KU LEUVEN

# Advanced Applied Econometrics

Shreekar Araveti - r0919044
Alessandro Salvatori - r0926525

*Group_07*

*MSc Statistics and Data Science*

$27^{th}$ May 2024

# Contents

**1 Introduction** **1**

**2 Data description** **1**

**3 Static Modeling** **2**

3.1 Pooled OLS . . . . . 3

3.2 Fixed Effects Model . . . . . 3

3.3 Random Effects Model . . . . . 4

3.4 Strict Exogeneity . . . . . 5

3.5 Model Comparison . . . . . 5

**4 Dynamic Modeling** **6**

4.1 Anderson-Hsiao Estimator . . . . . 6

4.2 GMM Estimator . . . . . 7

4.2.1 Difference GMM (Arellano-Bond) . . . . . 7

4.2.2 System GMM . . . . . 9

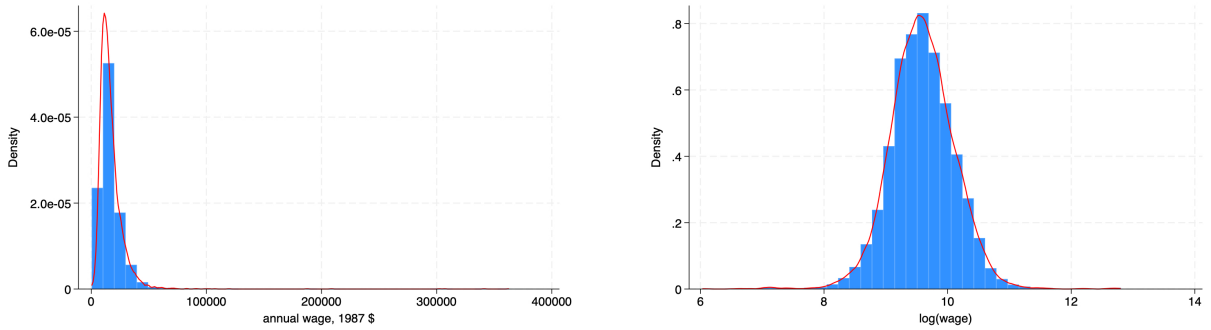
# 1 Introduction

Panel data analysis, which combines cross-sectional and time-series data, offers a comprehensive approach to understanding complex economic phenomena. In this paper, we focus on analyzing a dataset aimed at predicting wages, utilizing the strengths of panel data to uncover deeper insights. With panel data we can account for unobserved heterogeneity and capture the temporal dimension of wage changes, providing a more robust framework than purely cross-sectional or time-series analyses. In the first part of our analysis we will focus on static modeling, which focuses on cross-sectional aspects of panel data. In this section we conduct model specification by running static models like Fixed Effect and Random Effect and try to control for unobserved individual-specific effect. Additionally, we examine our regressors for any violations of the strict exogeneity assumption and we use instrumental variables to address potential issues arising from such violations. In the second part we will continue with a dynamic panel analysis, to inspect how also values of the past time points could affect the wage in the future. We first used IV-Anderson Hsiao estimator and then we proceed using the Arellano Bond (GMM difference) estimator and the System GMM estimator. The goal of this analysis is to find the right model that provides consistent but also efficient estimates.

# 2 Data description

The data used in this paper consist of an instructional panel dataset from the Boston College Department of Economics, proposed by Jeffrey M. Wooldridge, with 12,723 observations over a period of seven years (1981-1987). To ensure a perfectly balanced panel and enhance the reliability of our results, we retained only those observations that were present in all seven time periods, which constituted over 90% of the total data. As a result, we obtained a perfectly balanced panel dataset with each entity observed at every time period. In total there are 1738 individuals and 7 time periods of information for each. The dataset contains information on wages, years of education, and work experience across different types of occupations (white-collar, blue-collar, and services). The variables used as regressors are the number of years of schooling the individual has completed (`educ`), years of experience in a white collar, blue collar and service job (respectively `expwc`, `expbc`, `expser`), total years of work experience (`exper`) and the squared of the total years of experience (`expersq`). Regarding the dependent variable, wages often display a right-skewed distribution, where a small number of individuals earn significantly higher wages compared to the majority. Observing the histogram of the variable `wage` (Figure 1a) confirms this pattern. Therefore, we decided to use the logarithm of wage (`lwage`) as our dependent variable. The distribution plot of log(wage) in Figure 1b shows a

normal distribution, which helps in meeting the assumptions such as normally distributed errors and homoskedasticity, while also reducing the impact of outliers.




