The Effect of Health on Wages

Ricardo Vasques Moreno Marques

Masters in Economic and Corporate Decision Making (DEE)

January 2014

Supervisor:

Dr. Pierre Hoonhout

Abstract

The aim of this thesis is to investigate and identify the effect of health on wage. That is, to study how people´s own health can influence their own wages.

Although these variables are clearly related, as unhealthy people are less likely to be hired, it is not obvious what the effect of health on wage is after controlling a set of other variables. A variety of econometric methods are used in order to answer this question.

Careful attention was given to the choice of dataset, the econometric methods and the implementation using statistical software. Finally, having a “car” and “road” to make, we started an academic journey searching for possible solutions for the problems that we have encountered.

Contents

List of Tables

Acknowledgements

I must dearly thank the following people:

Firstly, to my supervisor, Dr. Pierre Hoonhout. Not only for having showed me the way to take a forward and important step in the study of the subject that I fell in love, but also, for having given me good advices at all stages of my research.

Secondly, to my family, especially to my parents and my brother, who encouraged me and made the possible and the impossible to give me a quiet and propitious environment, where I could do the best thesis possible.

Finally, I would like to thank very special people that have contributed that my thesis was written in correct English. Namely, my friends: Ana Catarina Lima, Marta Pinheiro, Lucian Bezuidenhout and Daniel Faes Graca.

Chapter 1

Introduction

This thesis examines the effect of health on wages. As described below, a credible estimate of this effect can only be obtained after taking into account the potential endogeneity of the regressor that represents health. The problem is described below, together with a short description of the solutions that we propose.

The aim this study is to find a credible estimate of the effect of health on wage. Therefore, in order to do that, a linear model to describe the relationship is used, where the dependent variable is wage and one of the regressors is an indicator of health.

The disturbance (or: error) in this regression model is unobserved, and will contain all the variables that cannot be included as regressors, because there is no data available on them. In the present case, this disturbance will include unobserved indicators that are related to the life-style of the subject (e.g. productivity). As the life-style of the subject will likely be related to wage as well,

this research encounters the so-called omitted variables problem.

As life-style is expected to be correlated with the regressor that is an indicator for the health of the subject, the omitted variable problem leads to endogeneity of the health regressor. Due to this endogeneity, several well-known estimators are useless. Estimates obtained by using these estimator will lead to conclusions that are misleading.

Panel data estimators are introduced to deal with this endogeneity problem. It is believed that they will lead to credible estimation results.

This thesis is organised as follows: Chapter 2 describes the problem, Chapter 3 reviews previous literature, Chapter 4 describes the data, Chapter 5 describes the model, Chapter 6 describes the empirical results and Chapter 7 summarizes and concludes.

> *"Man, sacrifices his health in order to make money. Then he sacrifices money to recuperate his health. And then he is so anxious about the future that he does not enjoy the present; the result being that he does not live in the present or the future; he lives as if he is never going to die, and then dies having never really lived."*
>
> *Dalai Lama*

Chapter 2

A Description of the Problem

Within the area of Econometrics, there exist several estimators (like the Ordinary Least Squares or the Generalized Least Squares), where at minimum, the error in each time period is hypothesized to be uncorrelated with the explanatory variables in the same time period. However, when the same cross section units can be observed at different points in time, that is, if a panel data set can be collected, this assumption can be relaxed. Actually, the main purpose for using panel data is that it allows us to take the omitted variables problem into account.

In other words, if a set of random variables can be selected, computing y as the possible dependent variable, and x as a set of the possible explanatory variables, the omitted variables problem appears when:

There are unobserved variables that are influencing y, after controlling for x, but, usually because of data unavailability, they cannot be included in a regression model.

This study will therefore regard this omitted variables problem, also called unobserved effects (α), treated here as random variables. Where, according with this framework, the major issue is to know whether the unobserved effect is uncorrelated or correlated with the explanatory variables.

That is, if α is included as a regressor along with x, a linear model can be written as:

$$E(y \mid x, \alpha) = \beta_0 + x\beta + \alpha \tag{2.1}$$

where the main interest lies in the $K \times 1$ vector β.

On one hand, if α is uncorrelated with each x_j , then α is just another unobserved factor influencing y. Which means that α is thought not to be systematically related to the observable explanatory variables. In contrast, if $Cov(x_j, \alpha) \neq 0$ for some j, putting α into the error term can lead to serious problems (Ordinary Least Squares is inconsistent). Which means that, without additional information, the vector β cannot be consistently estimated, neither it will be possible to determine whether there is a problem.

Therefore, facing the fact that this α is quite possibly correlated with the explanatory variables, a microencometric research has to be initialized.

Chapter 3

A Review of the Literature

In the past decades, several authors have developed the topic of health on econometrics, particularly microeconometrics. A recent textbook treatment is Jones and Balia [2007], which is, in part, a replication of the work done by Contoyannis and Rice [2001]. This book will be the basis of our study.

The idea that health may impact on wages had previously been studied by Mushkin [1962], Grossman and Benham [1974], Luft [1975] and Berkowitz et al. (1983). Where it was discovered that an increase in health reproduces an increase in productivity, which replicates an increase in wages. However, as this moderation happens, it leads to a possible endogeneity of the health regressor. Several authors had given their contribute in order to develop new and suitable estimators that could take into account this scenario, like Hausman and Taylor [1981], Amemiya and MaCurdy [1986], Breush and Shmidt [1989] - hereafter HT, AMC and BMS, respectively. Nevertheless, applications of these techniques to wage equations were an original contribution of

HT, who was the first to introduce a new type of estimator for this problem (Hausman and Taylor estimator), considering the possible endogeneity of the education variable in the wage equation. In the same research area, facing the multiple estimators alternatives, Cornwell and Rupert [1988] - hereafter CR - tried to compare all the implemented estimators, in order to seek for the best option. Additionally, Baltagi and Khanti-Akom [1990], tried to improve the CR study, and discover how more efficient were AMC and BMS in relation with HT estimator. However, in all of these studies the health indicator was: either assumed exogenous (HT) or not taken into account (CR and Baltagi and Khanti-Akom [1990]). Only in 2001, Contoyannis and Rice [2001] had finally included a health regressor, assuming its real endogeneity, considering the several estimators already discussed.

Several authors had already exploited the panel data available in the British Household Panel Survey (BHPS), in order to compute estimation models of wages and earnings, like Harkness [1996], Disney and Gosling [1998] and Hildreth [1999]. However, all these studies have not included a health indicator among the regressors. Again, Contoyannis and Rice [2001] introduce this measure, in the case of a developed economy.

Some other examples of work related to ours are Berkowitz et al. (1983) ,Lee [1982], Haveman et al. (1994), Sundberg [1996], Walker and Thompson [1996], Madden [1962], Deaton and Paxson [1998],Benzeval and Judge [2000], Salas [2002] and Contoyannis and Rice [2004].

The methodology is mainly of panel data econometrics, for which several book-length treatments exist. The main references are Baltagi [1995], Wooldridge [2001] and Cameron and Trivedi [2005].

Chapter 4

A Description of The Data

The data that is used in this dissertation comes from the British Household Panel Survey (BHPS). The British Household Panel Survey is a major government funded survey in the United Kingdom. It is a nationally representative panel survey, which has been supported continuously since 1991. This survey has been conducted by the ESRC UK Longitudinal Studies Centre (ULSC), together with the Institute for Social and Economic Research (ISER) at the University of Essex.

BHPS is an annual survey where, since from September 1991, each adult member (aged 16 years and over) of a nationally representative sample of more than 5,000 households is interviewed year after year, in successive waves. If this individuals leave their original households, all adult members of their new households are also interviewed.

However, the wording of health variable has changed at wave 9 question. Namely, for waves 1-8, the health indicator represents “health status over

the last 12 months", but, in the SF-36 questionnaire, the one that is included in wave 9, the health variable represents "general state of health", using the question: "In general, would you say your health is: excellent, very good, good, fair, poor?". This change can lead to a greater item non-response for this variable at wave 9 in comparison with the other waves, which can consequently complicate our analysis of attrition rates. Therefore, this study approach will focus only on the first eight waves, in order to have the same scenario for all time periods.

