The Importance of Individual Heterogeneity in the Decomposition of Measures of Socioeconomic Inequality in Health: An Approach Based on Quantile Regression

by

Andrew M. Jones*[a] and Angel López Nicolás[b]

June 2002

[a] Department of Economics and Related Studies, University of York, York YO10 5DD, UK

[b] Departament d'Economia i Empresa and CRES, Universitat Pompeu Fabra, 08005-Barcelona, Spain

Abstract

This paper shows how recently developed regression-based methods for the decomposition of health inequality can be extended to incorporate individual heterogeneity in the responses of health to the explanatory variables. We illustrate our method with an application to the Canadian NPHS of 1994. Our strategy for the estimation of heterogeneous responses is based on the quantile regression model. The results suggest that there is an important degree of heterogeneity in the association of health to explanatory variables which, in turn, accounts for a substantial percentage of inequality in observed health. A particularly interesting finding is that the marginal response of health to income is zero for healthy individuals but positive and significant for unhealthy individuals. The heterogeneity in the income response reduces both overall health inequality and income related health inequality.

JEL classification: D63, I12, C21

Keywords: Health inequalities; Unobserved heterogeneity; Quantile regression

* Corresponding author. Tel. +44 1904 433790. Fax: +44 1904 433759. E-mail: amj1@york.ac.uk.
This paper derives from the project "Economic determinants of the distribution of health and health care in Europe" (known as the ECuity II Project), which is funded in part by the European Community's Biomed II programme (contract BMH4-CT98-3352). We are grateful to the EC and to the Spanish Ministry of Education (grant PR2001-0295) and the Spanish DGES (project PB98-1058-C03-01) for financial support and to Statistics Canada for access to the NPHS data. We thank Eddy van Doorslaer for support and advice.

Introduction

Health economists have adopted the Gini coefficient and concentration indices to provide summary measures of inequalities of health within populations (see e.g. Wagstaff et al, 1989, 1991, 1994, and van Doorslaer et al. 1997). A recent contribution by Wagstaff et al. (2002) has shown how a linear regression approach can be used to decompose these indices into the contributions of different explanatory variables. The decomposition treats individual responses to these explanatory variables (the slope coefficients) as homogeneous across individuals. In this paper we show how the decomposition can be expanded to allow for individual heterogeneity and we illustrate the method with an application to the measurement of health inequality using the Canadian National Population Survey of 1994. This survey has been used recently by van Doorslaer and Jones (2002) for research that requires the application of Wagstaff et al. (2002) methodology and therefore provides a benchmark for our results. We find that the heterogeneity of individual responses accounts for 51% of the observed Gini coefficient and 18% of the observed concentration for health.

We allow for heterogeneity in individual responses by means of a method based on quantile regression. This technique is gradually becoming a standard econometric procedure in situations where estimation of the conditional mean function is not enough to capture the full pattern of associations between the dependent variable and the covariates over the distribution of the former (see e.g. Koenker and Basset, 1978 and Buchinsky, 1994). Recent work has used quantile regression in order to estimate a model of heterogeneous returns to schooling. Arias et al. (2001) argue that quantile regression allows more flexibility than a random coefficients model in these circumstances. The technique can also retrieve causal effects, as shown in the work of Abadie et al. (2002), who use quantile regression to capture the heterogeneous pattern of treatment effects of a youth training program. Despite its attractions, however, there are not many applications of quantile regression in health economics with the exception of Manning et al. (1995) and Abrevaya (2001). The latter provides evidence that illustrates the relevance of quantile regression in the context of our analysis of health inequality. Its object of study is the relationship of a health outcome (birthweight) with a series of demographic variables. As in Abrevaya's work, we find that the effect of explanatory variables varies systematically over the distribution of health. In particular, our analysis shows that income has a positive and significant marginal effect for individuals in the bottom of the health distribution and a zero marginal effect on healthy individuals.

The structure of the paper is the following. In section 2 we show how the decomposition of the Gini index and the concentration indices into the contributions of different explanatory variables in a regression model can be modified to incorporate individual heterogeneity in all the coefficients. In section 3 we illustrate how the quantile regression model allows the estimation of heterogeneous responses. Section 4 discusses the main features of the data set used in this study. Section 5 presents and discusses the estimates from the quantile regression model and section 6 reports the decomposition of the inequality measures using these estimates. Section 7 concludes.

2. Regression based decompositions of inequality

The departure point for our methodology is the decomposition of inequality measures into the contributions of different explanatory variables by means of a linear regression model (see e.g., Wagstaff et al., 2002). Suppose we are interested in calculating the Gini coefficient for a measure of health using individual data in a sample from the population of interest. Let y_i denote a measure of health for the i^{th} individual and R_i denote the cumulative proportion of the population ranked by y_i up to the i^{th} individual (the 'relative rank'). Ignoring, for expositional purposes, the fact that in general sampling weights will be necessary, the Gini coefficient, G, for health is given by (see e.g., Lambert, 1994 p.43, van Doorslaer and Koolman, 2000),

$$G = \left(\frac{2}{\bar{y}} \right) \mathrm{cov}(y_i , R_i) \tag{1}$$

Now let y_i be given by the following linear regression model

$$y_i = \beta_1 + \sum_{k=2}^{K} \beta_k x_{ki} + \varepsilon_i \tag{2}$$

By substituting this for y_i, the Gini index of y can be written as (see Wagstaff et al., 2002),

$$G = \sum_{k=2}^{K} \left(\beta_k \frac{\bar{x}_k}{\bar{y}} \right) C_k + \left(\frac{2}{\bar{y}} \right) \mathrm{cov}(\varepsilon_i , R_i) \tag{3}$$

where the first term in brackets is the elasticity of y with respect to x_k evaluated at the mean of the sample, and C_k is the concentration index of x_k on y. The latter expression can be easily modified to obtain the concentration index of y against another variable of interest. For instance, the concentration index, CI, of health against income would be computed according to

$$CI = \sum_{k=2}^{K} \left(\beta_k \frac{\bar{x}_k}{\bar{y}} \right) C'_k + \left(\frac{2}{\bar{y}} \right) \mathrm{cov}(\varepsilon_i , R'_i) \tag{4}$$

where C'_k denotes the concentration index of x_k against income and R'_i is the cumulative proportion of the population ranked by income up to the ith individual.

