LABOR DEMAND ESTIMATION

The relation we want to estimate is: ln = f(lw, ly) [labor demand function]

Introduction

The first part of the document will analyze the labor_data.dta dataset omitting observations that have an "id" included between 11 (included) and 20 (excluded). The aim of this first part is to estimate the labor demand function ln = f(lw, ly). The data frame includes a bunch of data about different firms.
In the first stage, a cross-section approach will be undertaken. Each variable (ln, lw, ly) will be singularly analyzed with the purpose of spotting the presence of outliers (severe and mild). Moreover, a normality test will be performed. Consequently, the labour demand function for the year 1997 will be estimated through an OLS (ordinary least squares) model. Test of heteroskedasticity, normality and reset test will be run on it.
Furthermore, the labor demand for year 1997 will be estimated through the IV method. In this case, the instrument variables will be the first-lagged data. A Hausman test will be done to compare the OLS and IV estimates.
In conclusion, data will be explored with a panel approach. In this context, the labor demand function will be estimated with two different approach: a pooled OLS model and a fixed-effects model. A comparison of these two models will be finally done.

A preliminary specification

First of all can be useful define the variables we are going to use in our analysis. *Employment* indicates the "average listed total employment (without part-timers and non-listed employees) for the reporting year".[1] The *unit wage cost* is the "total wage bill divided by employment in the same year".[2] Lastly, the *Output* is the "production output in current prices (without VAT, excise tax and special tax) for the reporting year". [3]
Especially, the variables employed are the logarithm of the variable above. In this sense, *ln* is the logarithm of the Employment, *lw* is the logarithm of the unit wage cost (the ratio between the total wage bill and the employment) and *ly* is the logarithm of the output. Please note that all the variables are measured for two years: 1997 and 1996.
Making some *ex-ante* consideration, it is possible to expect that the *ln* and the *lw* will be negatively correlated. An increase in unit wage cost probably will lead the employers to demand for less workers.
Furthermore, we expect that *ln* will be positively correlated with *ly*. If the output increase, it is needed more workforce to produce it. In this sense, the labor demand will increase and employment will increase too.

1. Cross-section approach

Table 1. Preliminary analysis of the dependent variable and the explanatory variables

	ln	lw	ly
N observations	1420	1420	1420
Median (50%)	4.9523	4.4736	8.3652
Mean (10% trim)	4.8942	4.4699	8.2984

[1] Jozef Konings and Hartmut Lehmann, "Marshall and Labor Demand in Russia: Going Back to Basics*", Journal of Comparative Economics* 30 (2002): 157.

[2] Jozef Konings and Hartmut Lehmann, "Marshall and Labor Demand in Russia: Going Back to Basics*", Journal of Comparative Economics* 30 (2002): 157.

[3] Jozef Konings and Hartmut Lehmann, "Marshall and Labor Demand in Russia: Going Back to Basics*", Journal of Comparative Economics* 30 (2002): 157.

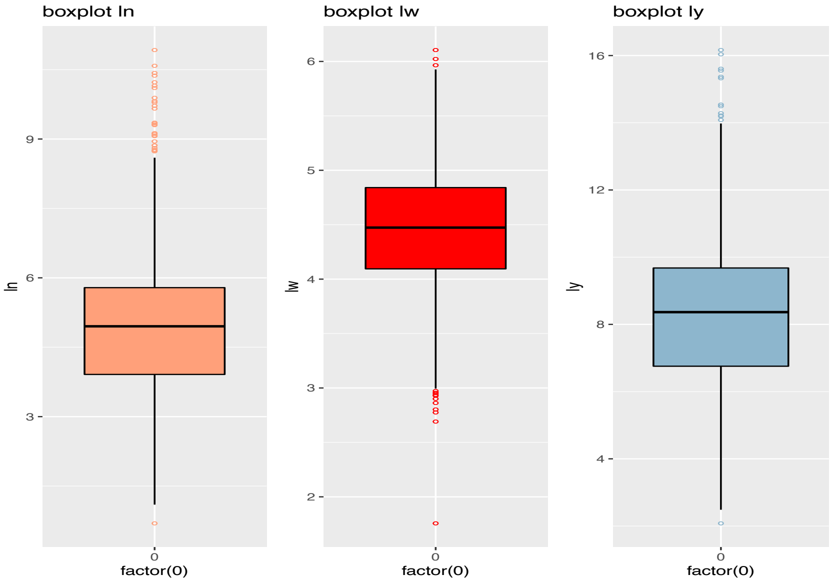
Mean	4.9602	4.4519	8.3345
Std. Deviation	1.4469	0.5501	2.0728
N mild outliers	28	13	13
N severe outliers	0	1	0
N total outliers	28	14	13
normality (Jarque-Bera test) [1]	< 2.2e-16***	< 2.2e-16***	< 2.2e-16***