After eliminating all of the missing cases, the sample consists of 39632 individuals after all waves (from September 1991 to April 1999). Definitions and statistics for the principal variables are provided in the tables 4.1 to 4.4. Of these, 3958 "survivors" are a subset of the original sample members who, not only completed the questionnaire at all waves, but also gave valid responses for all variables utilized in our model (see table 4.5). Furthermore, only the individuals, who were in either full-time or part-time employment and participated in each of the eight waves, have been considered. This was done with the purpose of constructing an average hourly wage, as the major issue of this study is to evaluate the effect of health on wages.

Although the data used on this study comes from the BHPS, it is related to a study from chapter eight of Jones and Balia [2007]'s book entitled "Applied Health Economics" . These authors take into account the relationship between wage and health, similarly to the work of Contoyannis and Rice (2001). In this report, in contrast with the Jones and Balia [2007], the principal attention goes to the health outcome which, in this dataset, is defined by a self-assessed response to the question: "Please think back over the last 12 months

about how your health has been. Compared to people of your own age, would you say that your health has on the whole been excellent/good/fair/poor/very poor?". This measurement can be interpreted as indicator of each individual´s health status having in mind the concept of "normal health" for their age group. This method has been widely used in previous studies about the relationship between health and socioeconomic status (e.g., Ettner, 1996; Deaton and Paxson, 1998; Smith, 1999; Benzeval et al., 2000; Salas, 2002; Adams et al., 2003; Contoyannis et al., 2004).

From the answers to the question above, likewise to the study of the Jones and Balia [2007], the health variable ($hlstat$), originally categorical, is converted into three dummy variables, coded to if an individual has excellent health ($sahex$), has good health ($sahgd$), or has fair or worse than fair health ($sahfp$) (which is going to be the baseline variable). It is assumed that health and wages have a positive relationship, that is, an increase in health leads to higher wages and so it is expected that the coefficient on excellent and good health will be positive with a larger coefficient on excellent health. Another important indicator related to health, that is worth to be included, in order to complement the information of each individual's health condition, is the General Health Questionnaire (GHQ). The GHQ was not only constructed as a screening instrument for psychiatric illness but also to be used as an indicator of subjective well being. Its main purpose is detecting psychiatric disorders that require clinical attention, among respondents in community settings and non-psychiatric settings. There are 12 individual elements to the shortened GHQ. Then, a four-point ordinal scale is used to indicate how the respon-

dents have recently felt with respect to the item in question. A Likert[1] scale ($hlghq1$) is then used to obtain an overall score by summing the responses to each question. This provides a variable ranging from 0 (no problems) to 36 (most problems). The coefficient on this variable is expected to be negative, and it is hypothesized that health status should be endogenous in our model of wages. Age and experience are included as a second-order polynomial, (age, $agesqrd$ and exp, $expsqrd$), with the objective of catching time-varying reports. Age should capture general labor market experience and tenure effects. Experience is calculated as the number of years for which an individual has been doing the same job with its current employer. It is expected that both variables have positive coefficients for the levels of each of these variables with their effects declining over the life cycle, leading to a concave function in both experience and age. (Mincer 1974). In order to take into account a possible geographical segmentation of wages it is also included several of dummy variables. ($SouthW$, $London$, $Midland$, $NorthW$, $NorthE$ and $Scot$). According to Grossman and Benham [1974], the sign of the deviation from national average prices should be a reflection of the sign of the coefficients on these dummies. Additionally, to consider the individual's work situation in the study, several variables were constructed, namely: (i) two binary variables to indicate the workforce sector, to distinguish between the public and private sectors ($jobpriv$). (ii) a variable that measures the number of employees at the individual's place to work ($jbsize$) (Harkness 1996). (iii) a variable demonstrating if the individuals had any kind of training introduction or

[1]The original idea for the Likert scale (often called a rating scale) is found in Rensis Likert´s 1932 article in Archive of Psychology titled, "A Technique for the Measurement of Attitudes." This idea was expanded by Likert´s 1934 Journal of Social Psychology article titled "A Simple and Reliable Method of Scoring the Thurstone Attitude Scales." It is a tool used in questionnaires in which participants are asked to respond to statements on a scale ranging from "bad feelings" to "good feelings" about some matter.

education related to their current employment ($ljtrain$), predicting that will have a positive coefficient. (iv) a variable to represent union status, which takes the value of one if the individual is member a union and zero otherwise ($covmem$). Following Hildreth (1999) is assumed that the impact of unionization is positive, with a larger effect being for members than non-members. And (v), a number of binary variables to indicate occupational status ($prof$, $manag$, $skillnm$, $skllm$). It is thought that the workforce variables should have an endogenous relationship in wages. (Disney and Gosling 1998). The marital status distinguishes between widow, divorced/separated, never married and married or living as a couple, which is the excluded category ($widow$, $divsep$ and $nvmar$). These indicators are included to access the household economies of scale and productivity effects that are not captured by any other variable. Furthermore, a variable that measures the number of children aged between 0 and 4 years of age ($kids04$) is included in the study. Previous work has found a positive and significant coefficient for the presence of children in the household for men and a negative and significant coefficient for women (Harkness 1996). Ethnic indicator is also included, but as an exogenous time-invariant variable, coded one if the respondent is white and zero otherwise (white). In addition, the education variable is measured by the highest educational qualification reached by the end of the sample period in descending order of attainment. Highest academic qualification is degree or higher degree, HND or equivalent, A level or is O level/CSE. (deg,$hndct$, $alevel$ and $ocse$, respectively). Previous work has found a differential in wages across educational attainment and so there is included this type of indicators (Harkness 1996). No qualifications (no academic qualifications) is the excluded category for the educational variable. It is hypothesized that educational attainment is

endogenous relatively to wages, according with previous research (e.g. Hausman and Taylor 1981; Cornwell and Rupert 1988; Baltagi and Khanti-Akom 1990). Since the BHPS does not contain an hourly wage variable, one had to be constructed (the method suggested by Jones and Balia [2007] seemed to be the best option). Firstly, an usual gross monthly payment derived from the main job of an individual was selected (using $paygu$ variable). This was then divided by the number of hours (including overtime) worked per month in their main job (derived from BHPS variables $jbhrs$ and$jbot$). Similarly, an hourly wage was also obtained in a secondary job (where $j2has$ is defined if the individual has a secondary job, $j2pay$ the pay that him receives and$j2hrs$ the hours worked) alike done in the main job. Then, an overall average of hourly wage ($wage$) was constructed**,** taking into account a ratio between the hourly wage in the main and secondary jobs and the proportions of total hours worked in their main and secondary jobs. Using this method it was obtained a measure of 'maximum average' productivity, that can be interpreted as: the individuals who earn relatively less money are more likely to search for another job, in order to seek a compensation for their salary. Where as, those who are comfortable in their main jobs, are less likely to look for an alternative job. With this method it was possible to take into account a measure of productivity of each individual, which it is thought to be one of the most important unobserved factor that is influencing the relationship between the variables health and wages.