Thus these inequality measures can be decomposed into an “explained part” and an “unexplained part” (see Wagstaff et al., 2002). The “explained” part can be usefully broken down into the contributions of individual explanatory variables. As for the “unexplained” part, it is a scaled measure of the covariance of the residuals in the regression model with the position of the individual in the distribution of the variable of interest. As such, the unexplained part should be zero if the regression model for the measure of health is specified in a way such that there is no systematic variation in unobserved heterogeneity in health according to the position of the individual in the distribution of the relevant variable.

However, the pervasive presence of unobserved heterogeneity in econometric models for cross sectional data, as reflected by low coefficients of determination, would lead to the suspicion that the unexplained part in these regression based inequality measures might be non-trivial. For example, Heckman (2001) argues that;

"..the most important discovery was the evidence on the pervasiveness of heterogeneity and diversity in economic behaviour…not only were intercepts variable but so were slopes.."

Indeed, recent work by Van Doorslaer and Jones (2002), using the Canadian National Population Health Survey of 1994, shows that while a regression model for health explains up to a 96% of the concentration index, only 48% of total inequality in health, as measured by the Gini index, can be explained by the same model.

We now propose a method that deals with unobserved heterogeneity, while retaining the useful summary information provided by the regression approach. A very general way to allow for individual heterogeneity is by means of a regression model for the health variable with heterogeneous parameters. Thus, the regression model can be modified to yield,

$$y_i = \beta_{i0} + \sum_{k=2}^{K} \beta_{ik} x_{ki} + u_i = \beta_{i1} + \sum_{k=2}^{K} \beta_{ik} x_{ki} = X_i^{'} \beta_i \tag{5}$$

where all the parameters in the model are individual specific. Note in particular that the intercepts in this model, $\beta_{i1,}$ comprise both unobserved systematic individual effects and unsystematic pure random errors. If we substitute equation (5) into (1) we obtain the following expression for the Gini index (see the Appendix for a full derivation),

$$G=\sum_{k=2}^{K}\beta_k^{OLS}\,\frac{\bar{x}_k}{\bar{y}}C_k+\left(\frac{2}{\bar{y}}\right)\sum_{k=2}^{K}\sum_{i}x_{ik}\left(\beta_{ik}-\beta_k^{OLS}\right)\left(R_i-1/2\right)+\left(\frac{2}{\bar{y}}\right)\mathrm{cov}\left(\beta_{i1},R_i\right)$$

(6)

The first term of this equation is exactly the same as the first term in equation (3) when model (2) is estimated by OLS. The residual term in equation (3) is now split into two components given by the second and third terms in equation (6). The second term is the contribution to overall inequality of the covariance (weighted by the values of x_k) of the slope parameters with the health rank. The third term is simply the covariance of the intercepts (centered at the OLS intercept coefficient) with the health rank.

Similarly, the concentration index for health can be written as,

$$CI=\sum_{k=2}^{K}\beta_k^{OLS}\,\frac{\bar{x}_k}{\bar{y}}C_k^{'}+\left(\frac{2}{\bar{y}}\right)\sum_{k=2}^{K}\sum_{i}x_{ik}\left(\beta_{ik}-\beta_k^{OLS}\right)\left(R_i^{'}-1/2\right)+\left(\frac{2}{\bar{y}}\right)\mathrm{cov}\left(\beta_{i1},R_i^{'}\right)$$

(7)

Each component has a similar interpretation to the Gini coefficient, with health rank, R, replaced by income rank, R’. The first term is identical to the first term in (4) and the second two terms decompose the generalised concentration index of the residual, allowing for heterogeneity.

3. Identification and quantile regression

The decompositions introduced in the previous section rely on individual specific β_i’s. To apply these in practice requires an estimator that allows for heterogeneous response. The approach we use is based on quantile regression.

Let θ_i denote the position that the i^th individual occupies in the distribution of health conditional on X_i. That is, if F(Y|X) is the CDF of the distribution of health conditional on observed characteristics,

$$\theta_i=F(y_i\mid X=X_i)$$

(8)

It follows that,

$$y_i = Q_{\theta_i}\left(Y \mid X = X_i\right)$$

(9)

Where $Q_\theta(Y|X)$ denotes the θ^{th} quantile of Y conditional on X. We now make the following identifying assumption,

$$\begin{aligned}
&\forall i, j, \quad if\ \theta_i = \theta_j = \theta, \quad then \\
&\exists Q_\theta(Y \mid X) \quad such \quad that \\
&Q_\theta(Y \mid X = X_i) = y_i\ \ and\ Q_\theta(Y \mid X = X_j) = y_j
\end{aligned}$$

(10)

In particular, if we assume the conditional quantile functions to be a linear combination of the regressors, the vector β_i is identified by the coefficients of the θ_i^{th} conditional quantile of function. That is,

$$\begin{aligned}
&\forall i, j, \quad if\ \theta_i = \theta_j = \theta, \quad and \quad Q_\theta\left(Y \mid X\right) = X^{'}\beta^\theta \quad then \\
&y_i = Q_{\theta_i}(Y \mid X = X_i) = X_i^{'}\beta^\theta = X_i^{'}\beta_i \\
&y_j = Q_{\theta_i}(Y \mid X = X_j) = X_j^{'}\beta^\theta = X_j^{'}\beta_j
\end{aligned}$$

(11)

It is important to note that, when making this assumption, we are interpreting the intercept terms as systematic unobserved heterogeneity. This could be problematic in the presence of pure random noise but in a cross section it is not possible to separate one from the other. Our approach is the polar case with respect to OLS, where the totality of the error term is assumed to be unsystematic.

