# Take home examination: Advanced econometrics

Candidate n° 10054

28.05.2021

# 1 Introduction

## 1.1 Context

The goal of this paper is to estimate how large was the union wage premium in the United-States of America between 1983 and 1987. A labor union is usually defined as an association of workers who bargain collectively with their employer regarding the term and conditions of employment. The main objective is to fix higher wages that rely on the job condition and not on individual skills. However, every employee is not unionized and wages between unionized and non-unionized workers should differ. The union wage premium refers to the degree in which union wages exceed non-union member wages, it is computed in terms of the percentage hourly wage difference.
There is a general consensus that unions do raise wages, at least in English-speaking countries which are characterized by fragmented collective bargaining. The forecast effect is usually between 10 to 20%, meaning that unionized workers have an hourly wage that is between 10 to 20% higher than non unionized workers.

Unions have different effects on the wage premium as shown by **Farber (2001)**, the first level is the direct effect of unions on the wages of workers in jobs where wages are set through collective bargaining. The second level is that the presence of unions in an economy can change the level and distribution of wages generally. Unions can raise the wages of their members, but the wages of nonunion workers can fall or rise as well. A standard competitive market model would imply a fall in the wage of nonunion workers. The mechanism is that unions, in raising wages of their members, cause a decline in employment in the union sector. The result is an increase in supply of workers to the nonunion sector as displaced union workers move to that sector, implying a drop in the nonunion wage rate, called the "spillover effect" of unions.

## 1.2 Methodology and results

Let us define statistically the union wage premium and his interpretation. The effect of union on wages for union and nonunion members is in fact not computable because it will need observable data on wages in a labour market without an union and this does not exist. However, we can compute the wage differential between union and nonunion workers on a market with a labor union, so the direct effect of unions.
The differential is $\Delta = \frac{W_u - W_n}{W_n}$ and as long as differential are small between these wages, this could be approximated by $\Delta = ln(W_u) - ln(W_n)$.
This is the difference in wages between union and nonunion members but not the effect of union membership on the hourly wage of one worker. This difference is the dependent variable we want to estimate to know how large is the union wage premium.

One main issue that arises in a lot of past studies is the possible endogeneity of the union dummy because the nature of the union membership is non random. Workers make choices regarding whether to seek union or non-union employment in part based on the wage they would receive in each situation. There is a reverse causality, union statut seems endogenous to wage with a selection bias problem. Thus the model is first estimated through pooled ordinary least squares and then the study focused on trying to control for the endogeneity issue with panel data estimation and simultaneous equation modelisation. For the simultaneous equation, the choice is made to use an equation that determines the choice of unionized in function of exogenous dummy variables that represent different job characteristics.

The result of this study is that the best estimation seems to be made with a fixed effect estimator on panel date, the union wage premium is estimate at 5,21% i.e. that unionized workers earn in average a hourly wage 5,21% larger than non unionized workers, all other variables being constants.

First, data used are present and describe thought usual statistics and graphics. Then the model based on **Lewis (1986)** research is present, it is first estimate to pooled ordinary least square, then to fight the endogeneity issue a fixed model estimator is used and another modelisation with simultaneous equations based on **Schmidt (1978)**. Finally limitations and extensions of modelisation and estimation are presented.

# 2 Data

## 2.1 Presentation of data

To work on our economic question, we have a panel dataset of observation with 789 individuals, 4 years time. It means that we have information on the same individuals across time, it makes it possible to control for unobserved individual characteristics and study variations between individuals and within the individual.
There are some classical independent variables used to estimate wages such as the age of the respondent, the number of years of schooling for the respondent, the number of years of experience and the year of tenure of the present job and a dummy variable that takes one if the respondent is black and zero otherwise. The increase of education should have a positive impact and the experiences should have at first a positive return and after a certain amount of years a decreasing one, following classical labor economic theory. For the dependent variable, there is one variable that accounts for the yearly income and another one for the yearly hours of work. They are used to compute an hourly wage that will be used as the dependent variable of the model.
However, the objective is to estimate the union wage premium in terms of the percentage hourly wage difference and this will be estimated through the estimated coefficient of a dummy variable that takes one if the worker is unionized and zero if he is not. The estimation should be between 10 and 20% according to past studies such as the one conducted by H. Gregg Lewis (1986).
Then, there are seven dummy variables that account for the position or job that the respondent has. For example, the variable sales take one if the respondent works in sales and zero otherwise. The intuition is that union will bargain different wage in function of this characteristics and not in function of individuals skills.

