UNIVERSITY OF GREENWICH

*Student: Thiago Prates Bossardi*

*University of Greenwich*

*MSc in Economics*

*Applied Econometrics*

*Professor Mehmet Ugur*

# Table of Contents

# Determinants of wage earnings

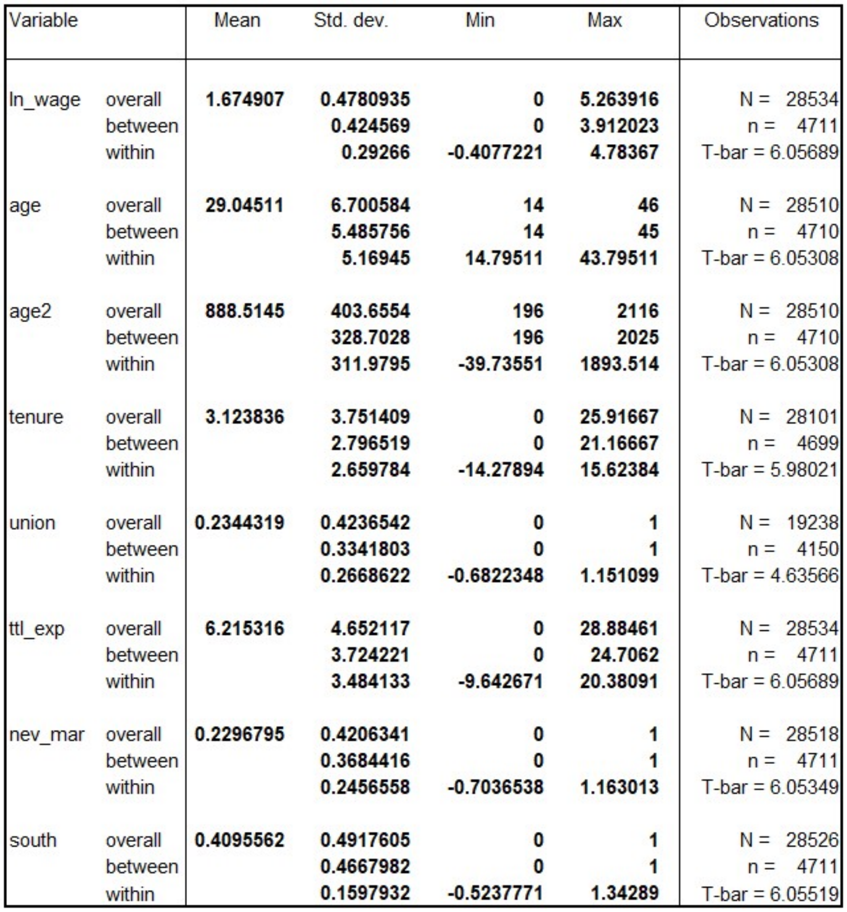
## Data

The dataset used was the ***Wage_at_work_panel***, a panel data extracted from the National Longitudinal Survey of Youth in the US. It contains data relative to the laborer's characteristics for the years between 1968 to 1988, which is unbalanced, meaning that the data is somewhat inconsistent and there are some gaps between and within the variables.

## Descriptive Statistics

Table 1 below provides a summary statistic of all variables from the panel data used for this study. It shows the number of observations, the mean, standard deviation, minimum and maximum values of each variable, and an individual or entity's within-group and between-group variances.

*Table 1 - Statistical Summary*



| Variable | | Mean | Std. dev. | Min | Max | Observations |
|---|---|---|---|---|---|---|
| ln_wage | overall | **1.674907** | **0.4780935** | **0** | **5.263916** | N = 28534 |
| | between | | **0.424569** | **0** | **3.912023** | n = 4711 |
| | within | | **0.29266** | **-0.4077221** | **4.78367** | T-bar = 6.05689 |
| age | overall | **29.04511** | **6.700584** | **14** | **46** | N = 28510 |
| | between | | **5.485756** | **14** | **45** | n = 4710 |
| | within | | **5.16945** | **14.79511** | **43.79511** | T-bar = 6.05308 |
| age2 | overall | **888.5145** | **403.6554** | **196** | **2116** | N = 28510 |
| | between | | **328.7028** | **196** | **2025** | n = 4710 |
| | within | | **311.9795** | **-39.73551** | **1893.514** | T-bar = 6.05308 |
| tenure | overall | **3.123836** | **3.751409** | **0** | **25.91667** | N = 28101 |
| | between | | **2.796519** | **0** | **21.16667** | n = 4699 |
| | within | | **2.659784** | **-14.27894** | **15.62384** | T-bar = 5.98021 |
| union | overall | **0.2344319** | **0.4236542** | **0** | **1** | N = 19238 |
| | between | | **0.3341803** | **0** | **1** | n = 4150 |
| | within | | **0.2668622** | **-0.6822348** | **1.151099** | T-bar = 4.63566 |
| ttl_exp | overall | **6.215316** | **4.652117** | **0** | **28.88461** | N = 28534 |
| | between | | **3.724221** | **0** | **24.7062** | n = 4711 |
| | within | | **3.484133** | **-9.642671** | **20.38091** | T-bar = 6.05689 |
| nev_mar | overall | **0.2296795** | **0.4206341** | **0** | **1** | N = 28518 |
| | between | | **0.3684416** | **0** | **1** | n = 4711 |
| | within | | **0.2456558** | **-0.7036538** | **1.163013** | T-bar = 6.05349 |
| south | overall | **0.4095562** | **0.4917605** | **0** | **1** | N = 28526 |
| | between | | **0.4667982** | **0** | **1** | n = 4711 |
| | within | | **0.1597932** | **-0.5237771** | **1.34289** | T-bar = 6.05519 |

