the variables. Some variables, such as age, sex, or ethnicity are not alterable by policy. Hence it seems appropriate to remove their influence from an index of income related inequality.[4] Call these variables *standardising* variables. The remaining variables are those which can be altered by policy. In what follows suppose that z is a standardising variable and x is a variable which can be altered by policy.

The distinction between the two types of variable depends on the circumstances. A variable may sometimes be regarded as standardising and sometimes as policy relevant. If we are using the index to examine the effect on income related inequality of improving the housing conditions of the poor, we should treat housing as a policy relevant variable and include the term $(\beta_x \mu_x / \mu_h) C_{xy}$ in the inequality measure. The policy is aimed at reducing income related inequality by reducing the correlation between income and housing quality and will affect the index via C_{xy}. Alternatively, if we are examining the effects of different methods of allocating health care resources to poor areas, then we can take the effect of housing on health and the correlation between income and housing as given, and treat housing conditions as a standardising variable.

[4] If inequalities in health between different ethnic groups or between men and women are of concern it seems appropriate to measure them directly. The relevant issue then would be whether say, men and women of a given age and income have the same health. Here we are concerned with whether individuals of given age and sex but different incomes have the same health. The health production function can provide a direct answer to both these questions. It can also be used to provide summary indices of both the extent of income related inequality, via the partial concentration index, and of the extent of inequality across sexes, via, for example, an Oaxaca decomposition (Oaxaca, 1973).

For example, whether policy can change the effect of education on health or on the relationship between education and income, may depend on the timescale considered, or whether one is constructing an inequality index to measure the effects of specific interventions which may or may not include educational policies.

To obtain a policy relevant measure of income related inequality the effects of the standardising variables must be removed from the overall concentration index. Only in cases in which there are no standardising variables, perhaps because we are examining income related inequality within in a highly specific population sub group defined by particular values of the standardising factors, will C_{hy} be a suitable summary measure of inequality.

The obvious way to measure income related inequality when there are standardising variables is to deduct the terms involving them from C_{hy}. If all non income variables are standardising (there are no non income policy relevant variables x in the health equation (2)) the inequality index is the *partial concentration index*

$$I_{hy} \equiv \frac{\beta_y \mu_y}{\mu_h} C_{yy} \qquad (5)$$

which is just the first term in (3). Since I_{hy} is the product of the elasticity of health with respect to income and the Gini, it reflects both the effect of income on health, holding all other factors constant and the extent of variation in income across the population. Plotting the two components of I_{hy} in (elasticity, Gini) space can yield useful figures for cross sectional comparisons (Dusheiko and Gravelle, 2001) or for showing the time path of inequality (Gravelle and Sutton, 2001).

If the standardising variable z was say age and had a negative effect on health ($\beta_z < 0$) and was positively correlated with income ($C_{zy} > 0$), then the partial concentration index I_{hy} will be greater than the concentration index of unstandardised health C_{hy}. Conversely if the rich are on average younger than the poor. We can illustrate the removal of the effects of the standardising variables from the concentration index of raw health using Figure 1. Suppose that only income and the standardising variable affect health. Then from (4) we have $C_{hy} = I_{hy} - C_{h^*y}$ where

$$C_{h^*y} = \frac{2}{\mu_h}\text{Cov}\,(\beta_0 + \beta_z z, F(y)) = \frac{2}{\mu_h}\text{Cov}\,(h^*, F(y))$$

and $h^* = \beta_0 + \beta_z z$ is the expected level of health after removing the effect of income. We can plot the concentration curve L^*(s) of h^* against income in Figure 1 and C_{h^*y} is twice the area between L^*(s) and the 45^o line. In the figure is it assumed either that the standardising variable is positively correlated with income and with health, or is negatively correlated with income and health. Hence L^*(s) lies below the 45^o line because the poor will have a disproportionately small share of h^*.

To get the true picture for income related inequality we need to remove the effects of the correlated standardising variable which in this case increase the concentration index of raw health. The partial concentration index $I_{hy} = C_{hy} - C_{h^*y}$ is the shaded area between the concentration curves $L^*(s)$ and $L(s)$. Since income is assumed to have a positive effect on health the partial concentration index is positive but it is less than the area between $L(s)$ and the 45^o line.

If there are other policy relevant variables x in addition to income, deduction of the direct influence of the standardising variable z from the decomposition of C_{hy} gives the *augmented partial concentration index*

$$I^A_{hy} = \frac{\beta_y \mu_y}{\mu_h} C_{yy} + \frac{\beta_x \mu_x}{\mu_h} C_{xy} \tag{6}$$

To bring out the factors underlying income related inequality we will often assume that the means of z and x conditional on y are linear in y. The assumption does not affect the substance of the results. Linearity of the conditional means implies

$$C_{xy} = \frac{2}{\mu_x}\text{Cov}(x, F(y)) = \frac{2}{\mu_x}\text{Cov}(E[x|y], F(y)) = \frac{2}{\mu_x} b_{xy}\text{Cov}(y, F(y)) = \frac{\mu_y}{\mu_x} b_{xy} C_{yy}$$

where b_{xy} is the coefficient from the linear regression of x on y. Substituting in the expression for the augmented partial concentration index gives

$$I^A_{hy} = \frac{\beta_y \mu_y}{\mu_h} C_{yy} + \frac{\beta_x \mu_x}{\mu_h}\left(\frac{\mu_y}{\mu_x} b_{xy} C_{yy}\right) = I_{hy}\left(1 + \frac{\beta_x}{\beta_y} b_{xy}\right) \tag{7}$$

3. Standardisation

3.1 Direct standardisation

There are two methods of direct standardisation for estimating the partial concentration indices. They differ in their treatment of the residuals from the estimated health equation but are asymptotically equivalent. We concentrate here on the one which is most immediately comparable with indirect standardisation. The first direct standardisation method for is to

(a) estimate the health production function (2) as

$$h = b_0 + b_y y + b_z z + b_x x + e \tag{8}$$

where e is the residual

(b) use the estimated coefficients b_z on the standardising variables to calculate expected health as though only the standardising variable affected health

$$h^b = b_0 + b_z z \tag{9}$$

(c) calculate the concentration index of h^b against income[5]

$$\hat{C}_{h^b y} = \frac{b_z \bar{z}}{\bar{h}^b} \hat{C}_{zy}\,, \tag{10}$$

(d) calculate the concentration index of unstandardised health against income $\hat{C}_{hy}$,

(e) multiply the concentration index of h^b by $\bar{h}^b / \bar{h}$ and subtract it from the concentration index of unstandardised health to get the directly standardised inequality index

$$I_{hy}^{D1} = \hat{C}_{hy} - \frac{\bar{h}^b}{\bar{h}} \hat{C}_{h^b y} \tag{11}$$

We have spelt out the procedure in some detail since it is, except for the first step, the same as the procedure for calculating inequality by indirect standardization. Equivalently, we could have reduced the last three steps to a single step by using the "convenient regression" procedure described in footnote 5. Running a single regression of $(h - h^D)2S_{FF} / \bar{h}$ on F gives I_{hy}^{D1} immediately.

[5] Since the concentration index of any variable w against income can be written as $2\mathrm{Cov}(w, F(y))/\bar{w}$, OLS regression of $w[2S_{FF}/\bar{w}]$ on F (where S_{FF} is the sum of squared deviations of F from its sample mean), yields a regression coefficient $\mathrm{Cov}(w[2S_{FF}/\bar{w}], F)/S_{FF} = 2\mathrm{Cov}(w, F)/\bar{w} = \hat{C}_{wy}$. See Kakwani et al (1997).

A third equivalent method of estimating directly standardised inequality is to estimate the health production function (8) as in step (a), calculate

$$h^D = h + b_z(z^o - z) = b_0 + b_y y + b_z z^o + b_x x + e \tag{12}$$

and regress $(h^D 2S_{FF} / \overline{h}$ on F to again give I_{hy}^{D1}. We will refer to h^D as directly standardized health. It is the value of health predicted for an individual with income y if we replace their actual standardizing characteristics z with a fixed value z^o. In the case of the linear production function considered so far the choice of z^o has no effect on I_{hy}^{D1}. We could choose it to be equal to average of z over all income groups in the population or indeed over some other reference population. (See the discussion of non-linear health production functions in section 4 and of grouped data in section 5.)