## 2.2 Description of data

From *table1*, it shows that the percentage of unionized workers on the dataset is 24,58% and 17% of workers are black people. Men in the dataset have an average age of 37 years old, they have done 13 years of studies and have 18 years of experiences with 8 years this the last time they changed of job.They earn $27809 and work 2 217 hours per years, that gives an average hourly wage of $12,70.
There are several working statuses in the dataset, 20% of the individuals are professionnels, 5% work in sales, 5 other percent work as clerks, 22% in a craft occupation, 8% as machine operators and around 4% in the service sector. This represents 63% of occupation, it means that the last 37% workers have another occupation as the one listed.

Table 1: Summarize data

| Variable | Mean | Std.Dev | Min | Max |
|---|---|---|---|---|
| age | 36,68 | 8,08 | 19 | 55 |
| educ | 13,10 | 2,12 | 5 | 16 |
| union | 0,2458 | 0,4307 | 0 | 1 |
| laborinc | 27809,51 | 19904,65 | 2478 | 352113 |
| hours | 2217,09 | 441,09 | 1016 | 1995 |
| tenure | 8,35 | 7,44 | 0 | 37 |
| exp | 18,57 | 8,40 | 2 | 44 |
| hwage | 12,70 | 8,94 | 2,03 | 191,36 |
| black | 0,1761 | 0,3810 | 0 | 1 |
| prof | 0,2284 | 0,4198 | 0 | 1 |
| manager | 0,1997 | 0,3998 | 0 | 1 |
| sales | 0,0495 | 0,2168 | 0 | 1 |
| clerk | 0,04968 | 0,2173 | 0 | 1 |
| operator | 0,0839 | 0,2773 | 0 | 1 |
| service | 0,0370 | 0,1888 | 0 | 1 |

Let us focus the data description on the hourly wage in function of the unionization, from *table2* the medium hourly wage for unionized workers is $12.10 whereas for non unionized workers the medium hourly wage is $10,80. This first statistic observation is in accordance with the economic consensus that being unionized allows for a higher hourly wage.
One striking point is that the maximum wage for the non unionized worker seems to be really high, it is $191 per hour. This is a hint that it is important to control for outliers in the dataset and decide to delete them or not. Using the first and last quantile it allows for an interval in which the hourly wage observations should be, the *table8* reports it for both union and non union members.The choice is made to keep only observations with hourly wage under $60 to the description of the dataset,because it is the maximum threshold for non unionized workers. However, **to keep a balanced panel data, outliers will be kept for the estimation section.**

Table 2: Hourly wage statistics in function of Union membership

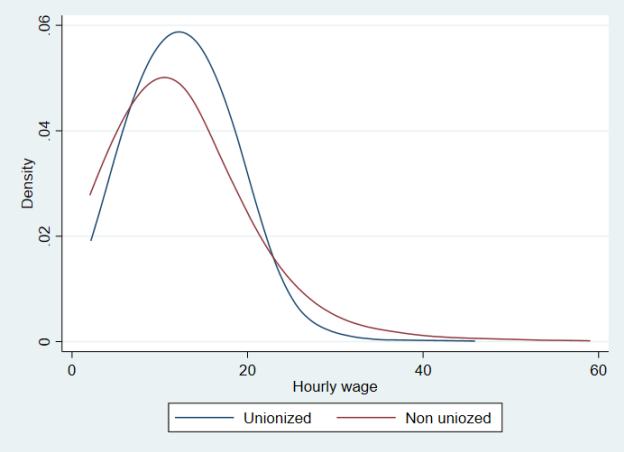
| Union dummy | Obs | Mean | Std.Dev | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|---|
| | | | | Hourly wage | | | | |
| 0 | 2975 | 12,77 | 9,99 | 2,02 | 6,88 | 10,81 | 15,74 | 191,37 |
| 1 | 970 | 12,48 | 4,32 | 2,14 | 9,82 | 12,10 | 14,60 | 45,93 |