In order to estimate the conditional quantile functions, note that for any θ we may define the θ^{th} quantile residual for the i^{th} individual as,

$$y_i - X_i^{'}\beta^{\theta} = \omega_{\theta i}$$

(12)

and search for the values of β^{θ} that minimise some criterion function of these residuals. In particular Koenker and Basset (1978) show that β^{θ} can be estimated consistently by the following algorithm based on the Least Absolute Deviation criterion,

$$\hat{\beta}^{\theta} = \arg min_{\beta} \sum_{i} \left|\omega_i^{\theta}\right| \left[l(\omega_i^{\theta} > 0)\theta + \left(1 - l(\omega_i^{\theta} > 0)\right)(1-\theta) \right]$$

(13)

Thus in theory we could estimate our model of heterogeneous parameters by first computing θ_i for each individual and subsequently estimate the conditional quantile function for each one of the different values of θ obtained in the first step. As this would require constructing cells for each set of unique values of the conditioning variables and choosing an arbitrary level of accuracy for θ, our practical strategy consists in first choosing randomly a large number of values for θ from the (0,1) interval. For each of these values we estimate the parameters of the conditional quantile function and assign to each individual the coefficients of the quantile function that, given his or her characteristics, minimises the absolute difference between the observed value of health and the predicted value of health. This criterion can be stated formally as,

$$\hat{\beta}_i = \hat{\beta}^{\theta}, \quad \theta = \arg min_{\theta} \left| y_i - X_i^{'}\hat{\beta}^{\theta} \right|$$

(14)

Therefore our estimated model of heterogeneous parameters can be written as,

$$y_i = X_i^{'}\hat{\beta}_i + \xi_i$$

(15)

where ξ_i is an estimation residual.

4. Data

The data used in this paper are taken from the first wave (in 1994-1995) of the Canadian *National Population Health Survey (NPHS).* The target population of the *NPHS* includes household residents in all provinces, with the exclusion of populations on Indian Reserves, Canadian Forces Bases and some remote areas of Ontario and Quebec. A total of 26,430 households were selected for the survey. In each household, a randomly selected household member, aged 12 years or older, was selected for a more in-depth interview. This interview included questions on health status, risk factors, and demographic and socio-economic information. The data were weighted using the survey weights to adjust for the complex multi-cluster sample design of the NPHS. Detailed information about the *NPHS* content and sample design has been published elsewhere (e.g. Tambay and Catlin, 1995) and the sample has been used in previous analyses of inequality in health by Humphries and van Doorslaer (2000) and van Doorslaer and Jones (2002).

A particular attraction of the NPHS is that it contains a continuous measure of health status that is suitable for regression and decomposition analysis. This is the McMaster Health Utility Index (HUI). Each respondent was assigned a HUI score based on their response to the questions of the eight-attribute Health Utility Index Mark III health status classification system. The HUI is a generic health status index, developed at McMaster University, that measures both quantitative and qualitative aspects of health (Feeny *et al*. 1995; Torrance *et al*. 1995, 1996). It provides a description of an individual's overall functional health, based on eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain. The HUI assigns a single numerical value, between zero and one, for all possible combinations of levels of these eight self-reported health attributes. A score of one indicates perfect health.

Total income before taxes and deductions is measured in the NPHS as a categorical variable with 11 response categories. The two lowest income groups- no income and less than $5,000- were combined into one group, thus reducing the number of income categories from 11 to 10. The midpoint of each income category was then attributed to all households in that category and subsequently divided by an equivalence factor equal to (number of household members)$^{0.5}$, to adjust for differences in household size. The income values assigned for the top and bottom groups were Can$2,500 and Can$87,500.00 respectively.

Other health determinants included in the analysis are the following. (i) Education level, the highest level of general or higher education completed is available at three levels: recognised third level education (ISCED 5-7), second stage of secondary level of education (ISCED 3) and less than second stage of secondary education (ISCED 0-2)); (ii) Marital status distinguishes between married, separated/divorced, widowed and unmarried (including co-habiting); (iii) Activity status includes employed, self-employed, student, unemployed, retired, housework and 'other economically inactive'.

The NPHS has a complex multi-stage stratified sampling design. In order to keep the sample representative of the Canadian adult population, sampling weights are used in all analyses.

5. Regression results

We estimate the conditional quantile functions at 75 different quantiles, chosen randomly over the interval (0,1), and also the conditional mean function by OLS. A convenient way to summarise the information provided by such a large set of parameters is by means of graphical display. Figure 1 shows the quantile regression and the OLS estimates for a selection of 16 variables from the right hand side in the HUI equation. For each of the selected variables we plot the values of the quantile regression coefficient, together with the upper and lower bounds of its 95% confidence interval, over the (0,1) range. The three horizontal lines in the graphs represent the OLS coefficient and the upper and lower bounds of its 95% confidence interval.

The graph for the intercept term (which estimates the quantiles of the distribution of health conditional on the characteristics of the reference individual) is in the top left panel of Figure 1. This shows that the quantile regression point estimates increase over the HUI distribution and reach the value of 1 (maximum health) at approximately θ=0.7. Note also that for θ<0.25 and θ>0.5 the confidence interval for the quantile regression estimate does not overlap with that of the OLS estimate.

Focus now on the graph for the coefficient on the logarithm of equivalised household income. Here we find that the point estimate decreases over the distribution of health and reaches a zero effect at approximately θ=0.5. This suggests that, for the healthiest half of the population, increases in household income are not associated to increases in health. Conversely, for θ<0.25 the quantile regression point estimate lies above the confidence interval for the OLS estimate (although the confidence intervals overlap) suggesting that the OLS estimate understates the effect of income on less healthy individuals. These features of the data are also represented in Figure 2, which plots the predicted relationship between health and equivalised household income for different parts of the health distribution. Note that the predicted schedule for θ=0.05 is steeper than the rest of the plotted schedules. In fact, the relationship implied by the OLS estimate becomes practically flat at a relatively low level of income. The horizontal line at the top of the graph corresponds to the predicted schedule for θ=0.75, again suggesting that for healthy individuals the marginal effect of income on health is zero.

Look now at the coefficients for the education dummy variables. The specification omits the highest educational category and the OLS coefficients are all negative and significantly different from zero (except educ3, whose confidence interval includes 0). This conforms with the intuitive idea that more educated individuals will have better health. The quantile regression coefficients are consistent with this idea only to a limited extent; they are zero for values of θ above 0.5 (educ2 and educ4) or 0.7 (educ1). Furthermore, the OLS coefficient understates the association of education and health at low levels of health.

The coefficient on the dummy for disability provides perhaps the best example of the limited ability of OLS on the level of HUI to capture the heterogeneous pattern of effects over the health distribution. Again, the OLS coefficient underestimates the (negative) effect of disability on the health measure in the bottom part of the distribution while it overestimates the effects at the top part of the distribution. In this case the confidence intervals only overlap for values of θ between 0.35 and 0.55.