[1] For the normality test please write only p-values

*Note. *p<0.1; **p<0.05; ***p<0.01. Source of the data: dati_esame.dta. Trimmed mean excludes 10% of the largest and smallest value.*

The presence of Outliers while doing an OLS estimator can lead the estimator to a bias. Regarding this, in the preliminary analysis it is useful to understand if they are present or not. In the data frame there are 54 mild outliers and just 1 severe outlier. Overall just the 3,87% of the observations can be considered as an outlier. This means that the 3rd least assumption of the OLS model (large outliers are unlikely) is respected. Moreover, in every variable mean and trimmed mean are quite closer, underlying the fact that extreme observations doesn't influence too much the computation (for a further proof, trimmed means for every variable excluding just the percentage of outliers are even closer to the variable's mean). Probably this is due to the high number of observations, thus a bunch of observations (even if much bigger or much smaller than the other) doesn't matter too much. Comparing the mean and the median, just the variable *ln* has a mean that is bigger than the median. However, just the variable *lw* has a negative skewness (-0.3211482). In any case, the normality test strongly reject the null-hypothesis that the variable distribution is normal for each one of the variable. So, neither *ln* nor *lw* and *ly* are distributed as a normal.

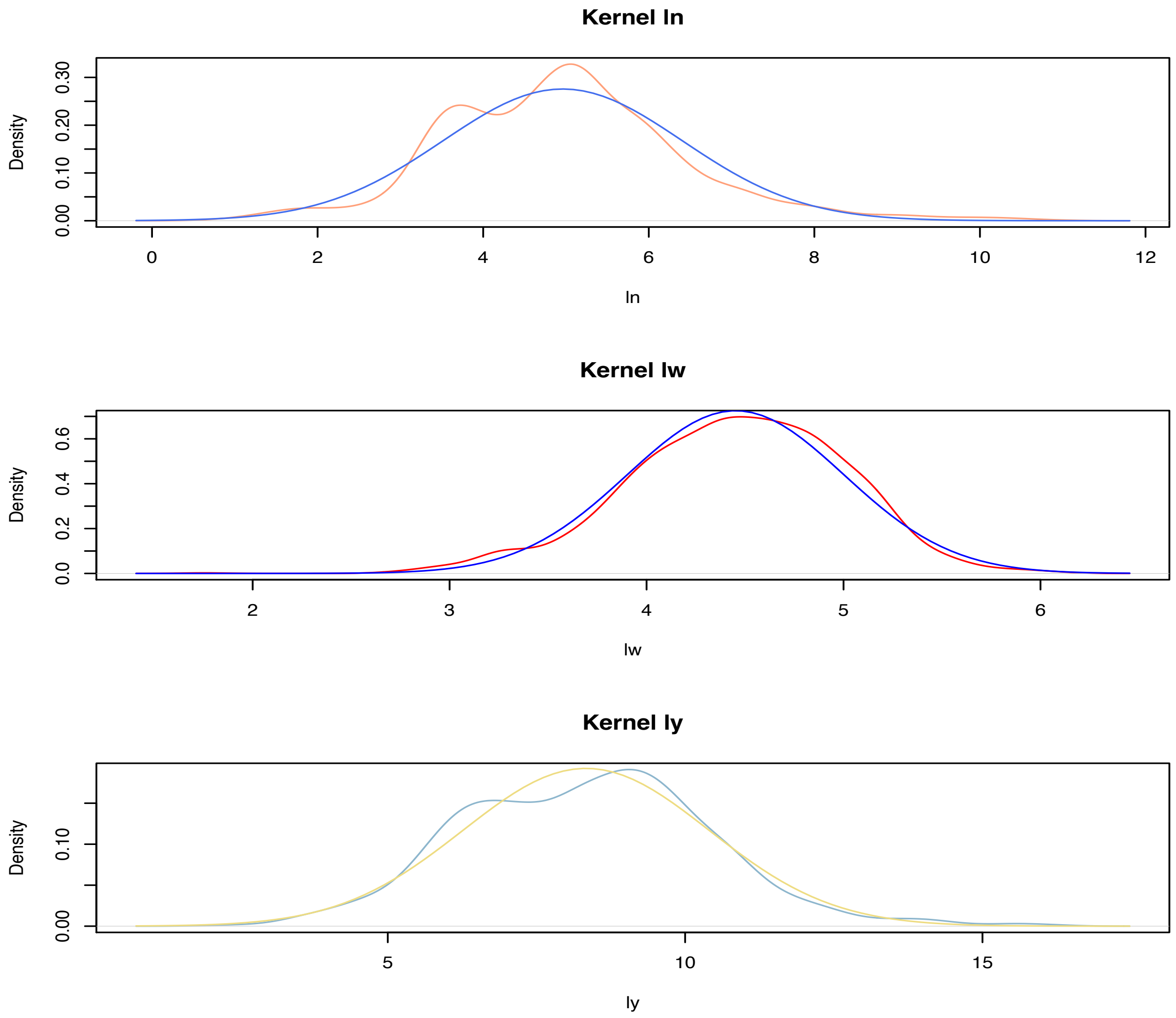
Figure 1. Boxplots of the dependent variable (*ln*) and the independent variables (*lw*, *ly*)



Note. Source of the data: dati_esame.dta.

Figure 1 graphically represent data of *table 1* for each variable considered, especially shows the median, the interquartile range and the corresponding outliers. Note that just the variable *lw* has a severe outlier (1.756118).

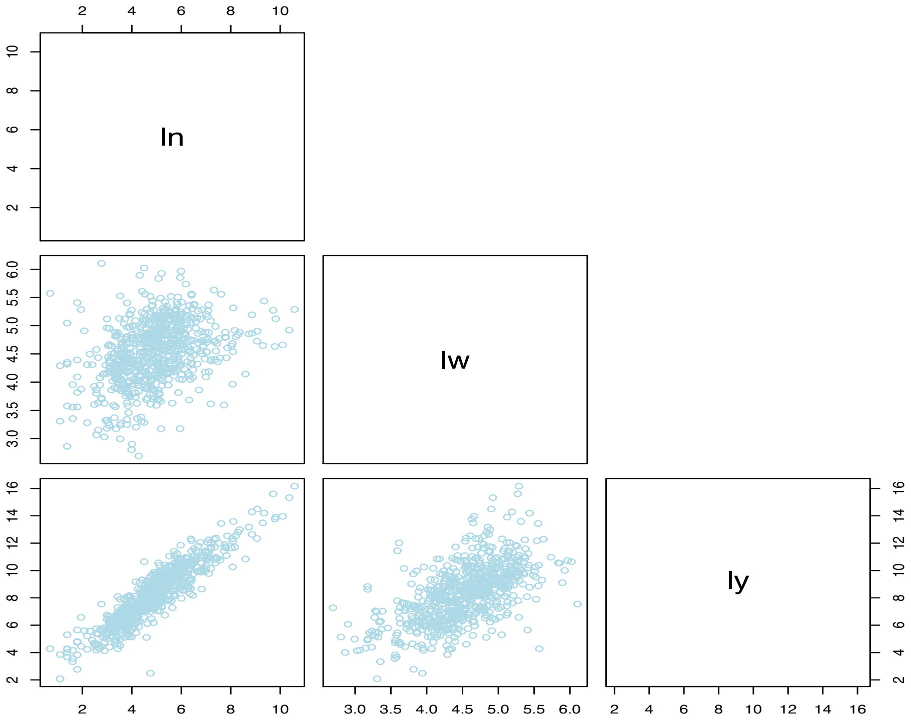
Figure 2. Density plot of the dependent variable (*ln*) and the independent variables (*lw*, *ly*)