Table 4.1: Data Set

Variable	Label
age	Age in years
$agesqrd$	Age Squared
exp	Duration of spell in current job in years
$expsqrd$	Experience Squared
$wage$	Average hourly wage
$lnwage$	Logarithm of average hourly wage
$jbsize$	Number of employees at workplace
$kids04$	Number of children in the household aged 0 to 4
$hlghq1$	General Health Questionnaire Score
$covmem$	Unionization indicator: 1=Covered union member
$jobpriv$	Sector indicator: 1=Employed in the private sector
$ljtrain$	Training indicator: 1=Received education/training related to current job
$Scot$	Regional indicator: 1=lives in Scotland
$Wales$	Regional indicator: 1=lives in Wales
$London$	Regional indicator: 1=lives in London
$NorthE$	Regional indicator: 1=lives in Northeast
$NorthW$	Regional indicator: 1=lives in Northwest
$Midland$	Regional indicator: 1=lives in Midlands
$SouthW$	Regional indicator: 1=lives in Southwest
$widow$	Marital status indicator: 1=Widowed
$divep$	Marital status indicator: 1=Divorced or Seperated
$nvrmar$	Marital status indicator: 1=Never married
$prof$	Occupation Indicator: 1=Professional
$manag$	Occupation Indicator: 1=Managerial
$skllnm$	Occupation Indicator: 1=Skilled non Manual
$skllm$	Occupation Indicator: 1=Skilled Manual
$jobpt$	Employment Indicator: 1=Part-time employee
$male$	Sex indicator: 1=Male
$sahex$	Health Indicator: 1=Self Assessed health reported as excellent
$sahgd$	Health Indicator: 1=Self Assessed health reported as good
$white$	Ethnicity indicator: 1=White
$ocse$	Education indicator: 1=Highest academic qualification is O level/CSE
$hndct$	Education indicator: 1=Highest academic qualification is HND
deg	Education indicator: 1=Highest academic qualification is degree or higher

Table 4.2: Panel Statistics

Variable		Mean	Std. Dev.	Min	Max	Observation
pid	Overall	2.01e+07	5.07e+07	1.00e+07	8.85e+07	N = 39632
	Between		2.32e+07	1.00e+07	8.85e+07	n = 9838
	Within		0	2.01e+07	2.01e+07	T = 4.027
wave	Overall	4.611	2.346	1	8	N = 39632
	Between		0	1.994	4.5	n = 9838
	Within		1.926	0.111	9.111	T = 4.027
lnwage	Overall	1.637	0.636	-1.725	6.431	N = 39632
	Between		0.592	-.3406	4.223	n = 9838
	Within		0.253	-0.497	4.715	T-bar = 1.983
age	Overall	36.829	12.428	15	84	N = 39632
	Between		13.209	15	84	n =9838
	Within		1.924	31.995	41.579	T-bar = 4.02755
agesqrd	Overall	1510.841	980.532	225	7056	N = 39632
	Between		1035.136	225	7056	n = 9838
	Within		151.924	940.041	2110.441	T-bar =4.02755
exp	Overall	5.136	5.819	0	52	N = 39632
	Between		5.324	0	49.8	n = 9838
	Within		2.804	-34.463	42.011	T-bar = 4.02755
expsqrd	Overall	60.251	153.473	0	2704	N = 39632
	Between		144.741	0	2481.8	n = 9838
	Within		71.282	-1960.403	1750.472	T-bar = 4.02755
jbsize	Overall	4.963	2.464	1	11	N = 39632
	Between		2.241	1	11	n = 9838
	Within		1.313	-3.369	12.534	T-bar = 4.02146
*kids*04	Overall	0.588	0.934	0	6	N = 39632
	Between		0.887	0	6	n = 9838
	Within		0.334	-2.286	4.874	T-bar = 4.02146
*hlghq*1	Overall	10.703	4.925	1	36	N = 39632
	Between		4.109	1	36	n = 9838
	Within		3.358	-8.725	34.578	T-bar = 4.02146
covmem	Overall	.2171241	0.412	0	1	N = 39632
	Between		0.329	0	1	n = 9838
	Within		0.273	-0.657	1.092	T-bar = 4.02146
jobpriv	Overall	.644	0.478	0	1	N = 39632
	Between		0.409	0	1	n = 9838
	Within		0.257	-0.231	1.519	T-bar = 4.02146
Scot	Overall	0.091	0.288	0	1	N = 39632
	Between		0.288	0	1	n = 9838
	Within		0.034	-0.783	0.966	T-bar = 4.02146
Wales	Overall	0.045	0.207	0	1	N = 39632
	Between		0.208	0	1	n = 9838
	Within		0.031	-0.829	0.920	T-bar = 4.02146
London	Overall	0.098	0.298	0	1	N = 39632
	Between		0.296	0	1	n = 9838
	Within		0.068	-0.776	0.973	T-bar = 4.02146

Table 4.3: Panel Statistics(cont.)

NorthE	Overall	0.096	0.295	0	1	N = 39632
	Between		0.282	0	1	n = 9838
	Within		0.046	-0.778	0.971	T-bar = 4.02146
NorthW	Overall	0.104	0.305	0	1	N = 39632
	Between		0.298	0	1	n = 9838
	Within		0.045	-0.752	0.979	T-bar = 4.02146
Midland	Overall	0.171	0.377	0	1	N = 39632
	Between		0.377	0	1	n = 9838
	Within		0.061	-0.703	1.046	T-bar = 4.02146
SouthW	Overall	0.091	0.288	0	1	N = 39632
	Between		0.276	0	1	n = 9838
	Within		0.057	-0.783	0.966	T-bar = 4.02146
widow	Overall	0.014	0.121	0	1	N = 39632
	Between		0.114	0	1	n = 9838
	Within		0.042	-0.862	0.889	T-bar = 4.02146
divep	Overall	0.092	0.289	0	1	N = 39632
	Between		0.275	0	1	n = 9838
	Within		0.121	-0.782	0.967	T-bar = 4.02146
nvrmar	Overall	0.317	0.465	0	1	N = 39632
	Between		0.474	0	1	n = 9838
	Within		0.133	-0.557	1.192	T-bar = 4.02146
prof	Overall	0.049	0.217	0	1	N = 39632
	Between		0.184	0	1	n = 9838
	Within		0.116	-0.825	0.924	T-bar = 4.02146
manag	Overall	0.285	0.451	0	1	N = 339632
	Between		0.387	0	1	n = 9838
	Within		0.225	-0.589	1.161	T-bar = 4.02146
skllnm	Overall	0.272	0.445	0	1	N = 39632
	Between		0.401	0	1	n = 9838
	Within		0.217	-0.602	1.147	T-bar = 4.02146
sklnm	Overall	0.186	0.389	0	1	N = 39632
	Between		0.349	0	1	n = 9838
	Within		0.202	-0.688	1.061	T-bar = 4.02146
jobpt	Overall	0.236	0.424	0	1	N = 39632
	Between		0.403	0	1	n = 9838
	Within		0.212	-0.638	1.111	T-bar = 4.02146
male	Overall	0.483	0.499	0	1	N = 39632
	Between		0.4999	0	1	n = 9838
	Within		0	0.493	0.493	T = 8
sahex	Overall	0.287	0.452	0	1	N = 39632
	Between		0.367	0	1	n = 9838
	Within		0.307	-0.587	1.162	T-bar = 4.02146
sahgd	Overall	0.506	0.499	0	1	N = 39632
	Between		0.379	0	1	n = 9838
	Within		0.379	-0.368	1.381	T-bar = 4.02146

Table 4.4: Panel Statistics(cont.)

white	Overall	0.9582805	0.199	0	1	N = 39632
	Between		0.199	0	1	n = 9838
	Within		0	0.958	0.958	T-bar = 4.02146
ocse	Overall	0.03	0.171	0	1	N = 39632
	Between		0.147	0	1	n = 9838
	Within		0.143	-0.469	0.905	T-bar = 4.02146
alevel	Overall	0.035	0.185	0	1	N = 39632
	Between		0.205	0	1	n = 9838
	Within		0.141	-0.464	0.911	T-bar = 4.02146
hndct	Overall	0.007	0.083	0	1	N = 39632
	Between		0.095	0	1	n = 9838
	Within		0.064	-0.493	0.882	T-bar = 4.02146
deg	Overall	0.004	0.051	0	1	N = 39632
	Between		0.055	0	1	n = 9838
	Within		0.039	-0.4974742	.8775258	T-bar = 4.02146
ljtrain	Overall	0.067	0.251	0	1	N = 39632
	Between		0.188	0	1	n = 9838
	Within		0.206	-0.682	0.942	T-bar = 4.02146

[2]This table is an exportation from the xtsum command for the statistical software Stata.

