The econometrics of inequality and poverty

Lecture 10: Explaining poverty and inequality using econometric models

Michel Lubrano

October 2016

Contents

1 Introduction

Up to now, we focussed on the description of the income distribution. We saw how to compare two distributions, either between two different countries or for the same country between two different points of time. But we stayed on a descriptive standpoint, we did not try to explain the formation of the income distribution or to explain poverty. In doing this, we followed the dichotomy that exists in the literature between measuring inequality and poverty and the theory of income formation. Household income can be divided in several parts: wages or earning (the most important part of income), rents and financial income and finally taxes and transfers. Labour economists examined the question of wage inequality and wage dispersion in the eighties, promoting for instance the dichotomy between skilled and unskilled labour. However, they have never tried to relate this question to household income inequality. We shall not try to fill up the gap in this chapter, asking the reader to refer to Atkinson (2003) for instance. We shall however try to present some econometric tools that are useful for decomposing a poverty index or for analysis the evolution of an income distribution.

2 Decomposing poverty and inequality

The idea is to split the inequality or the poverty measured by an index into different and mutuality exclusive groups. Which group in the population is more subject to poverty? This principle can be extended to the decomposition of inequality, most of the time wage inequality in the literature, between two groups. For instance is wage differential between male and females or black and white due to intrinsic differences or to a mere discrimination? For that, we need a wage equation, a model based on a regression and then to decompose the regression between different effects. This is the Oaxaca decomposition.

2.1 FGT indices

The index of Foster et al. (1984) is decomposable because of its linear structure. Let us consider the decomposition of a population between rural and urban. If X represents all income of the population, the partition of X is defined as $X = X^U + X^R$. Let us call p the proportion of X^U in X. Then the total index can be decomposed into

$$P_\alpha = p\frac{1}{n}\sum_{i=1}^{n_U}\left(\frac{z-x_i^U}{z}\right)^\alpha \mathbb{1}(x_i \leq z) + (1-p)\frac{1}{n}\sum_{i=1}^{n_R}\left(\frac{z-x_i^R}{z}\right)^\alpha \mathbb{1}(x_i \leq z) \quad (1)$$

$$= p\,P_\alpha^U + (1-p)\,P_\alpha^R. \quad (2)$$

where P_α^U is the index computed for the urban population and P_α^R the index computed for the rural population. So decomposition for a poverty index means that poverty for the total population can be expressed as a weighted sum of the same poverty index applied to each group. Inequality indices can also be decomposed. But here, as we have already seen in Chapter 7, decomposability means something else. Inequality within the total population is expressed as a weighted sum

of inequality within each group plus a remainder which is interpreted as inequality between the groups.

Table 1: Decomposing poverty in the 1996 UK

	Retired	Working	Unemployed	Others	Total
n	1806	2355	949	933	6043
%	0.299	0.390	0.157	0.154	1.000
P_0	4.23	0.46	16.13	2.50	4.36
$P_0 n_i/n$	1.27	0.18	2.53	0.39	4.36

We illustrate this decomposability using the FES data for 1996. We have defined a poverty line as 50% of the mean income for the total sample. We can divide this sample into mutual exclusive groups, depending on the status of the head of the household. In Table 1, we see that poverty is concentrated among the unemployed followed by the retired group. When the head of the household is working, there is only 0.5% chances that the household is classified as poor.

2.2 Generalised entropy indices

A decomposable inequality index can be expressed as a weighted average of inequality within subgroups, plus a remainder that is interpreted as inequality between the subgroups. More precisely, let $I(x, n)$ be an inequality index for a population of n individuals with income distribution x. $I(x, n)$ is assumed to be continuous and symmetric in x, $I(x, n) \geq 0$ with perfect equality holding if and only if $x_i = \mu$ for all i, and $I(x, n)$ is supposed to have a continuous first order partial derivative. Under these assumptions, Shorrocks (1980) defines additive decomposition condition as follows:

Definition 1. *Given a population of of any size $n \geq 2$ and a partition into k non-empty subgroups, the inequality index $I(x, n)$ is decomposable if there exists a set coefficients $\tau_j^k(\mu, n)$ such that:*

$$I(x, n) = \sum_{j=1}^{k} \tau_j^k I(x^j; n_j) + B,$$

where $x = (x^1, \ldots, x^k)$, $\mu = (\mu_1, \ldots, \mu_k)$ is the vector of subgroup means $\tau_j(\mu, n)$ is the weight attached to subgroup j in a decomposition into k subgroups, and B is the between-group term, assumed to be independent of inequality within the individual subgroups.

- Some inequality indices do not lead themselves to a simple decomposition depending only on group means, weights and group inequality. The relative mean deviation, the variance of logarithms, the logarithmic variance are standard examples. The Gini coefficient can be decomposed in this way only if groups do not overlap (the richer of one group is poorer than the immediate neighbouring group).

- The class of decomposable indices contains many examples. We can quote the inequality index of Kolm which has an additive invariance property (when usual indices have a multiplicative invariance property). The widest class of decomposable inequality indices is represented by the Generalised Entropy indices which contains as particular cases the Theil index, the mean logarithm deviation index and the Atkinson index.

We consider a finite discrete sample of n observations divided exactly in k groups. Each group has proportion p_i, size n_i and empirical mean μ_i. Inside a group, the generalised entropy index writes

$$I_{GE_i} = \frac{1}{c^2 - c}\left[\sum_{j=1}^{n_i} p_i \left(\frac{y_j}{\mu_i}\right)^c - 1\right]$$

Inequality between groups is measured as

$$I_{Between} = \frac{1}{c^2 - c}\left[\sum_{i=1}^{k} p_i \left(\frac{\mu_i}{\mu}\right)^c - 1\right]$$

where μ is the sample mean. Let us now define the income share of each group as

$$g_i = p_i \frac{\mu_i}{\mu}$$

Then inequality is decomposed according to

$$I_{Total} = \sum_{i=1}^{k} g_i^c p_i^{1-c} I_{GE_i} + I_{Between}$$

The Atkinson index is a non-linear function of the GE index. Consequently the decomposition of this index is ordinaly but not cardinally equivalent to the decomposition of the GE. For details of calculation, see Cowell (1995).

Table 2: Decomposing inequality in the 1996 UK

	Retired	Working	Unemployed	Others	Between	Total
n	1806	2355	949	933		6043
%	0.299	0.390	0.157	0.154		1.000
g_i	0.237	0.504	0.103	0.155		1.000
GE, $c = 0.5$	0.114	0.0986	0.134	0.132		
Weighted GE	0.0304	0.0437	0.0170	0.0204	0.0331	0.145
GE, $c = 1.5$	0.142	0.109	0.159	0.167		
Weighted GE	0.0300	0.0628	0.0133	0.0259	0.0325	0.165

g_i represents the income shares, while $\%$ are the percentages of individual per group. GE represents the inequality within each group and the weighted GE the weighted inequality that sums to the overall inequality.

We illustrate this decomposability using again the FES data for 1996. We have again divided the sample into mutual exclusive groups, depending on the status of the head of the household. In Table 2, we see that weighted inequality is concentrated among the working people according to both indices, followed by the retired. On the contrary, there is very little inequality among the unemployed. The between inequality is of the same importance as within inequality for the retired. This is just the reverse picture as for poverty.

2.3 Oaxaca decomposition

In the previous section, we have decomposed a poverty rate according to mutually exclusive groups of the population. But, we provided no explanation on the reason of this decomposition, what made a person belong to one of these groups. Oaxaca (1973) was the first to try to give an explanation on the sources, the causes of inequality, using a regression model. But note also the paper Blinder (1973) published the same year, so that the decomposition is often called the Blinder-Oaxaca decomposition.

Oaxaca (1973) took interest in wage inequality between males and females. Suppose that we have divided our sample in two groups, one group of males, one group of females. We want to explain the difference in average wage that there exist between males and females, with the main interrogation: is this wage differential simply due to differences in characteristics, for instance males are more educated or have more experience, or is this difference due to discrimination, e.g. the yield of experience is lower for females. In order to answer these questions, we estimate for each group a wage equation which relates the log of the wage to a number of characteristics, among which we find experience and years of schooling. Other variables can include regional location and city size for instance:

$$\log(W_i) = X_i\beta_i + u_i, \qquad i = m, f.$$

Once these two equations are estimated, we have a $\hat{\beta}_m$ for males and a $\hat{\beta}_f$ for females. We are going to try to explain wages differences between males and females as follows. We can say that a part of this difference can be explained by different characteristics. For instance if males have more experience or if females are more educated. These objective differences are measured by $X_h - X_f$. But another part of the wage differences can be explained simply by the different yield of these characteristics: for an identical experience, a female is paid less than a male. These differences in yields are at the root of the discrimination existing between males and females on the labour market.

In a regression model, the mean of the endogenous variable is given by

$$\log(\overline{W}_i) = \overline{X}_i\hat{\beta}_i,$$

because of the zero mean assumption on the residuals. Using this property, Oaxaca proposed the following decomposition:

$$\log(\overline{W}_m) - \log(\overline{W}_f) = (\overline{X}_m - \overline{X}_f)\hat{\beta}_m + \overline{X}_f(\hat{\beta}_m - \hat{\beta}_f).$$

In this decomposition, the difference in percentage between the average male and female wages is explained first by the difference in average characteristics. As a second term comes the difference in yield of female average characteristics expressed by $\hat{\beta}_m - \bar{\beta}_f$.

This decomposition is very popular in the literature. The original paper is cited more than 3171 times (using GoogleScholar). It gave birth to many subsequent developments. For instance, Juhn et al. (1993) generalised the previous result to the framework of quantile regression. Radchenko and Yun (2003) provide a Bayesian implementation that make easier significance tests.

There are more than one way of decomposing wage inequality. We have chosen Oaxaca (1973) decomposition. The decomposition promoted by Blinder (1973) is also possible. This dual decomposition can be imbedded in a single formulation where the difference in means is expressed as

$$\log(\overline{W}_m) - \log(\overline{W}_f) = (\overline{X}_m - \overline{X}_f)\beta_* + [\overline{X}_m(\hat{\beta}_m - \beta_*) + \bar{X}_f(\beta_* - \beta_f)]. \tag{3}$$

The first part is the explained part, while the term in squared brackets is the unexplained part. We recover the previous decomposition for $\beta_* = \hat{\beta}_m$ while the Blinder decomposition is found for $\beta_* = \hat{\beta}_f$. Other decomposition found in the literature choose β_* as the average between the two regression coefficients.

Of course, a natural question is to know if those differences are statistically significant. Jann (2008) proposes to compute standard errors for this decomposition. There are various ways of computing these standard deviations, the question being to know if the regressors are stochastic or not. If the regressors are fixed, then we have the simple result

$$\text{Var}(\bar{X}\hat{\beta}) = \bar{X}'\text{Var}(\hat{\beta})\bar{X}.$$

If the regressors are stochastic, but however uncorrelated, Jann shows that this variance becomes

$$\text{Var}(\bar{X}\hat{\beta}) = \bar{X}'\text{Var}(\hat{\beta})\bar{X} + \hat{\beta}'\text{Var}(\bar{X})\hat{\beta} + \text{tr}(\text{Var}(\bar{X})\text{Var}(\hat{\beta})).$$

From these expressions, he derives the variance of the Oaxaca decomposition. This is simple, but tedious algebra. So it is better to have a ready made program. A command exists in Stata. It was only very recently implemented in R with the package `oaxaca` (2014) by Marek Hlavac from Harvard (Hlavac 2014). It reproduces the estimation methods available in the Stat package, provide bootstrap standard deviations and also nice plots.