Note. Source of the data: dati_esame.dta. Kernel ln: orange line = empirical distribution; blue line = normal distribution; Kernel lw: red line = empirical distribution; blue line = normal distribution; Kernel ly: Lightblue line = empirical distribution; yellow line = normal distribution.

Figure 2 shows the empirical distributions of *ln*, *lw* and *ly*. Moreover, these charts show the normal distribution for each one of the considered variables. These Kernel density plots illustartes graphically the difference between the normal distribution and the empirical one for every variable. Interesting is to note that the empirical distribution of the *ln* and *ly* are clearly bimodal, maybe due to the fact that this observations are collected for two different years (1996 and 1997).

Figure 3. Scatter plots of ln, lw and ly for the year 1997



Note. Source of the data: dati_esame.dta. Year: 1997

Figure 3 illustrates a scatter plot for each variables' pair. Looking to the charts, it is evident how *ln* and *ly* have a strong relation. In this sense, it is possible to expect that their correlation will be quite high. On the contrary, the relation between *ln* and *lw* seems not to be so clear as the one above. A special look is reserved for the *ly*-*lw* relation: it is not a perfect linear correlation, and so lead our model to avoid perfect multicollinearity, respecting the 4th least assumption for the multivariate OLS model (precisely, *lw*-*ly* correlation is 0.5970348).

Table 2. Estimated OLS models for the labor demand in 1997

```
==================================================================================================
                                                 Dependent variable:
                    ------------------------------------------------------------------------------
                                                          ln
                              (1)                        (2)                        (3)
--------------------------------------------------------------------------------------------------
lw                          0.998***                                             -0.708***
                            (0.093)                                               (0.042)

ly                                                     0.636***                   0.747***
                                                       (0.010)                    (0.011)

Constant                     0.396                    -0.382***                   1.901***
                            (0.422)                    (0.090)                    (0.156)

--------------------------------------------------------------------------------------------------
Observations                  710                        710                        710
R2                           0.141                      0.839                      0.884
Adjusted R2                  0.140                      0.839                      0.884
Residual Std. Error     1.356 (df = 708)           0.587 (df = 708)           0.498 (df = 707)
F Statistic         116.151*** (df = 1; 708) 3,682.981*** (df = 1; 708) 2,704.389*** (df = 2;
707)
==================================================================================================
```

Note. *p<0.1; **p<0.05; ***p<0.01. Source of the data: dati_esame.dta. Model (1) = ln ~ lw; Model (2): ln ~ ly: Model (3): ln ~ lw + ly