Table 4.5: Missinig Values Description

wave1	wave2	wave3	wave4	wave5	wave6	wave7	wave8	Mv	Freq.
.	.	.	.	.	.	.	.	8	30375
+	+	+	+	+	+	+	+	**0**	**3958**
.	.	.	.	.	.	+	+	6	790
+	.	.	.	.	.	.	.	7	634
.	.	.	.	.	.	.	+	7	515
+	+	.	.	.	.	.	.	6	352
.	.	.	.	.	.	+	.	7	321
.	.	.	.	.	+	+	+	5	237
+	+	+	.	.	.	.	.	5	217
.	+	+	+	+	+	+	+	1	186
+	+	+	+	.	.	.	.	4	186
.	.	.	.	+	+	+	+	4	186
.	+	.	.	.	.	.	.	7	186
.	.	.	+	+	+	+	+	3	174
+	+	+	+	+	+	+	.	1	150
.	.	+	+	+	+	+	+	2	150
+	+	+	+	+	.	.	.	3	142
.	.	.	+	.	.	.	.	7	142
.	.	.	.	.	+	.	.	7	142
.	.	+	.	.	.	.	.	7	142
+	+	+	+	+	+	.	.	2	118
.	.	.	.	+	.	.	.	7	118
+	.	+	+	+	+	+	+	1	79
								Rest	274

[3]Rest: All others existing patterns, that have frequency less than 100; 274 is the sum of these other panels' frequncies. [4]This table is an exportation from the mvpatterns command for the statistical software Stata.

Chapter 5

The Econometric Model

5.1 Introduction

Taking into account our panel data set, letting t denote the time period and assuming an independent, identically distributed cross section observations $(X_i, y_i) : i = 1, 2, ..., N$, where X_i contains the explanatory variables appearing anywhere in the system, and y_i contains the dependent variable for all T time periods, the model can be written as

$$y_{it} = \beta_0 + \delta_2 d2_t + \delta_3 d3_t + ... + \delta_T dT_t + \beta_1 x_{1it} + \underline{\beta}' \underline{\mathbf{x}}_{it} + \underline{\gamma}' \underline{\mathbf{z}}_i + \alpha_i + u_{it} \qquad (5.1)$$

where y denotes the average of hourly wage in logarithm ($lnwage$), x_1 denotes the health indicator (our variable of interest, that in the estimations is represented by the regressors *sahex* and *sahgd*, respectively) and $\underline{\mathbf{x}}$ denotes a

vector of time-varying control variables. The vector $\underline{z}$ consists of control variables that do not change over time.

dT_t are dummy variables that equals zero when $t = T - 1$ and one when $t = T$; it does not change across i, which is why it has no i subscript. The base period, as always, is $t = 1$. The intercept for the second time period is $\beta_0 + \delta_2$, and so on.

Considering that the population may have different distributions in different time periods (it is also possible that the error variance changes over time), it is important to allow that the interception differs across periods. Furthermore, as this data set is in a circumstance of large values of N and small values of T, it is a good idea to allow for separate intercepts for each time period(wave's dummies), for all the waves excluding one, that is going to be the our base year. Doing this, the aggregate time effects that have the same influence on y_{it} can be successfully preserved. It is important for our study to include these waves dummies mainly due to the economic cycles, which can easily change the individual's wages. Additionally, some disease cycles that suddenly "attack" in some years, is another argument that have influenced our decision to take into account this dummies.

This model divides the unobserved factors that affect the dependent variable in two types: those that are constant and those that vary over time.

α_i captures all unobserved time-constant factors that affect y_{it} (The fact that a_i has no t subscript tells us that it does not change over time). It is generally called an unobserved effect. It is also common in applied work to find a_i referred to as a fixed effect, which helps us to remember that α_i is fixed over

time.

u_{it} is often called the idiosyncratic error or time-varying error, because it represents unobserved factors that change over time and affect y_{it}.

However, for calculating a precise estimation using panel data, it is necessary to infer a number of assumptions, such as:

It is assumed to have repeated observations on a cross section of N individuals. Also, in this research, we have the same time periods, denoted $t = 1, 2, ..., T,$ for each cross section observation. In other words, our panel is balanced, because the same time periods are available for all cross section units. Furthermore, it is assumed that we possess N-large asymptotics (N is sufficiently large relative to T), which it is convenient to view the cross section observations as independent, identically distributed draws from the population. This way our asymptotic analysis should provide suitable approximations. Moreover, for any cross section observation i—denoting a single individual—we denote the observable variables for all T time periods by $\{(y_{it}, x_{it}) : t = 1, 2, ..., T\}$. Because of the fixed T assumption, the asymptotic analysis is valid for arbitrary time dependence and distributional heterogeneity across t.

So, how should the parameter of interest, β_1, be estimated given T time periods of panel data?

5.2 Pooled Ordinary Least Squares

One possibility is to just pool the T waves and use the Pooled Ordinary Least Squares (POLS) estimator.

This well-know estimator, under certain assumptions, can be used to obtain consistent estimator of β_1 in model (5.1). The model can be written as

$$y_{it} = \beta_0 + \delta_2 d2_t + \delta_3 d3_t + ... + \delta_T dT_t + \beta_1 x_{1it} + \underline{\beta}' \underline{\mathbf{x}}_{it} + \underline{\gamma}' \underline{\mathbf{z}}_i + v_{it} \qquad (5.2)$$

Where, the unobserved effect and an idiosyncratic error are merged into one: the composite errors, v_{it}. (for each t, $v_{it} = \alpha_i + u_{it}$)

To make the pooled OLS a consistent estimator for this equation, a key assumption has to be assumed

Assumption POLS.1:

$$E(x_{it}`v_{it}) = 0 \qquad (5.3)$$

Which addresses the contemporaneous exogeneity assumption. In other words, the errors of each time cannot be correlated with the regressors of the respective time. Practically speaking, this estimator only works if there is no correlation between x_{it} and v_{it}. For instance, the regressors must not only be uncorrelated with the idiosyncratic error u_{it}, but also uncorrelated with the unobserved effect α_i.

Even if it is assumed that $E(x_{it}`u_{it}) = 0$ occurs, should it be true that $E(x_{it}`a_i) = 0$ also occurs?

Various factors including individuals' productivity, propensity to diseases and family background might be contained in α_i. Such factors may well be affecting the health coefficient, which is the main purpose within this study. For instance, a number of studies have found that an increase in health leads to an increase in productivity, which should be reflected in an increased wage rate (Mushkin [1962]; Grossman and Benham [1974]; Luft [1975] and Berkowitz et al. 1983).

Namely, it is thought that $Cov(x_{1it}, \alpha_i) \neq 0$. This means that health is possibly endogenous in our model, because the employer may perceive health to be correlated with unobservable attributes of an individual which affect productivity and accordingly over higher wages to healthier employees. It is assumed that these unobservable attributes are constant over time, and therefore part of the unobserved heterogeneity term α_i.

Thus, it is well-known that the Ordinary Least Squares estimator is inconsistent under these circumstances. Therefore, in order to contour this trouble, panel data estimators must emerge.

5.3 Panel Data

Within the panel data field, the three most famous estimators are the Random Effect (RE), the Fixed Effect (FE) and First Differences (FD). However due to the fact that the RE characteristics do not fit our model, mainly because α_i cannot be dependent of x_i, it will not be taken into consideration on this study.

5.3.1 The Fixed Effects Estimator

In many applications the main purpose of using panel data is to allow α_i to be arbitrarily correlated with the x_{it}. A fixed effects (FE) analysis achieves this point explicitly.

Assumption FE.1: $E(u_{it} \mid x_i, \alpha_i) = 0, t = 1, 2, ..., T$

The FE analysis supposes strict exogeneity assumption. In other words, the errors must, not only be uncorrelated with the regressors from the respective time, but also with all the other time periods.

It is important to highlight that this assumption does not impose anything about the unobserved effect α_i. Thus, there are no contradictions that could withhold these effects from truly influencing the regressors. And this relation is what we are ultimately looking for. The partial effects can be estimated, even in the presence of time-constant omitted variables that can be arbitrarily related to the observables x_{it}.