Jann illustrates his method for decomposing the gender wage gap on the Swiss labour market using the Swiss Labour Force Survey 2000 (SLFS; Swiss Federal Statistical Office). The sample includes Employees aged 20-62, working fulltime, having only one job. The dependent variable is the Log of hourly wages. The explanatory variables are the number of years of schooling, the number of years of experience, its square divided by 100, two dummy variables concerning Tenure and the gender of the supervisor. There are 3383 males and 1544 females. From the estimates reported in Table 3, we can compute the original Oaxaca decomposition with results displayed in Table 4. The bootstrap and the stochastic regressor assumption give very comparable standard deviations. Assuming fixed regressors under-evaluate the standard deviations. Wage

Table 3: Wage equations for Switzerland 2000

	Men		Women	
Log wages	Coef.	Mean	Coef.	Mean
Constant	2.4489		2.3079	
	(0.0332)		(0.0564)	
Education	0.0754	12.0239	0.0762	11.6156
	(0.0023)	(0.0414)	(0.0044)	(0.0548)
Experience	0.0221	19.1641	0.0247	14.0429
	(0.0017)	(0.2063)	(0.0031)	(0.2616)
Exp^2	-0.0319	5.1125	-0.0435	3.0283
	(0.0036)	(0.0932)	(0.0079)	(0.1017)
Tenure	0.0028	10.3077	0.0063	7.6729
	(0.0007)	(0.1656)	(0.0014)	(0.2013)
Supervisor	0.1502	0.5341	0.0709	0.3737
	(0.0113)	(0.0086)	(0.0193)	(0.0123)
R^2	0.3470		0.2519	

Table 4: Oaxaca overall decomposition for Switzerland 2000

	Value	Bootstrap	Stochastic	Fixed
Differential	0.2422	0.0122	0.0126	0.0107
Explained	0.1091	0.0076	0.0075	0.0031
Unexplained	0.1331	0.0113	0.0112	0.0111

differentials is more explained by discrimination than by differences in characteristics. These differences are significant. There are both differences in characteristics and discrimination.

Further developments: Bourguignon et al. (2008) explains, using a Oaxaca type decomposition differences between the income distribution of Brazil and of the USA. An idea would be to analyse the dynamics of income using the regression model of Galton-Markov and then compare and explain the differences in income dynamics between two countries. The ECHP could serve as data source.

2.4 Oaxaca in R

We are now going to explain how the main commands of the `R` package `oaxaca` are working. The package provides a data base `data("chicago")` concerning *Labour market and demographic data for employed Hispanic workers in metropolitan Chicago.* This a 2013 sample of Current Population Survey Outgoing Rotation Group. The data frame contains 712 observations and 9 variables:

1. age: the worker's age, expressed in years

2. female: an indicator for female gender
3. foreign.born: an indicator for foreign-born status
4. LTHS: an indicator for having completed less than a high school (LTHS) education
5. high.school: an indicator for having completed a high school education
6. some.college: an indicator for having completed some college education
7. college: an indicator for having completed a college education
8. advanced.degree: an indicator for having completed an advanced degree
9. ln.real.wage: the natural logarithm of the worker's real wage (in 2013 U.S. dollars)

The question is to know the impact of being born in a foreign country can explain wage differentials. So the main command is `oaxaca`. To interpret correctly the data, we have first to delete the rows of the data set that have NA. This is the case only in column 9, which corresponds to wages. Then we recreate a new data set

```
data("chicago")
id = !is.na(chicago[,9])
chicago = chicago[id,]
attach(chicago)
n = length(age)

age2 = age^2/100
lwage = ln.real.wage
wage = exp(lwage)
idf = foreign.born==1
idn = foreign.born==0
sum(idn)
sum(idf)
mean(wage[idf],na.rm=T)
mean(wage[idn],na.rm=T)

Chic = data.frame(wage,age,age2,female,college,
advanced.degree,foreign.born)
out = oaxaca(wage ~ age + age2 + female +
             college + advanced.degree | foreign.born,
             data = Chic, R = 30)
```

The wage equation is described in a formula framework while the group indicator variable is given after the vertical bar. Standard errors are estimated by bootstrap with $R = 30$ replications. It is not wise to try to print the whole object. It is better to print only some elements. `res` being

here the name of the object, we can first access the useful sample sizes by typing `res$n`. There are missing observations in the wage variable so that at the end there are only 666 observations left, with $n_A = 287$ natives and $n_B = 379$ foreign born.

Then, the mean values of the endogenous variable for the two groups is obtained with `res$y` mean wage is \$17.58 for natives and \$14.56 for foreign born. The difference is also indicated (3.02).

Several decomposition methods are available in the object. We shall focus only on what is called the twofold decomposition which corresponds to (3). For choosing the definition of the reference β_* in our notations, β_R in the package notations, several weights are proposed. With a zero weight, $\beta_* = \beta_B$, while with a unit weight $\beta_* = \beta_A$ (to be checked). The other weights are of a less interest as they are more difficult to interpret. We have access to the first two lines of the overall result with `out$twofold$overall[1:2,1:5]`. The decomposition at the variable level is obtained with `out$twofold$variables[1:2]`. This last decomposition can be visualised using `plot(out, decomposition = "twofold", weight = 0)`.

The overall decomposition gives The first line is obtained by setting $\beta_* = national$ and the

weight	coef(explained)	se(explained)	coef(unexplained)	se(unexplained)
0.00	-0.09	0.62	3.11	0.87
1.00	-1.65	0.52	4.67	0.85

second line with $\beta_* = foreign$. Standard errors are indicated as se. This table was obtained using `xtable(out$twofold$overall[1:2,1:5])`. The influence of each variable is more complicated to put into a table. So perhaps, the command `plot(out, decomposition = "twofold", weight = 0)` is more appropriate. Figure 1 provides a graphical represen-

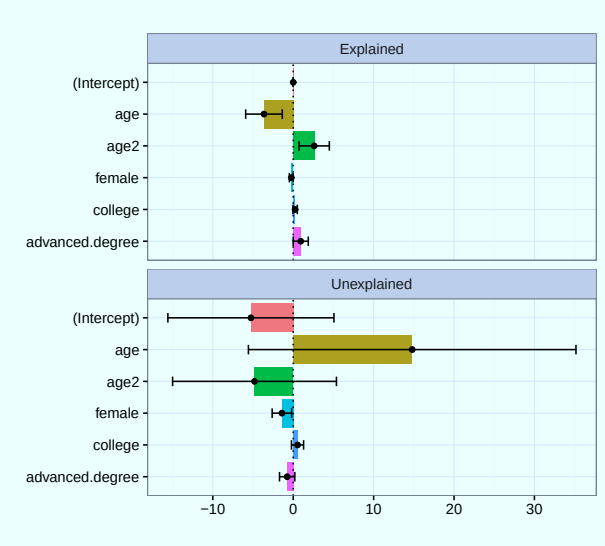

Figure 1: Graphs issued by the `R` command `oaxaca`

tation of each coefficient with an indication of a confidence interval.

2.5 Explaining the income-to-needs ratio

Let us consider a poverty line z and the income y_i of an household. The ratio y/z is known to be the **income-to-needs ratio** in the literature. It can be used to explain the probability that this household has of getting in a state of poverty. $\log(y_i/z)$ is negative if the household is poor, positive otherwise. We can then estimate a regression

$$\log(y_i/z) = x_i'\beta + u_i$$

where x_i is a set of characteristics of the household. If we suppose that u_i is normal, we can compute the probability that an household is poor by mean of

$$P_0 = \Pr(x_i\hat{\beta} < 0) = \Phi(-x_i\beta/\sigma)$$

where σ^2 is the variance of the residuals and $\Phi(\cdot)$ the normal cumulative distribution. When n tends to infinity, the estimated variance tends to zero so that this probability approaches the head count measure.

We can now extend the approach of Oaxaca to explain the difference that there exist of being poor between two groups: white and black households in the US or between Serbs and Albanian households in Kosovo. Yun (2004) propose a generalisation of Oaxaca decomposition for non-linear models and in particular for probit models. Let us call A and B the two groups we consider. The decomposition proposed by Yun (2004) is as follows:

$$\begin{aligned} P_A^0 - P_B^0 &= \overline{\Phi(-X_A\beta_A/\sigma_A)} - \overline{\Phi(-X_B\beta_B/\sigma_B)}] \\ &= [\overline{\Phi(-X_A\beta_A/\sigma_A)} - \overline{\Phi(-X_B\beta_A/\sigma_A)}] \\ &\quad + [\overline{\Phi(-X_B\beta_A/\sigma_A)} - \overline{\Phi(-X_B\beta_B/\sigma_B)}]. \end{aligned}$$

which corresponds to the difference between the characteristics and the difference between the coefficients. This is an overall decomposition, giving global figures. We could be interesting in detailing the influence of each variable in this decomposition. This is not straightforward, because we are in a non-linear model. Yun (2004) has proposed a method to circumvent this difficulty by defining a series of weights. Assuming that there are k characteristics or exogenous variables, we can write

$$\begin{aligned} P_A^0 - P_B^0 &= \textstyle\sum_{i=1}^k W_{\Delta X}^i [\overline{\Phi(-X_A\beta_A/\sigma_A)} - \overline{\Phi(-X_B\beta_A/\sigma_A)}] \\ &\quad + \textstyle\sum_{i=1}^k W_{\Delta\beta}^i [\overline{\Phi(-X_B\beta_A/\sigma_A)} - \overline{\Phi(-X_B\beta_B/\sigma_B)}] \end{aligned}$$

Of course the question is how to define those weights. The weights $W_{\Delta X}^i$ and $W_{\Delta\beta}^i$ are given in Bhaumik et al. (2006a) following a linearisation argument developed in Yun (2004). They are (to be checked):

$$W_{\Delta\beta}^i = \frac{\bar{X}_i(\hat{\beta}_i^A - \hat{\beta}_i^B)}{\sum_{i=1}^k \bar{X}_i(\hat{\beta}_i^A - \hat{\beta}_i^B)} \qquad W_{\Delta X}^i = \frac{\hat{\beta}_i^B(\hat{X}_i^A - \hat{X}_i^B)}{\sum_{i=1}^k \hat{\beta}_i^B(\hat{X}_i^A - \hat{X}_i^B)}.$$

As an alternative and simpler method, one could consult Bazen and Joutard (2013), which is based on a Taylor expansion.

Bhaumik et al. (2006b) use the 2001 Living Standards Measurement Survey (LSMS) data for Kosovo to decompose the difference in the average likelihood of poverty incidence between Serb and Albanian households. The survey, which was carried out between September and December of 2000, collected data from 2,880 households. After accounting for missing values, the survey provides information on 2101 Kosovo Albanian households and 416 Kosovo Serbian households. The ratio $R = y_i/z$ is computed using the World Bank poverty line for Kosovo. The differences in the average probability of being poor between groups A and B, ($\bar{P}_A - \bar{P}_B$), can be algebraically decomposed into two components which represent the characteristics and coefficients effects. The predicted poverty rate for Serbs is 55.98% while it is only of 45.41% for Albanian. There is a gap of 10.56%. How can we explain this gap? Bhaumik et al. (2006b) provide in their Table 2 (reproduced here) an estimation for the two equations. In their Table 3 (reproduced here), they analyse the differences in poverty between the two communities.

