U LISBOA | UNIVERSIDADE DE LISBOA

ISEG Lisbon School of Economics & Management Universidade de Lisboa

MASTER
QUANTITATIVE METHODS FOR DECISION-MAKING IN ECONOMICS AND BUSINESS

MASTER'S FINAL WORK
DISSERTATION

THIS DOCUMENT'S UNIQUELY PURPOSE IS FOR OBTAINING THE MASTER'S DEGREE

THE IMPACT OF THE GREAT RECESSION ON LABOR INCOME BETWEEN DEMOGRAPHIC GROUPS IN THE UNITED STATES: A PANEL DATA ANALYSIS

FELIPE CAETANO MAGRIN ORTIZ

SUPERVISION:
PROFESSOR PIERRE JOSEPH MARIA HOONHOUT

DECEMBER – 2021

GLOSSARY

CM – Construction and Manufacturing.

FE – Fixed Effects.

FIRE – Finance, Insurance and Real Estate.

GDP – Gross Domestic Product.

IPW – Inverse Probability Weighting.

LR – Likelihood Ratio.

PSID – Panel Study of Income Dynamics.

MAR – Missing at Random.

OFUM – Other Family Unit Member.

OLS – Ordinary Least Squares.

RE – Random Effects.

SRC – Survey Research Center.

WWII – World War II.

ABSTRACT

This dissertation provides insights on the estimation of a labor income (wage) function considering the impacts of the Great Recession on American workers, with a special focus to the different paths from the three generations of individuals that were the majority on the labor market by 2006 – Boomers, Gen X, and Millennials – regarding the labor outcome of interest. By using a subset of data from the Panel Study of Income Dynamics, covering 7 waves from 2007 to 2019, and controlling for some sociodemographic characteristics and other variables traditionally seen in any standard Mincerian equation, it was in fact inconclusive whether the presence of self-selection due to the effect of attrition on estimates, common to these types of datasets, may pose a hazard for the consistency of the econometric methods. Anyhow, methods to tackle this issue were discussed and a two-step estimation using Inverse Probability Weighting (IPW) was also considered. Indeed, the results of the weighted (IPW) and the unweighted estimations on the unbalanced panel were not much different, and both models point out to a better path for individuals that were born within the range of the Millennials generation, with blacks and women being the unlucky cohorts.

KEYWORDS: Attrition; Great Recession; Wage function; Panel data; PSID.

JEL CODES: C33; C55; C83; C87; G01; J31

RESUMO

Essa dissertação apresenta uma abordagem à estimação de uma função salarial que considera os impactos da Grande Recessão para os trabalhadores estadunidenses, com foco nas diferentes trajetórias das três gerações de indivíduos que eram a maioria no mercado de trabalho para o ano de 2006 – *Boomers*, Geração X, e *Millennials* –, em relação ao efeito laboral de interesse. Ao usar um subconjunto de dados do *Panel Study of Income Dynamics*, abarcando 7 ondas desde 2007 até 2019, e controlando para algumas características sociodemográficas e outras variáveis tradicionalmente relacionadas em qualquer equação Minceriana padrão, não se pôde concluir se a hipótese de presença de auto-seleção advinda do efeito do atrito nas estimativas, comum nesses tipos de conjunto de dados, constitui-se num risco para a consistência dos métodos econométricos. De toda forma, discutem-se métodos para atacar esse problema e uma estimação usando *Inverse Probability Weighting* (IPW) em duas etapas é também considerada. De facto, os resultados da estimação ponderada (IPW) e da não ponderada no painel de dados não balanceado não foram muito diferentes, e ambos apontam para uma trajetória mais positiva para os indivíduos que nasceram dentro do espaço temporal da geração *Millennials*, tendo também os afro-americanos e as mulheres como os grupos mais afetados negativamente.

PALAVRAS-CHAVE: Atrito; Grande Recessão; Função salarial; Dados em painel; PSID.

CÓDIGOS JEL: C33; C55; C83; C87; G01; J31.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

Since the downfall of the world's economy due to the subprime crisis in the last half of the 2000s, a multitude of research was undertaken by social science authors regarding the specific effects of this event over the world's society. The labor market was seriously impacted, as employees around the globe saw their jobs disappearing and "too big to fail" companies even failed. Although it has been more than a decade since the end of this chaotic period, for some groups of individuals the effects took too long to be vanished or are still in place.

The recession triggered by this financial crisis, the so-called Great Recession, occurred between the last quarter of 2007 and the second quarter of 2009, having hit the American economy harder than the other disastrous post World War II (WWII) financial crisis, in the early 1980s [e.g., Hoynes et al. (2012), NBER (2021)]. As an expected consequence, the labor market does not react equally for all workers, having some groups a higher likelihood of being more negatively impacted in recessionary times.