As a result, fixed effects analysis is more robust than random effects analysis. However, this is not done without a price. Without additional assumptions, we cannot include time-constant factors in x_i. Namely, if α_i can be arbitrarily correlated with each element of x_{it}, there is no way to distinguish the effects of time-constant observables from the time-constant unobservable α_i.

Thereby, the only barrier that is restraining this study from estimating consistently the coefficients is these unobserved effects, α_i.

To solve this problem, the FE estimator uses the "time-demanding" method. That is, from the first model (5.1) it is defined $\ddot{y}_{it} = y_{it} - \bar{y}_i$, where y_{it} is trans-

formed into $\ddot{y}_{it}$, a deviation of y_{it} from its time-average for each individual i. The model then implies that:

$$\ddot{y}_{it} = \beta_0 + \delta_2 d2_t + \delta_3 d3_t + ... + \delta_T dT_t + \beta_1 \ddot{x}_{1it} + \underline{\beta}' \underline{\ddot{\mathbf{x}}}_{it} + \ddot{u}_{it} \tag{5.4}$$

The transformation from y_{it} to $\ddot{y}_{it}$ is called the within-transformation (time-demeaned data on y, and similarly for $\ddot{x_{it}}$, $\underline{\ddot{x}}_{it}$ and $\ddot{u}_{it}$). This transformation achieves that the time-constant part of u_{it}, α_i, disappears from the equation, also as the time-constant observed regressors $\underline{\mathbf{z}}_i$.

As α_i was considered to be the only cause of the endogeneity, β_1 can be estimated consistently by pooled OLS in (5.4) . Nevertheless, to make the estimation with pooled OLS, it is mandatory to first make sure that its key assumption is assumed. That is,

$$E(\ddot{x}_{it}\text{`}\ddot{u}_{it}) = 0 \tag{5.5}$$

Only under (5.5) and FE.1, it is verified that the estimations are truly consistent. Actually, a lot more can be said about condition (5.5). Under Assumption FE.1, $E(\ddot{u}_{it} \mid x_i) = E(u_{it} \mid x_i) - E(\overline{u}_i \mid x_i)$, which in turn implies that $E(\ddot{u}_{it} \mid \ddot{x}_{i1}, \ddot{x}_{i2}, ..., \ddot{x}_{iT}) = 0$, since each $\ddot{x}_{it}$ is just a function of $x_i = (x_{i1}, ..., x_{iT})$. This result shows that the $\ddot{x}_{it}$ satisfies the conditional expectation form of the strict exogeneity assumption in the model (5.4). Among other things, this conclusion implies that the fixed effects estimator of β is actually unbiased under Assumption FE.1.

Thus, it can be concluded that the fixed effects (FE) estimator, denoted by $\hat{\beta}_{FE}$, is nothing but the POLS estimator for the model (5.4), under FE.1 and (5.5).

Still, in order for this estimator to give proper results asymptotically, it is necessary that the variables are independent between each other. In other words, there are no exact linear relationships among the regressors in the population (to avoid colinearity, simultaneity, etc)

Assumption FE.2: : $rank\left(\sum_{t=1}^{T} E(\ddot{x}_{it}`\ddot{x}_{it})\right) = rank\left[E(\ddot{X}_{it}`\ddot{X}_{it})\right] = K$

This standard assumption can fail if at least one of the regressors can be written as a linear function of the other regressors in the population.

Also, if x_{it} contains an element that does not vary over time for any i, then the corresponding element in $\ddot{x}_{it}$ is identically zero for all t and any draw from the cross section. Since $\ddot{X}_{it}$ would contain a column of zeros for all i, Assumption FE.2 could not be true.

Until here, a consistent estimator was finally achieved. $\hat{\beta}_{FE}$, under FE.1, FE.2 and (5.5), is that estimator. However, without any more assumptions it is not ensured that this estimator is efficient. To achieve efficiency, one important assumption should be made.

Assumption FE.3: $E(u_i u_i` \mid x_i, \alpha_i) = \sigma_u^2 I_T$

This Assumption FE.3 can be interpreted as having two parts.

1. Firstly, $E(u_i u_i` \mid x_i, \alpha_i) = E(u_i u_i`)$ is standard in system estimation contexts. If $E(u_i \mid X_i) = 0$, then assumption FE.3 is the same as assuming

$Var(u_i \mid X_i) = Var(u_i)$, which means that each variance and each covariance of elements involving u_i must be constant conditional on all of X_i. This is a very natural way of stating a system homoskedasticity assumption.

2. Secondly, the unconditional variance matrix $E(u_i u_i`)$ has the special form $\sigma_u^2 I_T$. This implies that the idiosyncratic errors u_{it} have a constant variance across t and are serially uncorrelated.

Under Assumptions FE.1–FE.3, multiple restrictions are most easily tested using an F statistic, providing that the degrees of freedom are appropriately computed.

At this point, it is known that under FE.1 and FE.2 the achieved estimator $\hat{\beta}_{FE}$ is normal and consistent. However, if the assumption FE.3 cannot be verified, this estimator variance ($Var(\hat{\beta}_{FE})$) can lead to misleading standard errors, which make our estimation improper to interpret.

As it was seen, the FE.3 can fail if:

1. $Var(u_i \mid X_i) \neq Var(u_i)$, the idiosyncratic errors are heteroskedastic.

2. $E(u_i u_i`) \neq \sigma_u^2 I_T$, the idiosyncratic errors are serial correlated.

Whilst the heterodasticity problem will always be a problem, as in practice homoskedasticity does not happen a lot, and so, this assumption is usually not valid. Correlation between the errors will be here the most important element.

However, it is essential to remember that, when using the FE method, nothing excludes serial correlation in $\{u_{it} : t = 1, ..., T\}$. While it is true that the presence of α_i dominates the observed serial correlation in the composite errors, $v_{it} = \alpha_i + u_{it}$, there are some cases, where u_{it} can have very strong serial dependence, which sometimes can lead to false FE standard errors obtained from the asymptotic estimate variance.

The main problem here is that the u_{it} cannot be estimated, because of the time demeaning used in FE, only the time-demeaned errors, $\ddot{u}_{it}$ can be estimated. The test is complicated by the fact that the $\{\ddot{u}_{it}\}$ is serially correlated under the null hypothesis. For testing, a repeated sampling distribution (RSD) of the test statistic is needed. Consequently, if the null-hypothesis is true, with serial correlation it is hard to obtain the RSD.

5.3.2 The First Differences Estimator

Other Panel Data estimator that deserves attention due to its characteristics is the First Differences (FD) estimator.

This estimator and the FE estimator are very similar methods, whereas both share the first assumption.

Assumption FD.1: Same as Assumption FE.1

Moreover, the only difference is how they achieve the disappearance of α_i. In the FD case, is defined $\Delta y_{it} = y_{it} - y_{i,t-1}$. This variable is called the first-difference of y_{it}.

In this estimator, the time period one is substrate from time period two and time period two from time period three, and so on. This gives:

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \alpha_4 d4_t + ... + \alpha_T dT_t + \beta_1 \Delta x_{1it} + \underline{\beta}' \Delta \underline{\mathbf{x}}_{it} + \Delta u_{it}. \tag{5.6}$$

where now the total number of periods on each unit i for the first-differenced (FD) equation is $T-1$. The total number of observations is $N(T-1)$. Note also that the time-constant observed regressors $\underline{z}_i$, once again disappeared.

Consequently, with the unobserved effect α_i out of the picture, the pooled OLS can again be putted into practice. Although, it is necessary to keep in mind that, in order for the pooled OLS to be consistent in this equation, its major assumption needs to be assumed. In the FD case, it will be:

$$E(\Delta x_{it}{}' \Delta u_{it}) = 0 \tag{5.7}$$

Solely under (5.7) and FD.1, the first-difference (FD) estimator, $\hat{\beta}_{FD}$, is consistent.

Therefore, the $\hat{\beta}_{FD}$ estimator is merely an pooled OLS estimation for the model (5.6), under (5.7) and FD.1.