Table 2
Determinants of Ratio of Per Capita Expenditure to Poverty Line (ML estimation)

	Albanians		Serbs	
	Estimate	S.E.	Estimate	S.E.
Constant	- 0.33***	(0.09)	- 1.10***	(0.21)
Demographic characteristics of households				
Proportion aged 15 or below	- 0.58***	(0.06)	- 0.17	(0.12)
Proportion aged above 65	- 0.10	(0.11)	- 0.06	(0.13)
Proportion of adults who are male	0.04	(0.09)	0.23	(0.16)
Households with male head	- 0.06	(0.05)	0.06	(0.09)
Education				
Proportion of adults with primary education	0.18**	(0.08)	0.31	(0.19)
Proportion of adults with secondary education	0.58***	(0.08)	0.92***	(0.20)
Proportion of adults with vocational training	0.52***	(0.10)	0.91***	(0.23)
Proportion of adults with tertiary education	0.75***	(0.10)	1.46***	(0.21)
Labor market characteristics				
Proportion of working adults	0.45***	(0.06)	0.22**	(0.11)
Proportion of households with members working in family farms & businesses	- 0.00	(0.07)	- 0.04	(0.11)
Wealth/Assets				
Acreage of land household owns (000)	0.17	(0.15)	0.01	(0.01)
Value of animals household owns (000 DM)	0.03	(0.02)	0.04	(0.03)
Transfers				
Households at least one of whose members has a disability card	0.02	(0.04)	- 0.10	(0.07)
Household at least one of whose members receive private transfers	0.09***	(0.02)	0.33***	(0.11)
Geographic Characteristics				
Urban households	0.05	(0.03)	0.06	(0.06)
Standard deviation of error term (σ)	0.46***	(0.01)	0.46***	(0.03)
Log-likelihood (L)	-150785.98		-19300.24	
Constrained Log-likelihood ($L0$)	-180607.61		-25300.89	
Number of households	2101		416	

Note: *, ** and *** indicate significance at the 10%, 5% and 1% levels, respectively. Weights are used in estimation. Standard errors which are robust to mis-specification are reported. Constrained log-likelihood is calculated only when constant and standard deviation of error term are estimated. Likelihood ratio test, $2*(L - L0)$, rejects the null hypothesis that coefficients except for the constant are zero for both Serbs and Albanians.

Table 3
Decomposing Difference in Poverty Rates of 10.56% between Serbs and Albanians using Estimates of Per Capita Expenditure Regression Equations

	Characteristics Effect		Coefficients Effect	
	Estimate	Share	Estimate	Share
Aggregate Effect	- 0.035	- 33.55	0.141***	133.55
Aggregate Effect Without Constants	- 0.035	- 33.55	- 0.429**	- 405.66
Constant			0.570***	539.21
Demographic characteristics of households	- 0.016	- 15.04	- 0.244**	- 231.24
Proportion aged 15 or below	- 0.021*	- 19.52	- 0.095***	- 90.19
Proportion aged above 65	0.003	2.90	- 0.001	- 0.97
Proportion of adults who are male	- 0.001	- 1.33	- 0.068	- 64.11
Proportion with male head	0.003	2.92	- 0.080	- 75.96
Education	- 0.113***	- 106.66	- 0.191	- 180.66
Proportion of adults with primary education	0.034*	32.64	- 0.044	- 41.81
Proportion of adults with secondary education	- 0.165***	- 155.94	- 0.076*	- 72.17
Proportion of adults with vocational training	0.006***	5.33	- 0.024	- 22.42
Proportion of adults with tertiary education	0.012***	11.32	- 0.047***	- 44.25
Labor market characteristics	- 0.008*	- 7.64	0.074**	70.41
Proportion of working adults	- 0.010*	- 9.67	0.067*	63.30
Proportion of households with members working in family farms & businesses	0.002	2.02	0.008	7.11
Wealth/Assets	0.003	2.46	0.003	2.88
Acreage of land household owns (000)	- 0.000	- 0.36	0.008	7.39
Value of animals household owns (000 DM)	0.003	2.82	- 0.005	- 4.51
Transfers	0.106**	100.00	- 0.068*	- 64.28
Proportion of households at least one of whose members has a disability card	- 0.000	- 0.03	0.009	8.65
Proportion of household at least one of whose members receive private transfers	0.106**	100.03	- 0.077**	- 72.92
Geographic Characteristics				
Urban households	- 0.007	- 6.67	- 0.003	- 2.77

Note: Share is the ratio of the contribution of each factor to the "predicted" overall difference in poverty rate (10.56%) between Serbs (55.98%) and Albanians (45.41%), in percentage terms. The observed overall difference in poverty rate are 11.87% between Serbs (57.38%) and Albanians (45.52%). The predicted poverty rate is computed using estimates from the per capita expenditure regression. The details of the computation using the per capita expenditure regression is explained in the main text. *, ** and *** indicate significance at the 10%, 5% and 1% levels, respectively.

The overall characteristics effect is -0.035. This means that of the 10.56 percentage point gap in poverty rate, -3.54 percentage points are due to the characteristics effect, or -3.54/10.56 = -33.55% of the gap in poverty incidence is due to characteristics differences. The overall coefficients effect (or discrimination effect) is 0.141. Of the 10.56 percentage point gap, 14.11

percentage points or 14.11/10.56 = 133.55% of the gap in poverty incidence.

In other words, Serbs would be worse off if the differences between their characteristics and those of the Albanian households disappear, and Serbs would be better off if there is no difference in the poverty mitigating effectiveness of those characteristics between the Serbian and Albanian households. When we look at detailed decomposition, it becomes clear that the main reason why Serbs have higher poverty incidence is due to coefficients effect of the constant term. Even though Serbs have better characteristics which can lower poverty incidence, and enjoy stronger poverty mitigating effect of these characteristics relative to Albanians, there is huge baseline gap in poverty incidence between the two ethnic groups, captured by the coefficients effect of the constant term.

2.6 A model for poverty dynamics

Household do not stay all the time in poverty. They have poverty spells, they enter into poverty and get out of it. Stevens (1999) got interest in explaining the duration of these poverty spells for the USA. In her paper, she proposes several models. We keep only one which explains again the logarithm of the income-to-needs ratio as a function exogenous variables but also of dynamic errors. The model is then used to make judgement about the persistence of poverty spells in the USA in order to evaluate the economic situation of an household. The income-to-needs ratio is computed by considering the household income which does not include transfers and by dividing it by the official poverty rate corresponding to the household composition. The basic model is as follows

$$\log\left(\frac{y_{it}}{z}\right) = x_{it}\beta + \delta_i + v_{it} \quad (4)$$

$$\delta_i \sim N(0, \sigma_\delta^2) \quad (5)$$

$$v_{it} = \gamma v_{it-1} + \eta_{it}. \quad (6)$$

The log of the income to needs ratio is explained by individual variables that are time independent as sex and education level, and by individual variables that are time varying. There is a random individual effect δ_i for unobserved heterogeneity. Parameter γ models a permanent effect common to all individuals. We can says that the individuals receive permanent shocks v_{it}. Under a normality assumption for δ_i and η_{it}, Stevens (1999) simulates this model for 20 years and compute the mean period spent in a poverty state. When estimating this model using the PSID data set, we find that the average period spent in a state of poverty is slightly longer if the head of the household is black or if it is a woman.

3 Models for income dynamics

In this section, we give some details about a new and recent concern in empirical work concerning the income distribution: its evolution over time, its dynamic behaviour. Several tools are available for that. We shall detail the approach based on Markov matrices and Markov processes.

In a first step we shall consider simple Markov matrices, detail the significance of income mobility and indicate how Markov matrices can be estimated. We propose some mobility indices together with their asymptotic distribution. We finally indicate how one can introduce explanatory variables for explaining income mobility using a dynamic multinomial logit model.

3.1 Income dynamics

In his presidential address to the European Society for Population Economics, Jenkins (2000) underlines that the income distribution in the UK has experienced great changes during the eighties, but that since 1991, this distribution seems to have remained relatively stable. If the poverty line is defined as half the mean income, the percentage of poor remains relatively stable, while if it is defined as half the mean of 1991 in real term, this percentage decreases steadily. The Gini coefficient remains extremely stable around 0.31-0.32. These figures characterise a cross-section stability in income.

However, since 1991, the UK started the British Household Panel Survey (BHPS). This means that the same household are interviewed between 1991 and 1996 each year. It then become possible to study income dynamics. Jenkins provide an estimation for a transition matrix between income groups at a distance of one year. These groups are defined by reference to a fraction of the mean, fraction taken between 0.5 and 1.5 In lines, we have groups for wave t, and

Table 5: Transition probabilities in percentage

	Period t					
Income group	< 0.5	0.5-0.75	0.75-1.0	1.0-1.25	1.25-1.5	> 1.5
Period $t-1$						
< 0.5	54	30	9	4	2	2
0.5-0.75	15	56	21	5	1	2
0.75-1.0	5	19	48	20	5	3
1.0-1.25	3	6	20	44	20	7
1.25-1.5	2	3	8	25	35	27
> 1.5	1	2	4	6	12	75

in columns groups for wave $t-1$. If we except the very rich who have a probability of 0.75 to remain rich, the other groups have in general a probability less than 0.50 to stay in their original group and a probability of going to the neighbouring group of 0.20 on average. Consequently, there was a large income mobility in dynamics. The percentage of poor remained the same, but the persons in a state of poverty were not the same along the 6 years of the panel.

3.2 Transition matrices and Markov models

How was the previous transition matrix computed? It characterises social mobility, the passage between different social states over a given period of time.

- There are k different possible social states.
- i is the starting state, j the destination state
- p_{ij} is the probability to move from state i to state j during the reference period.

We are in fact introducing a Markov process of order one. It can be used to model

- changes in voting behaviour
- changes of social status between father and son: Prais (1955).
- change in occupational status
- change in geographical regions
- Income mobility between different income classes over one or several years

Let us consider k different states (job status, occupational status, income class, etc...) such that an individual is assigned to only one state at a given time period. We let n_{ij}, $i, j = 1...k$ be the number of individuals initially in state i moving to state j in the next period. We define

$$n_{i.} = \sum_{j=1}^{k} n_{ij}$$

the initial number of people in state i and $n = \sum_{i=1}^{k} n_{i.}$ the total number of individuals in the sample. We define a transition matrix P as a matrix with independent lines which sum up to one, $P = [p_{ij}]$ where p_{ij} represents the conditional probability for an individual to move from state i to state j in the next period. We have $\sum_j p_{ij} = 1$.

Let us call $\pi^{(0)}$ the row vector of probabilities of the k initial states at time 0. The row vector of probabilities at time 1 is given by $\pi^{(1)}$. The relation between $\pi^{(0)}$ and $\pi^{(1)}$ is given by

$$\pi^{(1)} = \pi^{(0)} P,$$

by definition of the transition matrix. From the **Stationarity** Markov assumption we can derive that the transition matrix P is constant over time such that the distribution at time t is given by

$$\pi^{(t)} = \pi^{(0)} P^t.$$

We suppose that the transition matrix has k distinct eigenvalues $|\lambda_1| > |\lambda_2| > .... > |\lambda_m|$. Since P is a row stochastic matrix, its largest left eigenvalue is 1. Consequently, P^t is perfectly defined and converges to a finite matrix when t tends to infinity.

The stationary distribution $\pi* = (\pi_1^*, ..., \pi_k^*)^*$ is a row vector of non negative elements which sum up to 1 such that

$$\pi^* = \pi^* P.$$

This distribution vector is a normalised (meaning that the sum of its entries is 1) left eigenvector of the transition matrix associated with the eigenvalue 1. If the Markov chain is irreducible (it is possible to get to any state from any state) and aperiodic (an individual returns to state i can occur at irregular times), then there is a unique stationary distribution π^* and in this case P^t converges to a rank-one matrix in which each row is the stationary distribution π^*, that is

$$lim_{t\to\infty} P^t = \begin{pmatrix} \pi_1^* \cdots \pi_m^* \\ \cdots \\ \pi_1^* \cdots \pi_m^* \end{pmatrix} = \mathbf{i}'\pi^*$$

with **i** being the unity column vector of dimension k.