Provided a consistent estimator of the production function is used (and OLS is consistent under the assumptions made so far)

$$\begin{aligned}
\text{plim}\, I_{hy}^{D1} &= C_{hy} - \text{plim}\frac{\overline{h}^b}{\overline{h}}\hat{C}_{hy}^D = C_{hy} - \frac{\mu_{h^D}}{\mu_h}\frac{\beta_z \mu_z}{\mu_{h^D}} C_{zy} \\
&= \frac{\beta_y \mu_y}{\mu_h} C_{yy} = I_{hy} && \text{if } h = \beta_0 + \beta_y y + \beta_z z + \varepsilon \\
&= \frac{\beta_y \mu_y}{\mu_h} C_{yy} + \frac{\beta_x \mu_z}{\mu_h} C_{zy} = I_{hy}^A && \text{if } h = \beta_0 + \beta_y y + \beta_z z + \beta_z x + \varepsilon
\end{aligned}$$

Thus the directly standardised inequality index I_{hy}^{D1} is a consistent estimator of the partial concentration indices, whether we are interested in just the inequality generated by the direct effects of income on health, or are also concerned about its indirect effects via other variables x which affect health, are correlated with income, and amenable to policy.

The second procedure for estimating a directly standardised concentration index is to estimate the health production function (8) to yield the coefficient on income b_y, calculate the Gini coefficient $\hat{C}_{yy}$, mean health $\overline{h}$ and income $\overline{y}$ and so get

$$I_{hy}^{D2} = \frac{b_y \overline{y}}{\overline{h}} \hat{C}_{yy}$$

(If we want to estimate the augmented partial concentration index, we merely include the appropriate estimated coefficients, means and concentration indices on the additional policy relevant variables.) If the health production function is consistently estimated the second version of the direct standardisation procedure also produces a consistent estimate of the relevant partial concentration index. The two procedures for estimating the partial concentration index by direct standardisation are asymptotically equivalent. They differ for finite samples only because the first procedure leaves the residuals from the estimated health production function in the estimated inequality index.[6] We show in section 4, where we discuss the concentration index of residuals, that the probability limit of the difference between the two direct standardisation procedures is zero.

The first direct standardisation procedure is perhaps slightly easier to implement if one is interested in the augmented partial concentration index I_{hy}^{A}. The second has the merit of giving a decomposition of the partial concentration index I_{hy} as a product of the Gini coefficient and the income elasticity of health.

3.2 Indirect standardisation

Indirect standardisation has been suggested as a simple and convenient method of removing the confounding effects of health affecting policy irrelevant variables which are correlated with income (Wagstaff and van Doorslaer 2000a, 2000b; van Doorslaer and Koolman, 2000). The indirect standardisation procedure differs from the first procedure set out above for calculating I_{hy}^{D1} only in step (a): health is estimated by fitting a regression of health only on the standardizing variable *z*. This yields indirectly standardised health as

$$h^{N} = a_0 + a_z z \, . \tag{13}$$

The estimated concentration index for indirectly standardised health h^N is

---

[6] Using the additivity properties of the covariance: $I_{hy}^{D1} = \hat{C}_{hy} - \frac{\overline{h}^{b}}{\overline{h}}\hat{C}_{h^b y} = \frac{b_y \overline{y}}{\overline{h}}\hat{C}_{yy} + \hat{C}_{ey} = I_{hy}^{D2} + \hat{C}_{ey}$

where $\hat{C}_{ey} = 2\mathrm{Cov}(e, \hat{F}(y)) / \overline{h}$. See section 4.1.3 for a fuller discussion.

$$\hat{C}_{hy}^{N} = \frac{2}{\overline{h}^{N}}\mathrm{Cov}(h^{N}, F(y)) = \frac{2}{\overline{h}^{N}}\mathrm{Cov}(a_0 + a_z z, F(y)) = \frac{2a_z\overline{z}}{\overline{h}^{N}\overline{z}}\mathrm{Cov}(z, F(y))$$
$$= \frac{a_z\overline{z}}{\overline{h}^{N}}\hat{C}_{zy} \qquad (14)$$

and the indirectly standardised inequality index is (since $\overline{h} = \overline{h}^{N}$)

$$I_{hy}^{N} = \hat{C}_{hy} - \frac{\overline{h}^{N}}{\overline{h}}\hat{C}_{zy} = \hat{C}_{hy} - \hat{C}_{hy}^{N} = \hat{C}_{hy} - \frac{a_z\overline{z}}{\overline{h}^{N}}\hat{C}_{zy} \qquad (15)$$

To examine the properties of I_{hy}^{N}, consider first the case in which only income and the standardising variable affect health. Since the indirect standardising regression equation omits income:

$$\mathrm{plim}\, a_z = \beta_z + b_{yz}\beta_y$$

where b_{yz} is the regression coefficient of income on z. Hence the indirectly standardised inequality index has

$$\mathrm{plim}\, I_{hy}^{N} = \mathrm{plim}\left(\hat{C}_{hy} - \hat{C}_{hy}^{N}\right) = C_{hy} - \mathrm{plim}\, \hat{C}_{hy}^{N}$$
$$= \frac{\beta_y\mu_y}{\mu_h}C_{yy} + \frac{\beta_z\mu_z}{\mu_h}C_{zy} - \frac{(\beta_z + b_{yz}\beta_y)\mu_z}{\mu_h}C_{zy} \qquad (16)$$
$$= I_{hy} - \frac{b_{zy}\beta_y\mu_z}{\mu_h}C_{zy}$$

and I_{hy}^{N} is not a consistent estimator of I_{hy}.

Using the assumption of the linearity of the conditional mean of z gives

$$\mathrm{plim}\, I_{hy}^{N} = \frac{\beta_y\mu_y}{\mu_h}C_{yy} - \frac{\beta_y b_{yz}\mu_z}{\mu_h}\left(\frac{\mu_y}{\mu_z}b_{zy}C_{yy}\right) = I_{hy}\left(1 - b_{yz}b_{zy}\right) = I_{hy}\left(1 - r_{yz}^{2}\right) \qquad (17)$$

where r_{yz}^{2} is the squared correlation coefficient between income and the standardising variable.

We have allowed for only one standardising variable but we show in the appendix that the result in (17) carries over to the case of several standardising variables. The only difference is that the squared correlation coefficient between y and z is replaced by the coefficient of determination $R^{2}(y, \mathbf{z})$ from the multiple regression of income on the vector of standardising variables **z**.

The indirectly standardised inequality index tends to underestimate the partial concentration index I_{hy} irrespective of whether the standardising variable z has a positive or negative effect on health or whether the standardising variable is negatively or positively correlated with income.

A stark illustration of the potential problems with indirect standardisation is provided if the "standardising" variable z has no effect on health but is correlated with income.[7] (See section 5 for a numerical illustration with grouped data.) Since there are no confounding variables (and no policy relevant x variables) $C_{hy} = I_{hy}$ and there is no need to standardise. Since income is correlated with health and the standardising variable, C_{hy}^N is non zero and so $I_{hy}^N = C_{hy} - C_{hy}^N = I_{hy} - C_{hy}^N$ is clearly not a suitable estimate of I_{hy}. The example is extreme but it illustrates the difficulty that indirect standardisation by regression only on non-income variables tends to over correct for confounding. It removes both the direct effect of standardising variables and any indirect effect due to their correlation with income. But by removing the indirect influence of income via the standardising variables it also reduces the direct effect of income on health and hence tends to underestimate income related inequality.

Now suppose that health depends on income, the standardising variable z and a policy relevant variable x, so that we wish to measure the augmented partial concentration index I_{hy}^A. To estimate I_{hy}^N we again estimate the standardizing equation (12) and, because both income and x are omitted from the standardizing equation,

$$\text{plim}\, a_z = \beta_z + \beta_y b_{yz} + \beta_x b_{xz} \tag{18}$$

where b_{yz}, b_{xz} are the regression coefficients of y and x on z. Proceeding as before

$$\begin{aligned}\text{plim}\, I_{hy}^N &= C_{hy} - \text{plim}\, \hat{C}_{hy}^N \\ &= \frac{\beta_y \mu_y}{\mu_h} C_{yy} + \frac{\beta_z \mu_z}{\mu_h} C_{zy} + \frac{\beta_x \mu_x}{\mu_h} C_{xy} - \frac{(\beta_z + b_{yz}\beta_y + b_{xz}\beta_x)\mu_z}{\mu_h} C_{zy} \\ &= I_{hy}^A - \frac{(b_{yz}\beta_y + b_{xz}\beta_x)\mu_z}{\mu_h} C_{zy}\end{aligned} \tag{19}$$

Making use of the linearity of conditional means

---

[7] One might wish to compare age and sex specific income related health inequalities across different areas and standardise on ethnicity which might not have a direct effect on health but might be correlated with income.