Without outliers, the maximum wage for non unionized workers is still over the maximum hourly wage for unionized workers (table4), this can be explained by the fact that labor unions fight for a more fair and normal distribution of wages which relies on the job held and not on personal skills or characteristics. This intuition seems confirmed by figure1 that shows the hourly wage distribution across unionized and non unionized workers. The distribution of hourly wages for unionized workers is a normal distribution with a standard deviation of 4,32 and is centered around $12. But the distribution of hourly wages for non unionized workers is asymmetric with a higher standard deviation 7,34 , and it has a tail in the right direction. This is typical of distribution with a positive skewness that induces a median value lower than the average value and the conclusion is that hourly wages are not independent and identically distributed on the non unionized workers sample.

Table 3: Hourly wages statistics in function of Union membership without outliers

| Union dummy | Obs | Mean | Std.Dev | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|---|
| | | | | Hourly wage | | | | |
| 0 | 2952 | 12,22 | 7,34 | 2,02 | 6,86 | 10,77 | 15,50 | 59,04 |
| 1 | 970 | 12,48 | 4,32 | 2,14 | 9,82 | 12,10 | 14,60 | 45,93 |

Figure 1: Distribution of hourly wages in function of the Union membership




# 3 Presenation and estimation of the model

The model is based on the standard union effect from **Lewis (1986)** using an extended mincer equation with the add of a dummy variable to compute the union membership.

$$ln(W_{it}) = \beta_0 + \beta_1 X_{it} + \delta U_{it} + v_{it} \tag{1}$$

- $ln(W_{it})$ the hourly wage in logarithm at time t for an individual i
- $X_{it}$ a vector with individuals skill characteristics, to avoid collinearity the age variable is excluded
- $U_{it}$ the dummy variable that takes 1 if the worker is unionized and 0 otherwise, the estimate coefficient $\delta$ is the estimate value of the union wage premium since it represents the percentage hourly wage difference between unionized and non-unionized workers

- $v_{it}$ a composit error term that contain $u_{it}$ an idiosyncratic component of unobservable variables that varies across time and individual and an unit specific term $\alpha_i$ that is constant over time

And the following assumption should be true for this model:

- H1 : The model is correctly specified
- H2.1 : $E(\alpha_i \alpha_j | X_{it}) = 0$ if $i \neq j$ , the individual effect is uncorrelated across individuals.
- H2.2: $E(u_{it} u_{js} | X_{it}) = 0$ if $i \neq j$ and $t \neq s$, the exogeneity restriction, the error term is uncorrelated across individuals and time with explanatory variables.
- H3.1 : $Var(v_{it}|X) = \Theta^2$, homoscedasticity restriction, the variance should be constant across time and individuals.
- H3.2 : $Cov(v_{it} vjs | X_{it}) = 0$, no serial correlation restriction, errors term are not link with a reproduction processus across time or individuals.
- H4 : There is no strict multicollinearity between the explanatory variables.
  The matrix $X'_{it} X_{it}$ is of full rank and invertible

## 3.1 Pooled Ordinary Least Squares

The model is first estimate via OLS, and recall that for panel data two main assumptions to respect are :

$$E(\alpha_i \alpha_j | X_{it}) = 0$$

$$E(u_{it} u_{js} | X_{it}) = 0$$

In other words, if any of the two error component is correlated with one of the dependant variable, the estimations are biased and inconsistent.