Markov processes model the transition between mutually exclusive classes or states. In a group of applications, mainly those coming from the sociological literature, those classes are easy to define because they correspond to a somehow natural partition of the social space. We have for instance social classes, social prestige, voting behaviour or more simply economics job status as working, unemployed, not working. In fact those social statuses are directly linked to dichotomous variables. For studying income mobility, the problem is totaly different because income is a continuous variable that has to be discretised. And there are dozen of ways of discretising a continuous variable.

It is easier to detail the various aspects of Markov processes used to model social mobility, it is easier to start from the case where the classes are directly linked to a discrete variable. We shall investigate income mobility in a second step, detailing at that occasion the specific questions that are raised by discretisation.

3.3 Building transition matrices

When considering income as a continuous random variable, there are several ways to build income classes. Let us start by considering a joint distribution between two income variables $x \in [0, \infty)$ and $y \in [0, \infty)$ with a continuous joint cumulative distribution function $K(x, y)$ that captures the correlation between x and y. These correlations may be intergenerational if x is, say, the father and y the son or intra-generational if x and y are the same sample income given at two points in time. The marginal distribution of x and y are denoted $F(x)$ and $G(y)$ such that $F(x) = F(x, \infty)$ and $G(y) = G(\infty, y)$. We assume that $F(.)$, $G(.)$ and $K(.,.)$ are strictly monotone and the first two moments of x and y exist and are finite.

For m given income class boundaries $0 < \zeta_1 < ... < \zeta_{m-1} < \infty$ and $0 < \xi_1 < ... < \xi_{m-1} < \infty$, we can derive the income transition matrix P related to $K(x, y)$ such that each element p_{ij} could be written as

$$p_{ij} = \frac{Pr(\zeta_{i-1} \leq x \leq \zeta_i \text{ and } \xi_{j-1} \leq y \leq \xi_j)}{Pr(\zeta_{i-1} \leq x \leq \zeta_i)}, \tag{7}$$

where $\zeta_0 = \xi_0 = 0$ and $\zeta_m = \xi_m = \infty$.

Four approaches are recommended in Formby et al. (2004) to construct an income transition matrix.

The first one considers class boundaries as defined exogenously. The resulting matrix is referred as a size transition matrix. With this approach the class boundaries do not depend on a particular income regime or distribution. One major advantage of this method is that it reflects income movements between different income levels. Thus both the exchange of income positions as well as the global income growth are taken into account. In their comparison of mobility dynamics between the US and Germany during the eighties, Formby et al. (2004) set five earning classes and normalised German earning using the US mean earnings to compare mobility in the US to mobility in Germany. We shall see that, on average, there is more mobility in the US than in Germany.

The second approach is recommended when mobility is considered as a relative concept and we want to isolate the effects of global income growth from the effects of mobility. In this case mobility is considered as a re-ranking of individuals among income classes and we'll use quantile transition matrices. The main advantage of this approach is that the transition matrix is bi-stochastic ($\sum_{i=1}^{m} p_{ij} = \sum_{j=1}^{m} p_{ij} = 1$) and the steady state condition is always satisfied. Hungerford (1993) used quantile transition matrices to asses the changes in income mobility in the US in the seventies and the eighties.

The third and fourth approaches include both elements of the absolute and relative approaches to mobility. In fact, class boundaries are computed as percentages of the mean or the median. The resulting matrices are referred as mean transition matrices and median transition matrices. Using British data from the BHPS waves 1-6, Jenkins (2000) estimates mean transition matrices to show the importance of income mobility in the UK society. We have reproduced that matrix in the introduction of this section.

3.4 What is social mobility: Prais (1955)

According to Bartholomew (1982), Prais (1955) was the first paper in economics to study social mobility using a Markov model (see Feller 1950, chap 15, or Feller 1968 for a theory of Markov processes). Prais (1955) considered a random sample of 3500 males aged over 18 from the Social Survey in 1949. He studied mobility between father and son and produced the following Markov transition matrix reproduced in Table 6. The equilibrium distribution is given by

$$\pi'_* = \pi'_* P.$$

Once this distribution is reached, it will be kept for ever. Thus the equilibrium distribution is independent of the starting distribution. It is also independent of the time span. As if P relates the status of sons to that of fathers, the matrix relating that of grandsons to grandfathers is P^2.

There is perfect immobility if a family always stays in the same class. This would correspond to $P = I$. The more mobile is a family, the shorter the period it would stay in the same class.

Let us call n_j the number of families in class j at the beginning of the period. In the second generation, there will be $n_j p_{jj}$, then $n_j p_{jj}^2$ and so on. The average time (measured in number of

Table 6: The Social Transition Matrix in England, 1949
The j^{th} element of row i^{th} gives the proportion of fathers in the i^{th} class whose sons are in the j^{th} social class. Transition from i^{th} class to j^{th} class

		1	2	3	4	5	6	7
1	High Administrative	0.388	0.146	0.202	0.062	0.140	0.047	0.015
2	Executive	0.107	0.267	0.227	0.120	0.206	0.053	0.020
3	Higher grade supervisory	0.035	0.101	0.188	0.191	0.357	0.067	0.061
4	Lower grade supervisory	0.021	0.039	0.112	0.212	0.430	0.124	0.062
5	Skilled manual	0.009	0.024	0.075	0.123	0.473	0.171	0.125
6	Semi skilled manual	0.000	0.013	0.041	0.088	0.391	0.312	0.155
7	Unskilled manual	0.000	0.008	0.036	0.083	0.364	0.235	0.274

Table 7: Actual and equilibrium distributions
of social classes in England, 1949

Class	Fathers	Sons	Equilibrium
High Administrative	0.037	0.029	0.023
Executive	0.043	0.046	0.042
Higher grade supervisory	0.098	0.094	0.088
Lower grade supervisory	0.148	0.131	0.127
Skilled manual	0.432	0.409	0.409
Semi skilled manual	0.131	0.170	0.182
Unskilled manual	0.111	0.121	0.129

generations) is given by

$$1 + p_{jj} + p_{jj}^2 + \cdots = \frac{1}{1 - p_{jj}},$$

with standard deviation:

$$\frac{\sqrt{p_{jj}}}{1 - p_{jj}}.$$

In a perfectly mobile society, the probability of entering a social class is independent of the origin. The matrix P representing perfect mobility has all the elements in each column equal (each row in the notations of Prais). But of course, columns can be different.

We consider a particular society. We compute the equilibrium distribution. The perfectly mobile society that can be compared to it is characterised by a transition matrix that has all its rows equal to the equilibrium distribution π. In other words, from the introduction, this matrix is obtained as the limit of P^t when $t \to \infty$. The least mobile families are those belonging to the top executive (professional) class. The decimal part of the third column indicates the excess of immobility in percentage. Large self recruiting in the top group. The closer to perfect mobility are the Lower grade non-manual.

Table 8: Average number of generations spent in each social class

Class	England today	Mobile Society	Ratio	S.D.
Professional	1.63	1.02	1.59	1.02
Managerial	1.36	1.04	1.30	0.71
Higher grade non-manual	1.23	1.10	1.12	0.54
Lower grade non-manual	1.27	1.15	1.11	0.58
Skilled manual	1.90	1.69	1.12	1.30
Semi-Skilled manual	1.45	1.22	1.19	0.81
Unskilled manual	1.38	1.15	1.20	0.72

A mobility index was later given the name of Prais and is expressed as

$$M_P = \frac{k - tr(P)}{k - 1}$$

The reason is that Prais has shown that the mean exit time from class i (or the average length of stay in class i) is given by $1/(1-p_{ii})$. Since M_P can be rewritten as $M_P = \sum_i (1-p_{ii})/(k-1)$ it is the reciprocal of the harmonic mean of the mean exit times, normalised by the factor $k/(k-1)$.

Shorrocks (1978) gave an axiomatic content to the measurement of social mobility using Markov transition matrices. He studied the properties of existing mobility indices and looked at which axioms would be needed. The existing indices cannot satisfy all these axioms. The conflict comes from the definition of what is a perfectly mobile society when confronted to the requirement that a matrix P is more mobile than P' if some of its off diagonal elements are increased at the expend of the diagonal elements. We note $P \succ P'$. Here are the main available mobility indices. Some of them are posterior to the paper of Shorrocks. In this table,

Table 9: Main mobility indices

Measures	Sources
$M_1(P) = \frac{k - \text{tr}(P)}{k-1}$	Prais (1955), Shorrocks (1978)
$M_3(P) = 1 - \det(P)$	Shorrocks (1978)
$M_4(P) = k - \sum_i \pi_i^* p_{ii}$	Bartholomew (1982)
$M_5(P) = \frac{1}{k-1} \sum_i \pi_i^* \sum_j p_{ij} \lvert i-j \rvert$	Bartholomew (1982)

π^* represents the equilibrium vector of probabilities, the equilibrium distribution.

Shorrocks introduces several axioms that could be imposed over mobility indices and the needed restrictions over transition matrices that could help to insure the compatibility of those axioms.

N Normalisation: $0 \leq M(P) \leq 1$

M Monotonicity: $P \succ P' \Rightarrow M(P) > M(P')$

I Immobility: $M(I) = 0$

SI Immobility: $M(I) = 0$ iff $P = I$

PM Perfect mobility: $M(P) = 1$ if $P = i'x$ with $x'i = 1$.

The index of Bartholomew satisfies (**I**) but not (**SI**), (**N**), (**M**), or (**PM**). The reason is that the axioms (**N**), (**M**), and (**PM**) are incompatible. The basic conflict is thus between (**PM**) and (**M**). This conflict can be removed reasonably by considering transition matrices that are maximal diagonal

$$p_{ii} > p_{ij}, \qquad \forall i, j$$

or quasi maximal diagonal

$$\mu_i p_{ii} > \mu_j p_{ij}, \qquad \forall i, j \quad \text{and } \mu_i, \mu_j > 0.$$

With this last restriction, the Prais index satisfies (I), (SI), and (M).

3.5 Estimating transition matrices

Each row of a transition matrix P defines a multinomial process which is independent of the other rows. Anderson and Goodman (1957) or Boudon (1973, pages146-149) among others proved that the maximum likelihood estimator of each element of P is

$$\hat{P} = [\hat{p}_{ij}] = \left[\frac{n_{ij}}{n_i}\right].$$

This estimator $\hat{p}_{ij}$ is consistent and has variance

$$n_i p_{ij}(1 - p_{ij})/n_i^2 = p_{ij}(1 - p_{ij})/n_i.$$

When n tends to infinity, each row P_i of P tends to a multivariate normal distribution with

$$\sqrt{n_i}(\hat{P}_i - P_i) \xrightarrow{D} N(0, \Sigma_i),$$

where

$$\Sigma_i = \begin{pmatrix} \frac{p_{i1}(1-p_{i1})}{n_i} & \cdots & -\frac{p_{i1}p_{ik}}{n_i} \\ & \ddots & \\ -\frac{p_{ik}p_{i1}}{n_i} & \cdots & \frac{p_{ik}(1-p_{ik})}{n_i} \end{pmatrix}.$$

As each row of matrix P is independent of the others, the stacked vector of the rows P_i verifies:

$$\sqrt{n}(\mathbf{vec}(\hat{P}) - \mathbf{vec}(P)) \xrightarrow{D} N(0, \Sigma),$$

where

$$\Sigma = \begin{pmatrix} \Sigma_1 & \cdots & 0 \\ 0 & \ddots & \\ 0 & \cdots & \Sigma_k \end{pmatrix} \tag{8}$$

is a $k^2 \times k^2$ block diagonal matrix with Σ_i on its diagonal and zeros elsewhere.