$$C_{zy} = \frac{\mu_y}{\mu_z} b_{zy} C_{yy}, \tag{20}$$

where b_{zy} is the regression coefficient of z on y, we have[8]

$$\begin{aligned}
\text{plim}\, I_{hy}^N &= I_{hy}^A - b_{yz} b_{zy} \frac{\beta_y \mu_y}{\mu_h} C_{yy} - b_{xz} b_{zy} \frac{\beta_x}{\beta_y} \frac{\beta_y \mu_y}{\mu_h} C_{yy} \\
&= I_{hy}^A - I_{hy} \left[r_{yz}^2 + r_{yz}^2 \frac{\beta_x}{\beta_y} b_{xy} - r_{yz}^2 \frac{\beta_x}{\beta_y} b_{xy} + b_{xz} b_{zy} \frac{\beta_x}{\beta_y} \right] \\
&= I_{hy}^A \left(1 - r_{y1}^2\right) - I_{hy} \left[b_{xz} b_{zy} - r_{yz}^2 b_{xy} \right] \frac{\beta_x}{\beta_y} \qquad (21) \\
&= I_{hy}^A \left(1 - r_{y1}^2\right) - I_{hy} \left[b_{xz} - b_{yz} b_{xy} \right] \frac{\beta_x}{\beta_y} b_{zy} \\
&= I_{hy}^A \left(1 - r_{yz}^2\right) - I_{hy} \left(1 - r_{yz}^2\right) \frac{\beta_x}{\beta_y} b_{xz.yz} b_{zy}
\end{aligned}$$

where $b_{xz.yz}$ is the coefficient on z from a multiple regression of the omitted x variable on the variables y, z included in the health equation (2).

The indirectly standardised inequality index I_{hy}^N is an inconsistent estimator of the partial augmented concentration index I_{hy}^A. There are now two types of problem. The first arises from the correlation between the standardising variable and income and always leads to underestimates of inequality. The second arises if, in addition, the standardising variable is partially correlated at given income with the policy relevant variable x. The direction of the second effect depends on whether the correlations are of the same or of opposite signs. If they are of the same sign then the estimate of inequality is further depressed. The appendix shows that the same problems arise when both z and x are vectors.

---

[8] The last step uses follows from $b_{xz} - b_{xy} b_{yz} = (1 - r_{yz}^2) b_{xz.y}$. With S_{yz} etc denoting the sum of the products of the deviations of y and z etc from their means, $(1 - r_{yz}^2) = (S_{yy} S_{zz} - S_{zy}^2) / S_{yy} S_{zz}$ and $b_{xz.yz} = (S_{xz} S_{yy} - S_{xy} S_{yz}) / (S_{yy} S_{zz} - S_{zy}^2)$. Multiplying the two expressions and canceling gives $S_{xz} / S_{zz} - (S_{xy} / S_{yy})(S_{yz} / S_{zz}) = b_{xz} - b_{xy} b_{yz}$.

4. Omitted variables and non-linearities

4.1 Omitted variables

Data sets may lack information on variables influencing health. Consider two polar cases: omission of standardising variables and omission of policy relevant variables.

4.1.1 Omitted standardising variables

Assume that the health equation is

$$h = \beta_0 + \beta_y y + \beta_{z1} z_1 + \beta_{z2} z_2 + \beta_x x + \varepsilon \tag{22}$$

and that the standardising variable z_2 is not observed. We want to estimate the augmented partial concentration index I^A_{hy}. To do so by direct standardisation we estimate the health equation as $h = b_0 + b_y y + b_{z1} z_1 + b_x x + u$. The effect of omitted variable bias is

$$\operatorname{plim} b_y = \beta_y + \beta_{z2} b_{2y.y1x}, \quad \operatorname{plim} b_{z1} = \beta_{z1} + \beta_{z2} b_{21.y1x}, \quad \operatorname{plim} b_x = \beta_x + \beta_{z2} b_{2x.y1x} \tag{23}$$

where $b_{2y.y1x}$, for example, is the coefficient on y from the regression of z_2 on the included variables y, z_1 and x. And so[9]

$$\begin{aligned}\operatorname{plim} I^D_{hy} &= C_{hy} - \operatorname{plim}\left(\frac{b_{z1}\bar{z}}{\bar{h}}\hat{C}_{1y}\right) \\ &= I^A_{hy} + \frac{\beta_{z2}\mu_{z2}}{\mu_h} C_{2y} - \frac{\beta_2 b_{21.y1x}\mu_{z2}}{\mu_h} C_{1y} \\ &= I^A_{hy} + I_{hy}\frac{\beta_{z2}}{\beta_y}\left[b_{2y} - b_{21.y1x} b_{1y}\right] = I^A_{hy} + I_{hy}\frac{\beta_{z2}}{\beta_y}\left[b_{2y.y1x} + b_{2x.y1x} b_{xy}\right]\end{aligned} \tag{24}$$

The penultimate expression shows that deduction of the effects of the observed standardising variable z_1 from the overall concentration index fails to remove all the effects of the unobserved standardising variable z_2, even if the two standardising variables are correlated (conditional on the non standardising variables y and x included in the estimated health equation).

---

[9] Since the two procedures using direct standardisation are asymptotically equivalent we will use the symbol I^D_{hy} for both henceforth where no ambiguity results. The last step in (24) follows from $b_{2y} = b_{2y.y1x} + b_{21.y1x} b_{1y} + b_{2x.y1x} b_{xy}$

We could equivalently have calculated the directly standardised inequality index using the second procedure sketched in section 3, using from the estimates b_y, b_x to derive the last expression in (24). This formulation indicates that direct standardisation leads to inconsistency because the variables y and x contained in I_{hy}^D are partially correlated with the omitted standardising variable z_2 whose effects ought not to appear in I_{hy}^A.

The coefficient in the indirect standardising health equation, which omits y, x by design, and z_2 because of lack of data, has

$$\text{plim}\, a_{z1} = \beta_{z1} + \beta_y b_{y1} + \beta_{z2} b_{21} + \beta_x b_{x1} \tag{25}$$

where for example b_{21} is the coefficient from the bivariate regression of z_2 on z_1. The estimated indirectly standardised inequality index has

$$\begin{aligned}
\text{plim}\, I_{hy}^N &= C_{hy} - \text{plim}\, \hat{C}_{hy}^N = C_{hy} - \text{plim}\, \frac{a_{z1}\bar{z}_1}{\bar{h}} \hat{C}_{1y} \\
&= I_{hy}^A + \frac{(\beta_{z1} - \text{plim}\, a_{z1})\mu_{z1}}{\mu_h} C_{1y} + \frac{\beta_{z2}\mu_{z2}}{\mu_h} C_{2y} \\
&= I_{hy}^A - \frac{(\beta_y b_{y1} + \beta_{z2} b_{21} + \beta_x b_{x1})\mu_{z1}}{\mu_h} \frac{\mu_y}{\mu_{z1}} b_{1y} C_{yy} + \frac{\beta_{z2}\mu_{z2}}{\mu_h} \frac{\mu_y}{\mu_{z2}} b_{2y} C_{yy} \\
&= I_{hy}^A - I_{hy}\left[\left(b_{y1} b_{1y} + \frac{\beta_x}{\beta_y} b_{x1} b_{1y}\right) + \frac{\beta_{z2}}{\beta_y}\left(b_{2y} - b_{21} b_{1y}\right)\right]
\end{aligned} \tag{26}$$

Using the same procedure used to derive (21) gives

$$\text{plim}\, I_{hy}^N == I_{hy}^A\left(1 - r_{y1}^2\right) - I_{hy}\left(1 - r_{y1}^2\right)\frac{\beta_x}{\beta_y} b_{x1.y1} b_{1y} + I_{hy}\left(1 - r_{y1}^2\right)\frac{\beta_{z2}}{\beta_y} b_{2y.y1} \tag{27}$$

Compared to the case in which there was no omitted standardising variable, there is now a third problem. Indirect standardisation with an omitted standardising variable fails to directly remove the influence of the omitted variable from the inequality index. The problem is worse the greater the extent to which the omitted standardising variable is partially correlated with income, controlling for the other included standardising variable which has been deducted to derive I_{hy}^N.

4.1.2 *Omitted policy relevant variables*

Now suppose that the omitted variable is the policy relevant variable *x* rather than a standardising variable. We estimate the augmented partial concentration index by deducting the concentration index of standardised health from the estimated overall concentration index. The estimated health equation used for direct standardisation has $\operatorname{plim} b_z = \beta_z + b_{xz}\beta_x$ and so

$$\begin{aligned}\operatorname{plim} I_{hy}^{D} &= C_{hy} - \operatorname{plim}\frac{b_z \bar{z}}{\bar{h}}\hat{C}_{zy} = I_{hy}^{A} + \frac{(\beta_z - \operatorname{plim} b_z)\mu_z}{\mu_h} C_{zy} \\ &= I_{hy}^{A} - \frac{\beta_x}{\beta_y} b_{xz.yz} \frac{\beta_y \mu_y}{\mu_h} b_{zy} C_{yy} = I_{hy}^{A} - I_{hy}\frac{\beta_x}{\beta_y} b_{xz.yz} b_{zy}\end{aligned} \tag{28}$$

Direct standardisation is inaccurate to the extent that the omitted policy variable is partially correlated with the included standardising variable at given income.