Figure 1 contains graphs for other variables whose association with health varies over the distribution such as the dummy controlling retirement or the dummy controlling marital status, for which the plot suggests a positive association with health in the low part of the health distribution. In other cases, such as the dummy controlling the female 75 to 79 age group, the confidence interval for the OLS includes almost the whole series of quantile regression estimates. The overall conclusion to be drawn from Figure 1 is that a model that imposes homogeneous coefficients does not capture many important features of the data. Moreover, there seems to be a systematic relationship between the coefficients and the health rank, which suggests that parameter heterogeneity has a role in the explanation of health inequality.

Recall that, before proceeding to compute and decompose inequality measures, we need to assign a particular conditional quantile function to each individual according to expression (14). It is useful then at this stage to evaluate the "goodness of fit" of this procedure against the benchmark provided by the OLS predictions. Figures 3 and 4 present the model predictions against the actual values of HUI using the OLS estimates and the quantile regression estimates respectively. The 45^{o} line traces the actual values of HUI and the scatter of points around it correspond to predicted values. The comparison is, not surprisingly, favourable to the model with heterogeneous parameters. Indeed, the unadjusted R-squared from the OLS predictions is 22% while that derived from the quantile regression model predictions is 95.3%.

6. Decomposition analysis

We now use the parameter estimates discussed in the previous section in order to calculate and decompose the Gini coefficient and concentration indices for HUI. Table 1 presents the results for the decomposition of the two measures of inequality into:

i) the contribution of the product of the OLS elasticities and the concentration indices of the regressors on health rank (or income rank in the case of the CI),

ii) the contribution of the covariance of the slope parameters with the health rank (or income rank in the case of the CI),

iii) the contribution of the covariance of the intercept parameters with the health rank (or income rank in the case of the CI) and

iv) a residual corresponding to the covariance of the approximation errors in the heterogeneous parameters model with the health rank (or income rank in the case of the CI).

Table 1. Summary of decomposition analysis of Gini and concentration indices

Actual	OLS		Heterogeneous Slopes		Heterogeneous Intercepts		Residual	
	Contrib.	% Contrib.	Contrib.	% Contrib.	Contrib.	% Contrib.	Contrib.	% Contrib.
G=0.0678	0.0151	22.26%	-0.0347	-51.19%	0.0830	122.44%	0.0044	6.48%
CI=0.0141	0.0135	95.83%	-0.0026	-18.47%	0.0033	23.44%	-0.0002	-1.15%

In Tables 2 and 3 we present the contribution of each explanatory variable to the inequality measures.

Table 2. Contribution of explanatory variables to Gini index.

Regressor	OLS		Heter. Parameters		Total for regressor	
	Contrib	%Contrib	Contrib.	%Contrib	Contrib.	%Contrib
Lincome	0.0005	0.77%	-0.0540	-79.68%	-0.0535	-78.91%
Educ1	0.0015	2.23%	0.0009	1.28%	0.0024	3.51%
Educ2	0.0003	0.47%	0.0013	1.97%	0.0017	2.44%
Educ3	0.0000	-0.07%	0.0004	0.65%	0.0004	0.58%
Educ4	-0.0001	-0.16%	0.0020	2.97%	0.0019	2.81%
Househ	0.0002	0.28%	0.0024	3.58%	0.0026	3.86%
Student	0.0000	-0.03%	0.0000	0.04%	0.0000	0.01%
disabled	0.0062	9.11%	0.0019	2.78%	0.0081	11.89%
unemploy	0.0000	0.01%	0.0003	0.37%	0.0003	0.38%
Retired	0.0016	2.43%	0.0041	6.00%	0.0057	8.43%
Other	0.0000	0.05%	0.0003	0.43%	0.0003	0.47%
Married	0.0001	0.21%	-0.0032	-4.66%	-0.0030	-4.45%
Div_wid	0.0002	0.31%	0.0004	0.66%	0.0007	0.97%
m20_24	0.0001	0.13%	0.0000	-0.05%	0.0001	0.08%
m25_29	0.0000	-0.03%	0.0001	0.20%	0.0001	0.17%
m30_34	-0.0001	-0.17%	0.0005	0.79%	0.0004	0.62%
m35_39	-0.0001	-0.09%	0.0004	0.59%	0.0003	0.50%
m40_44	-0.0001	-0.11%	0.0004	0.66%	0.0004	0.55%
m45_49	0.0000	0.02%	0.0006	0.82%	0.0006	0.85%
m50_54	0.0001	0.15%	0.0003	0.46%	0.0004	0.61%
m55_59	0.0002	0.23%	0.0001	0.18%	0.0003	0.41%
m60_64	0.0001	0.20%	0.0002	0.34%	0.0004	0.54%
m65_69	0.0001	0.21%	0.0001	0.17%	0.0003	0.39%
m70_74	0.0003	0.41%	0.0002	0.36%	0.0005	0.76%
m75_79	0.0002	0.25%	0.0001	0.20%	0.0003	0.45%
m80_	0.0006	0.92%	0.0003	0.47%	0.0009	1.39%
f15_19	0.0000	0.00%	0.0001	0.09%	0.0001	0.09%
f20_24	0.0000	-0.07%	0.0002	0.36%	0.0002	0.29%
f25_29	-0.0001	-0.08%	0.0002	0.36%	0.0002	0.29%
f30_34	-0.0001	-0.11%	0.0005	0.80%	0.0005	0.69%
f35_39	-0.0001	-0.13%	0.0006	0.96%	0.0006	0.83%
f40_44	-0.0001	-0.14%	0.0008	1.21%	0.0007	1.07%
f45_49	0.0003	0.45%	0.0007	1.00%	0.0010	1.45%
f50_54	0.0002	0.32%	0.0004	0.64%	0.0007	0.96%
f55_59	0.0003	0.40%	0.0003	0.42%	0.0006	0.83%
f60_64	0.0002	0.32%	0.0001	0.10%	0.0003	0.42%
f65_69	0.0004	0.64%	0.0001	0.16%	0.0005	0.80%
f70_74	0.0004	0.63%	0.0001	0.22%	0.0006	0.85%
f75_79	0.0005	0.67%	0.0003	0.44%	0.0007	1.10%
f80_	0.0011	1.62%	0.0003	0.48%	0.0014	2.10%
Total slopes	0.0151	22.26%	-0.0347	-51.19%	-0.0196	-28.93%
Intercept			0.0830	122.44%	0.0830	122.44%
Total Parameters					0.0634	93.51%
Residual					0.0044	6.48%
Actual					0.0678	100.00%

Table 3. Contribution of explanatory variables to Concentration Index.