Table 4: Estimatiom by Pooled OLS

| Variables | lhwage | se | tstat |
|---|---|---|---|
| educ | 0.1215 | (0.0037) | 32.628 |
| exp | 0.0336 | (0.0040) | 8.376 |
| exp2 | -0.0004 | (0.0001) | -4.338 |
| tenure | 0.0306 | (0.0029) | 10.443 |
| tenure2 | -0.0007 | (0.0001) | -6.199 |
| black | -0.2091 | (0.0195) | -10.706 |
| **union** | **0.1286** | (0.0174) | **7.369** |
| Constant | 0.1846 | (0.0608) | 3.033 |
| Observations | 3,945 | | |
| R-squared | 0.351 | | |

This estimation of the model gives a **union wage premium of 12,85%**, the unionized workers earn at average an hourly wage that is 12,85% higher than non unionized workers, in a labor market that evolved with labor union.
However, using a White test the estimation is heteroscedastic (*table9*) it means that the error term is dependant of at least one of the independant variable and thus the variance is not constant across time and or individuals. A robust estimation could be used to correct the variance matrix. But since membership is suspect to be non-random there is likely to be non-zero correlation between membership and the unit specific term, the model as an endogenous variable and estimations are biased.

## 3.2 Non random selection and endogeneity of the union variable

Non random selection arises because an individual's membership decision is probably based on characteristics that may also affect earnings. If so, and it is not possible to control for all characteristics affecting wage and union variable simultaneously, some correlation between the error term and membership can be expected.

### 3.2.1 Fixed effect estimator

A first solution is to use the fixed effect estimator to eliminate the individual effects and estimate a unbisaed and consistent union wage premium as **Hildreth (1999)** use the fixed effect to reduce issue link to non random selection in panel data.

$$ln(W_{it}) - ln(\overline{W_i}) = \beta_1(X_{it} - \overline{X_i}) + \delta(U_{it} - \overline{U_i}) + (u_{it} - \overline{u_i}) \quad (2)$$

Table 5: Estimation by fixed effect

| Variables | lhwage | se | tstat |
|---|---|---|---|
| o.educ | - | | - |
| exp | 0.0552 | (0.0066) | 8.401 |
| exp2 | -0.0007 | (0.0002) | -3.977 |
| tenure | 0.0178 | (0.0032) | 5.476 |
| tenure2 | -0.0005 | (0.0001) | -3.573 |
| o.black | - | | - |
| union | 0.0521 | (0.0236) | 2.207 |
| Constant | 1.5314 | (0.0654) | 23.412 |
| Observations | 3,945 | | |
| R-squared | 0.069 | | |
| Number of id | 789 | | |

*Education* and *race* variables have been omitted because they do not vary from their respective average value across time. An agent will not be educated more through time and his skin will not change color. With this estimation method, **the union wage premium is 5,21%**, this is still in accordance with literature even if it is lower than with the first biased OLS estimation method. Thus it means that on average a unionized worker will earn a wage 5,21% higher than a non unionized worker. Stata gives an estimation of the correlation between independent variables and the unit specific term of 0,069 so different from 0 which confirms the hypothesis of bias in the first estimation method.

The *experience* and *tenure* variables have as expected a positive coefficient and their square have a negative coefficient because of decrease in the marginal effect across time. Thus one more year of **experience** induces an increase of **5,38%** $(0,0552 - 2 * 0,0007 = 0,0538)$ in the hourly wage, other variables being constant. For the tenure variable, i.e. the number of year spend with the actual job, one more year of **tenure** induces an increases of **1,68%** $(0,0178 - 2 * 0,0005 = 0,0168)$ in the hourly wage, other variables being constant.

### 3.2.2 Simultaneous equation model

Another solution is to use a simultaneous equation model since there is a reverse causality between unionized decision and hourly wages. The decision of unionized is endogenous to the level of hourly wages and vice versa.

$$ln(W_{it}) = \beta_0 + \beta_1 X_{it} + \delta U_{it} + v_{it} \tag{3}$$
$$U_{it} = \alpha_0 + \alpha_1 Z_{it} + \omega ln(W_{it}) + e_{it} \tag{4}$$

- $ln(W_{it})$ and $U_{it}$ exogenous variables
- $X_{it}$ a vector of exogenous variables with individuals skill characteristics, to avoid collinearity the age variable is excluded
- $Z_{it}$ a vector of exogenous variables with job characteristics link to the union decision