(a) Wage (b) Log(Wage)

Figure 1: Distributions

# 3 Static Modeling

In this section we will apply different static models to our dataset. With static models we examine the direct impact of factors such as education and occupational experience on wage levels ignoring the influence of past values of the dependent variable. We also control for individual-specific effects that do not change over time, providing a clear picture of cross-sectional relationships. We will begin by running a Pooled OLS regression, followed by an analysis of two important static models: the Fixed Effects and the Random Effects model. Our goal is to determine which model best fits our data.

We specify our general model equation as follows:

$$y_{it} = X'_{it}\beta + c_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \text{IID}(0, \sigma^2_\varepsilon) \tag{1}$$

where $i$ stands for the total number of individuals and $t$ denotes the measurement time points. Matrix $X$ contains the predictor variables `educ`, `expwc`, `expbc`, `expser`, `exper`, `expersq`, and potentially their lags and/or differences. We are primarily interested in estimating the coefficients $\beta$, which are assumed to be constant across all entities $i$ and time periods $t$. Additionally, we aim to determine whether the individual-specific effect $c_i$ should be considered as unknown parameters that capture the specific characteristics of the $i$-th individual or as random draws from an underlying distribution that is independent of the covariates in $X$, based on what estimator is used.

## 3.1 Pooled OLS

In the Pooled OLS model we don't make use of our panel structure, therefore the unobserved heterogeneity is ignored and the dataset is treated like cross sectional data, taking each observation separately regardless of the time or the identifier variable. Consequently, the model equation becomes:

$$y_{it} = X'_{it}\beta + \varepsilon_{it}, \tag{2}$$

with $i$ being the total number of observations in the dataset.

This model does not account for individual-specific effects, but it can be easily estimated using Ordinary Least Squares (OLS). It relies on the assumption of contemporaneous exogeneity of the regressors ($X$). This means that the regressors are uncorrelated with the error term at each specific time point.

$$E(\varepsilon_{it}|X_{it}) = 0 \tag{3}$$

Using this estimator we can also test if there is presence of heteroskedasticity using the Breusch–Pagan/Cook–Weisberg test, where the null hypothesis states that the errors are homoskedastic. Looking at the result of the test we obtain a p-value of 0.5359, so we do not reject the null hypothesis, thus there is no need for standard error correction to control for heteroskedasticity. After running the model, we obtained significant p-values for all the coefficients (estimates are shown later in Table 1) other than the variable `expser` that was omitted due to collinearity, and an R-squared of 0.23. However, this model does not account for unobserved heterogeneity. Therefore, to address this issue, we need to run a Fixed Effects model, which will help us control for and check the presence of unobserved heterogeneity.

## 3.2 Fixed Effects Model

Unlike Pooled OLS, this model controls for unobserved individual-specific characteristics $c_i$ that are constant over time to better capture the variability in the data. It also assumes that $c_i$ is correlated with $X_{it}$. There are two specific methods within the fixed effects framework: the "within" transformation (or "demeaning") and the "first-difference" (FD) transformation.

### Within Estimator

The within transformation approach involves removing individual-specific effects by demeaning the data, in other words, centering the data around the mean of each individual. The model equation then becomes:

$$(y_{it} - \overline{y_i}) = (X_{it} - \overline{X_i})'\beta + (\varepsilon_{it} - \overline{\varepsilon_i}) \tag{4}$$

This model requires stricter assumptions compared to Pooled OLS. Specifically, it assumes strict exogeneity, meaning that the error term must be uncorrelated with the independent variables across all time periods. This assumption is also required for the Random Effects model.

$$E(\varepsilon_{it}|X_{i1}, X_{i2}, ..., X_{iT}) = 0 \tag{5}$$

When we run this model, we also test for unobserved heterogeneity, and the F-test is used to determine whether the fixed effects (entity-specific effects) in the panel data model are jointly significant. Based on the F-test results, we obtain a significant p-value, leading us to reject the null hypothesis. This indicates that the fixed effects are significant, and there is unobserved heterogeneity among the individuals. Consequently, Pooled OLS is not an appropriate model for our data.