In general, unemployment is the outcome that depicts these churning moments better, as firms tend to cut costs to work at the margin. Anyhow, although nominal wages have a kind of rigidity, due to some form of enforcement by the public authorities and to some labor unions' bargain power, these might also be affected.

Considering all that above, this dissertation brings to the discussion the following question: which age groups were the most impacted (either negatively or positively) by the financial collapse that took place in the last quarter of 2007, in terms of labor income, in the United States? Furthermore, what other sociodemographic groups were also significantly affected by this event?

In order to answer these questions, a labor income (wage) function was devised which includes some sociodemographic variables – race, gender, etc. – and that also controls for, amongst other factors, the profiles of the three generations of individuals that were the majority of the labor force by the time of the year right before the recession, 2006 – Boomers, Gen X and Millennials. To estimate that, data from the seven waves starting in 2007 of the Panel Study of Income Dynamics (PSID) were used.

In practice, a fixed effects (FE) model was assumed for the estimation since this type of estimator makes more sense in the context of a wage equation, but also because the Hausman test for the difference between this and a random effects (RE) model rejected the absence of correlation between the explanatory variables and the individual effects.[1] Also, the individual effects, or the unobserved heterogeneity, are indeed allowed to be correlated with at least one of the observed variables, education; but the correlation between this and the idiosyncratic errors is not going to be accounted here as this would require methods that are beyond of the scope of this work, such as instrumental variable's techniques. Further, an Inverse Probability Weighting (IPW) model was devised to tackle the problem of possible self-selection due to attrition (which was tested but no undisputable conclusive results could be delineated).

In effect, differences between attritors and non-attritors were tested for some selected features for the initial wave, where all the sample was observed according to the rules implemented, and some of these were significantly different. By any means, the main conclusions from the estimations of the objective function point to just a few differences between the weighted (IPW) and the unweighted estimated coefficients, mainly related to their significance powers. Therefore, a formal testing procedure was not considered. In fact, both estimations indicated the Millennials generation as having the most positive path from the pre-recession until the recovery years, but the rate of such improvements has been slowing down. Also, women and blacks were the expected losers, having the last group showed statistically significant and negative coefficients for the recovery years after 2012, while the gender variable was only significantly different from zero for the unweighted estimation.

Considering the main covariates of a standard Mincerian equation – working experience and education –, the raw variables in the PSID that are related to these are not so well-documented, i.e., instead of bringing updates after each wave, these are mostly asked as part of a kind of background section. This poses a risk for any estimation using just the raw values. So, some adjustments had to be undertaken in order to use both variables. In short, the estimated coefficients of the experience variable for the objective function are plausibly acceptable; yet, the same cannot be applied to the education

[1] The "phtest" from the "plm" package was used here for the Hausman test (Croissant and Millo, 2008). The Chi-squared statistics is 453.1, with 46 degrees of freedom, and p-value close to zero.

variable in full, as the results show some opposite direction when compared to the economic literature reviewed; but these coefficients are not significantly different from zero, and these effects could be due to some correlation to the idiosyncratic errors, something not covered in this work.

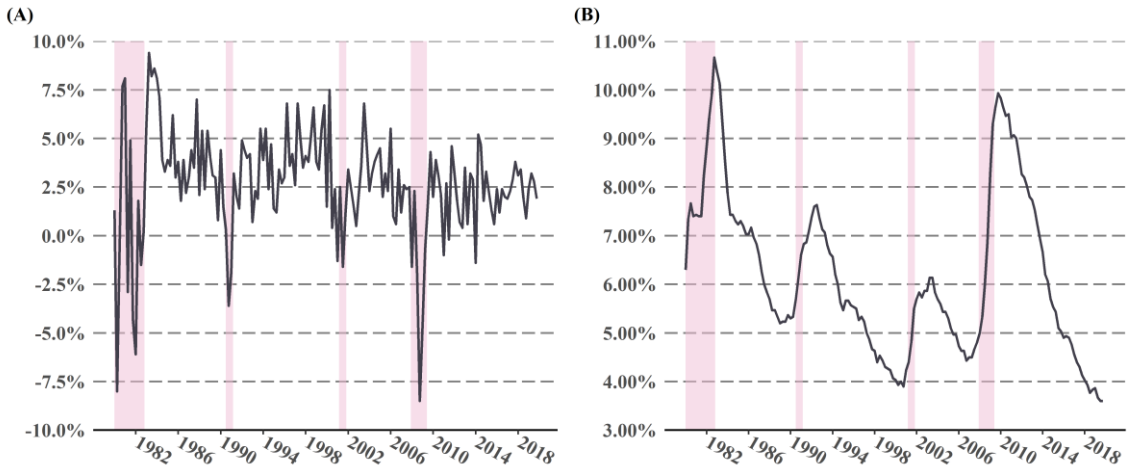
So, the next chapter provides a literature review on the subjects of the Great Recession in the context of the labor market in the U.S. and on the econometric methods needed to perform the analyses using the PSID dataset. Chapter 3 touches the main characteristics of the PSID variables to be included in the objective function and provides discussions about the tests and estimations undertaken. Chapter 4 analyses the results of the estimations and the main differences found for attritors and non-attritors. Finally, chapter 5 provides some conclusions and ideas for future works.