## Methodology

The theoretical foundation behind all work done in this essay relates to the wage and employment determination theory. To do this, we used the papers of MacCurdy and Pencavel (1986), Angrist and Newey (1991), and Williams (1991). The first proposes competing models for wage and employment determination in unionized markets, the second discusses the over-identification tests in earnings functions with fixed effects, and the third re-examines the relationship between wages, tenure, and experience.

Together, these papers suggest that, at least, a regression model for wage determination should consider the role of unions, tenure, and experience. Therefore, following the guidelines of the task proposed, the model in this essay is based on the following equation:

*Equation 1*

$$\ln(wage) = \beta 0 + \beta 1\ (age) + \beta 2\ (age2) + \beta 3\ (tenure) + \beta 4\ (union) + \beta 5\ (ttl\_exp) + \varepsilon$$

Where:

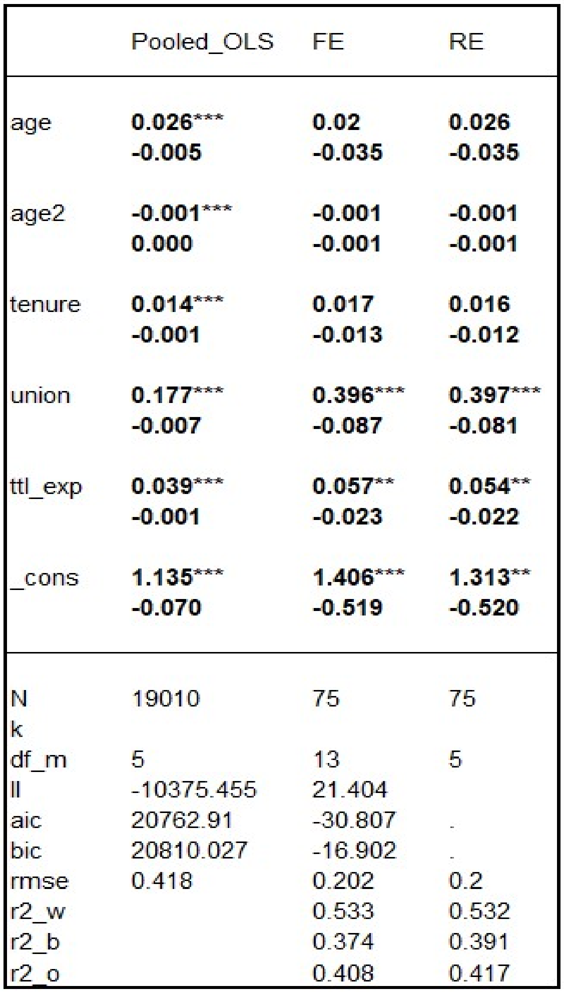
- lnwage: the natural logarithm of wage.
- age: the age of the employee in the current year.
- age2: the squared age of the employee in the current year.
- tenure: the number of years the employee has worked in their current job.
- union: dummy variable (1 if union member)
- ttl_exp: total work experience.
- $\beta 0$, $\beta 1$, $\beta 2$, $\beta 3$, $\beta 4$ and $\beta 5$: regression coefficients to be estimated
- $\varepsilon$: the error term, which captures the variation in lnwage that the independent variables cannot explain.