These models are OLS models to estimate the labor demand function. Number *(1)* and *(2)* are bivariate models, otherwise model *(3)* is a multivariate model. *Model (1)* shows us that the regressor *lw* is relevant in explaining *ln*. However, this model goes against our *ex-ante* intuition of the negative relation between these two variables. In *model (2)* just the regressor *ly* is considered to explain *ln*. In this case we see that *ly* is relevant in explaining *ln*, and in this model the constant became relevant too. However, it is evident that the most appropriate OLS model to estimate the labor demand for 1997 is *Model (3)*. Within it, all the regressor are singularly significant for an alpha of 0.01. It confirms the initial intuition: *lw* and *ln* are negatively correlated while *ln* and *ly* have a positive relation. Moreover, another confirmation is given by changes in the coefficient: these are pretty high, and this lead us to think that in *model (1)* and *(2)* there was an omitted variable bias.
In conclusion, an F-test on the joint significance of the regressors reject the null-hypothesis of not statistically significant relevance of the regressor (p-value < 2.2e-16). That is, regressors are statistically jointly relevant in explaining the dependent variable.

Table 3. Breusch-Pagan test for heteroskedasticity

```
 Breusch Pagan Test for Heteroskedasticity
=====================================================================================
 Ho: the variance is constant
 Ha: the variance is not constant

             Data
-------------------------------------------------------------------------------------
 Response : ln
 Variables: fitted values of ln

         Test Summary
-------------------------------------------------------------------------------------
 DF             =    1
 Chi2           =    12.60316
 Prob > Chi2    =    0.0003850961
=====================================================================================
```

Note. Source of the data: dati_esame.dta.

The Breusch-Pagan test verifies the null-hypothesis that the variance of the residual is constant (presence of homoskedasticity). However, the null-hypothesis is rejected for every value of alpha (p-vale: 0.0003850961) . That is, heteroskedasticity is present. Hence the residuals are heteroskedastics: the variance of its conditional distribution depends on the regressor. Since heteroskedasticity is present, our OLS estimator is not the BLUE (best linear conditionally unbiased estimator) and a robust to heteroskedasticity formula is needed to compute the standard errors. For this purpose, the White formula is used.

Table 4. Estimated OLS model with robust standard errors for the labor demand

```
=====================================================================================
Coefficients:
            Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF

(Intercept)   1.9007    0.16987   11.19  7.137e-27   1.5672   2.2342 707
lw           -0.7076    0.05019  -14.10  5.895e-40  -0.8061  -0.6091 707
ly            0.7469    0.01608   46.44  6.067e-217   0.7153   0.7785 707
-------------------------------------------------------------------------------------
Multiple R-squared:  0.8844 ,    Adjusted R-squared:  0.8841
F-statistic:  1393 on 2 and 707 DF,  p-value: < 2.2e-16
-------------------------------------------------------------------------------------
              (Intercept)          lw          ly
```

```
const          0.1564488     0.04235461     0.01107644
HC0            0.1689816     0.04991342     0.01599615
HC1            0.1693397     0.05001920     0.01603005
HC2            0.1698678     0.05019083     0.01608268
HC3            0.1707620     0.05047082     0.01616988
-----------------------------------------------------------------------------------------
Wald Intervals:
                                 2.5 %     97.5 %
(Intercept)                  1.5682593  2.2331973
lw                          -0.8057991 -0.6093912
ly                           0.7154163  0.7783607
=========================================================================================
```

Note. Source of the data: dati_esame.dta. White formula (HC1) has been used to computed robust standard errors.

This is the computation of the robust standard errors made by the White formula. The standard errors values give another proof of the presence of heteroskedasticity: indeed, they differ from the ones computed with the homoskedasticity-only formula. Furthermore, are reported the Wald intervals for the regressors and the constant with a confidence level of the 95%.

Table 5. Normality test of the residuals and RESET test

```
=========================================================================================
Title:
 Jarque - Bera Normalality Test
-----------------------------------------------------------------------------------------
Test Results:

  STATISTIC:                 X-squared: 989.742
  P VALUE:                   Asymptotic p Value: < 2.2e-16

=========================================================================================
=========================================================================================
RESET test
-----------------------------------------------------------------------------------------
data:  OLS_labdem_1
RESET = 5.4459, df1 = 3, df2 = 704, p-value = 0.00105
=========================================================================================
```

Note. Source of the data: dati_esame.dta.