In addition, since the assumption of strict exogeneity is hold in the first difference equation $E(\triangle u_{it} \mid \triangle x_{i2}, \triangle x_{i3}, ..., \triangle x_{iT}) = 0$, the FD estimator is actually unbiased conditional on X.

Such as in the FE case, it must be ensured that the explanatory variables are independent. Meaning that perfect independence must be imposed.

Assumption FD.2: $rank\left(\sum_{i=1}^{T} E\left(\triangle x_{it}`\triangle x_{it}\right)\right) = K$

Again, as mentioned in the FE.2, equation FD.2 makes it clear why the elements of x_{it} must be time varying (for at least some cross section units). Since, otherwise Δx_{it} will have its entire elements identical to zero for all i and t.

Additionally, not only Assumption FD.2 rules out time-constant explanatory variables, but also it excludes the perfect collinearity among the time-varying variables. This means, there are subtle ways in which perfect collinearity can arise in Δx_{it}. For example, as we have variables that increase only by one wave, after wave for every person in the sample, like the individual's experience, and knowing that our model need waves dummies to control external cycles, it is easily seen that this type of variables (like $\Delta exper_{it}$) are perfectly collinear with the waves dummies. Therefore, this is the reason why we have to control them.

According to what has been seen so far, under FE.1-FE.3, the FE is an efficient estimator in the class of estimators using the strict exogeneity assumption.Thus, it can be inferred that the FD is not as efficient as FE, under the same assumptions.

The big difference here is that FE achieves efficiency with its assumption FE.3, in which its residuals are considered to be homoskedastic and serial uncorrelated. However, in the FD case, this serial uncorrelation within the u_{it} can be a very strong assumption.

A suitable alternative to this is ensuring that the first difference of the idiosyncratic errors,$\{e_{it} \equiv \triangle u_{it} : t = 2, ..., T\}$, are serial uncorrelated and homoskedastic, instead of the u_{it} itself.

Assumption FD.3: $E(e_i e_i` \mid x_{i1}, x_{i2}, ..., x_{iT}) = \sigma_e^2 I_{T-1}$, where e_i is the $(T-1) \times 1$ vector containing e_{it}, $t = 2, ..., T$

Under Assumption FD.3 it can be wrote $u_{it} = u_{it-1} + e_{it}$. So that no serial correlation in the e_{it}, implies that u_{it} is a random walk. A random walk has substantial serial dependence, and so Assumption FD.3 represents an opposite extreme from Assumption FE.3.

Under Assumptions FD.1-FD.3, not only it can be shown that the FD estimator is most efficient in the class of estimators using the strict exogeneity assumption FE, but also, it can be assumed that all statistics reported from the pooled regression on the first-differenced data are asymptotically valid, including F statistics based on sums of squared residuals.

Moreover, with the variables not having suffered any kind of "time-demanding" transformation, and since the strict exogeneity assumption holds, the e_{it} can effectively be tested in order to discover if they are really correlated or not. Based on $T-2$ time periods the test is made as:

1. Obtain the residuals from the FD estimation as $e_{it} = \Delta y_{it} - \Delta x_{it}\hat{\beta}_{FD}$
2. AR(1) on this residuals, $\hat{e}_{it} = \hat{\rho}\hat{e}_{it-1} + error_{it}, t = 3, ..., T; i = 1, ..., N$
3. The test statistic is the usual $t - statistic$ on $\hat{\rho}$.

So, if the idiosyncratic errors $\{u_{it} : t = 1, ..., T\}$ are uncorrelated to begin with $\{e_{it} \equiv \triangle u_{it} : t = 2, ..., T\}$ will be autocorrelated.

5.3.3 Panel Data Estimators Comparison

So far, setting aside pooled OLS, it was seen two competing methods for estimating unobserved effects models. One that involves time-demeaning transformation and another that involves a differencing procedure. So, how do we know which one to use?

Since both FE and FD are consistent and unbiased under FE.1 and each pooled OLS assumption (5.5 for the FE and 5.7 for the FD), neither one of these properties can be used for comparison.

For large values of N and small of T, what deserves to be investigated is the efficiency property (FE.3 and FD.3), defined by the existence of no correlation between the idiosyncratic errors (assuming that they are homoskedastic, since efficiency comparisons require homoskedastic errors).

As testing the u_{it} from the FE procedure is very difficult, what should be tested is the e_{it}, as described before.

When u_{it} are thought to be serially uncorrelated, fixed effects should be the best option, as it is more efficient than first differencing (and the standard errors reported from fixed effects are valid). On the other hand, if u_{it} follows a random walk, the difference Δu_{it} is serially uncorrelated, and so the first differences model is the one that should be adopted.

Chapter 6

The Empirical Results

6.1 Pooled OLS Estimation

From OLS (see table 7.1), if the estimated equation is casually interpreted, it implies that an increase in the health rate lowers the wage rate, at least for the individuals who consider their own health being good $(sahgd)$. This is certainly not what was expected to find. As it was emphasized throughout the problem description, this simple regression equation is likely to suffer from omitted variable problems. While the estimated coefficient on psychological health is negative, which reflects an increase in ill health related to a decrease in wages, however, the coefficient fails to attain statistical significance.

Relatively to the occupational status variables, it is important to observe that, besides all being very significant, they clearly show a gradient association with the increased wages as their status moves from skilled manual, through

skilled non-manual and managerial to professional occupational. The baseline category for these variables represents unskilled, part-skilled and the armed forces. In contrast, the difference in working in the private sector or in the public (baseline category) is slightly significant. Accordingly to this estimator, if an individual works in the private sector, he will earn less salary comparing to an individual that works in the public sector.

The number of employees in the workplace reveals a positive relationship with wages, as suspected, as equally does unionization. Apparently, an employee who belongs to a workunion attracts higher wages, comparing to the ones that do not belong to any work association. Additionally, its coefficient achieves statistical significance. In contrast, the $ljtrain$ variable exhibits an unexpected negative coefficient, but like the union variable, it's coefficient not significant at any suitable level.

Looking at the educational variables, it is difficult to understand what is happening. Namely, only the variables $alevel$ and $hndct$ display the expected signal. These regressors can also be suffering from omitted variables problem. It is important to check if a differential change happens in the panel data estimations. Still, none coefficient seems to be significant, which is a very surprising fact.

Evaluating regions, it can be seen that the South East (baseline category) workers, with the exception of London, command higher wages rates than the other regions.

It is important to highlight that the wave‘s dummies exhibit a positive evolution, presumably reflecting wage inflation over the observed period. In addition, all pass the test for statistical significance.

As expected, both quadratic relations in age and experience represent a significant concave relationship with the logarithm of hourly wages.

The coefficients on the marital status variables suggest that comparing with the baseline of married or living with a partner, individuals who are divorced or separated, together with individuals that have never been married, tend to receive lower wages. However, the presence of kids with less than 4 years old, has a negative relation with wage rate. Which means, that if an individual has a kid in his household, is expected to earn less money.

Hereupon, it can be seen that the first possible solution presented to the omitted variables problem, the POLS estimation, can be put aside, as the estimations for the main variables, $sahex$ and $sahgd$, still reveal some bias from omitted variables, after the inclusion of more explanatory regressors.

6.2 Fixed Effects Estimation

From the fixed effects estimation (see table 7.2), in an overall look, it can easily be seen that all variables have changed. However, they did not necessarily change to better interpretations, as almost all variables are no longer significant. Despite the exclusion of the of the unobserved effects α_i, both health variables continue to suffer from a type of bias, since its coefficients still have a negative signal.

Again, expectedly, the estimate of psychological well-being ($hlghq1$) maintains the same relation to wage rates. However, it still remains non-significant.

Another surprising fact is that, now, being a part of a job union, is not good if a person wants to earn more money. Its coefficient changed to negative, in contrast to the $ljtrain$ variable. This has suffered a transformation in the opposite way. Now, according with the FE estimations, individuals who had training in their work earn more money. Although, it still fails the statistical significance.