## Panel Data Estimators

Three main estimators were employed to test the effect of the independent variables on the dependent variable; see Table 2. The first one is the Pooled OLS estimator, which assumes that there are no individual-specific effects and that all variables have the same impact across individuals. Therefore, we have used Least Squared Dummy Variables (LSDV) estimator to test for individual specific effects. This approach creates dummy variables for each individual and adds them to the regression model as explanatory variables. By doing this, the dummy variable method captures the impact of the categorical variable on the outcome variable, and the LSDV method guarantees that each individual has a unique intercept. In this case, by applying the LSDV, we can discard the Pooled OLS estimator because each individual in did has its own intercept, as we can see in Figure 1, where the P-value is less than 5%, meaning that we reject the null hypothesis that there is no difference between individuals, so we should not run the regression without controlling for individual specific effects. Moreover, we can also discard the LSDV estimator because it is computationally intensive with large panels, which leads us to our second estimator.

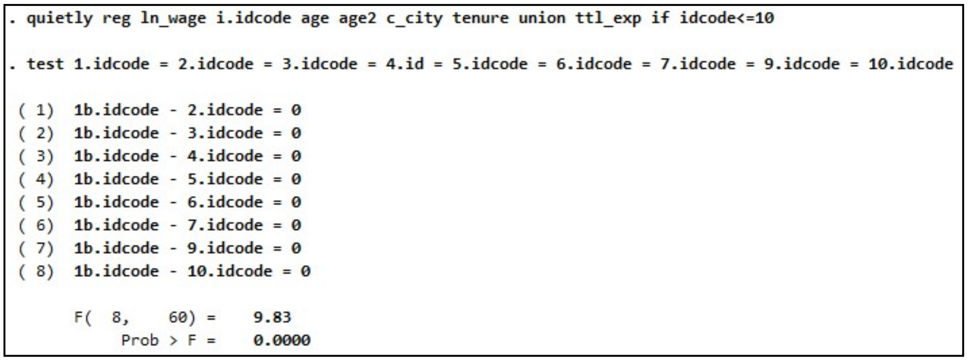
The second estimator is the Fixed Effects (FE), which controls for individual-specific effects by subtracting the induvial specific means for each variable. This method eliminates any time-invariant heterogeneity, but it assumes that the individual-specific effects are time-invariant, which may not be accurate in all cases. The last estimator used in this essay is the Random Effects (RE), which allows individual-specific effects to correlate with the regressors but it assumes that the individual specific effects are uncorrelated whit the error term. This estimator is more efficient than the FE estimator, but it can produce biased estimates if the individual-specific effects are correlated with the error term.

*Table 2 - Panel data estimators*



| | Pooled_OLS | FE | RE |
|---|---|---|---|
| age | **0.026*****<br>**-0.005** | **0.02**<br>**-0.035** | **0.026**<br>**-0.035** |
| age2 | **-0.001*****<br>**0.000** | **-0.001**<br>**-0.001** | **-0.001**<br>**-0.001** |
| tenure | **0.014*****<br>**-0.001** | **0.017**<br>**-0.013** | **0.016**<br>**-0.012** |
| union | **0.177*****<br>**-0.007** | **0.396*****<br>**-0.087** | **0.397*****<br>**-0.081** |
| ttl_exp | **0.039*****<br>**-0.001** | **0.057****<br>**-0.023** | **0.054****<br>**-0.022** |
| _cons | **1.135*****<br>**-0.070** | **1.406*****<br>**-0.519** | **1.313****<br>**-0.520** |
| N | 19010 | 75 | 75 |
| k | | | |
| df_m | 5 | 13 | 5 |
| ll | -10375.455 | 21.404 | |
| aic | 20762.91 | -30.807 | . |
| bic | 20810.027 | -16.902 | . |
| rmse | 0.418 | 0.202 | 0.2 |
| r2_w | | 0.533 | 0.532 |
| r2_b | | 0.374 | 0.391 |
| r2_o | | 0.408 | 0.417 |

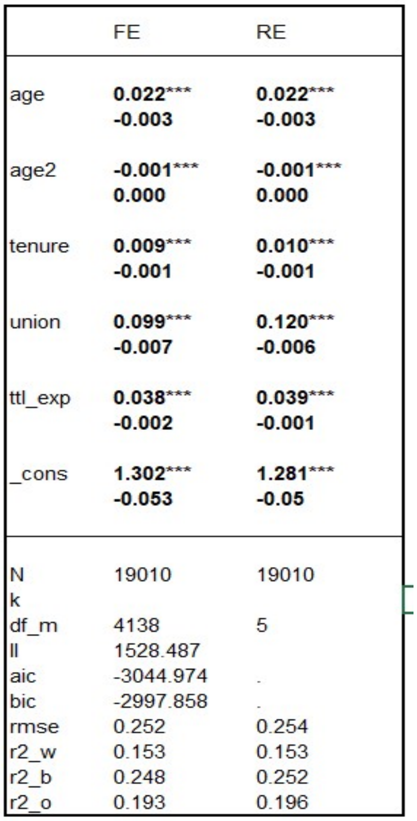
*Figure 1 - LSDV individual  intercept check*



```
. quietly reg ln_wage i.idcode age age2 c_city tenure union ttl_exp if idcode<=10

. test 1.idcode = 2.idcode = 3.idcode = 4.id = 5.idcode = 6.idcode = 7.idcode = 9.idcode = 10.idcode

 ( 1)  1b.idcode - 2.idcode = 0
 ( 2)  1b.idcode - 3.idcode = 0
 ( 3)  1b.idcode - 4.idcode = 0
 ( 4)  1b.idcode - 5.idcode = 0
 ( 5)  1b.idcode - 6.idcode = 0
 ( 6)  1b.idcode - 7.idcode = 0
 ( 7)  1b.idcode - 9.idcode = 0
 ( 8)  1b.idcode - 10.idcode = 0

       F(  8,    60) =    9.83
            Prob > F =    0.0000
```