The indirectly standardising health equation omits both the policy relevant variable (whether it is measurable or not) and income and so

$$\operatorname{plim} I_{hy}^{N} = I_{hy}^{A}\left(1 - r_{yz}^{2}\right) - I_{hy}\left(1 - r_{yz}^{2}\right)\frac{\beta_x}{\beta_y} b_{xz.y} b_{zy} \qquad \text{(21) (repeated)}$$

Indirect standardisation suffers both from the over-correction due to the correlation of the standardising variable with income and from omitted variable bias.

4.1.3 *Concentration index of residuals as an indicator of omitted variable bias*

Analyses which present a decomposition of the estimated inequality index $\hat{C}_{hy}$ as the weighted sum of the concentration indices of the variables affecting health sometimes include in the estimated health equation include a term

$$\hat{C}_{ey} = \frac{2\operatorname{Cov}(e, \hat{F}(y))}{\bar{h}} \tag{29}$$

showing the contribution of the concentration index of the residuals from the regression to overall inequality.[10] It is suggested (Wagstaff, van Doorslaer and Watanabe, 2000; van Doorslaer and Koolman, 2000) that the term shows the

[10] Multiplying and dividing by $\bar{e}$ would bring out the analogy between this expression for the proportionate contribution of the residuals to the overall estimated concentration index and the contribution of the regressors, but since the average residual $\bar{e}$ from OLS regression is always zero we omit this step. Wagstaff, van Doorslaer and Watanabe (2000) refer to the expression in the text as a generalised concentration index.

importance of factors which are correlated with income but which are not accounted for by the variables in the decomposition.

van Doorslaer and Koolman (2000) note that their estimates of $\hat{C}_{ey}$ are typically very small and statistically insignificant. They suggest that the result is not surprising since the health equation used to derive the decomposition contains an income term. In fact the implications of the fact that income is in the health equation are stronger and undermine the case for using $\hat{C}_{ey}$ as a diagnostic statistic.

The empirical covariance between the residual and any explanatory variable is precisely zero for every OLS regression. The concentration index of residuals is proportional to the empirical covariance of the residual and the cumulative frequency of income *F*(*y*) and is typically non-zero. But this does not mean that $\hat{C}_{ey}$ conveys any pertinent information about the decomposition of income related inequality. It is not useful either for showing how much of $\hat{C}_{hy}$ is unexplained nor for indicating possible misspecification due to omission of variables from the estimated health equation underlying the decomposition. If the true health equation is linear then the expected value of the empirical covariance of *e* and $\hat{F}(y)$ is zero, even if the estimated health equation omits variables affecting health and correlated with income.

The expected value of the empirical covariance between the residuals and the cumulative relative frequency of income is

$$\begin{aligned} \mathrm{E}\,\mathrm{Cov}\,(e,\hat{F}(y)) &= \mathrm{E}\,n^{-1}\sum e_i\hat{F}(y_i) = n^{-1}\sum \mathrm{E}\,e_i\hat{F}(y_i) = n^{-1}\sum \mathrm{Cov}\,(e_i,\hat{F}(y_i)) \\ &= n^{-1}\sum \mathrm{Cov}\left(\mathrm{E}\left[e_i \middle| \hat{F}(y_i)\right], \hat{F}(y_i)\right) = n^{-1}\sum \mathrm{Cov}\left(\mathrm{E}\left[e_i \middle| y_i\right], \hat{F}(y_i)\right) \end{aligned} \tag{30}$$

Suppose that the true health equation is

$$\dot{h}_i = \beta_y \dot{y}_i + \beta_z \dot{z}_i + \varepsilon_i \tag{31}$$

where dots above variables indicate that they are mean deviations. Suppose also that $\dot{z}$ is omitted from the estimated health equation. (As the Appendix shows the result is unaffected with more variables included in the true health equation and omitted from the estimated equation.) If the conditional expectation of the omitted variable is linear in income then $\mathrm{E}\left[\dot{z}_i \middle| y\right] = b_{zy}\dot{y}_i$ and

$$\mathrm{E}\left[b_y \middle| y\right] = \frac{\mathrm{E}S_{hy}}{S_{yy}} = \frac{\sum\left(\beta_y \dot{y}_i + \beta_z \mathrm{E}\left[\dot{z}_i \middle| y\right] + \mathrm{E}\left[\varepsilon_i \middle| y\right]\right)\dot{y}_i}{\sum \dot{y}_i^2} = \beta_y + \beta_z b_{zy} \tag{32}$$

So

$$\begin{aligned}\mathrm{E}\left[e_i \middle| y\right] &= \mathrm{E}\left[\left(h_i - \hat{h}_i\right)\middle| y\right] = \beta_y \dot{y}_i + \beta_z \mathrm{E}\left[\dot{z}_i \middle| y\right] + \mathrm{E}\left[\varepsilon_i \middle| y\right] - \mathrm{E}\left[b_y \middle| y\right]\dot{y}_i \\ &= \beta_y \dot{y}_i + \beta\ b_{zy}\dot{y}_i + 0 - \left(\beta_y + \beta_z b_{zy}\right)\dot{y}_i = 0\end{aligned} \tag{33}$$

Hence the expected value of the empirical covariance of the residuals and the cumulative relative frequency of income is zero. The concentration index of residuals $\hat{C}_{ey}$ thus has a probability limit of zero even if there are omitted variables, provided that the conditional mean of the omitted variables are linear in the included variables.

$\hat{C}_{ey}$ has limited appeal as a test for misspecification due to omitted variables: it can only detect omitted variables if they have conditional means which are non-linear in the included variables. If there are no omitted variables with non linear conditional means it is also useless as a measure of unexplained inequality since its probability limit (zero) is unaffected by how much inequality is accounted for by the variables in the decomposition. It is better to assess the overall fit and specification of the health equation underlying the decomposition with standard tests applied directly to the health equation rather than using the potentially misleading $\hat{C}_{ey}$.

4.2 Non-linear health equation

4.2.1 Inessential non-linearity

The assumption of linearity of the health equation (2) necessary for the decomposition (3) is less restrictive than it appears. Even though the true health equation may be non-linear it will often be possible to estimate it as a linear form by suitable transformations of the right or left hand side variables. Transformations involving the standardising variables (for example including power terms in age) have no effect on the interpretation of the partial concentration indices since the parts of the overall concentration index involving the transformed standardising variables are removed to derive the partial concentration index.

Income is generally thought to have a positive but decreasing effect on health and health equations frequently use powers of income or the log of income to allow for the concavity of health income relationship (Ettner, 1996; Backland, Sorlie and Johnson, 1996). Such transformations make no difference to the calculation of the partial concentration indices obtained by deducting the effects of the standardising variables. Thus if the health equation includes powers of income say to the cube and no other policy relevant variables

$$I_{hy} = \frac{\beta_y \mu_y}{\mu_h} C_{yy} + \frac{\beta_y \; \mu_{y^2}}{\mu_h} C_{y^2 y} \; \frac{\beta_{y^3} \mu_{y^3}}{\mu_h} C_{y^3 y} \tag{34}$$

I_{hy} still picks up all, and only, the effects of income on health and the distribution of income. Similarly, if income is entered in logarithmic form, the partial concentration index is

$$I_{hy} = \frac{\beta_y \mu_{\ln y}}{\mu_h} C_{(\ln y) y}$$

Some care may be called for in interpretation, since for example, the Ginis for income and for the log of income are not monotonically related. However, the decomposition allows for changes in income related inequality to be traced either to changes in the distribution of income or to the health production function.

Similarly, if the health variable is transformed to yield a linear estimating equation it is possible to estimate the partial concentration index of transformed health against income by applying the procedures outlined in section 2. Such transformations will usually only present problems for comparisons of inequality when the form of the health equation differs across the areas or time periods being compared, since then we are in effect comparing inequalities in different transformed variables.

4.2.2 Interaction effects

It is not implausible that the health equation should contain interaction terms to reflect the fact that say the effects of education or gender are different for high and low income groups. Such interactions are inconsequential when they concern only the standardising variables since they are removed to derive the partial concentration indices. If they concern policy relevant variables only then calculation of the augmented partial concentration is also unaffected, though the interpretation of the decomposition again requires care.