3.6 Distribution of indices

A mobility index $M(.)$ is a function of P. Thus its natural estimator will be:

$$\hat{M}(P) = M(\hat{P}),$$

i.e. a function of the estimated transition matrix. A standard deviation for that estimator will be given by a transformation of the standard deviation of the estimators of each elements of the transition matrix P. As the transformation $M(.)$ is most of the time not linear, we will have to use the Delta method to compute it. Let us recall the definition of Delta method in the multivariate case.

Definition 2. *Let us consider a consistent estimator b of $\beta \in \mathbb{R}^m$ such that :*

$$\sqrt{n}(b - \beta) \xrightarrow{D} N(0, \Sigma).$$

Let us consider a continuous function g having its first order derivatives. The asymptotic distribution of $g(\beta)$ is given by

$$\sqrt{n}(g(b) - g(\beta)) \xrightarrow{D} N(0, \nabla g(\beta)' \Sigma \nabla g(\beta)),$$

where $\nabla g(\beta)$ is the gradient vector of g evaluated in β.

Let's verify that the mobility index $M(.)$ fulfills the Delta method assumptions. First we have shown previously that $\hat{P}$ is a consistent estimator of P. Then, from Trede (1999) we have that the asymptotic distribution of $\hat{P}$ is normal with independent rows: each row follows a multinomial distribution, hence for $n \to \infty$

$$\sqrt{n}(\mathbf{vec}(\hat{P}) - \mathbf{vec}(P)) \xrightarrow{D} N(0, \Sigma),$$

where Σ is defined in (8).

Therefore the delta method is applicable and we can derive then that

$$\sqrt{n}(M(\hat{P}) - M(P)) \to N(0, \sigma_M^2),$$

with

$$\sigma_M^2 = (DM(P))\Sigma(DM(P))'$$

Moreover,

$$DM(P) = \frac{\partial M(P)}{\partial \mathbf{vec}(P')'}$$

is a m^2 vector and $\mathbf{vec}(P)$ is the row vector emerging when the rows of P are put next to each other.

Trede (1999) has computed the derivation $DM(P)$ for several mobility indices and has summarised them in Table 2 to make easy asymptotic estimation of these mobility indices.

Obviously, $DM(P)$ and Σ are unknown and need to be estimated using the estimation of the matrix $\hat{P} = [\hat{p}_{ij}]$ and $\hat{p}_i$. Therefore we replace each element p_{ij} in $DM(P)$ and in Σ by its estimator $\hat{p}_{ij}$. Thus an estimation of σ_M would be $\hat{\sigma}_M^2 = (DM(\hat{P}))\hat{\Sigma}(DM(\hat{P}))'$.

Table 10: Transition matrix mobility measures and their derivative

Index	DM(P)
M_P	$-\frac{1}{m-1}vec(I)'$
M_E	$-\frac{1}{m-1}vec(\sum_i \check{P}_\lambda^{,})'$
M_D	$-sign(det(P))vec(\widetilde{P}')'$
M_2	$-vec(\check{P}'_{\lambda_2})'$
M_B	$\left[(\sum_{ij} p_{ij}\pi_s(z_{ti}-z_{mi})\lvert i-j\rvert) + \pi_s(\lvert s-t\rvert - \lvert s-m\rvert)\right]_{s,t=1\ldots m}$
M_U	$-\frac{m}{m-1}\left[(\sum_i \pi_s(z_{ti}-z_{mi})(1-p_{ii})) - (\delta_{st}\pi_s - \delta_{sm}\pi_m\right]_{s,t=1\ldots m}$

$\tilde{P}$ is the matrix of cofactors of P, $\check{P}_\lambda = \frac{\partial|\lambda|}{\partial P}$, Z is a fundamental matrix of P, $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

3.7 Modelling individual heterogeneity using a dynamic multinomial logit model

To introduce observed heterogeneity, we have to consider a dynamic multinomial logit model which explains the probability that an individual i will be in state k when he was in state j in the previous period as a function of exogenous variables. Using the model of Honoré and Kyriazidou (2000) and Egger et al. (2007), but without individual effects, the unobserved propensity to select option k among K possibilities for individual i at time t can be modelled as:

$$s^*_{kit} = \alpha_k + x_{it}\beta_k + \sum_{j=1}^{K-1} \gamma_{jk} \mathbb{1}\{s_{i,t-1} = j\} + \epsilon_{kit}. \tag{9}$$

The observed choice s_{it} is made according to the following observational rule

$$s_{it} = k \text{ if } s^*_{kit} = \max_l(s^*_{lit}).$$

If the ϵ_{kit} are identically and independently distributed as a Type I extreme value distribution (also Gumbel distribution), then the probability that individual i is in state k at time t when he was in state j at time $t-1$ has a simple analytical expression:

$$\Pr(s_{it} = k | s_{i,t-1} = j, x_{it}) = \frac{\exp(\alpha_k + x_{it}\beta_k + \gamma_{jk})}{\sum_{l=1}^{K} \exp(\alpha_l + x_{it}\beta_l + \gamma_{jl})}, \tag{10}$$

where x_{it} are explanatory the variables. α_k is a category specified constant common to all individuals. γ_{jk} is the coefficient on the lagged dependent variable attached to the transition between state j to state k. As the probabilities have to sum to 1, we must impose a normalisation. We can chose $\alpha_K = \gamma_K = 0$, $\beta_K = 0$. This model can be estimated using the package `VGAM` in `R`. Recall that the Extreme value distribution is:

$$f(x) = \frac{1}{\sigma}\exp(-(x-\mu)/\sigma)\exp(\exp(-(x-\mu)/\sigma)).$$

We present in Table 11 the estimation of a model explaining the dynamic transitions between job statuses in the UK using the BHPS over 1991-2008. This is an unbalanced panel and the software can deal this aspect of the data. We chose "non-participating" as the baseline. The

Table 11: Estimation of a dynamic Multinomial Logit model for job status transitions

			Marginal effects	
Destination status	Working	Unemployed	Working	Unemployed
Origin: Working	4.395 (0.031)	2.004 (0.058)	0.191	-0.064
Origin: Unemployed	1.855 (0.054)	3.080 (0.068)	0.018	0.036
intercept	3.950 (1.725)	19.245 (2.177)		
$\log age$	−2.204 (0.974)	−10.709 (1.241)	0.172	-0.245
$(\log age)^2$	0.333 (0.137)	1.444 (0.176)	-0.021	0.032
high educ	0.757 (0.038)	−0.164 (0.055)	0.047	-0.025
mid educ	0.464 (0.035)	−0.155 (0.048)	0.030	-0.017
gender	−2.111 (0.048)	−2.501 (0.057)	-0.049	-0.013
N. Obs	115 991			
log-likelihood	-32 019 without time dummies			
log-likelihood	-31 965 with time dummies			

We used the routine `vglm` of the package `VGAM` of `R` to estimate this equation. Observations are pooled. Standard errors in parentheses.

interpretation of these coefficients is complex. It is much easier to compute marginal effects. Marginal effects are defined as the derivative of the base probability $\Pr(s_{it} = k|s_{i,t-1} = j, x_{it})$ with respect to each exogenous variable. A marginal effect does not have necessarily the same sign as the coefficient of the variable. Marginal effects are computed as

$$\frac{\partial \Pr(s = k)}{\partial x} = \Pr(s = k)[\beta_k - \sum_j \Pr(s = j)\beta_j],$$

where the mean value is taken for $\Pr(s = j)$ and that probability is computed using (10).

In this example, marginal effects are documented in the last two column of Table 11. Age has an U-shaped effect on the probability of being employed while it has an inverted U-shaped effect on the probability of being unemployed. Education has a positive effect on the probability of working and obviously a negative effect on the probability of being unemployed. But females have both lower probability of being employed or unemployed, which means that they mostly prefer to stay at home.

3.8 Transition matrices and individual probabilities

A Markov transition matrix is usually estimated by maximum likelihood which is shown to correspond to (see the seminal paper of Anderson and Goodman 1957 or the appendix in Boudon 1973):

$$\hat{p}_{ij}(t) = \frac{n_{ij}(t)}{\sum_j n_{ij}(t)},$$

where $n_{ij}(t)$ the number of individuals in state i at time $t-1$ moving to state j at time t. When there are more than two periods and if the process is homogenous, the maximum likelihood estimator is obtained by averaging the $\hat{p}_{ij}^t$ obtained between two consecutive periods. Of course, this estimator is not at ease when the panel is incomplete.

The dynamic multinomial logit model can be seen as an alternative to estimate a Markov transition matrix. We can exploit the conditional probabilities given (10) that we recall here

$$\Pr(s_{it} = k | s_{i,t-1} = j, x_{it}) = \frac{\exp(\alpha_k + x_{it}\beta_k + \gamma_{jk})}{\sum_{l=1}^{K} \exp(\alpha_l + x_{it}\beta_l + \gamma_{jl})},$$

to reconstruct the first $K-1$ lines of the transition matrix P and using the identification restrictions $\alpha_K = \gamma_K = 0$, $\beta_K = 0$ for the last line. The last column of the matrix is found using the constraint that each line sums up to 1. Of course, in order to obtain a single probability, we have to take the covariates at their sample mean. Using the estimated model as reported in Table 11, we derived two transition matrices computed at the mean value of the exogenous variables (except for gender), one for males, one for females. We report the results in Table 12. If the average of these two matrices look pretty the same as the marginal one given in Table 12, there are huge differences between males and females for the unemployed and not working lines. Males are almost always participating. Their only alternative is between working or being unemployed. Females mostly do not stay unemployed. They either go back to work or leave the labour market. When they have left the labour market, they have a strong tendency to stay in this state.

Table 12: Implicit conditional transition matrices

	Working	Unemployed	Non-particip.
	Males		
Working	0.973	0.021	0.005
Unemployed	0.533	0.429	0.038
Non-particip.	0.591	0.038	0.263
	Females		
Working	0.943	0.015	0.043
Unemployed	0.466	0.264	0.269
Non-particip.	0.207	0.269	0.757

4 Introducing and illustrating quantile regressions

A regression model gives the link, either linear or non-linear, that exists between an endogenous variable and one or more exogenous variables. In the most simplest case, the regression line represents a linear conditional expectation. A non-parametric regression explains a conditional expectation in a non-linear way, without specifying a predefined non-linear relationship. But none of these regressions is designed to explain the quantiles of the conditional distribution of the endogenous variable. The quantile regression was introduced in econometrics by Koenker and Basset (1978); it gives the adequate tool to explain the compete evolution of the conditional distribution of the endogenous variable. The basic principle of the quantile regression is simple, but its numerical implementation is more complex. In particular, standard deviations are not easy at all to compute. So most of the time, bootstrap is used to obtain numerical values.

4.1 Classical quantile regression

Let us consider a linear regression model expressed as

$$y = x'\beta + e.$$

In the usual linear regression the assumption is $\mathrm{E}(e|x) = 0$. And no other special assumption considering the distribution of y is needed.