Regressor	OLS		Heter. Parameters		Total for regressor	
	Contrib	%Contrib	Contrib.	%Contrib	Contrib.	%Contrib
Lincome	0.0040	28.55%	-0.0033	-23.13%	0.0008	5.42%
educ1	0.0017	12.20%	-0.0004	-2.58%	0.0014	9.62%
educ2	0.0008	5.58%	-0.0002	-1.68%	0.0005	3.90%
educ3	0.0000	0.10%	0.0000	-0.01%	0.0000	0.09%
educ4	-0.0001	-0.88%	0.0001	0.74%	0.0000	-0.15%
Househ	0.0007	5.29%	0.0002	1.57%	0.0010	6.86%
Student	0.0000	0.14%	0.0000	-0.13%	0.0000	0.00%
Disabled	0.0029	20.46%	0.0002	1.37%	0.0031	21.83%
Unemploy	0.0002	1.10%	0.0000	0.12%	0.0002	1.22%
Retired	0.0012	8.62%	0.0005	3.88%	0.0018	12.50%
Other	0.0001	0.66%	0.0001	0.61%	0.0002	1.27%
Married	0.0004	2.93%	-0.0002	-1.22%	0.0002	1.72%
div_wid	0.0002	1.74%	0.0000	0.06%	0.0003	1.80%
m20_24	0.0000	0.02%	0.0000	0.06%	0.0000	0.08%
m25_29	0.0000	-0.05%	0.0000	0.05%	0.0000	0.00%
m30_34	-0.0001	-0.40%	0.0001	0.36%	0.0000	-0.03%
m35_39	0.0000	-0.09%	0.0000	0.07%	0.0000	-0.03%
m40_44	-0.0001	-0.60%	0.0000	0.11%	-0.0001	-0.48%
m45_49	-0.0003	-2.14%	0.0001	0.59%	-0.0002	-1.55%
m50_54	-0.0002	-1.75%	0.0000	-0.13%	-0.0003	-1.88%
m55_59	-0.0002	-1.14%	0.0000	-0.25%	-0.0002	-1.40%
m60_64	0.0000	0.29%	0.0000	0.05%	0.0000	0.34%
m65_69	0.0001	0.74%	0.0000	-0.33%	0.0001	0.41%
m70_74	0.0001	0.87%	0.0000	-0.04%	0.0001	0.83%
m75_79	0.0002	1.28%	0.0001	0.36%	0.0002	1.64%
m80_	0.0003	2.43%	0.0000	0.34%	0.0004	2.77%
f15_19	0.0000	0.02%	0.0000	0.08%	0.0000	0.10%
f20_24	0.0001	0.52%	0.0000	0.03%	0.0001	0.55%
f25_29	0.0000	0.10%	0.0000	0.05%	0.0000	0.15%
f30_34	0.0000	0.15%	0.0000	-0.08%	0.0000	0.07%
f35_39	0.0000	0.09%	0.0000	0.20%	0.0000	0.28%
f40_44	-0.0001	-0.77%	0.0001	0.54%	0.0000	-0.23%
f45_49	-0.0003	-2.33%	0.0001	0.95%	-0.0002	-1.38%
f50_54	-0.0002	-1.60%	0.0001	0.71%	-0.0001	-0.89%
f55_59	-0.0001	-0.59%	0.0000	-0.07%	-0.0001	-0.66%
f60_64	0.0002	1.24%	0.0000	0.03%	0.0002	1.26%
f65_69	0.0003	2.28%	-0.0001	-0.60%	0.0002	1.68%
f70_74	0.0003	2.47%	0.0000	-0.32%	0.0003	2.15%
f75_79	0.0004	2.56%	0.0000	-0.25%	0.0003	2.32%
f80_	0.0008	5.75%	-0.0001	-0.42%	0.0008	5.33%
Total slopes	0.0135	95.83%	-0.0026	-18.30%	0.0109	77.53%
Intercept			0.0033	23.44%	0.0033	23.44%
Total Parameters					0.0142	100.97%
Residual					-0.0002	-1.15%
Actual					0.0141	100.00%

Recall the expressions for the decompositions of the Gini coefficient and the concentration index in (6) and (7). As demonstrated by Wagstaff et al. (2002), the contributions to the "OLS decomposition" depend on the product of the elasticity of health with respect to each explanatory variable and the concentration index for each variable, which in turn depend on the scaled covariance between the variable and the relative rank of health or income. The OLS decomposition treats the elasticity as homogeneous across individuals as it is evaluated at the OLS estimation of β_k and the mean of y and x_k. In table 1 the OLS components account for 26% of the observed Gini coefficient and 96% of the concentration index of health on income. These contributions show how the variation in the explanatory variables influences the Gini coefficient or the concentration index when the slope coefficients are constant. These figures are comparable to the results obtained by van Doorslaer and Jones (2002, Table 3a) in their decomposition of the Gini index and the concentration index using the same data and explanatory variables. In fact, they report an explained Gini of 0.0326 rather than our equivalent figure of 0.0151 because their computation uses the rank for predicted HUI rather than actual HUI. Thus the decomposition method that does not allow for heterogeneity in individual responses performs well for the concentration index of health on income, but it offers a less complete picture for overall health inequality as measured by the Gini index.

The second component of the decomposition shows how heterogeneity in the slope coefficients modifies the contribution to inequality of the explanatory variables. The figures in table 1 show that heterogeneity in responses reduces the Gini coefficient by 51% and the concentration index by 18%. The impact of income is of particular interest in this reduction of health inequality. Tables 2 and 3 show that heterogeneity in the income effect result in the Gini coefficient and the concentration index being smaller than what they would be if everyone had the average income slope coefficient all else held equal. In particular tables 2 and 3 suggests that the Gini coefficient and the concentration indices would be, respectively, 80% and 23% greater if the marginal effect of income was homogeneous in the population. This makes intuitive sense. The OLS results in Figure 2 show that the use of the logarithm of income captures concavity in the relationship between HUI and income, supporting the notion of diminishing marginal returns on average. However, allowing for variation around the conditional mean function due to individual heterogeneity shows that there is “excess” curvature (i.e. additional concavity). The individual heterogeneity inherent in the data implies greater concavity in the relationship between individual income and health than in a world where all individuals have the same response to income, with the elasticity given by the OLS estimates. This extra concavity reduces health inequality. Indeed, as Contoyannis and Forster (1999) show, for a given level of income inequality, the more concave is the relationship between income and health, the smaller is the level of health inequality.