It is also important to notice, now, only the individuals who live on the North and London part of the United-Kingdom earn more than those who live in the South-East region. Not only the region variables change its course, but they also lost its significance.

Relatively to the marital status, it can easily be seen that our situation did not improve. Namely, now it appears that the widow individuals earn more money than the married ones, which is a very strange fact. However, with this estimation, all the marital status variables lost their significance.

Another interesting point is the coefficients on the occupational status variables, where expected gradient is lost, partly due to the lost of significance of all occupation variables.

In terms of additional work, there is a big difference too. That is, not only the part-time job variable loses its extremely significance, but also it changes its relation to wages. Now it is good to have a part time job. However, it fails the significance test.

The education variables also changed. $alevel$ is no more significant, neither has a positive partial effect. As a result, it seems that another strong endogeneity problem might be affecting this variables. An additional interesting study should be take into account, not only the endogeneity of the health indicator, but also the educational measures.

The interpretation that can be collected from the waves is also interesting, as the results are almost the same of the pooled OLS estimation. Nevertheless, now all these dummies are non-significative.

Therefore, it is concluded that it will not be this estimator that will answer the so desired question of this thesis. Namely, now, having not the existence of the unobserved effect α_i, the results still seem to suffer from bias, where a possible cause for this situation might be that the time demanding transformations are not strong enough to make consistent estimations. In order words, after this procedure the idiosyncratic errors might still be correlated with the explanatory variables, leading to an inconsistent estimation by the pooled OLS, under FE.1-FE.3.

Since the test for correlation in the FE method is very difficult to achieve, as earlier discussed, we have one possible alternative. This is, to try to estimate with the FD estimator. This happens because the FD estimator has similar assumptions to the FE one, and additionally it posses a stronger method to make the α_i disappear. In addition, after achieving the FD estimations results, it can also be attempted a test for serial correlation, in order to know for sure if we are in the right track.

Notice that, some variables disappeared from the model, as $Scot$, $Wales$, $white$

and $male$. They have been excluded because, according to our dataset they do not change over time.

6.3 First Differences Estimation

From the FD estimator (see table 7.3), it is instantly seen that its results do not differ very much from the fixed effect estimator, which is not a big surprise, as they are both very similar. However, with this estimation, the expected relationship between wage and health is finally achieved. Namely, both coefficients of the variables $sahex$ and $sahgd$ have a positive signal. In other words, ceteris paribus, an increase in an individual´s health leads to a higher wage‘s rates. In words words, if an individual is healthy,there is a higher probability that he will earn more money than an unhealthy person. However, like it has been the case in the previous estimations, this two estimators fail to accomplish statistical significance. Furthermore, with the FD estimation, not only the $hlghq1$ keeps its negative relationship with wages, but also now, it achieves statistical significance.

The majority of the other variables retain similar interpretations to the FE estimates, excluding the variables deg and $covmem$, that with this estimation finally possess the expected positive coefficient. Nonetheless, the time dummies do not possess a specific trend. Which might be what had really happened with the United Kingdom economy. Anyway, they fail to achieve statistical significance, as in the FE estimation, fail to achieve statistical significance.

6.3.1 Test for Autocorrelation

Table 6.1: Test for Autocorrelation

F(1, 354) =	0.021
Prob > F =	0.8844
H0: no first order autocorrelation	

Having as the null hypothesis of no serial correlation, this test suggests that there is no evidence for autocorrelation. In other words, the Δu_{it} are thought to be serially uncorrelated, which implies that the u_{it} are, probably, a random walk.

Therefore, according with this test, the method that should be adopted is the first differences estimator, subscribing the conclusion from the estimations results.

Chapter 7

Conclusion

7.1 My Conclusions

This thesis considered the effect of health on individual nominal hourly net income using a longitudinal data from the British Household Panel Survey. To understand this relation, the pooled OLS, the fixed effects and first differences estimators were employed in a single-equation.

Arguing that health might well be an endogenous regressor, for being correlated with unobservable attributes of an individual, this report proposed some alternatives to seek for the best possible model in order to achieve our goal. Firstly, it has tried to estimate an pooled OLS model, however it is well-known that the Ordinary Least Squares (OLS) estimator is inconsistent under this study circumstances. Secondly, when looking for ways to eliminate the latter problem, two models were estimated using panel data estimators. These are the fixed effect estimator and the first differences estimator, which are both

very similar. Here, it was attempted to remove the unobserved error α_i , because it was thought to be correlated with the x_i, and consequently making the regressor-variable health an endogenous regressor.

Hereupon, the major conclusions from this study are: (i) the pooled OLS-results are not reliable and (ii) the best suitable estimations seem to came from the FD estimator.

From the FD estimations, although it can be observed there is no evidence for an effect of health on wage (as the effect fails to attain statistical significance), after controlling for all the regressors that have been also included, if more data had been collected, an evidence might have been found. Further, given the positive sign of the effect in FD, this evidence would be more likely to have a positive effect rather than a negative one.

However, according to these results, an individual who considers himself as possessing an excellent health, ceteris paribus, is thought to earn 2% more than an unhealthy person (which is the based line). Similarly, a persona that considers its own health being very good/good, will only earn 0,8% more than an individual who has a poor or very poor health.

Still, it is strange that the FE estimates are so different from FD. They are based on different assumptions, so it is probable that FE-assumptions are invalid, given the estimation results.

Nevertheless, some interpretation difficulties still remain. As it could be easily seen during this study, trying to control endogeneity of a variable such as health, is not very easy. A variable with such dimension as this, although it is very easily to measure, has an effect that is easily influenced, as there are a

lot of variables that might indeed be correlated with this health variable. As it was seen, even with some panel data models, what can happen is that the estimations are yet suffering for bias. A strong and powerful method must be taken into account if we really want to know this effect.

7.2 Recommendations For Future Research

After finishing all this study, some recommendations for future research need to be passed. In my perspective, there are two main recommendations that should be made.

Firstly, the major recommendation for a future researcher is to try to obtain the best number of possible variables related to health. It is true that it is not easy to collect such variables like these, because it is a delicate subject. However for a study like this one, it will be much more interesting and appropriate to have more observable variables that can influence an individual‘s health. Doing so, the size of possible unobserved effects α_i would drastically decrease, which would make the estimations, after the disappearance of the α_i, much more suitable to the real world.

Secondly, as the main objective of the research was only achieved with the first difference model, and still the effect is not entirely significant, in future studies more waves should be taken into account.

Table 7.1: Pooled OLS Estimation

Variable	Coefficient	Std.Error
age	0.0613***	(0.00455)
agesqrd	-0.000694***	(5.69e-05)
exp	0.0153***	(0.00416)
expsqrd	-0.000501***	(0.000182)
jbsize	0.0403***	(0.00316)
*kids*04	-0.00353	(0.00938)
*hlghq*1	-0.00231	(0.00156)
covmem	0.0600***	(0.0215)
jobpriv	-0.0344*	(0.0181)
Scot	-0.0123	(0.0294)
Wales	-0.0380	(0.0365)
London	0.158***	(0.0277)
NorthE	-0.0875***	(0.0282)
NorthW	-0.0637**	(0.0292)
Midland	-0.101***	(0.0223)
SouthW	-0.148***	(0.0273)
widow	-0.169**	(0.0746)
divep	-0.0623**	(0.0264)
nvrmar	-0.0248	(0.0258)
prof	0.702***	(0.0393)
manag	0.507***	(0.0229)
skllnm	0.165***	(0.0215)
skllm	0.109***	(0.0253)
jobpt	-0.0640***	(0.0191)
male	0.212***	(0.0179)
sahex	0.00461	(0.0225)
sahgd	-0.0316	(0.0200)
white	-0.0621**	(0.0292)
ocse	-0.00470	(0.0507)
alevel	0.132***	(0.0488)
hndct	0.0286	(0.0817)
deg	-0.101	(0.101)
ljtrain	-0.00356	(0.0299)
2.*wave*	0.0875**	(0.0416)
3.*wave*	0.120***	(0.0417)
4.*wave*	0.126***	(0.0410)
5.*wave*	0.178***	(0.0408)
6.*wave*	0.174***	(0.0403)
7.*wave*	0.237***	(0.0374)
8.*wave*	0.278***	(0.0399)
Constant	-0.137	(0.105)
Observations: 3,958	R-squared: 0.468	
*** p<0.01, ** p<0.05, * p<0.1		