A quantile regression model considers a similar linear regression, but adds the fact that this regression can be estimated for every predefined quantile τ of the endogenous variable. So for the τ^{th} quantile, we have now the new regression:

$$y_i = x_i'\beta_\tau + e_{i\tau}, \tag{11}$$

where the parameter to be estimated are the $\beta_\tau' = (\beta_{0_\tau}, \cdots, \beta_{k_\tau})$. A coherent definition of this regression requires no longer that $\mathrm{E}(e_i|x_i) = 0$, but that the τ^{th} quantile of e is equal to zero. If $f(.)$ is the density of e, this means that

$$\int_{-\infty}^{0} f_\tau(e_i|x) d\, e_i = \tau. \tag{12}$$

In other words, if the distribution is $F(.)$, let us note $q_\tau(x)$ the quantile of level τ that we define as

$$q_\tau(y) = F^{-1}(\tau).$$

A quantile regression explains this quantile by a linear combination of the x

$$q_\tau(y) = x'\beta.$$

We shall first note that if F is a cumulative normal, this model will provide no valid new information, because first the mean and the median are identical for this distribution and second that its conditional quantiles are straight lines. We have to get out of this traditional framework in

order to get a valid and interesting model. A first possibility is to have heteroscedastic errors, for instance normal errors but with a non-constant variance σ_i^2; for instance this last parameter can be function of exogenous variables. A more radical assumption is simply to have distributional restriction for F and thus to use a semi-parametric framework. For this, we define the error function

$$\rho_\tau(u) = \begin{cases} u\tau & \text{if } u > 0, \\ u(\tau - 1) & \text{if } u \leq 0. \end{cases}$$

We then look for the value of β that minimises, not a quadratic distance of the error term, but the more peculiar function

$$\hat{\beta}_\tau = \text{argmin} \sum_i \rho_\tau(y_i - x_i'\beta).$$

This has to be solved using quadratic programming. This approach was first proposed by Koenker and Basset (1978). It is very difficult to compute standard errors.

4.2 Bayesian inference

Other routes are possible to define a quantile regression. Using a Bayesian framework, Yu and Moyeed (2001) show that estimating the quantile of y is equivalent to estimating the localisation parameter of an asymmetric Laplace distribution. This leads easily to writing the likelihood function as:

$$L_\tau(\beta; y, x) \propto \tau^n (1 - \tau)^n \exp\{-\sum_i \rho_\tau(y_i - x_i'\beta)\},$$

which is used by Yu and Moyeed (2001) to evaluate the posterior density of β. In this framework, it becomes easy to estimate standard deviations and compute confidence intervals.

4.3 Quantile regression using R

There are several packages in R for computing quantile regressions. Different approaches are possible.

The package `library(quantreg)` contains all the necessary tools for semi-parametric quantile regression. The basic command is `rq`. This package corresponds to the original method of Koenker and Basset (1978).

For a fully non-parametric approach, we need the general package `library(np)`. Then the routine is `npqreg`. There is an example using an Italian income panel which should be investigated seriously.

In a Bayesian approach, the package is `library(MCMCpack)`, and then we can use `MCMCquantreg`. The prior density for β is normal. The quantile has to be given. By default $\tau = 0.5$. There should be as many runs as quantiles needed.

4.4 Analysing poverty in Vietnam

Nguyen et al. (2007) use the Vietnam Living Standards Surveys from 1993 and 1998 to examine inequality in welfare between urban and rural areas in Vietnam. Their measure of welfare is the log of real per capita household consumption expenditure (RPCE), presumably because it is easier to have better data on consumption than on income. Their basic quantile regression for quantile τ is

$$q_\tau(y|x) = \beta_\tau^0 + x'\beta_\tau + urban(\gamma_\tau^0 + X\gamma_\tau) + south(\delta_\tau^0 + X\delta_\tau^0) + urban \times south\theta_\tau^0,$$

where y is the log of RPCE. They first run a regression with including only regional and urban dummies to highlight the differences. They will include the other explanatory variables later on. The coefficients labelled *base* are estimates of log RPCE for the base case: a northern

Table 13: Estimates of the urban-rural gap at the mean and at various quantiles

		OLS	5th	25th	50th	75th	95th
1993	base	7.25	6.62	6.98	7.24	7.51	7.96
	p-value	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	urban	0.52	0.34	0.42	0.51	0.59	0.74
	p-value	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	south	0.20	-0.02	0.15	0.22	0.29	0.36
	p-value	(0.00)	(0.83)	(0.00)	(0.00)	(0.00)	(0.00)
1998	base	7.56	6.85	7.26	7.53	7.84	8.35
	p-value	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	urban	0.72	0.60	0.64	0.72	0.79	0.93
	p-value	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	south	0.15	-0.05	0.12	0.17	0.21	0.22
	p-value	(0.00)	(0.54)	(0.00)	(0.00)	(0.00)	(0.00)

Bootstrapped standard errors were computed on 1000 replications and account for the effects of clustering and stratification. The p-values are for two-sided tests based on asymptotic standard normal distributions of the z-ratios under the null hypothesis that the corresponding coefficients are zero.

rural household. There is an increased dispersion for the urban households. The quantiles are not linear. The 95th quantile is much higher than expected. This dispersion is even increased in 1998 compared to 1993. On the contrary, the dispersion between north and south is much smoother and tends to decrease over time. These differences are significant for all the quantiles, except for the 5th which is not significant. Poverty is the same in both regions as the 5th quantile for the dummy south is not significant.

When the model is estimated in full, including all covariates, the apparent advantages of the south shown in the Table 13 disappear. So the differences are fully explained by these covariates.

The increase in the rural-urban gap over the period (as shown in Table 13) is due to changes in the distributions of the covariates and in changes in the returns of the covariates.

Nguyen et al. (2007) do not report in their main text the full regression, but simply comment on some covariates such as education. These comments are as follows:

Returns to education across the quantiles vary between the North and South. The returns to education show a marked increase at the upper quantiles in the South in 1993 for urban households. A comparable pattern is not seen in the North. In 1998, the upward sloping returns to education in the South are evident in both the urban and rural sectors. The North in 1998 continues to show a more stable pattern of returns across the quantiles with a huge blip up for the very top urban households. Finally, returns to education increased, most substantially in the South, over the five-year period covered by our data.

The urban - rural gap is thus mainly due to differences in education for the smallest quantiles. However, for the higher quantiles of the income distribution, the rural - urban gap is mainly due to differences in the yield of education. We can conclude that fighting against poverty goes through developing education in rural areas.

5 Marginal quantile regressions

In this section, we present another view of the quantile regression which was promoted by Firpo et al. (2009). Using a transformation of the endogenous variable, these authors manage to define a new concept of quantile regression, the marginal quantile regression, which proves to be very useful for computing an Oaxaca decomposition, which otherwise is quite difficult to define for the usual conditional quantile regression. This section draws on Lubrano and Ndoye (2012).

5.1 Influence function

The Influence Function (IF), first introduced by Hampel (1974), describes the influence of an infinitesimal change in the distribution of a sample on a real-valued functional distribution or statistics $\nu(F)$, where F is a cumulative distribution function. The IF of the functional ν is defined as

$$IF(y,\nu,F) = \lim_{\epsilon\to 0}\frac{\nu(F_{\epsilon,\Delta_y}) - \nu(F)}{\epsilon} = \frac{\partial \nu(F_{\epsilon,\Delta_y})}{\partial \epsilon}|_{\epsilon=0} \tag{13}$$

where $F_{\epsilon,\Delta_y} = (1-\epsilon)F + \epsilon\Delta_y$ is a mixture model with a perturbation distribution Δ_y which puts a mass 1 at any point y. The expectation of IF is equal to 0.

Firpo et al. (2009) make use of (13) by considering the distributional statistics $\nu(.)$ as the quantile function $(\nu(F) = q_\tau)$ to find how a marginal quantile of y can be modified by a small change in the distribution of the covariates. They make use of the Recentered IF (RIF), defined as the original statistics plus the IF so that the expectation of the RIF is equal to the original statistics.

Considering the τ^{th} quantile q_τ defined implicitly as $\tau = \int_{-\infty}^{q_\tau} dF(y)$, Firpo et al. (2009)

show that the IF for the quantile of the distribution of y is given by

$$IF(y, q_\tau(y), F) = \frac{\tau - \mathbb{1}(y \leq q_\tau)}{f(q_\tau)},$$

where $f(q_\tau)$ is the value of the density function of y evaluated at q_τ. The corresponding RIF is simply defined by

$$RIF(y, q_\tau, F) = q_\tau + \frac{\tau - \mathbb{1}(y \leq q_\tau)}{f(q_\tau)}, \tag{14}$$

with the immediate property that

$$\mathrm{E}\left(RIF(y, q_\tau)\right) = \int RIF(y, q_\tau) f(y) dy = q_\tau.$$

5.2 Marginal quantile regression

The illuminating idea of Firpo et al. (2009) is to regress the RIF on covariates, so the change in the marginal quantile q_τ is going to be explained by a change in the distribution of the covariates by means of a simple linear regression:

$$\mathrm{E}[RIF(y, q_\tau|X)] = X\beta + \epsilon. \tag{15}$$

They propose different estimation methods: a standard OLS regression (RIF-OLS), a logit regression (RIF-Logit) and a nonparametric logit regression. The estimates of the coefficients of the unconditional quantile regressions, $\hat{\beta}_\tau$ obtained by a simple Ordinary Least Square (OLS) regression (RIF-OLS) are as follows:

$$\hat{\beta}_\tau = \left(X'X\right)^{-1} X' \widehat{RIF}(y; q_\tau). \tag{16}$$

The practical problem to solve is that the RIF depends on the marginal density of y. Firpo et al. (2009) propose to use a non-parametric estimator for the density and the sample quantile for q_τ so that an estimate of the RIF for each observation is given by

$$\widehat{RIF}(y_i; q_\tau) = \hat{q}_\tau + \frac{\tau - \mathbb{1}(y \leq \hat{q}_\tau)}{\hat{f}(\hat{q}_\tau)}.$$

Standard deviations of the coefficients are given by the standard deviations of the regression. In Lubrano and Ndoye (2012), we propose a Bayesian approach to this problem where the marginal density of y is estimated using a parametric mixture of densities.

References

Anderson, T. W. and Goodman, L. A. (1957). Statistical inference about Markov chains. *The Annals of Mathematical Statistics*, 28(1):89–110.

Atkinson, A. (2003). Income inequality in oecd countries: Data and explanations. Working Paper 49: 479–513, CESifo Economic Studies, CESifo.

Bailar, B. A. . (1991). Salary survey of u.s. colleges and universities offering degrees in statistics. *Amstat News*, 182:3–10.

Bartholomew, D. (1982). *Stochastic Models for Social Processes*. Wiley, London, 3rd edition.

Bauwens, L., Lubrano, M., and Richard, J.-F. (1999). *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, Oxford.

Bazen, S. and Joutard, X. (2013). The Taylor decomposition: A unified generalization of the Oaxaca method to nonlinear models. Technical report, AMSE.

Bhaumik, S. K., Gang, I. N., and Yun, M.-S. (2006a). A note on decomposing differences in poverty incidence using regression estimates: Algorithm and example. Working Paper 2006-33, Department of Economics, Rutgers University, Rutgers University.

Bhaumik, S. K., Gang, I. N., and Yun, M.-S. (2006b). A note on decomposing differences in poverty incidence using regression estimates: Algorithm and example. Discussion Paper No. 2262, IZA, Bonn, Germany. Available at SSRN: http://ssrn.com/abstract=928808.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455.

Boudon, R. (1973). *Mathematical Structures of Social Mobility*. Elsevier, Amsterdam.

Bourguignon, F., Ferreira, F. H. G., and Leite, P. G. (2008). Beyond oaxaca-blinder: Accounting for differences in household income distributions. *Journal of Economic Inequality*, 6(2):117–148.

Cowell, F. (1995). *Measuring Inequality*. LSE Handbooks on Economics Series. Prentice Hall, London.

Egger, P., Pfaffermayr, M., and Weber, A. (2007). Sectoral adjustment of employment to shifts in outsourcing and trade: evidence from a dynamic fixed effects multinomial logit model. *Journal of Applied Econometrics*, 22(3):559–580.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. Wiley Series in Probability and Statistics. Wiley and Sons, New-York, 3rd edition.