The normality test for the residuals strongly rejects the null-hypothesis that the residuals are normally distributed. The Ramsey Regression Specification Error Test (RESET) null-hypothesis is correctly specified and so there is not an omitted variable bias. Our Reset test reject the null-hypothesis for every value of alpha. That is, our model suffers from misspecification and of omitted variable bias. Note that the RESET test is also used to catch if the model misses some important nonlinearities. In this sense, the model has missed important nonlinearities.

Table 6. Estimated OLS model for 1997 labor demand with homoskedasticity-only standard error (1) and robust standard errors (2)

```
===========================================================
                                  Dependent variable:
                              ---------------------------
                                           ln
                                   (1)            (2)
-----------------------------------------------------------
lw                              -0.708***      -0.708***
                                (0.042)        (0.050)

ly                               0.747***       0.747***
                                (0.011)        (0.016)

Constant                         1.901***       1.901***
                                (0.156)        (0.169)

-----------------------------------------------------------
Observations                       710            710
R2                                0.884          0.884
Adjusted R2                       0.884          0.884
Residual Std. Error (df = 707)    0.498          0.498
F Statistic (df = 2; 707)      2,704.389***  2,704.389***
===========================================================
```

*Note. *p<0.1; **p<0.05; ***p<0.01. Source of the data: dati_esame.dta. model (1) = homosekdasticity only standard errors; model (2) = robust standard errors*

Model (1) and *Model (2)* estimate the labor demand, the first has non-robust standard errors while the latter has standard errors computed with the White formula. The regressor remain exactly the same for both models. Moreover, the standard errors of every regressor are bigger in *model (2).* However, as expected, all the regressors remain highly significant in explaining *ln*. Starting from there, the model considered is *Model (2).*

Table 7. Instrumental variable estimation for the 1997 labor demand

```
=====================================================================================
Call:
ivreg(formula = ln ~ lw + ly | lwlag + lylag, data = .)
-------------------------------------------------------------------------------------
Residuals:
    Min      1Q  Median      3Q     Max
-1.5513 -0.2822 -0.0125  0.2434  3.9125
-------------------------------------------------------------------------------------
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.20902    0.18099   12.21   <2e-16 ***
lw          -0.84044    0.05007  -16.78   <2e-16 ***
ly           0.78210    0.01188   65.81   <2e-16 ***
-------------------------------------------------------------------------------------
Residual standard error: 0.502 on 707 degrees of freedom
Multiple R-Squared: 0.8824,      Adjusted R-squared: 0.882
Wald test:  2712 on 2 and 707 DF,  p-value: < 2.2e-16
=====================================================================================
```

*Note. *p<0.1; **p<0.05; ***p<0.01. Source of the data: dati_esame.dta. As instruments have been used the 1 period lagged data.*

In this case, the labor demand for 1997 has been estimated using as instrument the one period lagged data of the regressors. In practice, this means that have been used *lw* and *ly* data in the year 1996. As it is possible to see, *lw* still has a negative coefficient while *ly* has a positive one. Intuitively, seeing that instruments are lagged data of the regressor it is probably that will be correlated to the variables (*lw-lwlag* correlation = 0.8893117; *ly-lylag* correlation = 0.9823900) . That is, our instruments are relevant (first condition for a valid instrument is met).