**First Differences**

We then proceed by applying the First Differences (FD) approach, which eliminates individual-specific effects by differencing the data over time. This method uses one less time period (since differencing loses one time point per individual) but can be more robust to certain types of serial correlation in the error terms. The model equation then becomes:

$$\Delta y_{it} = \Delta X'_{it}\beta + \Delta\varepsilon_{it}, \tag{6}$$

Examining the results from both models, the within estimator drops the variable `exper` due to collinearity, while in the First Differences (FD) model the variables `educ` and `expwc` are dropped for the same reason. Comparing the results, we observe that the estimates from the two different approaches of the Fixed Effects model differ significantly. This finding supports the results of the strict exogeneity test because if the assumption were valid, the different formulations of the model would produce similar estimates. The estimates are shown later in Table 1.

## 3.3 Random Effects Model

In the Random Effects (RE) model $c_i$ is treated as random variables drawn from a distribution with mean zero and variance $\sigma_c^2$. This model is estimated using Generalized Least Squares (GLS) and makes the strong assumption that $c_i$ is uncorrelated with $X_{it}$, which is rarely met in real scenarios. It might be more efficient than the FE model if the assumptions hold, because it uses both within-entity and between-entity variations.

After running the model, we notice that the variable `exper` is dropped due to collinearity. Furthermore, we can compare the estimated coefficients with those from the Fixed Effects (FE) model and analyze the differences using the Hausman test. The low p-value obtained after running the Hausman test leads us to reject the null hypothesis, indicating a systematic difference between the fixed and random effects models. This suggests that the Fixed Effects model is more appropriate for our data.

## 3.4 Strict Exogeneity

We continue our analysis by testing the assumption of strict exogeneity, which is required for both the FE and RE models. To test for strict exogeneity in a FE model we use the approach suggested by Wooldridge, where leads of the independent variables are included in the regression and we test their significance. The test's criterion states states that if the lead values of the predictors are significant in the regression, the strict exogeneity assumption is violated for those variables.

After running the test we obtain that the lead values for the variables `expwc` and `expbc` are significant, meaning that they violate the assumption. This results in making both the random and the fixed effect models not appropriate.

According to Anderson/Hsiao (1981) we can use instrumental variables for the regressors that violate the strict exogeneity assumption, and as we have panel data, we have natural candidates for instrumental variables. They suggested to use the FD estimator and use lagged levels or differences of the independent variable as instruments. Various options were explored to determine the best instruments for the regressors: firstly, using only lags of levels, then only lagged differences, and lastly both lags and lagged differences. The F-value in the first stage regression was checked, yielding very high values for all models, indicating the significance of the instruments. The model with lagged differences of the regressors as instruments was selected, since it was the one with the lowest robust standard errors. In this model the variables `educ` and `expser` are dropped due to collinearity.

## 3.5 Model Comparison

In Table 1 the estimates and the standard errors (values in the brackets) of the coefficients of every static model used in this study is reported (* = significance at the 95% confidence level).

We compare the results of different static models and find that the pooled OLS and RE models provided the most efficient estimates. However, the assumptions underlying these estimators were violated. The same issue is observed for the fixed effects models. Based on

| | Pooled OLS | FE(Within) | FE(FD) | RE | IV-AH |
|---|---|---|---|---|---|
| `educ` | 0.089* | 0.183* | - | 0.092* | - |
| | (0.003) | (0.038) | - | (0.005) | - |
| `expwc` | 0.076* | 0.189* | - | 0.152* | -0.099* |
| | (0.008) | (0.025) | - | (0.009) | (0.047) |
| `expbc` | 0.070* | 0.173* | -0.013 | 0.138* | -0.107* |
| | (0.007) | (0.026) | (0.013) | (0.008) | (0.046) |
| `expser` | - | 0.178* | 0.025 | 0.111* | - |
| | - | (0.028) | (0.018) | (0.011) | - |
| `exper` | 0.077* | - | -0.434 | - | -0.100 |
| | (0.011) | - | (0.342) | - | (0.421) |
| `expersq` | -0.006* | -0.006* | -0.007* | -0.006* | -0.007* |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.002) |