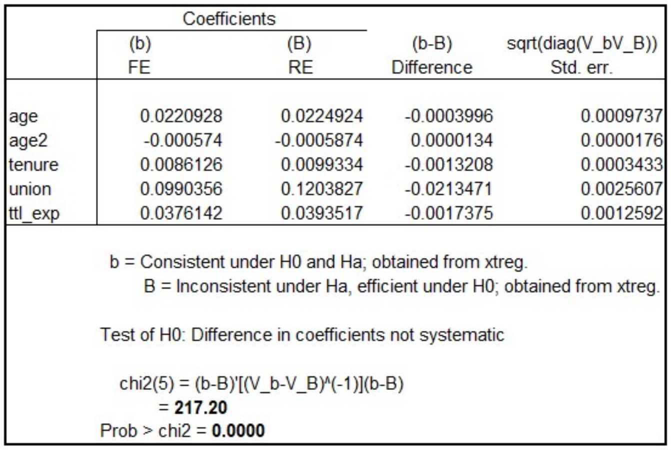
So, to see which estimator would be the best for our data, we run only the two lasting estimators, FE and RE (Table 3), and then perform the Hausman test. The Hausman test is a statistical test used to determine whether the coefficients in a FE or RE model are consistent. It tests whether the RE assumption is valid by comparing the estimates of the two models. If the p-value is less than 0.05 (5%), the null hypothesis that the coefficients are consistent with the random effects assumption is rejected, and the FE model is preferred, which is our case, as seen in Figure 2.

*Table 3 - Fixed Effects vs. Random Effects*



|  | FE | RE |
|---|---|---|
| age | **0.022*****<br>**-0.003** | **0.022*****<br>**-0.003** |
| age2 | **-0.001*****<br>**0.000** | **-0.001*****<br>**0.000** |
| tenure | **0.009*****<br>**-0.001** | **0.010*****<br>**-0.001** |
| union | **0.099*****<br>**-0.007** | **0.120*****<br>**-0.006** |
| ttl_exp | **0.038*****<br>**-0.002** | **0.039*****<br>**-0.001** |
| _cons | **1.302*****<br>**-0.053** | **1.281*****<br>**-0.05** |
| N | 19010 | 19010 |
| k |  |  |
| df_m | 4138 | 5 |
| ll | 1528.487 |  |
| aic | -3044.974 | . |
| bic | -2997.858 | . |
| rmse | 0.252 | 0.254 |
| r2_w | 0.153 | 0.153 |
| r2_b | 0.248 | 0.252 |
| r2_o | 0.193 | 0.196 |