Both equations are over-identified since the number of exogenous variables not present in the other model is higher than one. For the wage estimates equation, we have an over-identification of 6 but if all the exogenous estimated coefficients of the union equations are statically different from zero then the equation is empirically identified. The union estimates equation has an over-identification of 5 but once more if the estimated exogenous coefficients of the wage equation are statically different from zero the equation will also be empirically identified.
In other terms, the instruments have to be correlated with the endogenous variables to be consistent. Let us regress the endogenous regressors on the exogenous one:

$$ln(W_{it}) = \omega_0 + \beta_1 X_{it} + \alpha_1 Z_{it} + v_{it} \tag{5}$$
$$U_{it} = \omega_0 + \beta_1 X_{it} + \alpha_1 Z_{it} + e_{it} \tag{6}$$

After testing the model, the dummy variables *clerk* and *service* are removed because they are not significant at 5% following a student distribution. Then once this variables are removed from the $Z_{it}$ vector, equations are supposed to be empirically identified.

The OLS estimator will be biased for those equations because there is endogeneity

$$Cov(ln(W_{it}); e_{it}) \neq 0$$
$$Cov(U_{it}; v_{it}) \neq 0$$

The usual method is to use two-stage least square estimator. This is an estimation method in two parts, first it needed to determine a form equation and estimate it by OLS and in a second step, use the prediction of stage one as a substitute for the endogenous variable in the structural equation. Let us determine the union reduced form equation:

$$U_{it} = \pi_{0u} + \pi_{1u} * X_{it} + \pi_{2u} * Z_{it} + v_{it} \tag{7}$$

The union variable is estimates and then use to estimate the hourly wage. This gives the following results.

Table 6: Two-stage least square

| Variable | lhwage | se | tstat |
|---|---|---|---|
| **union** | **-0.3267** | (0.0568) | **-5.749** |
| educ | 0.1116 | (0.0042) | 26.560 |
| exp | 0.0259 | (0.0044) | 5.828 |
| exp2 | -0.0003 | (0.0001) | -2.639 |
| tenure | 0.0449 | (0.0036) | 12.496 |
| tenure2 | -0.0010 | (0.0001) | -7.657 |
| black | -0.1613 | (0.0219) | -7.367 |
| Constant | 0.4199 | (0.0715) | 5.874 |
| Observations | 3,945 | | |
| R-squared | 0.239 | | |

The estimate coefficient of the difference hourly wages between unionized and non unionized workers is negative, it would induces that non unionized worker have a larger income then unionized workers. This is not in accordance with literature and previous estimations. This issue arise from an estimation problem, a two-least square estimator is apply on a dummy endogenous variable so it means that the dummy union is estimate with an oridnary least square estimator in the first stage. This can lead to prediction under 0 or over 1 that do not have sens. Let us study the estimation of union in the first stage.

Table 7: Study of union prediction at first stage

| non sens | Freq | Percent |
|---|---|---|
| 0 | 3464 | 87.81% |
| 1 | 481 | 12.19% |
| Total | 3945 | 100% |

This allows to affirm that 12,19% of the sample is wrong estimate when using an ordinary least square estimator to estimate a dummy dependant variable. Several searcher worked on developping a so called "Simultaneous logit" estimator to resolved this issue, **Schmidt et al. (1975)** and then in **Schmidt (1978)** he focused on the estimation of a simultaneous equations model with jointly dependent continuous and qualitative variables to access the union earning question.

*The model could be rewriting as:*

$$log\frac{P(U_{it}=1)}{P(U_{it}=0} = \alpha_0 + \alpha_1 Z_{it} + \omega[(ln(W_{it})|U_{it}=1) - (ln(W_{it})|U_{it}=0)] + e_{it} \quad (8)$$

$$ln(W_{it}) = \beta_0 + \beta_1 X_{it} + \delta U_{it} + v_{it} \quad (9)$$

However, this modelisation will not be use in this study.

# 4 Limitations and model extension

In his review of the literature, **Lewis (1986)** concluded that, due to the deficiencies of simultaneous equation and panel techniques, the most appropriate way to estimate the impact of unions on wages is using simple linear regression with cross section data. The idea is that the estimation of the union wage premium suffers from omission of control variables correlated with the union status.