Table 1: Model Coefficients

these findings, we chose the First Differences IV Estimator for our static model. This estimator addresses the issue of strict exogeneity by using instruments, making it a more suitable choice for modeling our response. The resulting model equation from our static analysis is:

$$\Delta lwage_{it} = \Delta expwc_{it} \cdot (-0.099) + \Delta expbc_{it} \cdot (-0.107) + \Delta expersq_{it} \cdot (-0.007) \tag{7}$$

# 4 Dynamic Modeling

Dynamic models incorporate lagged values of the dependent variable as explanatory variables, capturing the temporal dependencies in the data.

$$y_{it} = \alpha(y_{t-1}) + (X'_{it})\beta + c_i + \varepsilon_{it} \tag{8}$$

We will first use IV-Anderson Hsiao estimator and then proceed using the Arellano-Bond (GMM) estimator and conclude our analysis with the System GMM estimator.

## 4.1 Anderson-Hsiao Estimator

In this model, the lagged value $lwage_{t-1}$ is correlated with the individual-specific effect. The Anderson-Hsiao estimator addresses this problem by transforming the model and using instruments for the lagged values of `lwage` to eliminate the endogeneity. We used both $lwage_{t-2}$ and $lwage_{t-2} - lwage_{t-3}$ as instruments.

$$\Delta lwage_{i,t} = \Delta lwage_{i,t-1} + \Delta expwc_{i,t} + \Delta expbc_{i,t} + \Delta expser_{i,t} + \Delta \epsilon_{i,t} \tag{9}$$

The lagged second difference IV variable of `lwage` has a positive effect on `lwage` in differences, as the coefficient is significant. The first differences of the explanatory variables do not seem to be statistically significant. The null hypothesis for the Kleibergen-Paap rk LM test is that the instruments are not correlated with the endogenous regressors, meaning the model is underidentified. With a statistic of 8.67 and a p-value of the 0.0032 the null has to be rejected, implying that model may be identified. Cragg-Donald Wald F Statistic of 34.457 (rule of thumb $> 10$) is realized, indicating strong evidence that the instruments are not weak when assuming homoskedastic errors. The Kleibergen-Paap rk Wald F Statistic value of -10.603 (rule of thumb $> 10$), indicates that the instruments are adequately strong even when accounting for potential heteroskedasticity.

Although the Anderson-Hsiao (AH) estimator is consistent, it is not efficient because it does not utilize all the available information in the data. Typically, more moment conditions are available, which is why we move to the Generalized Method of Moments (GMM) estimator.

## 4.2 GMM Estimator

### 4.2.1 Difference GMM (Arellano-Bond)

To see how regressors relate with past, contemporary or future error terms the Generalized-Method-of-Moments (GMM) model is fit. At first the one step GMM model is fit (referred to as GMM1 in table2) on the regressand `lwage`, without including a constant, and the following variables are used as regressors: `expbc`, `expwc`, `exper`, `expser` and `year`. These variables are considered to be exogenous, since the lags of the regressand `lwage` variable isn't exogenous, instruments are checked from $lwage_{i,t-2}$ to $lwage_{i,1}$ to see if they are valid.

The realized results indicate that the regressors and instruments are significant, however, the Arellano-Bond test for AR(2), that checks the auto-correlation, indicates that there is some auto-correlation, meaning that the instruments are not valid, even though the Sargan test suggests otherwise with a p-value of 0.625. It is to be noted that the GMM tends to under estimate the standard errors making the results unreliable. Worth noting that due to multicollinearity `exper` and `year` are dropped.