Fergusson, T. S. (1967). *Mathematical Statistics: a decision theoretic approach*. Academic Press, New York.

Firpo, S., Fortin, N. M., and Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, 77(3):953–973.

Formby, J. P., Smith, W. J., and Zheng, B. (2004). Mobility measurement,transition matrices and statistical inference. *Journal of Econometrics*, 120:181–205.

Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, 52:761–765.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hlavac, M. (2014). oaxaca: Blinder-Oaxaca decomposition in R. Technical report, Harvard University.

Honoré, B. E. and Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68(4):839–874.

Hungerford, T. L. (1993). U.S. income mobility in the seventies and eighties. *Review of Income and Wealth*, 31(4):403–417.

Jann, B. (2008). The blinder-oaxaca decomposition for linear regression models. *The Stata Journal*, 8(4):453–479. Standard errors for the Blinder-Oaxaca decomposition. Handout for the 3rd German Stata Users Group Meeting, Berlin, April 8 2005.

Jenkins, S. P. (2000). Modelling household income dynamics. *Journal of Population Economics*, 13:529–567.

Juhn, C., Murphy, K. M., and Pierce, B. (1993). Wage inequality and the rise in returns to skill. *The Journal of Political Economy*, 101(3):410–442.

Koenker, R. and Basset, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Lubrano, M. and Ndoye, A. J. (2012). Bayesian unconditional quantile regression: An analysis of recent expansions in wage structure and earnings inequality in the u.s. 1992-2009. Technical report, GREQAM.

Nguyen, B. T., Albrecht, J. W., Vroman, S. B., and Westbrook, M. D. (2007). A quantile regression decomposition of urban–rural inequality in vietnam. *Journal of Development Economics*, 83(2):466–490.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14:693–709.

Prais, S. J. (1955). Measuring social mobility. *Journal of the Royal Statistical Society, Series A, Part I*, 118(1):56–66.

Radchenko, S. I. and Yun, M.-S. (2003). A bayesian approach to decomposing wage differentials. *Economics Letters*, 78(3):431–436.

Shorrocks, A. F. (1978). The measurement of mobility. *Econometrica*, 46(5):1013–1024.

Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. *Econometrica*, 48(3):613–625.

Stevens, A. H. (1999). Climbing out of poverty, falling back in. *The Journal of Human Ressources*, 34(3):557–588.

Trede, M. (1998). Making mobility visible: a graphical device. *Economics Letters*, 59:77–82.

Trede, M. (1999). Statistical inference for measures of income mobility. *Jahrbucher fur Nationaloknomie und Statistik*, 218:473–490.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics and Probability Letters*, 54:437–447.

Yun, M.-S. (2004). Decomposing differences in the first moment. *Economics Letters*, 82(2):275–280.

6 Appendix

A Quantile regressions in full

A.1 Introduction

Considérons un échantillon d'une variable aléatoire Y et sa densité $f(y)$. On va définir la moyenne comme

$$\hat{\mu} = \int y f(y) dy$$

Si $F(.)$ est la distribution de Y, alors la médiane sera

$$q_{0.50}(y) = F_y^{-1}(0.50)$$

On peut définir de la même manière les autres quantiles.

Considérons maintenant un échantillon bivarié de deux variables aléatoires Y et X distribués conjointement selon $f(y, x)$. Si $f(y|x)$ est la distribution conditionnelle de y si x, alors l'espérance conditionnelle $\mathrm{E}(y|x)$ se définit comme

$$\mathrm{E}(y|x) = \int y f(y|x) dy$$

qui va prendre autant de valeurs différentes que x. Il s'agit donc d'une fonction. Si $F(.)$ est la distribution Normale de moyenne (μ_y, μ_x) et de variance $\sigma_y^2, \sigma_{yx}, \sigma_x^2$, alors la fonction de régression se note simplement par propriétés de Normale bivariée

$$\mathrm{E}(y|x) = \mu_y + \frac{\sigma_{yx}}{\sigma_x^2}(x - \mu_x).$$

On exprime donc l'espérance conditionnelle comme une fonction linéaire de x. Si l'on s'intéresse maintenant aux quantiles conditionnels dans cette même normale

$$q_p(x) = \mu_y + \frac{\sigma_{yx}}{\sigma_x^2}(x - \mu_x) + \Phi^{-1}(p)\sqrt{\sigma_y^2 - \frac{\sigma_{x,y}^2}{\sigma_x^2}}.$$

Le cas de la Normale est très particulier car $q_{0.50}(x) = \mathrm{E}(y|x)$ et les autres fonctions quantiles sont des droites parallèles étant donné que $\Phi^{-1}(0.50) = 0$. Le quantile conditionnel est la moyenne conditionnelle corrigée par la valeur du quantile de la normale standardisée multiplié par la racine carrée de la variance conditionnelle.

L'intérêt de la régression quantile introduite par Koenker and Basset (1978) c'est que dès que l'on sort du cadre normal, les fonctions quantiles ne sont plus des fonctions linéaires de X. On prendra comme exemple le modèle hétéroskédastique

$$\begin{aligned} Y_t &= 2 + X_t + \exp(-X_t)\epsilon_t \\ & \quad X \sim N(0,1) \\ & \quad \epsilon_t \sim N(0,1) \end{aligned}$$

que l'on a utilisé pour simuler un échantillon. Alors on peut comparer les deux types de régression dans le cas normal et dans le cas hétéroskédastique sur un échantillon simulé On a des résultats

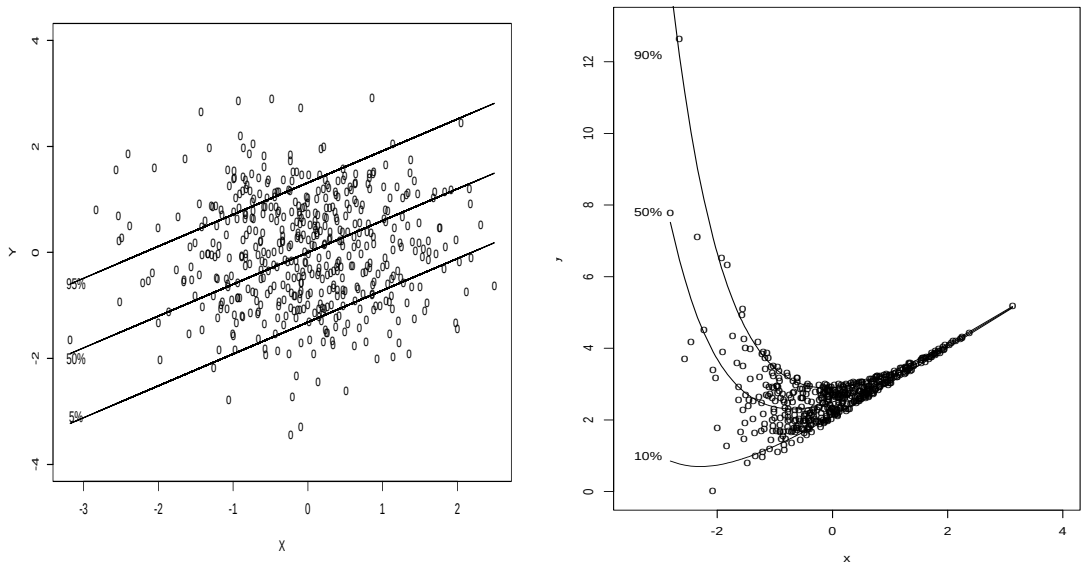

Figure 2: Quantile in a standard and in a heteroskedastic regression

analytiques dans des cas particuliers comme

$$Y = m(t) + m(t)\epsilon_t \qquad \epsilon_t \sim N(0,1).$$

Alors

$$q_p(t) = m(t) + m(t)\Phi^{-1}(p).$$

On peut remarquer que si le modèle n'était pas hétéroskédastique tous les quantiles seraient parallèles.

A.2 Applications

L'analyse de la dispersion des salaires en économie du travail et l'analyse de la distribution des revenus. La raison c'est que l'influence d'une variable, ne serait-ce que le temps, peut être très différente sur les groupes à faible ou fort salaire/revenus. Implémenter une politique fiscale ou sociale quand on veut cibler certains groupes.

Bailar (1991) a étudié l'évolution du salaire de 459 professeurs de statistique en prenant comme variable explicative le nombre d'années depuis laquelle ils avaient la tenure. On constate que les plus riches (quantile 0.75) sont devenu plus riches au cours du temps alors que les autres ont eu un revenu plus stationnaire.

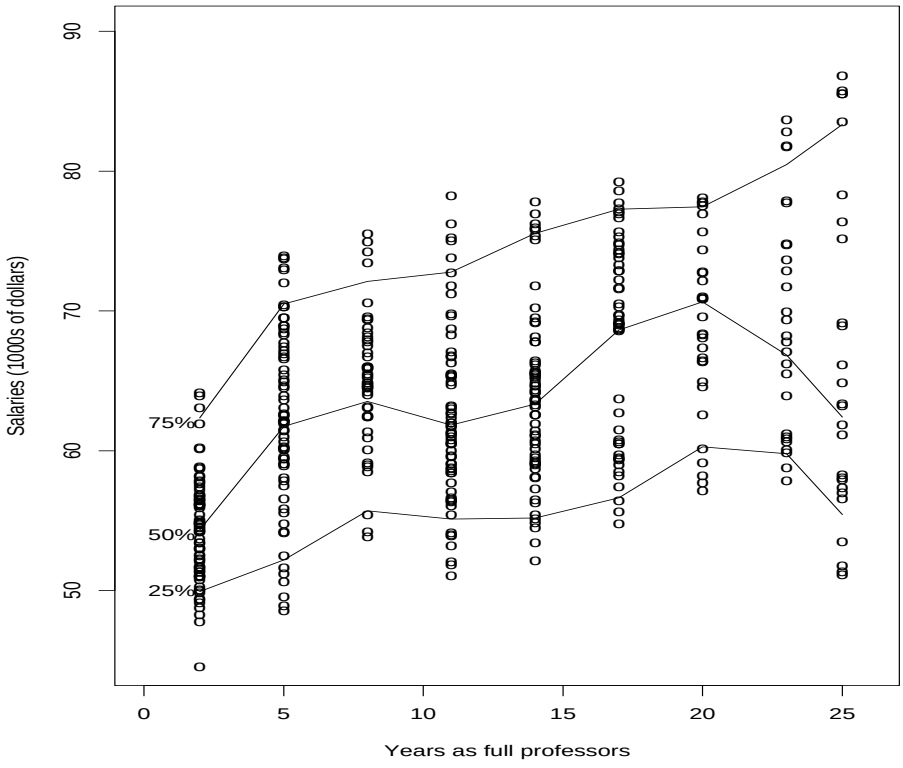

Figure 3: Wage distribution as a function of experience since tenure

Trede (1998) make use of a non-parametric quantile regression for explaining the income distribution of year t as a function of the income distribution of year $t-1$. The two case study are Germany and the USA.

Trede has considered three sample period for each country. The samples concern household income in both countries. The reference year is 1984. It serves to normalise the other years, using the median. If the distribution does not change, all the quantiles will be identical to the 45° line. The distribution of the first period completely determines the distribution in the second period. If on the contrary, the distribution of the second period is independent of the distribution in the first period, the quantile will be horizontal. This is a sign of income mobility.