The contribution of the heterogeneous intercepts is interpreted as the effect of unobserved heterogeneity. If the explanatory variables were not related to the level of health, these figures tell us that the Gini index would be 22% greater than the actual value but, on the other hand, the concentration index would be 77% smaller than the actual value. Table 3 suggests that, apart from income itself, the main correlates of income related health inequality are disability and retirement.

7. Summary and conclusion

In this paper we have shown how the regression based methods for the decomposition of health inequality developed by Wagstaff et al. (2002) can be extended to incorporate individual heterogeneity in the responses of health to the explanatory variables. We have illustrated our proposal with an application to the Canadian NPHS of 1994. Our strategy for the estimation of heterogeneous responses is based on the quantile regression model.

The results suggest that there is an important degree of heterogeneity in the association of health to explanatory variables which, in turn, accounts for a substantial percentage of the inequality in observed health. A particularly interesting finding is that the marginal effect of income on health is zero for healthy individuals but positive and significant for unhealthy individuals. The heterogeneity in the income response reduces both overall health inequality and income related health inequality. This suggests that considering the possibility of heterogeneity in health responses, which can be done in situations where a continuous measure of health is available, is likely to provide a fuller picture of both overall and income related health inequality than assuming homogeneity.

References

Abadie, A., Angrist, J., and Imbens, G., 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70, 91-117.

Abrevaya, J., 2001. The effect of demographics and maternal behaviour on the distribution of birth outcomes. *Empirical Economics* 26, 247-257.

Arias, O., Hallock, K., and Sosa-Escudero, W., 2001. Individual heterogeneity in the returns to schooling: Instrumental variable quantile regression using twins data. *Empirical Economics* 26, 7-40.

Buchinsky, M., 1994. Changes in U.S. wage structure 1963-1987. An application of quantile regression. *Econometrica* 62. 405-458

Contoyannis, P., and Forster, M. 1999. The distribution of health and income: a theoretical framework. *Journal of Health Economics* 18, 605-622.

Feeny, D., Furlong, W., Boyle, M., Torrance, G., 1995. Multi-attribute health status classification systems: Health Utilities Index. *Pharmacoeconomics* 7, 490-502.

Heckman, J. 2001 Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture, *Journal of Political Economy*, 109, 673-748.

Humphries, K and E van Doorslaer, 2000, Income-related inequalities in health in Canada, *Social Science and Medicine*, 50, 663-671

Koenker, R., and Bassett, G., 1978. Regression quantiles. *Econometrica*. 46, 33-50.

Lambert, P., 1993. The distribution and redistribution of income. A mathematical analysis. 2nd Edition. Manchester University Press. Manchester.

Lerman, R.I., Yitzhaki, S., 1989. Improving the accuracy of estimates of Gini coefficients. *Journal of Econometrics* 42, 43-47.

Manning, W., Blumberg, L., and Moulton, L., 1995. The Demand for Alcohol: The differential response to price. *Journal of Health Economics* 14, 123-148.

Tambay J-L., Catlin, G., 1995. Sample Design of the National Population Health Survey. *Health Reports* 7, 29-38.

Torrance, G.W., Furlong, W., Feeny, D., Boyle, M., 1995. Multi-attribute preference functions: Health Utilities Index. *Pharmacoeconomics* 7, 503-520.

Torrance, G.W., Feeny, D., Furlong W.J., Barr, R.D., Zhang, Y., Wang, Q., 1996. Multiattribute Utility Function for a Comprehensive Health Status Classification System. *Medical Care* 34, 702-722.

Van Doorslaer, E, and Jones, A., 2002. “The determinants of inequalities in self reported health: a validation of a new approach to measurement. Ecuity II Project. Working Paper # 2

Van Doorslaer, E, Koolman, X., 2000, *Income-related inequalities in health in Europe: evidence from the European Community Household Panel*, Ecuity II Project, Working Paper #1, Erasmus University, Rotterdam.

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.

Wagstaff, A., van Doorslaer, E., 1994. Measuring inequalities in health in the presence of multiple-category morbidity indicators. *Health Economics 3*, 281-291.

Wagstaff, A., van Doorslaer, E., Paci, P., 1989. Equity in the finance and delivery of health care: some tentative cross-country comparisons. *Oxford Review of Economic Policy* 5, 89-112.

Wagstaff, A., Paci, P., van Doorslaer, E., 1991. On the measurement of inequalities in health. *Social Science and Medicine 33*, 545-557.

Wagstaff, A, van Doorslaer, E., Watanabe, N., 2002. On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam, *Journal of Econometrics*, (forthcoming).

Appendix

Substituting equation (5) into (1) we obtain,

$$G = \left(\frac{2}{\bar{y}}\right) \mathrm{cov}(y_i, R_i) =$$

$$G = \left(\frac{2}{\bar{y}}\right) \sum_i \left(y_i - \bar{y}\right)\left(R_i - 1/2\right) =$$

$$G = \left(\frac{2}{\bar{y}}\right) \sum_i \left(\beta_{i1} + \sum_{k=2}^{K} \beta_{ik} x_{ik} - \bar{y}\right)\left(R_i - 1/2\right)$$

Since

$$\bar{y} = \beta_1^{OLS} + \sum_{k=2}^{K} \beta_k^{OLS} \bar{x}_k$$

we can write after some manipulation,

$$G = \left(\frac{2}{\bar{y}}\right)\left[\sum_i \left(\sum_{k=2}^{K} \beta_k^{OLS}\left(x_{ik} - \bar{x}_k\right) + \sum_{k=2}^{K}\left(\beta_{ik} - \beta_k^{OLS}\right) x_{ik} + \left(\beta_{i1} - \beta_1^{OLS}\right)\right)\left(R_i - 1/2\right)\right]$$