Interactions matter when the marginal effect of income (or other policy relevant variables) on health depends on the values of the standardising variables and it is not possible or not sensible to linearise the estimated equation by transforming the health variable. The concentration index of raw health against income C_{hy} now suffers from an additional problem. Income related inequality arises because incomes are not equal and income affects health. But if the effect of income on health depends on age or sex or other standardising variables, the amount of income related inequality measured by C_{hy} depends on the distribution of the standardising variables, even if the standardising variables are not correlated with income.

Suppose the health production function is $h = \beta_0 + \beta_y y + \beta_z z + \beta_{yz} yz + \varepsilon$. Deducting the direct effect of the standardising variable from the raw concentration index gives

$$\begin{aligned} C_{hy} - \frac{\beta_z \mu_z}{\mu_h} C_{zy} &= \frac{\beta_y \mu_y}{\mu_h} C_{yy} + \frac{\beta_{yz} \mu_{yz}}{\mu_h} C_{(yz)y} \\ &= I_{hy} + \frac{\beta_{yz}\left(\mu_z - b_{zy}\mu_y\right)\mu_y}{\mu_h} C_{yy} + \frac{\beta_{yz} b_{zy} \mu_{y^2}}{\mu_h} C_{y^2 y} \\ &= I_{hy}\left[1 + \frac{\beta_{yz}}{\beta_y}\left(\mu_z - b_{zy}\mu_y\right)\right] + \frac{\beta_{yz} b_{zy} \mu_{y^2}}{\mu_h} C_{y^2 y} \end{aligned} \tag{35}$$

since

$$\mathrm{Cov}\,(zy, F) = \mathrm{Cov}\,(yEz|y, F) = \mathrm{Cov}\,(y[\mu_z + b_{zy}(y - \mu_y)], F) \tag{36}$$

It is clear from the presence of the bivariate regression slope b_{zy} in (35) that deducting the direct effects of the standardising variable still leaves a measure of inequality contaminated by the correlation of the standardising variable with income and by its interaction with income. Note that even if the standardising variable was not correlated with income ($b_{zy} = 0$) (35) would still contain the interaction effect β_{yz}.

The direct standardisation approach is to calculate the concentration index for health with the standardising variables fixed across individuals. The procedure deals the problem of the contamination due to the correlation of the standardising variables and income, and although it does not solve the interaction problem completely, it makes it explicit and amenable to sensitivity analysis. Suppose that health equation is

$$h = h(\boldsymbol{\beta}; y, z, x, \varepsilon)$$

and consider

$$I_{hy}(z^o) \equiv \frac{\mu_{h^o}}{\mu_h} C_{h^o y} = \frac{\mu_{h^o}}{\mu_h} \frac{2\,\mathrm{Cov}\,(h(\boldsymbol{\beta}; y, z^o, x, \varepsilon), F)}{\mu_{h^o}} = \frac{2\,\mathrm{Cov}\,(h(\boldsymbol{\beta}; y, z^o, x, \varepsilon), F)}{\mu_h} \quad (37)$$

where $C_{h^o y}$ is the concentration index of $h^o = h(\beta; y, z^o, x, \varepsilon)$ for some arbitrary value of the standardizing variable and $\mu_{h^o} = \mathrm{E}h^o$. Since the standardising variable is held constant, h^o varies across individuals only because of differences in income and in unobserved factors uncorrelated with income. Hence $I_{hy}(z^o)$ is a measure of income related inequality in health which is not affected by the correlation of the confounding variables and income. The partial concentration index (5) defined in section 2 for the case of a linear health production function is a special case of (35): since the covariance is a linear operator it does not matter whether the confounding variable is fixed at zero or some value.

But in general the choice of z^o does affect the value of the inequality measure unless

$$\frac{\partial \mathrm{Cov}\,(h(\boldsymbol{\beta}; y, z^o, x, \varepsilon), F)}{\partial z^o} = \mathrm{E}\,\{[h_z(\beta; y, z^o, x, \varepsilon) - \mathrm{E}h_z(\beta; y, z^o, x, \varepsilon)](F - \mathrm{E}F)\} = 0$$

which is true if and only if the marginal effect of the standardizing variable h_z is independent of income, other policy relevant variables and the error. If, as in sections 2 and 3, the health production function is the sum of a function of the standardizing variables and a function of income, other policy relevant variables and the error, then the general version $I_{hy}(z^o)$ of the partial concentration index reduces to the measure defined in section 2.

Calculating the more general inequality measure for the case where $h = \beta_0 + \beta_y y + \beta_z z + \beta_{yz} yz + \varepsilon$ gives

$$I_{hy}(z^o) = \frac{\beta_y \mu_y}{\mu_h} C_{yy} + \frac{\beta_{yz} \mu_y z^o}{\mu_h} C_{yy} = I_{hy}\left(1 + \frac{\beta_{yz} \mu_y z^o}{\beta_y}\right) \quad (38)$$

which brings out clearly the dependence of the inequality measure on the fixed value of the standardizing variable when the health production function is not separable in the standardizing variable. The degree of income related inequality depends on the marginal effect of income on health since at any given income the expected marginal effect of an increase in income is $\beta_y + \beta_{yz} z^o$.

We can estimate $I_{hy}(z^o)$ by estimate the health production function as $h(\mathbf{b}; y, z, x, e)$ and then regressing $h(\mathbf{b}; y, z^o, x, e) 2S_{FF} / \overline{h}$ on F. The procedure is identical to the third method of calculating I_{hy}^{D1} outlined for the linear case in section 3.1. Provided that the production function is consistently estimated, this direct standardisation procedure yields a consistent estimate of $I_{hy}(z^o)$. The obvious choice for the fixed level of the standardising variable is its mean in some suitable reference population.

The indirect standardisation procedure is to regress health on the standardising variable and since income is omitted from the standardising regression

$$\text{plim}\, a_z = \beta_z + \beta_y b_{yz} + \beta_{yz} b_{(yz)z}$$

where $b_{(yz)z}$ is the coefficient from the linear regression of yz on z. Utilising (35) and (38) we get

$$\text{plim}\left[I_{hy}(\mu_z) - I_{hy}^{N}\right] = I_{hy}\left[\frac{\beta_{yz} b_{zy}}{\beta_y}\left(\mu_y + b_{(yz)z}\right) + r_{yz}^2\right] - \frac{\beta_{yz} b_{zy} \mu_{y^2}}{\mu_h} C_{y^2 y}$$

so that for large enough samples indirectly standardised inequality index may be larger or smaller than $I_{hy}(\mu_z)$ and the directly standardised index evaluated at the population mean of the standardising variable.

5. Inequality measurement with grouped data

Direct and indirect standardisation are epidemiological techniques developed to compare population health across areas, time periods or occupations (Armitage, 1971). The epidemiological techniques use grouped data: the only information on individuals is their which demographic sub group.

In the current instance the populations to be compared are income groups and subgroups may be defined by age, sex, education, etc. We wish to know how the average health of individuals in the different income groups compares after adjusting for the different mix of subgroups across income groups. In what follows it assumed that only income and standardising or policy irrelevant factors affect health. The arguments apply if there are policy relevant factors in addition to income: it is only necessary to define population sub-groups with respect to these factors as well.

5.1 Additive health model

To simplify the exposition restrict attention to two income groups, rich (r) and poor (p), and a single standardising variable, gender. The only information on individuals is to which of the four sub-groups they belong. Assume initially that income and gender have additive effects, so that the health of an individual is

$$H_{gs} = \beta_0 + \beta_y y_g + \beta_z z_s + \varepsilon \qquad (39)$$

where z_s is a dummy variable denoting sex with $z_s = 0, 1$ as s is m (male) or f (female). All individuals in the same income group g have the same income y_g.[11] The error term has zero mean. The expected health of subgroup gs is $h_{gs} = \mathrm{E}H_{gs}$. The proportion of individuals in income group g is d_g and the proportion of income group g of sex s is d_{gs}. The proportion of the total population who are of sex s is $d_s = d_r d_{rs} + d_p d_{ps}$ and the proportion of individuals of sex s who are income group g is $\delta_{sg} = d_g d_{gs}/d_s$. The average income of a person of sex s is $y_s = \delta_{sr} y_r + \delta_{sp} y_p$.

The expected health of an individual in income group g is

$$h_g = d_{gm} h_{gm} + d_{gf} h_{gf} = \beta_0 + \beta_y y_g + \beta_z d_{gf} \qquad (40)$$

and of a person of sex s is

$$h_s = \delta_{sr} h_{rs} + \delta_{sp} h_{ps} = \beta_0 + \beta_y y_s + \beta_z z_s \qquad (41)$$

where the average income of person of sex s is $y_s = \delta_{sr} y_r + \delta_{sp} y_p$.