One assumption is that some of the wage premium attributed to union membership is, in fact, attributable in part to the characteristics of members, their jobs and their employers which would give them higher wages than nonmembers in any case even in a labor market without union. **Bryson (2002)** follows this idea and try to estimate how much of the wage differential between unionized and non unionized workers is attributable to union membership, and how much is due to differences in personal, job and workplace characteristics. He showed that the addition of workplace-level data to the individual data substantially reduces the size of any membership effect, suggesting that some of the union effect attributed to membership in analyses based on individual or household data is actually related to the 'better' paying workplaces that members enter, thus personal characteristics.

There is no incentive to believe that the union wage premium is constant and similar in every industry, for every size of firm and between private and public sectors. Possible extension of the model is to add more control variables to estimate different union wage premium for different contexts. Here is a sum up of research that studied those topics.
**Bratsberg (2002)** explains that the effects of deregulation, heightened import competition, and technological change are felt unevenly across industries. At the aggregate level the union premium has become more stable at the end of the 90' but there is a question as to whether this stability is masking divergent trends at the industry level. For example, such industries as durable-goods manufacturing and communications and utilities, the union premium has trended upward, while unions in other industries such as construction and wholesale and retail trade, have seen their union wages premium decreased.
**Da Silva (2012)** studied the variation of wage premium across different sizes of firm and it appears that some of the observed skills; namely, education, age, and tenure have high returns in large firms. On the other hand, the price of non-observed skills is reduced as firm size increases. This finding is consistent with explanations based on the premise that large employers have more difficulty monitoring workers, which therefore leads them to monitor less closely. Then the wage premium should increase with the size of the firm.
Robinson (1984) exerges that union status appears to be strongly affected by the expected wage gain from joining the unionized sector. There is some evidence of larger union gains in the public sector than in the private sector.

A last point is that since the beginning of the 2000' labor union decreases and thus so is the union wage premium, less and less workers are unionized. One of the explanations of this situation is called the "threat effect",it results from non-union employers raising the wages of their workers in order to avoid becoming unionized **Rosen (1969)**. The less workers are unionized, the less is the bargaining power of the labor union and it becomes difficult to negotiate for a significant increase of hourly wages.

# 5 Conclusion

The goal of this study was to estimate how large was the union wage premium in USA between 1983 and 1987. Here, the wage premium is the difference between the hourly wages for unionized and non-unionized workers. In accordance with literature, depending on the estimation method, the wage premium seems to be between 5% and 12%. It means that unionized workers earn an hourly wage between 5% and 12% higher than non-unionized workers. However, this research fail to estimate a sensitive coefficient using the simultaneous equation method that will maybe lead to more significant estimation.
Moreover, there are an incentive to think that the estimation of the union wage premium suffers from omission of control variables correlated to the union status such as the size of the firm, the industries sector and so on. And it may have bias in the model because it do not take enough in account private characteristics of workers.

# 6 References


Belfield, C. R., Wei, X. (2004). Employer size–wage effects: evidence from matched employer–employee survey data in the UK. Applied Economics, 36(3), 185–193. https://doi.org/10.1080/0003684042000175316

Bratsberg, B., Ragan, J. F. (2002). Changes in the Union Wage Premium by Industry. Industrial and Labor Relations Review, 56(1), 65. https://doi.org/10.2307/3270649

Cerejeira, J., Guimarães, P. (2012). The price of unobservables and the employer-size wage premium. Economics Letters, 117(3), 878–880. https://doi.org/10.1016/j.econlet.2012.06.042

Hildreth, A. (1999). What Has Happened to the Union Wage Differential in Britain in the 1990s? Oxford Bulletin of Economics and Statistics, 61(1), 5–31. https://doi.org/10.1111/1468-0084.00114

Hirsch, B. T., Schumacher, E. J. (2001). Private Sector Union Density and the Wage Premium: Past, Present, and Future. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.283203

Lemieux, T. (1998). Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection. Journal of Labor Economics, 16(2), 261–291. https://doi.org/10.1086/209889

Robinson, C., Tomes, N. (1984). Union Wage Differentials in the Public and Private Sectors: A Simultaneous Equations Specification. Journal of Labor Economics, 2(1), 106–127. https://doi.org/10.1086/298025

Rosen, S. (1969). Trade Union Power, Threat Effects and the Extent of Organization. The Review of Economic Studies, 36(2), 185. https://doi.org/10.2307/2296836