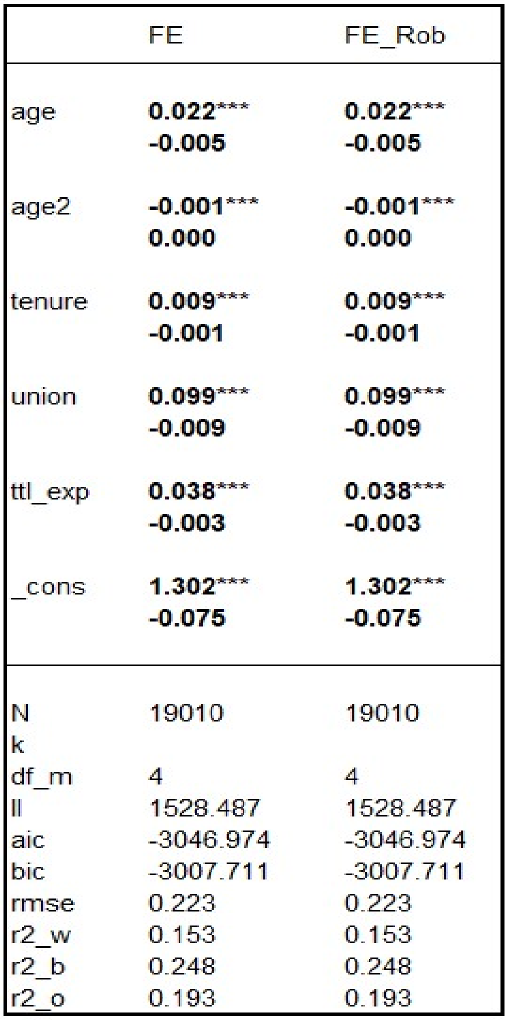
*Figure 2 - Hausman test: FE vs RE*



```
                   Coefficients
           (b)              (B)              (b-B)      sqrt(diag(V_bV_B))
           FE               RE             Difference        Std. err.

age        0.0220928        0.0224924      -0.0003996        0.0009737
age2       -0.000574       -0.0005874       0.0000134        0.0000176
tenure     0.0086126        0.0099334      -0.0013208        0.0003433
union      0.0990356        0.1203827      -0.0213471        0.0025607
ttl_exp    0.0376142        0.0393517      -0.0017375        0.0012592

            b = Consistent under H0 and Ha; obtained from xtreg.
                 B = Inconsistent under Ha, efficient under H0; obtained from xtreg.

           Test of H0: Difference in coefficients not systematic

              chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      = 217.20
           Prob > chi2 = 0.0000
```

Now that the FE estimator has been preferred, we must consider the heteroscedasticity and autocorrelation of error. This is done by using an adjusted standard error; this option specifies, in our case, that standard errors should be clustered by "*idcode*", which is a variable that identifies individuals in the dataset. This is because the unadjusted standard error can lead to an overstatement of the estimated result, which can be misleading when concluding, which is not the case for our data, as seen in Table 4, where all coefficients and significances are the same.