However, when running a robust one-step GMM estimator (referred to as GMM2 in Table 2), with strictly exogenous variables and accounting for heteroskedasticity, the instruments are found to be valid, as proved by Arellano-Bond test for AR(2) with a p-value of 0.161, and a Sargan and Hansen test statistics respectively of 11.77 with a p-value of 0.625 and 9.88 with a p-value of 0.771, indicating that the instruments are valid. It is seen that the standard errors of all the instruments used have increased while the coefficient value remain the same, however all the instruments remain significant. Due to multicollinearity `exper` and `year` are

dropped. It should also be noted that the GMM estimator yields a more significant values for the instruments as compared to the AH estimated results.

Now, a two-step GMM estimator (referred to as GMM3 in Table 2) is run, taking heteroskedasticity into account. The results show that the coefficients have similar, though not identical, values to those obtained from the one-step GMM, and all the instruments remain significant. The Arellano-Bond test AR(2) is significant (p-value of 0.159) indicating no auto-correlation, and the Sargan and Hansen test statistics have 11.77 (p=0.625) and 9.88(p=0.771) respectively, showing that the instruments are valid. Due to multicollinearity `exper` and `year` are dropped. However, the two-step GMM estimator tends to underestimate the standard error for small samples, hence a Windmeijer (robust) correction will have to be applied to circumvent this.

Upon performing the finite sample correction of standard errors on the two-step GMM estimation (robust) (referred to as GMM4 in table2), the coefficients realized on the instruments remain the same. However, the standard errors are increased and the z-value realized is lessened, in spite of this the instruments remain significant. The Arellano-Bond test AR(2) is significant with a p-value of 0.159 indicating no auto-correlation, while the Sargan and Hansen tests have 11.77 (p=0.625) and 9.88(p=0.771) statistic respectively, indicating that the instruments are valid. Due to multicollinearity `exper` and `year` are dropped.

The GMM models until now have been applied under the assumption of strictly exogenous condition, however, this condition may not always hold true, therefore a predetermined variable form is applied (referred to as GMM5 in Table 2). The same variables are used as in the previous models. The `exper` and `year` variables have been dropped due to collinearity. The Arellano-Bond test for AR(2) shows a p-value of 0.163 indicating no auto-correlation, and the Sargan and Hansen tests have p-values of 0.604 and 0.732 respectively, indicating instrument validity, however the `exper` instruments do not seem to be significant owing to a higher p-value.

The finite sample correction ie. Windmeijer correction (robust) is applied (referred to as GMM6 in table2) to ensure that the standard errors are not deflated. The realized results of Arellano-Bond test for AR(2) indicates no auto-correlation owing to a high p-value of 0.171. The Sargan and Hansen tests have p-values of 0.694 and 0.831 indicating instrument validity. However, on accounting for the finite sample correction, the p-value of `expser` has gone up and have been rendered insignificant, similar to the GMM model without accounting for finite sample correction.

Another GMM model with `expwc` and `expbc` as predetermined and `expser` as exogenous instruments has been fit. (referred to as GMM7 in table2). The Arellano-Bond test for AR(2) suggests no autocorrelation. Though the Sargan test is on the fence with overidentification restriction (pval=0.049), Hansen test has a high p-value of 0.327, indicating instrument validity.

The instruments are all significant including `expser`. Running it with finite sample correction the same values for Arellano-Bond, Sargan and Hansen tests are realized indicating instrument validity. However, the affect of the instruments of `expser`, though significant, are on the border with a p-value of 0.047. GMM7 model has 76 instruments and 902 groups.

| | GMM1 | GMM2 | GMM3 | GMM4 | GMM5 | GMM6 | GMM7 |
|---|---|---|---|---|---|---|---|
| lwageL1. | 0.4148(0.00) | 0. 4148(0.00) | 0.3972(0.00) | 0.3972(0.00) | 0.40525(0.00) | 0.05839(0.00) | 0.3560 |
| expbc | 0.03107(0.00) | 0.03107(0.00) | 0.0328(0.00) | 0.0328(0.00) | 0.04288(0.00) | 0.008274(0.00) | 0.04276(0.00) |
| expwc | 0.0497(0.00) | 0.0497(0.00) | 0.0484(0.00) | 0.0484(0.00) | 0.05431(0.00) | 0.01307(0.00) | 0.05385(0.00) |
| expser | 0.0825(0.00) | 0.0825(0.00) | 0.0805(0.00) | 0.0805(0.00) | -0.01857(0.412) | 0.02436(0.446) | 0.03789(0.047) |
| AR(2) pval | 0.015 | 0.161 | 0.159 | 0.159 | 0.163 | 0.163 | 0.139 |
| Sagan test pval | 0.525 | 0.625 | 0.625 | 0.625 | 0.604 | 0.604 | 0.049 |
| Hansen test pval | na | 0.771 | 0.771 | 0.771 | 0.732 | 0.732 | 0.327 |
| Instrument validity | no | yes | yes | yes | yes | yes | yes |