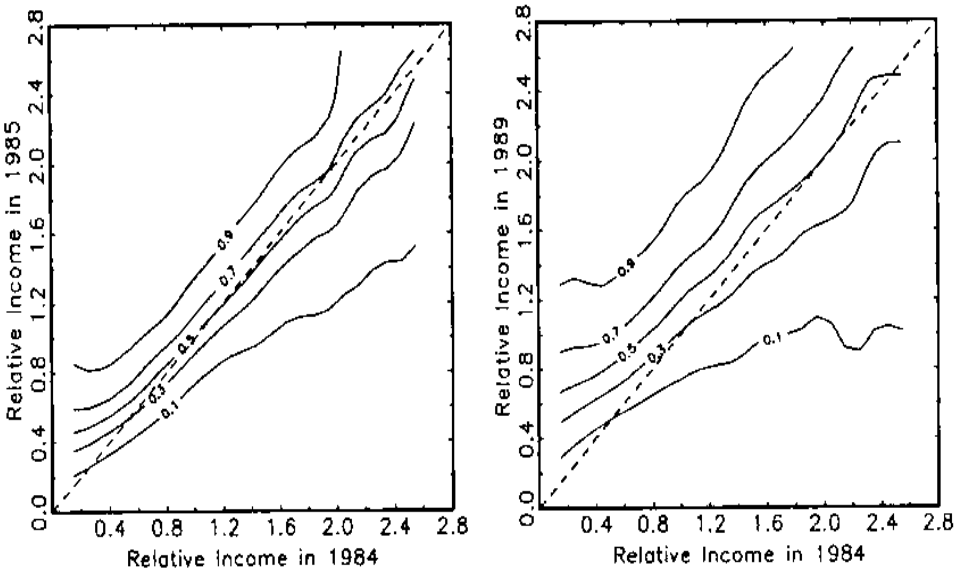

Fig. 1. Quantile regression of (relative) income in 1985 and 1989 on (relative) income in 1984, Germany.

Figure 4: Income mobility in Germany: 1984-1985 and 1984-1989

The first conclusion is that income mobility is more important in the long term than in the short term. This is a natural finding. What is more unexpected is that income mobility is greater in Germany than what it is in the USA.

B Statistical inference

Dans le papier original de Koenker and Basset (1978), l'expression des quantiles d'une distribution est tirée d'un exercice de Fergusson (1967) qui demande de préciser les paramètres d'une fonction de perte en valeur absolue dont la perte espérée associe est minimum pour le quantile. C'est un exercice classique dans la littérature Bayésienne.

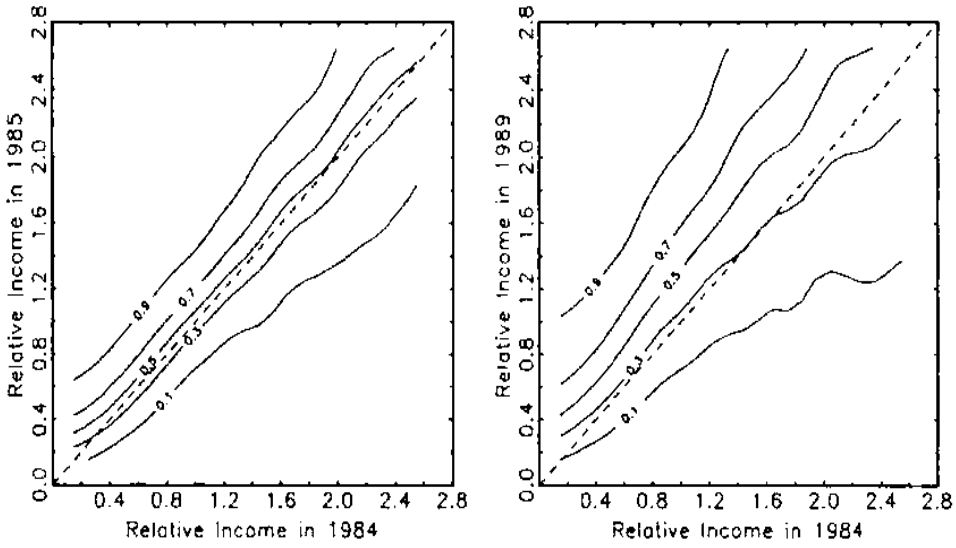

Fig. 2. Quantile regression of (relative) income in 1985 and 1989 on (relative) income in 1984, United States.

Figure 5: Income mobility in the USA: 1984-1985 and 1984-1989

Il est bon de rappeler qu'un estimateur Bayésien est un estimateur qui minimise la perte espérée a posteriori. Si la fonction de perte est quadratique, l'estimateur Bayésien sera l'espérance a posteriori. Si la fonction de perte est de la forme

$$l(x,\theta) = \begin{cases} c_1(x-\theta) & \text{si } x \geq \theta \\ c_2(\theta - x) & \text{si } x < \theta \end{cases}$$

l'estimateur Bayésien sera le fractile $c_2/(c_1 + c_2)$ de la distribution a posteriori de θ. Voir Bauwens et al. (1999).

Dans l'article de Koenker et Bassett on a la définition suivante pour calculer la valeur b du quartile p:

$$Q_p(y) = \underset{ArgMin}{b}[\sum_{y_t \geq b} p|y_t - b| + \sum_{y_t < b} |y_t - b|]$$

On cherche donc bien un estimateur b qui va minimiser une fonction de perte espérée. Pour la régression quantile, on généralise cette expression à:

$$R_p(y) = \underset{ArgMin}{b}[\sum_{y_t \geq x_t'b} p|y_t - x_t'b| + \sum_{y_t < x_t'b} |y_t - x_t'b|]$$

Le seul problème, et il est de taille, c'est que maintenant on doit faire de l'optimisation en dimension k, la taille de b alors que la fonction de perte n'est pas différentiable. Le premier résultat

de Konker et Basset c'est que cet estimateur est unique quand X est de rang plein. Le second, c'est que l'estimateur trouvé a une distribution asymptotique Normale qui est tirée de celle des estimateurs naturels des quantiles.

La littérature a ensuite parlé de check function. Ainsi dans l'article qui nous occupe, de la fonction de perte $\rho_p(z) = p|z|$, on passe à l'expression équivalente

$$\begin{aligned} \rho_p(z) &= pz\mathbb{1}(z > 0) - (1-p)z\mathbb{1}(z \leq 0) \\ &= z \times (p - \mathbb{1}(z < 0)) \end{aligned}$$

L'estimation du fractile p se fera en minimisant la perte espérée

$$\mathrm{E}_{y|x}[\rho_p(y - x_t'\beta)]$$

ou

$$\min_{\beta} \sum \rho_p(y_t - x_t'\beta)$$

Comme la fonction $\rho_p(z)$ n'est pas différentiable en zero, on est obligé de passer par un programme linéaire de la forme

$$\begin{aligned} z &= \text{ArgMin } c'z \\ Az &= y \\ z &\geq 0 \end{aligned}$$

qui n'est pas très efficace quand on a un grand nombre d'observations. Le problème vient de la contrainte de positivité. On va alors remplacer le programme original par

$$\underset{x}{\text{Min}}\ c'x - \mu \sum_m \ln x_m$$

Ces procédures sont implémentées dans la librairie QreG de Thierry Roncalli sous Gauss. Il semble que les écart-types ne soient pas disponibles.

B.1 Inférence Bayésienne

Une régression linéaire s'écrit

$$y = \theta(x) + \epsilon \qquad \mathrm{E}(\epsilon) = 0$$

Estimer $q_p(x)$ (le p-quantile de la distribution conditionelle $f(y|x)$) revient à estimer la régression

$$y = \theta(x) + \epsilon \qquad q_p(\epsilon) = 0$$

On se rend compte alors que pour conduire l'inférence dans ce type de modèle, il suffit de considérer une distribution asymétrique adéquate pour ϵ. Yu and Moyeed (2001) adoptent une distribution de Laplace asymétrique qui conduit à la fonction de vraisemblance

$$f_p(y|\beta) = p^n(1-p)^n \exp\{-\sum \rho_p(y_i - x_i'\beta)\}$$

L'inférence Bayésienne conduit à intégrer cette fonction sous une a priori possiblement uniforme. Ceci se fait très bien par Monte Carlo et peut se généraliser à des fonctions non-linéaires pour les quantiles.

B.2 Non-parametric inference

Les solutions classiques que l'on a décrites reposent sur la programmation linéaire. Les quantiles sont des fonctions linéaires de x. Une généralisation que l'on souhaite immédiatement apporter, c'est que ceux-ci soient des fonctions non-linéaires de x. On peut aussi vouloir adopter une approche non-paramérique.

B.2.1 L'estimation nonparamétrique des quantiles

Considérons un échantillon X de distribution F. On définit l'α quantile $Q(\alpha)$ par l'inverse à gauche de F

$$Q(\alpha) = inf\{x : F(x) \geq \alpha\}$$

L'estimateur traditionnel que l'on note SQ_α est défini à partir des statistiques d'ordre

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

qui n'est rien d'autre que l'échantillon ordonné. Le quantile empirique est obtenu au moyen de

$$SQ_\alpha = X_{[n\alpha]+1}$$

Cet estimateur n'est pas très efficace à cause de la variance d'échantillonnage qui fait varier les statistiques d'ordre. On va donc chercher à lisser par un kernel pour réduire cette variation. L'estimateur générique par noyau est

$$KQ_\alpha = \sum_{i=1}^{n} [\int_{i-1/n}^{i/n} K_h(t-\alpha)\, dt]\, X_i$$

On peut simplifier cet estimateur de différente manières dont la plus intuitive est

$$KQ_{\alpha.1} = \sum_{i=1}^{n} [\frac{1}{n \times h} K_h(i/n - \alpha)]\, X_i$$

Il est bien entendu que cette formule suppose que l'échantillon est ordonné, comme pour l'estimateur naturel. On rappelle qu'un noyau est une densité qui s'intègre à 1. On peut prendre un noyau uniforme, triangulaire ou normal. Par exemple

$$K_h(i/n - \alpha) = \phi(\frac{i/n - \alpha}{h})$$

avec $h = c * \sigma / n^{1/5}$ et ϕ la normale standardisée.

B.2.2 Régression quantile non-paramétrique

On va commencer par exposer une solution simple, celle d'Abberger (1997). Soient y_1 deux variables aléatoires. On considère la densité jointe $f(y_1, y2)$ et la densité conditionnelle de y_2 si y_1 $f(y_2|y_1) = f(y_1, y_2)/f(y_1)$. On peut alors définir la cumulative de y_2 conditionnelle à y_1

$$F(y_2|y_1) = \int_{-\infty}^{y_2} \frac{f(y_1, t)}{f(y_1)} dt$$

Le α quantile de cette distribution qui dépend de y_1 peut se calculer comme solution en y_2 de

$$F(q_\alpha(y_2)|y_1) = \alpha$$

Sous certaines conditions de régularité, on peut inverser F et calculer alors directement

$$q_\alpha(y_1) = F^{-1}(\alpha|y_1)$$

Mais ici il faut tout d'abord estimer cette densité conditionnelle de façon non-paramétrique. L'estimateur le plus simple est

$$\hat{F}(y_2|y_1) = \frac{\sum K_h(y_1 - y_{1,i}) \mathbb{1}(y_{2,i} \leq y_2)}{\sum K_h(y_1 - y_{1,i})}$$

Cette solution conduit à un estimateur qui n'est pas très lisse. Il peut conduire à trouver des quantiles conditionnels qui se coupent. Considérons un estimateur plus élaboré, où la fonction indicatrice est remplacé par un noyau

$$G(z) = \int_{-\infty}^{z} K(t)\, dt$$

ce qui conduit à

$$\hat{F}(y_2|y_1) = \frac{\sum K_h(y_1 - y_{1,i}) G_h(y_2 - y_{2,i})}{\sum K_h(y_1 - y_{1,i})}$$

Il est sans doute plus facile d'inverser numériquement l'estimateur lissé que l'autre.