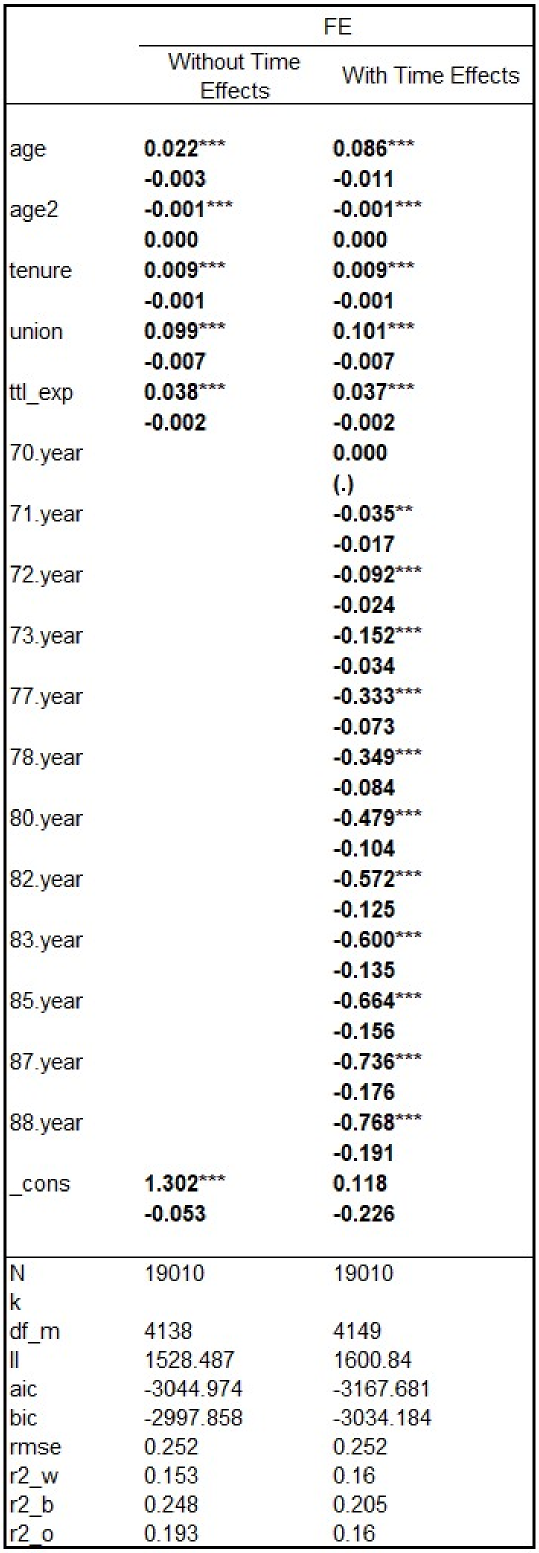
*Table 4 - FE standard error check*



| | FE | FE_Rob |
|---|---|---|
| age | **0.022*****<br>**-0.005** | **0.022*****<br>**-0.005** |
| age2 | **-0.001*****<br>**0.000** | **-0.001*****<br>**0.000** |
| tenure | **0.009*****<br>**-0.001** | **0.009*****<br>**-0.001** |
| union | **0.099*****<br>**-0.009** | **0.099*****<br>**-0.009** |
| ttl_exp | **0.038*****<br>**-0.003** | **0.038*****<br>**-0.003** |
| _cons | **1.302*****<br>**-0.075** | **1.302*****<br>**-0.075** |
| N | 19010 | 19010 |
| k | | |
| df_m | 4 | 4 |
| ll | 1528.487 | 1528.487 |
| aic | -3046.974 | -3046.974 |
| bic | -3007.711 | -3007.711 |
| rmse | 0.223 | 0.223 |
| r2_w | 0.153 | 0.153 |
| r2_b | 0.248 | 0.248 |
| r2_o | 0.193 | 0.193 |

When controlling for the time effect on FE, the negative and statistically significant coefficients of time control presented in the FE model regression (Table 5) indicate that time has a significant impact on the dependent variable and that this impact is negative. This means that as time increases, the value of the dependent variable decreases, and all other factors held constant. Overall, the negative coefficient of time control suggests that underlying trends or factors may affect the dependent variable over time.

*Table 5 - FE with time effect control*



| | FE | |
|---|---|---|
| | Without Time Effects | With Time Effects |
| age | **0.022*****<br>**-0.003** | **0.086*****<br>**-0.011** |
| age2 | **-0.001*****<br>**0.000** | **-0.001*****<br>**0.000** |
| tenure | **0.009*****<br>**-0.001** | **0.009*****<br>**-0.001** |
| union | **0.099*****<br>**-0.007** | **0.101*****<br>**-0.007** |
| ttl_exp | **0.038*****<br>**-0.002** | **0.037*****<br>**-0.002** |
| 70.year | | **0.000**<br>**(.)** |
| 71.year | | **-0.035****<br>**-0.017** |
| 72.year | | **-0.092*****<br>**-0.024** |
| 73.year | | **-0.152*****<br>**-0.034** |
| 77.year | | **-0.333*****<br>**-0.073** |
| 78.year | | **-0.349*****<br>**-0.084** |
| 80.year | | **-0.479*****<br>**-0.104** |
| 82.year | | **-0.572*****<br>**-0.125** |
| 83.year | | **-0.600*****<br>**-0.135** |
| 85.year | | **-0.664*****<br>**-0.156** |
| 87.year | | **-0.736*****<br>**-0.176** |
| 88.year | | **-0.768*****<br>**-0.191** |
| _cons | **1.302*****<br>**-0.053** | **0.118**<br>**-0.226** |
| N | 19010 | 19010 |
| k | | |
| df_m | 4138 | 4149 |
| ll | 1528.487 | 1600.84 |
| aic | -3044.974 | -3167.681 |
| bic | -2997.858 | -3034.184 |
| rmse | 0.252 | 0.252 |
| r2_w | 0.153 | 0.16 |
| r2_b | 0.248 | 0.205 |
| r2_o | 0.193 | 0.16 |

## Coefficients Interpretation

β0 is the intercept, which represents the expected wage value when all the independent variables equal zero. In this case, it represents the expected wage value for a worker who is not in a union, has zero years of experience, and has zero years of tenure, regardless of age.

β1 represents the effect of age on wage, holding all other variables constant. The coefficient of 0.086 indicates that a one-year increase in age is associated with an average increase in wage of 8.6%, all else being equal.

β2 represents the effect of the squared age variable on wage, holding all other variables constant. Squaring the age suggests that the impact of age on wages may be non-linear, meaning that age can have an adverse effect after a certain point. The coefficient of -0.001 indicates that a one-unit increase in the squared age variable is associated, after a certain point, with an average decrease in the wage of 0.1%.

β3 represents the effect of tenure on wages. The coefficient of 0.009 indicates that a one-year increase in tenure is associated with an average increase in wage of 0.9%, all else being equal.

β4 represents the effect of union membership on wages, holding all other variables constant. The coefficient of 0.101 indicates that union members have an average wage that is 10.1%higher than non-union members.