Directly standardised health for income group g is the weighted average of the sex and income group specific healths, where the weights are the population shares of some reference population, not the shares in the income group. Directly standardized health is the expected health that a person in income group g would have if group g had the same demographic characteristics as the reference population. Suppose for the moment that the reference population is the population for whom the inequality index is to be calculated. Directly standardised health for income group g is

$$h_g^D = d_m h_{gm} + d_f h_{gf} = \beta_0 + \beta_y y_g + \beta_z d_f \qquad (42)$$

[11] To ease comparison with earlier sections we interpret the income variable as non-categorical but we could equivalently interpret y_g as a dummy variable with $y_p = 0$, $y_r = 0$ and rescale the parameter β_y.

Since directly standardised health for income group g does not depend on the demographic composition of group g, h_g^D can be cumulated against income to get a measure of income related inequality which is not contaminated by the correlation of the standardising variable (sex) and income. (See the third method of calculating I_{hy}^{D1} outlined in section 3.1.) The resulting concentration index is a consistent estimate of the partial concentration index I_{hy}.

h_{gs} can be estimated by averaging H_{gs} over the sample of individual from the subgroup. Or, to further bring out the analogy with the direct standardisation procedures applied to individual level data in section 3.1, we could estimate the health equation (39) using dummy variables for income group and gender (van Vliet and van de Ven, 1985).

Indirectly standardised health for income group g is the sum of population sex specific health weighted by the demographic proportions in income group g:

$$\begin{aligned} h_g^N &= d_{gm}h_m + d_{gf}h_f = d_{gm}\left(\beta_0 + \beta_y y_m\right) + d_{gf}\left(\beta_0 + \beta_y y_f + \beta_z\right) \\ &= \beta_0 + \beta_y\left(d_{gm}y_m + d_{gf}y_f\right) + \beta_z d_{gf} \end{aligned} \tag{43}$$

Subtracting h_g^N from unstandardised health in group g gives

$$\begin{aligned} h_g - h_g^{\mathbf{N}} &= \beta_0 + \beta_y y_g + \beta_z d_{gf} - \beta_0 + \beta_y\left(d_{gm}y_m + d_{gf}y_f\right) + \beta_z d_{gf} \\ &= \beta_y\left[y_g - \left(d_{gm}y_m + d_{gf}y_f\right)\right] \end{aligned} \tag{44}$$

The last term in (44) is independent of the demographic composition of income group g if and only if the demographic make up of the income groups is identical and the average incomes of the sexes are equal:

$$y_m = y_f \Leftrightarrow \frac{d_r d_{rm} y_r + d_p d_{pm} y_p}{d_r d_{rm} + d_p d_{pm}} = \frac{d_r d_{rf} y_r + d_p d_{pf} y_p}{d_r d_{rf} + d_p d_{pf}} \Leftrightarrow d_{rm} = d_{pm} \tag{45}$$

so that $y_m = y_g = \bar{y}$.

Hence constructing a concentration index of $h_g - h_g^N$ against income or equivalently subtracting the concentration index of indirectly standardised health h_g^N from the concentration index for unstandardised health h_g does not in general give a measure which is free of contamination from the correlation of income and demographic factors. For both grouped and individual level data indirect standardisation yields a measure of income related inequality which is inconsistent.

Expected sex specific health can be estimated by averaging health over all individuals of given sex in the sample or by regressing individual health on gender dummy variables. Again note the analogy with procedure labeled "indirect standardisation" in earlier sections.

5.2 Interaction effects

If there is an interaction between income and the standardising variable so that

$$H_{gs} = \beta_0 + \beta_y y_g + \beta_z z_s + \beta_{yz} y_g z_s + \varepsilon \tag{46}$$

directly standardised and indirectly standardised health are

$$h_g^D = d_m h_{gm} + d_f h_{gf} = \beta_0 + \beta_y y_g + \beta_z d_f + \beta_{yz} d_f y_g \tag{47}$$

$$h_g^N = d_{gm} h_m + d_{gf} h_f = \beta_0 + \beta_y \left(d_{gm} y_m + d_{gf} y_f\right) + \beta_z d_{gf} + \beta_{yz} d_{gf} y_f \tag{48}$$

Subtracting indirectly standardised health (48) from h_g does not purge the demographic effects across income groups unless all demographic groups have the same income and the problem is now worse than when there was no interaction term.

As (47) shows, and we noted in section 4.2.2, the choice of reference population affects the concentration index of directly standardized health. It is however possible to investigate the sensitivity of results to the choice of reference population and since direct standardization yields a consistent estimate of the partial concentration index when the standardizing variables have a purely additive effect it seems sensible to employ direct rather than indirect standardization.

Table 1 presents simple numerical examples of the use of the direct and indirect standardisation procedures for grouped data to calculate income related inequality.[12] The proportions of the population in the four income/gender groups are held constant whilst the income/gender specific healths vary across the seven examples. Although there are equal numbers of men and women, men are over-represented amongst the minority of the population who are rich. In column 1 income does not affect health, so

[12] Sutton (2001) uses a similar device though for slightly different purposes.

that there is no income related inequality. Since gender does affect health and is correlated with income, the concentration index of raw health is not zero. The indirectly standardised inequality index and the directly standardised inequality index are zero. In column 2 income affects health but gender does not. There is therefore no need to allow for any confounding effect of gender and inequality is measured equally well by the raw concentration index and the directly standardised inequality index. However, the indirectly standardised inequality index underestimates income related inequality because in attempting to correct for confounding by gender (needlessly in this case), it removes part of the effect of income which is correlated with gender.

In columns 3, 4, and 5 health is affected additively by both income and gender so that standardisation is required. Direct standardisation yields an accurate measure of inequality. In column 3 men are richer and healthier than women and in column 4 women are poorer and healthier. Despite the fact that the effect of gender on health is in the opposite direction in the two cases, the indirectly standardised inequality index under estimates income related inequality in both cases. In column 5 males are healthier than women but, somewhat unrealistically, the rich have worse health than the poor. There is now pro-poor inequality and the inequality indices are negative. But again indirect standardisation yields a smaller absolute inequality score.

In columns 6 and 7 income and gender interact. In such circumstances, as we noted in section 4.2.2, it is no longer necessarily the case that indirect standardisation yields a lower estimate of inequality than direct standardisation. In column 6 income and gender are reinforcing and the indirectly standardised inequality index is less than the directly standardised index. In column 7 they are offsetting and the indirectly standardised index is less than the directly standardised index.

6. Conclusion

When the marginal effects of standardising, policy irrelevant variables, are independent of the levels of income and other policy relevant variables, the partial concentration index and the augmented partial concentration index are intuitively appealing measures of measures of income related inequality. When the standardising variables are not additively separable from the other variables in the health production function, the partial concentration indices can be generalised in an obvious way by fixing the standardising variables at their mean value in some reference population. In both cases the resulting inequality index can be consistently estimated as the concentration index of directly standardised health.

We have shown that indirect standardisation leads to inconsistent estimates of income related inequality. The shortcomings of indirect standardisation for epidemiological work, as compared with direct standardisation, have been pointed out over many years (Freeman and Holford, 1980; Julious, Nicholl and George, 2001; Kilpatrick, 1959; Yule, 1934.) The criticisms have had little effect on epidemiological practice and, for example, use of the standardised mortality ratio which relies on indirect standardisation continues to be widespread.

Three defences can be made for indirect standardisation: lack of data, ease of computation, and problems with direct standardisation. Indirect standardisation does not require income and demographic specific health data: all that is required is average health by demographic group at population level and the demographic composition of the income groups. Direct standardisation by contrast requires the average health by demographic categories within each income group. Thus, on occasion, indirect standardisation may be all that is possible with grouped data.

It has been suggested (Kakwani et al, 1997; Wagstaff and van Doorslaer, 2000) that direct standardisation requires the use of grouped data. But sections 3 and 4 show that it is unnecessary to aggregate to group level to estimate a health production function for direct standardisation. With individual level data it is always possible to directly standardise.

It is has also been suggested that indirect standardisation is computationally easier when there is grouped data since population level demographic specific health can be estimated by a single regression with dummy variables for the demographic factors. Direct standardisation requires demographic specific health within each income group. But there is no reason why this cannot be done with a single regression with dummy variables for income groups as well as demographic factors.

Direct standardisation does not solve problems arising from omitted variables and is sensitive to the choice of reference population when the health production function is not additively separable in the standardizing variables. But indirect standardisation also suffers from omitted variable bias and, in addition, fails to properly remove the effect of confounding policy irrelevant variables in the estimation of income related inequality, irrespective of the form of the health production function. If the data permit, and they always do at individual level, direct standardisation is preferable to indirect standardisation in the estimation of income related inequality.

References

Armitage, P. (1971). *Statistical Methods in Medical Research*, Blackwell, Oxford.