Collecting terms and changing the order of summation,

$$G = \sum_{k=2}^{K} \left(\frac{2}{\bar{y}}\right) \beta_k^{OLS} \sum_i \left(x_{ik} - \bar{x}_k\right)\left(R_i - 1/2\right) +$$

$$+ \left(\frac{2}{\bar{y}}\right) \sum_{k=1}^{K} \sum_i x_{ik}\left(\beta_{ik} - \beta_k^{OLS}\right)\left(R_i - 1/2\right) + \left(\frac{2}{\bar{y}}\right) \sum_i \left(\beta_{i1} - \beta_k^{OLS}\right)\left(R_i - 1/2\right)$$

Noting that the concentration index of x_k on y is given by,

$$C_k = \frac{2}{\bar{x}_k}\operatorname{cov}(x_{ik}, R_i)$$

And considering β_1^{OLS} a measure of central tendency for β_{i1}, we finally obtain equation (6) in the main text,

$$G = \sum_{k=2}^{K} \beta_k^{OLS} \, \frac{\bar{x}_k}{\bar{y}} C_k + \left(\frac{2}{\bar{y}}\right) \sum_{k=2}^{K} \sum_{i} x_{ik} \left(\beta_{ik} - \beta_k^{OLS}\right)\left(R_i - 1/2\right) + \left(\frac{2}{\bar{y}}\right) \operatorname{cov}\left(\beta_{i1}, R_i\right)$$

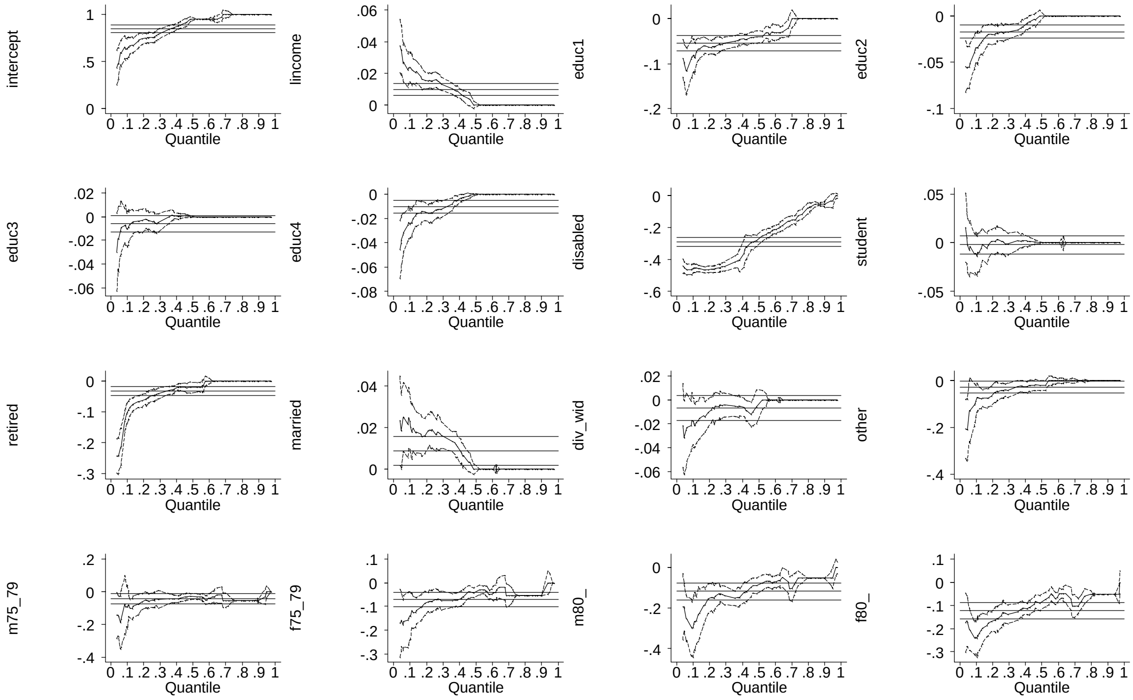

Figure 1. ***Quantile regression coefficients and OLS coefficients with 95% confidence intervals***

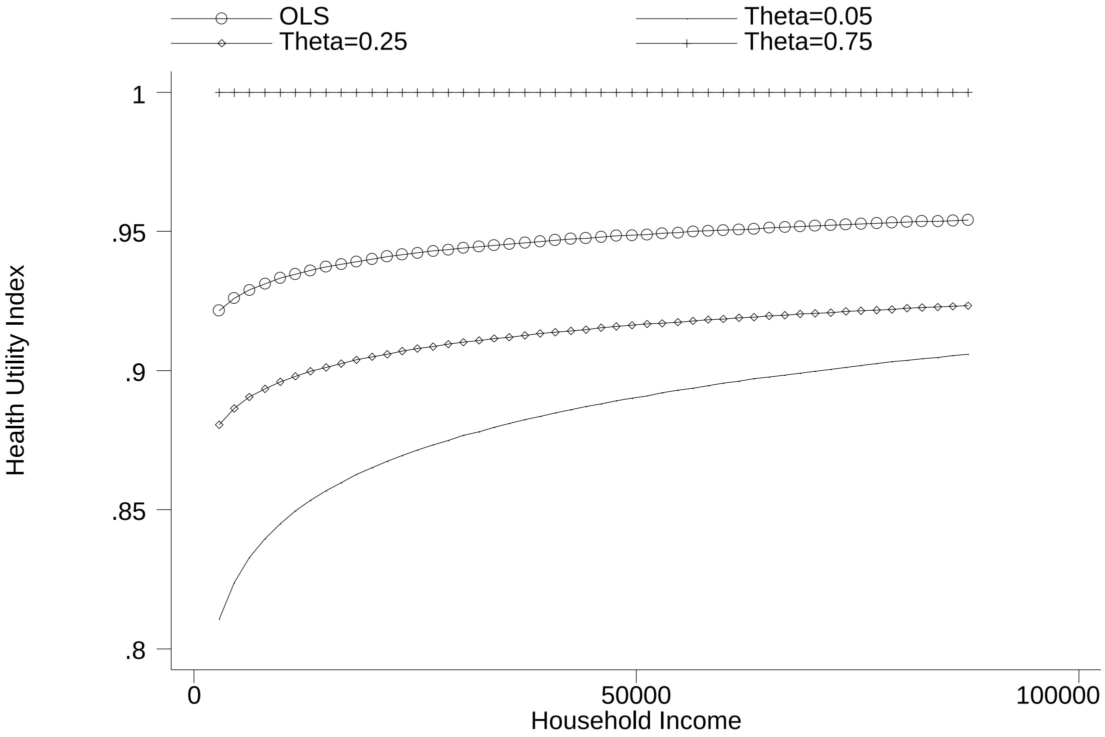

Figure 2. ***Health-income relationships implied by the quantile regression estimates and the OLS estimates.***