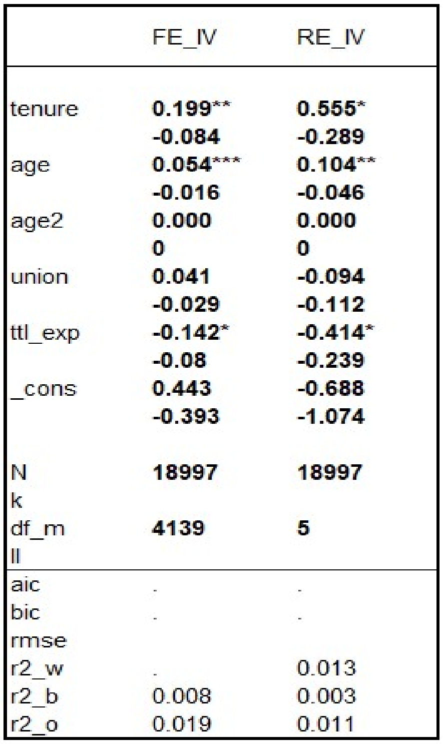
β5 represents the effect of experience on wages. The coefficient of 0.037 indicates that a one-year increase in experience is associated with an average increase in wage of 3.7%, all else being equal.

## Endogeneity

Now assuming that tenure is endogenous and using newly married (*nev_mar*) and living in the south (*south*) as instrumental variables (IV), we re-estimate the wage model with fixed and random effects, and we can see the results in Table 6.

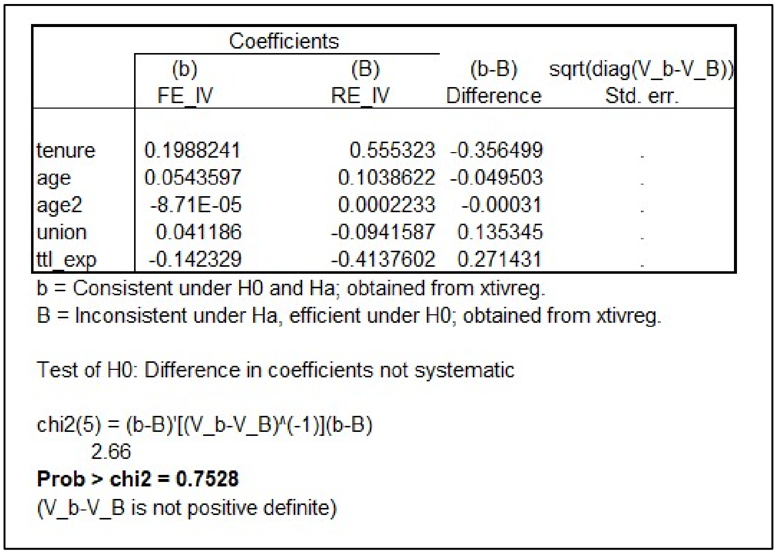
With the inclusion of IV, our new FE (FE_IV) and RE (RE_IV) present, as expected, different results from our previous estimations. The coefficients show a considerable difference between the estimators in this new setup, with the RE_IV model showing the highest coefficients but with the lowest significances. So, as before, we use the Hausman test (Figure 3) to decide which model is preferred, and if the p-value is less than 0.05, the null hypothesis is rejected, indicating that the difference between the two sets of coefficients is statistically significant, therefore FE_IV model is preferred. On the other hand, if the null hypothesis cannot be rejected, the RE_IV estimator should be selected, which is the case.

*Table 6 - Instrumental Variable test*



| | FE_IV | RE_IV |
|---|---|---|
| tenure | **0.199**** | **0.555*** |
| | **-0.084** | **-0.289** |
| age | **0.054***** | **0.104**** |
| | **-0.016** | **-0.046** |
| age2 | **0.000** | **0.000** |
| | **0** | **0** |
| union | **0.041** | **-0.094** |
| | **-0.029** | **-0.112** |
| ttl_exp | **-0.142*** | **-0.414*** |
| | **-0.08** | **-0.239** |
| _cons | **0.443** | **-0.688** |
| | **-0.393** | **-1.074** |
| | | |
| N | **18997** | **18997** |
| k | | |
| df_m | **4139** | **5** |
| ll | | |
| aic | . | . |
| bic | . | . |
| rmse | | |
| r2_w | . | 0.013 |
| r2_b | 0.008 | 0.003 |
| r2_o | 0.019 | 0.011 |

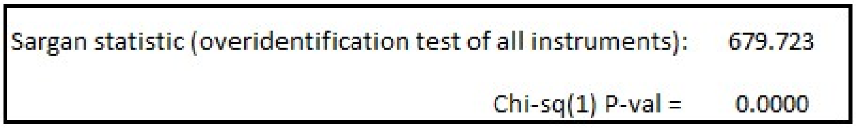
*Figure 3 - Hausman test with Instrumental Variable*



```
                    Coefficients
           (b)                  (B)          (b-B)     sqrt(diag(V_b-V_B))
           FE_IV                RE_IV        Difference      Std. err.

tenure     0.1988241            0.555323   -0.356499            .
age        0.0543597           0.1038622   -0.049503            .
age2       -8.71E-05           0.0002233    -0.00031            .
union       0.041186          -0.0941587    0.135345            .
ttl_exp    -0.142329          -0.4137602    0.271431            .
b = Consistent under H0 and Ha; obtained from xtivreg.
B = Inconsistent under Ha, efficient under H0; obtained from xtivreg.

Test of H0: Difference in coefficients not systematic

chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
        2.66
Prob > chi2 = 0.7528
(V_b-V_B is not positive definite)
```