Schmidt, P. (1978). Estimation of a Simultaneous Equations Model with Jointly Dependent Continuous and Qualitative Variables: The Union-Earnings Question Revisited. International Economic Review, 19(2), 453. https://doi.org/10.2307/2526312

Schmidt, P., Strauss, R. P. (1975). Estimation of Models with Jointly Dependent Qualitative Variables: A Simultaneous Logit Approach. Econometrica, 43(4), 745. https://doi.org/10.2307/1913083

Union Relative Wage Effects: A Survey by Lewis Gregg Lewis H. Gregg (1986–02-01) Hardcover. (1986). Academic Service.

# 7 Appendices

## 7.1 Table of contents

# Contents



## 7.2 Annexes

Table 8: Outliers: observations interval

| Group | lowest threshold | highest threshold |
|---|---|---|
| Non Unionized | -37,42 | 60,40 |
| Unionized | -14,08 | 38,5 |

Table 9: White test on Pooled OLS

| | Chi2 | df | p |
|---|---|---|---|
| Heteroscedasticity | 622.37 | 31 | 0.000 |

## 7.3 Do files

```
cd "D:\AMSE\Cours\Mag2\S2_NTNU\Econometrics\Exam\2021"

/* Open database */
clear all
use data

/*****************************************************************************/
/* DESCRIPTION DATA */
gen hwage = laborinc/hours
gen lhwage = log(hwage)
gen exp2=exp*exp
gen tenure2=tenure*tenure

/* Declare panel data */
xtset id year

describe
summarize


/*Analyse du salaire horaire en fonction de Union*/
table union, c(n hwage mean hwage sd hwage)
table union, c( min hwage p25 hwage median hwage p75 hwage max hwage )


/* Outliers */
/* After manually computing outlier thresholds, we delete some observations */
drop if hwage >60

table union, c(n hwage mean hwage sd hwage)
table union, c( min hwage p25 hwage median hwage p75 hwage max hwage )

graph twoway (kdensity hwage if union == 1, bw(5))
(kdensity hwage if union == 0, bw(5)), xtitle("Hourly wage") ytitle("Density")
 legend(label(1 "Unionized") label(2 "Non uniozed"))

/*Compare to a normal distribution for iid*/
/*Weak iid*/
qnorm hwage if union==0
/*Strong iid*/
qnorm hwage if union==1

/*****************************************************************************/

/* REGRESSION */

        /* Pooled OLS */
        reg lhwage educ exp exp2 tenure tenure2 black union
        est store Pooled
        predict v_hat, residual
                outreg2 [Pooled] using Table, bdec(4) sdec(4) noaster tex(frag)
                replace stat(coef se tstat)  sideway

        /*Correlation between suspect endogenous and residuals*/
```

```
        correlate union v_hat
        /*No correlation*/

        /*White test for heteroscedasticity*/
        estat imtest, white
        /*H0 is rejected, heteroscedasticity*/

        /*Robust estimation*/
        reg lhwage educ exp exp2 tenure tenure2 black union, robust


        /*Fixed effect*/
        xtreg lhwage educ exp exp2 tenure tenure2 black union, fe
                est store FE
                outreg2 [FE] using Table, bdec(4) sdec(4) noaster tex(frag)
                replace stat(coef se tstat)  sideway

/*************************************************************************/

/*SIMULTANEOUS EQUATION*/
        /* Identification of instruments*/
        reg lhwage educ exp exp2 tenure tenure2 black prof manager Sales craft
operator
        reg union educ exp exp2 tenure tenure2 black prof manager Sales craft
operator


        /*Reduce and structural form for 2SLS*/
        ivregress 2sls lhwage educ exp exp2 tenure tenure2 black
        (union=prof manager Sales craft operator )

        reg union educ exp exp2 tenure tenure2 black prof manager Sales craft
operator
        predict union_hat
        gen nonsens=0
        replace nonsens=1 if union_hat <0 | union_hat>1
        tab nonsens
        est store nonsens
        outreg2 [nonsens] using Table

        reg lhwage educ exp exp2 tenure tenure2 black union_hat
```