Table 8. Instrumental variable model (1) and OLS model (2) for the labor demand in 1997

```
======================================================================
                                       Dependent variable:
                              ---------------------------------------
                                               ln
                              instrumental           OLS
                                variable
                                  (1)                (2)
----------------------------------------------------------------------
lw                               -0.840***          -0.708***
                                  (0.050)            (0.050)

ly                                0.782***           0.747***
                                  (0.012)            (0.016)

Constant                          2.209***           1.901***
                                  (0.181)            (0.169)

----------------------------------------------------------------------
Observations                        710                710
R2                                 0.882              0.884
Adjusted R2                        0.882              0.884
Residual Std. Error (df = 707)     0.502              0.498
F Statistic                                 2,704.389*** (df = 2; 707)
======================================================================
```

*Note. *p<0.1; **p<0.05; ***p<0.01. Source of the data: dati_esame.dta. model (1) = instrumental variables; model (2) = robust standard errors OLS model*

Table 8 shows a comparison between the *IV model* and the OLS one. Both the model's coefficient of *lw* are negative, however the instrumental model shows a bigger negative effect of a change in a *lw* on the labor demand. In the same way, the instrumental model coefficient of *ly* is slightly bigger than the one in the OLS model. In any case, all the coefficient are significant in both models. The instrumental model shows how one-lag period data are relevant in explaining labor demand.

Table 9. Wu-Hausman test comparing OLS and IV estimation

```
======================================================================================
Diagnostic tests:
--------------------------------------------------------------------------------------
                      df1 df2 statistic p-value

Weak instruments (lw)   2 707   1369.30  <2e-16 ***
Weak instruments (ly)   2 707   9773.16  <2e-16 ***
Wu-Hausman              2 705     75.74  <2e-16 ***
Sargan                  0  NA        NA      NA
--------------------------------------------------------------------------------------
Residual standard error: 0.502 on 707 degrees of freedom
Multiple R-Squared: 0.8824,      Adjusted R-squared: 0.882
Wald test:  2712 on 2 and 707 DF,  p-value: < 2.2e-16
======================================================================================
```

*Note. *p<0.1; **p<0.05; ***p<0.01. Source of the data: dati_esame.dta.*

Weak instrument have a low correlation with the endogenous explanatory variable. In this case, the null-hypothesis of weak instrument is strongly rejected both for *lw* and *ly*, and this means that the instrument is relevant. This confirm our previous intuition.

On the other hand, the null-hypothesis of the Wu-Hausman test states that IV model is just consistent as OLS model, and since OLS is more efficient, it would be preferable. In this case, the null-hypothesis is strongly rejected (p-value: < 2e-16) and so the preferred model (assuming that our instrument are valid) is the IV one (that is unbiased and consistent). In conclusion, note that the Sargan test for overidentified restriction can not be done because the model is not overidentified (number of endogenous variables (2) = number of instruments(2)). Due to this, it is not possible to test exogeneity (the second condition of validity) of our instruments. This is left to our judgment.

2. Panel approach

[i] pooled OLS;

Table 10. Pooled OLS models for the labor demand in 1997

```
================================================================================
                                        Dependent variable:
--------------------------------------------------------------------------------
                                                ln
                       (1)                     (2)                       (3)
--------------------------------------------------------------------------------
lw                   0.910***                                        -0.755***
                     (0.066)                                          (0.029)

ly                                           0.638***                 0.755***
                                             (0.008)                  (0.008)

Constant             0.907***               -0.356***                 2.028***
                     (0.294)                 (0.065)                  (0.105)

--------------------------------------------------------------------------------
Observations           1,420                  1,420                    1,420
R2                     0.120                  0.835                    0.889
Adjusted R2            0.119                  0.835                    0.889
F Statistic  192.981*** (df = 1; 1418) 7,176.894*** (df = 1; 1418) 5,686.178*** (df = 
2; 1417)
================================================================================
```

*Note. *p<0.1; **p<0.05; ***p<0.01. Source of the data: dati_esame.dta.*

The Pooled OLS model is a type of model that has constant coefficients, referring both intercepts and slopes. In this model, all the observations have been pooled. *Model (1*) goes against our intuition that the coefficient of *lw* is negative. Moreover, adding a new regressor *ly* the value of *lw* coefficient change a lot: this is an indication of an omitted variable bias. In the same way, *model (2)* suffers of an omitted variables bias: if *lw* is added to this model, the constant changes a lot and the change in the coefficient of *ly* is quite big. For the reasons above, the better pooled OLS model is *model (3).*