Now that we have decided to proceed with the RE_IV model, we must validate the instrumental variables. The Sargan-Hansen test (Figure 4) is a test for instrument validity in instrumental variable regression. It was proposed by Sargan (1958) and compared the difference between the predicted and actual values of the endogenous variable using the instrument, checking for any correlation with the instrument. It is essential because it allows researchers to ensure that the instrumental variable affects the endogenous variable in the intended way. If the test is significant (i.e., if the p-value is less than 0.05), we reject the null hypothesis that the instrument is valid.

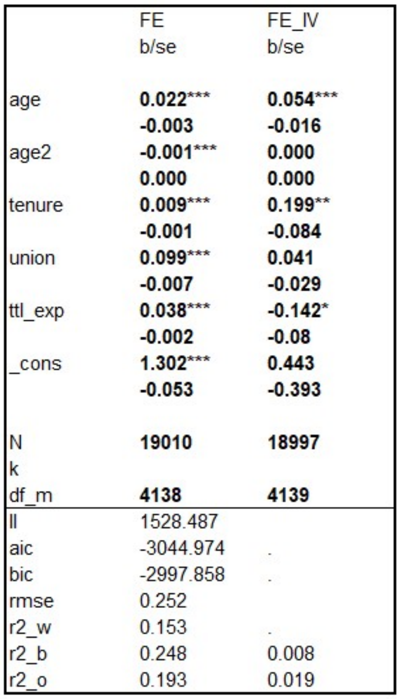
*Figure 4 - Sargan-Hansen test*



```
Sargan statistic (overidentification test of all instruments):     679.723

                                              Chi-sq(1) P-val =     0.0000
```

In this case, we rejected the null hypothesis that the instrument is valid, which got us back to testing with FE. So, the last step to be done here will be comparing the FE with FE_IV (Table 7), enabling us to see the differences between the models and then choose the preferred one.

*Table 7 - Fixed Effects vs. Fixed Effects with Instrumental Variables*



|  | FE | FE_IV |
|---|---|---|
|  | b/se | b/se |
| age | 0.022*** | 0.054*** |
|  | -0.003 | -0.016 |
| age2 | -0.001*** | 0.000 |
|  | 0.000 | 0.000 |
| tenure | 0.009*** | 0.199** |
|  | -0.001 | -0.084 |
| union | 0.099*** | 0.041 |
|  | -0.007 | -0.029 |
| ttl_exp | 0.038*** | -0.142* |
|  | -0.002 | -0.08 |
| _cons | 1.302*** | 0.443 |
|  | -0.053 | -0.393 |
| N | 19010 | 18997 |
| k |  |  |
| df_m | 4138 | 4139 |
| ll | 1528.487 |  |
| aic | -3044.974 | . |
| bic | -2997.858 | . |
| rmse | 0.252 |  |
| r2_w | 0.153 | . |
| r2_b | 0.248 | 0.008 |
| r2_o | 0.193 | 0.019 |

Looking at Table 7 and taking into account the invalidity of the instrumental variables and the focus on the coefficients' significance, we can conclude that the FE model without IV would be preferred.

## Conclusion

The study employed a regression model that included age, age squared, tenure, union status, and total work experience as independent variables to determine their effect on the natural logarithm of wages. Three estimators, Pooled OLS, Fixed Effects, and Random Effects, were used to test the impact of the independent variables on the dependent variable. Based on the Hausman test, the Fixed Effects estimator was preferred. The results show that tenure and union status positively impact wages. The study contributes to the literature on wage determination and provides insight into factors that influence wages in the US labor market.

## References:

MaCurdy, T. E., & Pencavel, J. H. (1986). Testing between competing models of wage and employment determination in unionized markets. Journal of Political Economy, 94(3, Part 2), S3-S39.

Angrist, J. D., & Newey, W. K. (1991). Over-identification tests in earnings functions with fixed effects. Journal of Business & Economic Statistics, 9(3), 317-323.

Williams, N. (1991). Reexamining the wage, tenure, and experience relationship. Review of Economics and Statistics, 73(3), 512-517

Sargan, J.D. (1958). The Estimation of Economic Relationships using Instrumental Variables. *Econometrica*, 26(3), p.393. doi:https://doi.org/10.2307/1907619.