Backlund, E., Sorlie, P. D. and Johnson, N. J. (1996). "The shape of the relationship between income and mortality in the United States: evidence from the national longitudinal study", *Annals of Epidemiology*, 6, 12-20.

Dusheiko, M. and Gravelle, H. (2001). "Measuring income related inequality in health at practice level", *Centre for Health Economics*, *Technical Paper* No. 22, University of York, August.

Ettner, S. L. (1996). " New evidence on the relationship between income and health", *Journal of Health Economics*, 15, 67-85.

Freeman, D. H. and Holford, T. R. (1980). "Summary rates", *Biometrics*, 36, 195-205

Gravelle, H. and Sutton, M. (2001). "Using the partial concentration index to examine trends in income related inequality in health: Britain 1979-95", August.

Julious, S. A., Nicholl, J. and George, S. (2001). "Why do we continue to use standardised mortality ratios for small area comparisons?", *Journal of Public Health Medicine*, 23, 40-46.

Kakwani, N., Wagstaff, A., and van Doorslaer, E. (1997). "Socioeconomic inequalities in health: measurement, computation, and statistical inference", *Journal of Econometrics*, 77, 87-103.

Kilpatrick, S. J. (1959). "Mortality comparisons in socio-economic groups", *Applied Statistics*, 12, 65-86.

Lambert, P. (1993). *The Distribution and Redistribution of Income*, 2nd Edition, Manchester University Press, Manchester.

Oaxaca, R. "Male-female wage differentials in urban labor markets", *International Economic Review*, 14, 693-703.

Propper, C. and Upward, R. (1992). "Need, equity and the NHS: the distribution of health care expenditure", *Fiscal Studies*, 13, 1-21.

Rao, V. (1969). "Two decompositions of the concentration ratio", *Journal of the Royal Statistical Society*, *Series A*, 132, 418-425.

Silcock, H. "The comparison of occupational mortality rates", *Population Studies* 1962; 16, 183-192.

Sutton, M. (2001). "Socio-economic inequities in health care use in Scotland: the importance of morbidity and need measurement", paper presented to the Tenth

European Workshop on Econometrics and Health Economics, London: Institute of Fiscal Studies, 5-7 September 2001

Urbanos-Garrido, R. M. (2001). "Measurement of inequity in the delivery of public health care: evidence from Spain (1997)", Universidad Complutense de Madrid, Documento de Trabajo 2001-15.

van Doorslaer, E. and Koolman, X. (2000). "Income related inequities in health in Europe: evidence from the European Community Household Panel", *Ecuity II Project Working Paper*, No. 1, March.

van Doorslaer, E., Wagstaff, A., Bleichrodt, H. et al. (1997). "Income related inequalities in health: some international comparisons", *Journal of Health Economics*, 16, 93-112.

van Doorslaer, E, Wagstaff, A., Calonge, S. et. al. (1992). "Equity in the delivery of health care: some international comparisons", *Journal of Health Economics*, 11, 389-411.

van Doorslaer, E., Wagstaff, A, van der Burgh, H. et al. (2000). "Equity in the delivery of health care in Europe and the US", *Journal of Health Economics*, 19, 553-594.

van Vliet, R. C. J. A. and van de Ven, W. P. M. M. (1985). "Differences in medical consumption between publicly and privately insured in the Netherlands: standardization by means of multiple regression", paper presented at the International Meeting on Health Econometrics of the Applied Econometric Association, Rotterdam, 16-17 December, 1985.

Wagstaff, A. and van Doorslaer, E. (2000a). "Equity in health care finance and delivery", in Culyer, A. J. and Newhouse, J. (eds.), *Handbook of Health Economics*, 1804-1862, Elsevier, Amsterdam.

Wagstaff, A. and van Doorslaer, E. (2000b). "Measuring and testing for inequity in the delivery of health care", *Journal of Human Resources*,

Wagstaff, A., van Doorslaer, E. and Watanabe, N. (2000). "On decomposing the causes of health sector inequalities with an application to malnuitrition inequalities in Vietnam", December.

Yule, G. U. (1934). "On some points relating to vital statistics, more especially statistics of occupational mortality", *Journal of the Royal Statistical Society*, 97, 1-84.

Appendix

Inconsistency of indirect standardization

The health equation is $h = \beta_0 + \beta_y y + \boldsymbol{\beta}_{\mathbf{z}}\mathbf{z} + \boldsymbol{\beta}_{\mathbf{x}}\mathbf{x} + \varepsilon$ and the estimated standardising equation is $h = a_0 + a_{\mathbf{z}}\mathbf{z}$, where $\mathbf{z}$ is K dimensional vector of standardising variables and $\mathbf{x}$ is a J dimensional vector of policy relevant variables. The OLS standardising equation estimates satisfy

$$\begin{bmatrix} Ea_1 \\ \vdots \\ Ea_K \end{bmatrix} = \begin{bmatrix} \beta_{z1} \\ \vdots \\ \beta_{zK} \end{bmatrix} + \begin{bmatrix} b_{y1.z} & b_{11.z} & \cdots & b_{J1.z} \\ \vdots & \vdots & \ddots & \vdots \\ b_{yK.z} & b_{1K.z} & \cdots & b_{JK.z} \end{bmatrix} \begin{bmatrix} \beta_y \\ \beta_{x1} \\ \vdots \\ \beta_{xJ} \end{bmatrix}$$
$$= \boldsymbol{\beta}_z + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} \quad \mathbf{x_1} \quad \cdots \quad \mathbf{x_J}) \begin{bmatrix} \beta_y \\ \boldsymbol{\beta}_{\mathbf{x}} \end{bmatrix} \tag{A1}$$

$b_{yk.z}$ is the partial regression coefficient on z_k from the multiple regression of y on all the included standardising variables. Similarly $b_{jk.z}$ is the regression coefficient of x_j on z_k from the regression of x_j on all the z. $\mathbf{Z}$ is an $n \times K$ matrix and $\mathbf{x_j}$ is $n \times 1$.

To reduce notation we henceforth interpret the variables y, z_k, x_j as deviations from their means. Following the procedure used in the text to derive (21) we get

$$\text{plim}\, I_{hy}^N = \frac{\beta_y \mu_y}{\mu_h} C_{yy} \left[1 - \sum_k b_{yk.z} b_{ky} + (\beta_y)^{-1} \left\{ \sum_j \beta_{xj} \left(b_{jy} - \sum_k b_{jk.z} b_{ky} \right) \right\} \right] \tag{A2}$$

where b_{ky}, b_{jy} are regression coefficients from the bivariate regressions of z_k, x_j on y. Now

$$\begin{aligned} \sum_k b_{yk.z} b_{ky} &= (b_{1y} \quad \cdots \quad b_{Ky}) \begin{bmatrix} b_{y1.z} \\ \vdots \\ b_{yk.z} \end{bmatrix} = (\mathbf{y}'\mathbf{y})^{-1}\mathbf{y}'(\mathbf{z_1} \quad \cdots \quad \mathbf{z_K}) \begin{bmatrix} b_{y1.z} \\ \vdots \\ b_{yk.z} \end{bmatrix} \\ &= (\mathbf{y}'\mathbf{y})^{-1}\mathbf{y}'\mathbf{Z}\mathbf{b}_{\mathbf{yz.z}} \\ &= (\mathbf{y}'\mathbf{y})^{-1}\mathbf{y}'\hat{\mathbf{y}} = (\mathbf{y}'\mathbf{y})^{-1}(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{e}) = (\mathbf{y}'\mathbf{y})^{-1}(\mathbf{y}'\mathbf{y} - \mathbf{e}'\mathbf{e}) \\ &= R^2(y, \mathbf{z}) \end{aligned} \tag{A3}$$

since $\mathbf{Z}\mathbf{b}_{\mathbf{yz.z}}$ yields the predicted $\hat{\mathbf{y}}$ from a regression of $\mathbf{y}$ on $\mathbf{Z}$ and all the variables are measured as deviations from their mean.