[ii] fixed effects;

Table 11. fixed-effects models for the labor demand in 1997

```
===============================================================================
                                     Dependent variable:
            -------------------------------------------------------------------
                                             ln
                       (1)                     (2)                     (3)
-------------------------------------------------------------------------------
```

```
lw                 -0.267***                                      -0.402***
                    (0.028)                                        (0.025)

ly                                           0.236***               0.328***
                                             (0.021)                (0.018)

-------------------------------------------------------------------------------------
Observations         1,420                    1,420                  1,420
R2                   0.111                    0.156                  0.384
Adjusted R2         -0.779                   -0.689                 -0.234
F Statistic  88.627*** (df = 1; 709) 130.965*** (df = 1; 709) 220.728*** (df = 2; 708)
=====================================================================================
```

*Note. *p<0.1; **p<0.05; ***p<0.01. Source of the data: dati_esame.dta.*

The fixed-effects model considers the fixed-effects, so effects that are different for different statistical unit but are fixed in time. Comparing *model (*1), *model (2)* and *model (3)* coefficient changes a lot while regressors are added. This lead us to think that some variables have been omitted in *model (*1) and *(2).* In this sense, *model (3)* seems not to suffer from this bias and also both the coefficients are significant for an alpha = 0.01. For this reasons, *model (3)* is the best fixed-effects model for explaining our dependent variable. Note that in all the models the constant is not include to avoid the dummy variable trap.

Table 12. pooled OLS model and fixed-effects models for the labor demand in 1997

```
Linear Panel Regression Models of demand function
=================================================================
                             Dependent variable:
             ----------------------------------------------------
                                      ln
                      OLS model             Fixed-effect model
-----------------------------------------------------------------
lw                    -0.755***                 -0.402***
                       (0.044)                   (0.039)

ly                     0.755***                  0.328***
                       (0.015)                   (0.029)

Constant               2.028***
                       (0.148)

-----------------------------------------------------------------
Observations            1,420                     1,420
R2                      0.889                     0.384
Adjusted R2             0.889                    -0.234
F Statistic  5,686.178*** (df = 2; 1417) 220.728*** (df = 2; 708)
=================================================================
```

*Note. *p<0.1; **p<0.05; ***p<0.01. Source of the data: dati_esame.dta.*

OLS model and the fixed effects model shows very different coefficients. At a first glance is evident how by adding fixed-effects both coefficient of the regressors change a lot. Especially, both decreased (the *ly* coefficient more than an half). Moreover, standard errors changes too. In any case, in all the model both the coefficient remain significant at a level of 99%.

Table 13. F-test for the individual effects in the fixed-effects model

```
=====================================================================================
F test for individual effects
-------------------------------------------------------------------------------------
data:  ln ~ lw + ly
```

```
--------------------------------------------------------------------------------------
F = 26.007, df1 = 709, df2 = 708, p-value < 2.2e-16

alternative hypothesis: significant effects
======================================================================================
```

Note. Source of the data: dati_esame.dta

The F-test for the individual effects rejects the null-hypothesis that individual effects are not statistically significant. That is, individual effects are relevant in explaining the dependent variable. This F-test confirms the fact that fixed-effects should be kept in our model.

[iii]. Coefficients interpretation

Interpreting the coefficient of the models, it is evident how the fixed effects model estimate a more modest effect of our regressor on the dependent variable. In any case, the *lw* coefficient remain negative and *ly* remain positive. This confirms our starting intuition. Moreover, the F-test shows that fixed-effects matter in explaining the labor demand. This can be a consequence of the different nature of firms (private, mixed or state), of the different sector of firms, or maybe of the produced products. Moreover, the productive process and the degree of the automation in the process are other factor that, changing from firm to firm, affects the labor demand. The pooled OLS model does not take into consideration these difference between firms.