Table 7.2: Fixed Effect Estimation

Variable	Coefficient	Std.Error
age	0.105***	(0.0400)
agesqrd	-0.00106***	(0.000193)
exp	-0.00156	(0.00592)
expsqrd	2.42e-05	(0.000274)
jbsize	0.00946*	(0.00536)
*kids*04	-0.0366*	(0.0213)
*hlghq*1	-0.00173	(0.00176)
covmem	-0.00135	(0.0277)
jobpriv	-0.00827	(0.0299)
London	0.276*	(0.146)
NorthE	0.0548	(0.187)
NorthW	0.375	(0.393)
Midland	-0.0942	(0.144)
SouthW	-0.232	(0.198)
widow	0.156	(0.233)
divep	-1.06e-05	(0.0619)
nvrmar	-0.0371	(0.0659)
prof	-0.0115	(0.0709)
manag	0.0649	(0.0441)
skllnm	0.0131	(0.0391)
skllm	0.00926	(0.0380)
jobpt	0.0249	(0.0311)
sahex	-0.0236	(0.0301)
sahgd	-0.00777	(0.0253)
ocse	-0.0437	(0.0561)
alevel	-0.00471	(0.0610)
hndct	0.0400	(0.114)
deg	-0.0573	(0.114)
ljtrain	0.033	(0.035)
2.*wave*	0.0783	(0.0518)
3.*wave*	0.0940	(0.0807)
4.*wave*	0.0982	(0.116)
5.*wave*	0.175	(0.153)
6.*wave*	0.175	(0.188)
7.*wave*	0.200	(0.225)
8.*wave*	0.251	(0.261)
Constant	-0.730	(1.184)
Observations: 3,958	R-squared: 0.123	
*** p<0.01, ** p<0.05, * p<0.1		

Table 7.3: First Differences Estimation

Variable	Coefficient	Std.Error
$D.age$	0.154***	(0.0466)
$D.agesqrd$	-0.00167***	(0.000489)
$D.exp$	-0.000574	(0.00719)
$D.expsqrd$	3.15e-05	(0.000322)
$D.jbsize$	-0.000314	(0.00652)
$D.kids04$	-0.0502	(0.0369)
$D.hlghq1$	-0.00323*	(0.00189)
$D.covmem$	0.0360	(0.0344)
$D.jobpriv$	-0.0477	(0.0327)
$D.London$	0.107	(0.233)
$D.NorthE$	-0.152	(0.338)
$D.Midland$	0.0282	(0.209)
$D.SouthW$	-0.0402	(0.339)
$D.widow$	0.120	(0.261)
$D.divep$	0.0321	(0.0917)
$D.nvrmar$	-0.000999	(0.0912)
$D.prof$	-0.0116	(0.0779)
$D.manag$	0.0697	(0.0531)
$D.skllnm$	0.0459	(0.0499)
$D.skllm$	-0.00940	(0.0439)
$D.jobpt$	0.0908**	(0.0404)
$D.sahex$	0.0190	(0.0336)
$D.sahgd$	0.00822	(0.0283)
$D.ocse$	-0.0770	(0.0708)
$D.alevel$	-0.00867	(0.0770)
$D.hndct$	-0.161	(0.162)
$D.deg$	0.0580	(0.155)
$D.ljtrain$	0.031	(0.040)
$3.wave$	0.0257	(0.0460)
$4.wave$	0.00320	(0.0441)
$5.wave$	0.0552	(0.0445)
$6.wave$	0.0203	(0.0406)
$7.wave$	0.0155	(0.0403)
$8.wave$	0.0628	(0.0396)
Observations: 1,581	R-squared:0.039	
*** p<0.01, ** p<0.05, * p<0.1		

Bibliography

Hurd M.D. McFadden D. Merrill A. Adams, P. and T. Ribeiro. Healthy, wealthy and wise? tests for direct causal paths between health and socioeconomic status. *Journal of Econometrics*, 112:3 Ð 56, 2003.

Amemiya and MaCurdy. Instrumental variable estimation of an error-components model. *Econometrica*, 54:869–880, 1986.

B.H Baltagi. *Econometric Analysis of Panel Data*. Wiley, West Sussex, 1995.

B.H. Baltagi and S. Khanti-Akom. On efficient estimation with panel data: An empirical comparison of instrumental variables estimators. *Journal of Applied Econometrics*, 5:401 Ð 406, 1990.

Taylor J. Benzeval, M. and K. Judge. Evidence on the relationship between low income and poor health: Is the government doing enough? *Fiscal Studies*, 21:375 Ð 399, 2000.

et al Berkowitz. The optimal stock of health with endogenous wages: application to partial disability compensation. *Journal of Health Economics*, 2: 139–147, 1983.

Mizon Breush and Shmidt. Efficient estimation using panel data. *Econometrica*, 57:695–700, 1989.

A. Colin Cameron and Pravin K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005. ISBN 9780521848053.

Jones A.M. Contoyannis, P. and N. Rice. The dynamics of health in the british household panel survey. *Journal of Applied Econometrics*, 19:473 Ð 503, 2004.

P. Contoyannis and N. Rice. The impact of health on wages: Evidence from the british household panel survey. *Empirical Economics*, 26:599 Ð 622, 2001.

C. Cornwell and P. Rupert. Efficient estimation with panel data: An empirical comparison of instrumental variables estimators. *Journal of Applied Econometrics*, 3:149 Ð 155, 1988.

A.S. Deaton and C.H. Paxson. Ageing and inequality in income and health. *American Economic Review, Papers and Proceedings*, 88:248 Ð 253, 1998.

R. Disney and A. Gosling. Does it pay to work in the public sector? *Fiscal Studies*, 19:347 Ð 374, 1998.

S. Ettner. New evidence on the relationship between income and health. *Journal of Health Economics*, 15:67 Ð 85, 1996.

Grossman and Benham. *Health, hours and wages*. Macmillian and Co, London, 1974.

S. Harkness. The gender earnings gap: evidence from the uk. *Fiscal Studies*, 17:1 Ð 36, 1996.

J. Hausman and W. Taylor. Panel data and unobservable individual effects. *Econometrica*, 49:1377 Ð 1398, 1981.

et al Haveman. Market work, wages and men's health. *Journal of Health E*, 13:163–182, 1994.

A. Hildreth. What has happened to the union wage differential in britain in the 1990s? *Oxford Bulletin of Economics and Statistics*, 61:5 Ð 31, 1999.

N. Bago dÕUva T. Jones, A. Rice and S. Balia. *Applied Health Economics*. Routledge, 2007.

Lee. Health and wage: A simultaneous equation model with multiple discrete indicators. *International Economic Review*, 23:199–221, 1982.

Luft. The impact of poor health on earnings. *Review of Ec*, 57:43–57, 1975.

Madden. *Labour market discrimination on the basis of health: An application to U.K data*. PhD thesis, University College Dublin, 1962.

J. Mincer. *Schooling, Experience and Earnings*. NBER, 1974.

Mushkin. Health as an investment. *Journal of Political Economy*, 1:129–157, 1962.

C. Salas. On the empirical association between poor health and low socioeconomic status at old age. *Health Economics*, 11:207 Ð 220, 2002.

J.P. Smith. Healthy bodies and thick wallets: the dual relationship between health and economic status. *Journal of Economic Perspectives*, 13:145 Ð 166, 1999.

Sundberg. *Essays on Health Economics*. PhD thesis, Department of Economics, Uppsala University, 1996.

Walker and Thompson. *Disability, wages and labour force participation: Evidence from UK panel data*. PhD thesis, Department of Economics, Keele University, 1996.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2001. ISBN 9780262232197.