Table 2: GMM models estimation

Other GMM models were fit with exogenous variable assumptions and predetermined variables assumption with and without capping the number of lags. It is found that the previous model is a good fit, with instruments being valid and significant. The following equation is realized:

$$\Delta lwage_{it} = 0.356\Delta lwage_{i,t-1} + 0.0428\Delta expbc_{it} + 0.0538\Delta expwc_{it} + 0.0378\Delta expser_{it} \quad (10)$$

Though the difference GMM makes better used of the information present in the data, it falls short to take the relative value at levels. The difference GMM estimator uses differences as regressors to predict the regressand. Regressors at different levels having the same differences will have the same effect, and this does not hold true in most cases. Hence there is a need to apply the system GMM in this case.

### 4.2.2 System GMM

The system GMM can be thought of as a random effects model in the dynamic case. It applies the GMM approach to the first differences (as in the previous section) and to the levels. The system GMM estimator is especially useful in case of persistent data with little variation. The data in hand seems to fit this, hence the system GMM approach is taken. The level equation is of the following form:

$$lwage_{it} = \alpha lwage_{i,t-1} + \beta_1 exper_{it} + \beta_2 expwc_{it} + \beta_3 educ_{it} + \beta_4 expser_{it} + \beta_5 year_{it} + \eta_i + \epsilon_{it} \quad (11)$$

The first difference equation is of the following form:

$$\Delta lwage_{it} = \alpha \Delta lwage_{i,t-1} + \beta_1 \Delta exper_{it} + \beta_2 \Delta expwc_{it} + \beta_3 \Delta educ_{it} + \beta_4 \Delta expser_{it} + \beta_5 \Delta year_{it} + \Delta \epsilon_{it} \tag{12}$$

Running the system GMM with `educ, exper, expwc, expbs, expser, year` and the lagged effect of `lwage` the following results are realized. `expbc` has been dropped due to collinearity:

| Variable | Coefficient | Corrected std. err. | z | P> $\|z\|$ | [95% conf. interval] |
|---|---|---|---|---|---|
| `lwage`L1. | 0.6421957 | .062308 | 10.31 | 0.000 | [.5200742 , .7643171] |
| `educ` | 0.0362161 | .0066122 | 5.48 | 0.000 | [.0232564 , .0491757] |
| `exper` | 0.023187 | .0073787 | 3.14 | 0.002 | [0087251 , .037649] |
| `expwc` | 0 .0058086 | .0053169 | 1.09 | 0.275 | [−.0046124 .0162296] |
| `expser` | −.0143392 | .0077614 | −1.85 | 0.065 | [−.0295512 , .0008729] |
| `year` | −.0028635 | .0041754 | 0.68 | 0.497 | [−.0053471,.0110201] |
| _cons | −2.772198 | .6404345 | 4.33 | 0.00 | [1.516961 ,4.027418] |

Table 3: System GMM

The Arellano-Bond test for AR(2) system GMM has a p-value of 0.245 indicating that the effect of autocorrelation is not much. Additionally, the Sargan and Hansen tests have p-values of 0.160 and 0.298 respectively, suggesting that the instruments are valid. The realized estimates suggest that there is a significant effect of the lagged variable of `lwage`, `educ` and `exper`. While `year` and `expwc` are not significant. The `expser` variable is on the fence with a p-value of 0.065. Therefore the log value of wage is determined by education, total experience, possibly the effect of experience in service job and the lagged effect of log wage.