Similarly

$$
\begin{aligned}
b_{jy} - \sum_k b_{jk.z} b_{ky} &= (\mathbf{y'y})^{-1}\mathbf{y'x_j} - \left((\mathbf{y'y})^{-1}\mathbf{y'Z}\right)\mathbf{b_{jz.z}} \\
&= (\mathbf{y'y})^{-1}\mathbf{y'x_j} - \left((\mathbf{y'y})^{-1}\mathbf{y'Z}\right)\left((\mathbf{Z'Z})^{-1}\mathbf{Z'x_j}\right) \\
&= (\mathbf{y'y})^{-1}\mathbf{y'}\left(\mathbf{I} - \mathbf{Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'}\right)\mathbf{x_j} \\
&= (\mathbf{y'y})^{-1}\left(\mathbf{y'Mx_j}\right)
\end{aligned}
\tag{A4}
$$

Consider the multiple regression of x_j on y,**z**. We can obtain the coefficient on y by first regressing x_j on **z** and then regressing the residuals on the residuals from a regression of y on **z**. $\mathbf{M} = \mathbf{I} - \mathbf{Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'}$ is the symmetric idempotent ($M'M = M$) residual maker matrix for OLS regressions. Hence the coefficient of x_j on y is

$$
b_{jy.z} = \left((\mathbf{My})'\mathbf{My}\right)^{-1}(\mathbf{My})'\mathbf{Mx_j} = (\mathbf{y'M'My})^{-1}\mathbf{y'M'Mx_j} = (\mathbf{y'My})^{-1}\left(\mathbf{y'Mx_j}\right) \tag{A5}
$$

Thus

$$
\begin{aligned}
b_{jy} - \sum_k b_{jk.z} b_{ky} &= (\mathbf{y'y})^{-1}\left(\mathbf{y'Mx_j}\right) = (\mathbf{y'y})^{-1}(\mathbf{y'My})b_{jy.z} = (\mathbf{y'y})^{-1}(\mathbf{y'M'My})b_{jy.z} \\
&= (\mathbf{y'y})^{-1}\left((\mathbf{yM})'\mathbf{My}\right)b_{jy.z} = (\mathbf{y'y})^{-1}(\mathbf{e'e})b_{jy.z} \\
&= \left[1 - R^2(y,\mathbf{z})\right]b_{jy.z}
\end{aligned}
\tag{A6}
$$

Using (A3), (A6) in (A2) we have

$$
\begin{aligned}
\operatorname{plim} I_{hy}^N &= \frac{\beta_y \mu_y}{\mu_h} C_{yy}\left(1 + (\beta_y)^{-1}\left\{\sum_j \beta_{xj} b_{jy.x}\right\}\right)\left[1 - R^2(y,\mathbf{z})\right] \\
&= I_{hy}\left(1 + (\beta_y)^{-1}\left\{\sum_j \beta_{xj} b_{jy.x}\right\}\right)\left[1 - R^2(y,\mathbf{z})\right] \\
&= I_{hy}^A\left[1 - R^2(y,\mathbf{z})\right]
\end{aligned}
\tag{A7}
$$

Even if all variables are observed indirect standardisation yields an inconsistent estimate of the partial concentration indices. If all non-income variables $(\mathbf{z},\mathbf{x})$ are standardising so that I_{hy} is the inequality measure to be estimated but the standardising regression omits $\mathbf{x}$, the penultimate line applies. If the $\mathbf{x}$ variables are policy relevant, so that I_{hy}^A is the measure to be estimated and the standardising regression is run on $\mathbf{z}$, then the last line is applies.

Concentration index of residuals

To economise on notation let the health equation be $h = \boldsymbol{\beta}_\mathbf{x}\mathbf{x} + \boldsymbol{\beta}_\mathbf{z}\mathbf{w} + \varepsilon$ where $\mathbf{x}$ = $(y, x_2, \ldots, x_K)$ are the variables included in the estimated equation, $\mathbf{w} = (w_1, \ldots, w_T)$ are the variables excluded and $x_2,\ldots,x_K$, $\mathbf{w}$ may be standardizing or policy relevant variables. All variables are in mean deviation form. The errors are uncorrelated with $\mathbf{x}$ and $\mathbf{w}$ and we assume that the conditional means of the omitted variables are linear functions of the included variables: $\mathrm{E}[\mathbf{W}|\mathbf{X}] = \mathbf{X}\mathbf{b}_{\mathbf{wx}}$ where $\mathbf{b}_{\mathbf{wx}}$ is a $(K \times T)$ matrix of the regression coefficients of $\mathbf{w}$ on $\mathbf{x}$.

Since the estimated concentration index of residuals is twice the covariance of the residuals e_i from the estimated health equation with the cumulative frequency of income $\hat{F}(x_{1i}) = \hat{F}(y_i)$ we are interested in

$$\begin{aligned} \mathrm{E}\,\mathrm{Cov}\,(e,\hat{F}(x_1)) &= n^{-1}\sum Ee_i\hat{F}(x_{1i}) = n^{-1}\sum \mathrm{Cov}\,(e_i,\hat{F}(x_{1i})) \\ &= n^{-1}\sum \mathrm{Cov}\left(\mathrm{E}\left[e_i\middle|\hat{F}(x_{1i})\right], \hat{F}(x_{1i})\right) = n^{-1}\sum \mathrm{Cov}\left(\mathrm{E}\left[e_i|x_{1i}\right], \hat{F}(x_{1i})\right) \end{aligned} \tag{A8}$$

Using the residual maker matrix $\mathbf{M} = (\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'})$ and remembering that $\mathbf{MX} = \mathbf{0}$

$$\begin{aligned} \mathrm{E}[\mathbf{e}|\mathbf{X}] &= \mathrm{E}[\mathbf{Mh}|\mathbf{X}] = \mathrm{E}[\mathbf{M}(\mathbf{X}\boldsymbol{\beta}_\mathbf{x} + \mathbf{W}\boldsymbol{\beta}_\mathbf{w} + \boldsymbol{\varepsilon}|\mathbf{X}] \\ &= \mathrm{E}[(\mathbf{MX})\boldsymbol{\beta}_\mathbf{x}|\mathbf{X}] + \mathrm{E}[(\mathbf{MW})\boldsymbol{\beta}_\mathbf{w}|\mathbf{X}] = \mathbf{M}\mathrm{E}[\mathbf{W}|\mathbf{X}]\boldsymbol{\beta}_\mathbf{w} \\ &= (\mathbf{MX})\mathbf{b}_{\mathbf{wx}}\boldsymbol{\beta}_\mathbf{w} = \mathbf{0} \end{aligned} \tag{A9}$$

so that $\mathrm{E}[e_i|\mathbf{x_1}] = \mathrm{E}[e_i|\mathbf{y}] = 0$ and $\mathrm{E}[e_i|y_i] = 0$. Hence $\mathrm{E}\,\mathrm{Cov}(e,\hat{F}(y))$ and $\mathrm{plim}\,\hat{C}_{ey} = \mathrm{plim}\left(\mathrm{Cov}\,(e,\hat{F}(y)/\bar{h}\right) = 0$

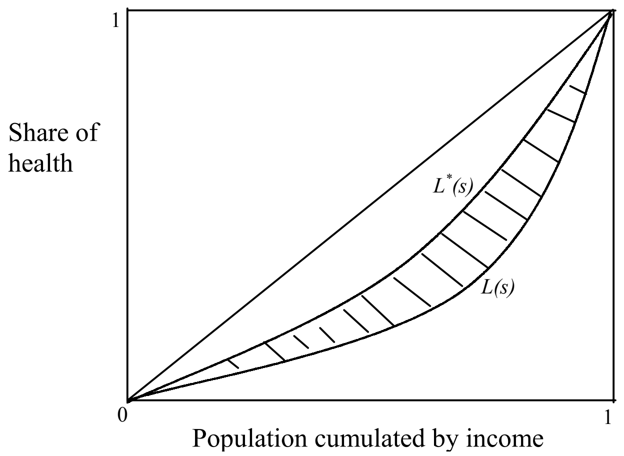

Figure 1. Concentration curves for raw health ($L(s)$) and for health with effect of income removed ($L^*(s)$). The partial concentration index I_{hy} is twice the shaded area and indicates pro-rich inequality if $L^*(s)$ lies above $L(s)$.

Table 1. Directly and indirectly standardized income related inequality

	Population proportion	*Gender and income specific health*						
		No inequality	Gender does not affect health	Additive: males healthier	Additive: females healthier	Additive: poor & males healthier	Gender & income reinforcing	Gender & income offsetting
Strata		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rich men	0.20	0.90	0.90	0.90	0.45	0.45	0.90	0.90
Poor men	0.30	0.90	0.25	0.60	0.15	0.70	0.60	0.60
Rich women	0.05	0.25	0.90	0.50	0.90	0.30	0.50	0.65
Poor women	0.45	0.25	0.25	0.20	0.60	0.55	0.15	0.55
Concentration indices								
Concentration index: raw health C_{hy}		0.085	0.295	0.182	0.050	-0.109	0.203	0.082
Concentration index: indirectly standardised health C_{hy}^{N}		0.085	0.035	0.077	-0.060	-0.018	0.089	0.019
Indirectly standardised inequality: $I_{hy}^{N} = C_{hy} - C_{hy}^{N}$		0.000	0.260	0.104	0.110	-0.091	0.114	0.063
Directly standardised inequality index: I_{hy}^{D}		0.000	0.295	0.118	0.125	-0.095	0.134	0.060