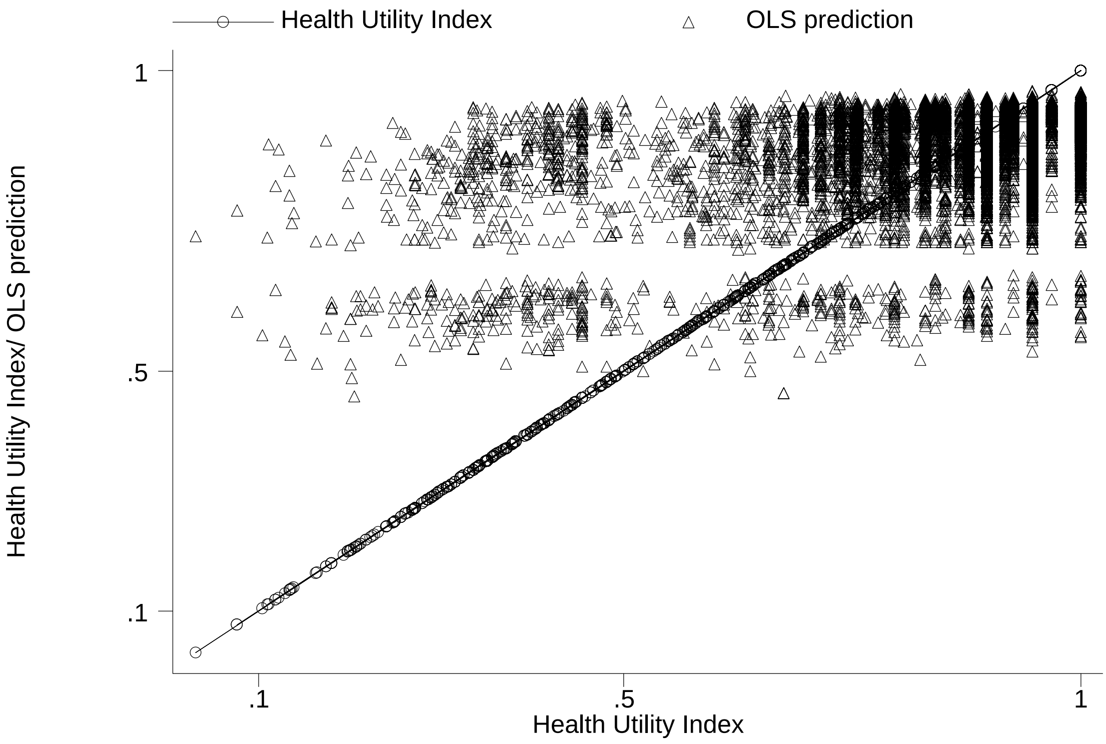

Figure 3. ***Actual and predicted values for HUI. Homogeneous (OLS) regression model.***

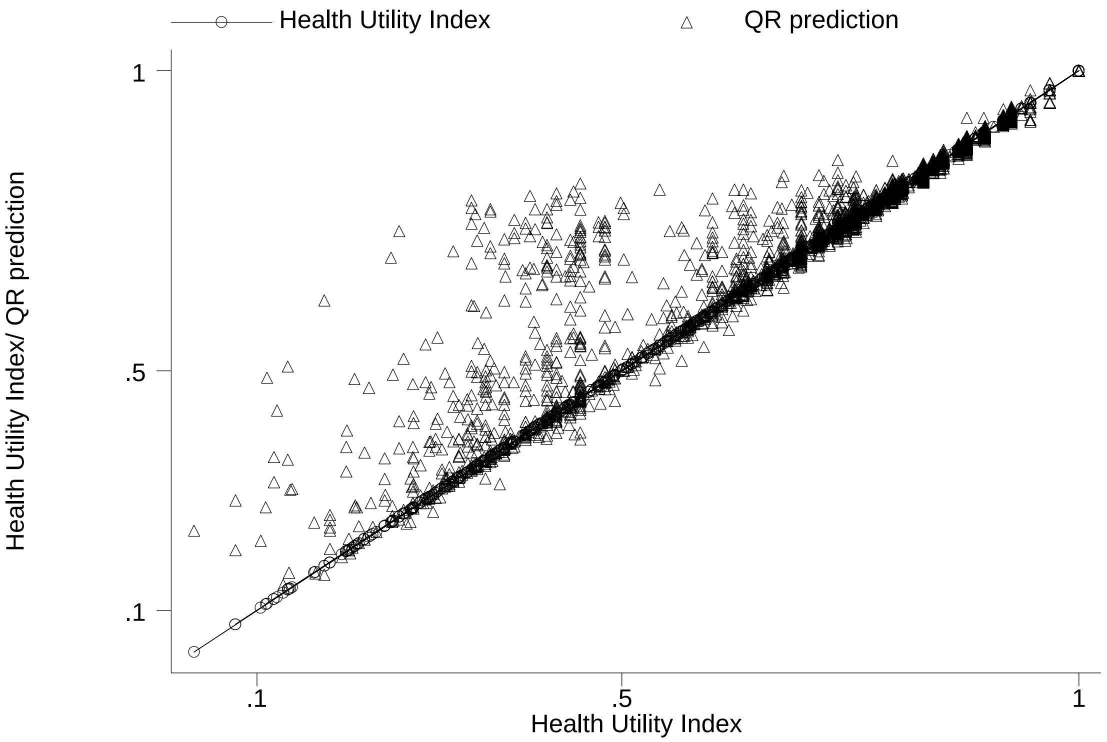

Figure 4. ***Actual and predicted values for HUI. Heterogeneous (quantile) regression model.***