2. LITERATURE REVIEW

In this chapter, the main literature that regards the subjects of the Great Recession on the U.S. labor market, and the econometrics to effectively analyse a complex survey dataset will be presented. Despite the major interest of this work being the quantitative analysis of the set of data retrieved from the PSID, it is indeed highly preponderant to understand what has been done in terms of academic research about the effects of the Great Recession on the U.S. labor market. This will be of great importance when drawing conclusions about the results of the estimated models.

2.1. On the effects of the Great Recession on the U.S. labor market

Figure 1 depicts the quarterly change, in percentage points, of the GDP (A) and the registered unemployment rate for each quarter (B) in the U.S. from 1980 to 2019. As one can conclude, considering the last four decades, the recession triggered by the last financial crisis, before the 2020 pandemic, was the most striking both in terms of production and unemployment. Something also stated by Hoynes et al. (2012), whose analysis also pointed that, except for Hispanic men, virtually all demographic groups showed worse outcomes for this last recession when compared to the 80s recession, in terms of unemployment.



Source: A: BEA: Data Tools (2021), B: BLS Data[2] (2021).

FIGURE 1 – Quarterly change of GDP (A) and unemployment rate (B) in the U.S., from 1980 to 2019, with official cycles highlighted on the background.

[2] Monthly unemployment rate was retrieved, with the quarterly rate being the average value for each of the three months representing each quarter.

Erken et al. (2015) analysed national accounts and found that profits fell right away due to the financial crisis, but this pattern reverted similarly fast while in the recovering period. Also, the authors pointed to some sort of differential outcomes between countries, with the U.S. being included in the group which showed a limited level of long-term unemployment, but a relatively large decline in real wages when compared to labor productivity.

Now, regarding the individuals who started their careers in a context of a recessionary period, considering the last four decades, there is evidence of a higher likelihood of worse labor outcomes. In fact, workers aged 18-34 and holding at least a high school diploma who entered the labor market in 2009 (recession) performed worse when compared to those who entered in 2006 (pre-recession) and in 2012 (post-recession), in terms of median annual wages, considering data from 2006 to 2017 (Atherwood and Sparks, 2019). From a broader perspective, entering the labor market in a high unemployment conjecture turns out to have long-lasting effect. Schwandt and von Wachter (2019) analysed several databases concerning labor market entrants from 1976 to 2015 and concluded that for a raise of 3 points in the unemployment rate, cumulated earnings are predicted to be 60% less of a year of earnings, and that these initial effects are due to employment and wage reductions, while long-term effects are due to persistent declines in wages.

Early research on the matter of the Great Recession on the U.S. labor market mostly presented similar results regarding the negative impacts on some demographic groups, especially on young people. Dickens and Triest (2012) used data from the 2004 and 2008 panels of the Survey of Income and Program Participation and concluded that the Great Recession indeed played a crucial role in the likelihood of an involuntary job transition, although did not greatly change the relative likelihoods related to different types of workers, having the young, less educated, and short-tenured workers more likely to be displaced both before and during the recession. A similar conclusion about youth unemployment was devised by Bell and Blanchflower (2011), who explain that young workers are more likely to be dismissed from their positions due to be less skilled or because they put less pressure on statutory redundancy payments, while they also face an experience trap, as firms rather hire more experienced individuals. The authors analysed this issue on the context of the years of the Great Recession and found that youth

unemployment in virtually all OECD countries had increased the gap compared to adult unemployment.

A different point of view was devised by Sironi (2017) and Atherwood and Sparks (2019) concerning the path of young adults. The trajectories that young people go through to adulthood have changed in the last decades, compared to their counterparts on 1950s and 1960s. By that time, a normal transition would have the following steps: graduate from school, move out from the parental home, start working, marry, and have children. All of this yet on their early twenties. However, recently, young people have shown more diverse patterns, postponing these steps to late twenties or thirties. Also, it is not quite clear when they start or end a formal step (education, work, marriage, etc.), especially in recessionary times. As Sironi (2017) also explains, although youngsters were part of the group that was hit hardest by the last recession, the societal structure of the country where they live might dictate a better or worse path into the labor market.

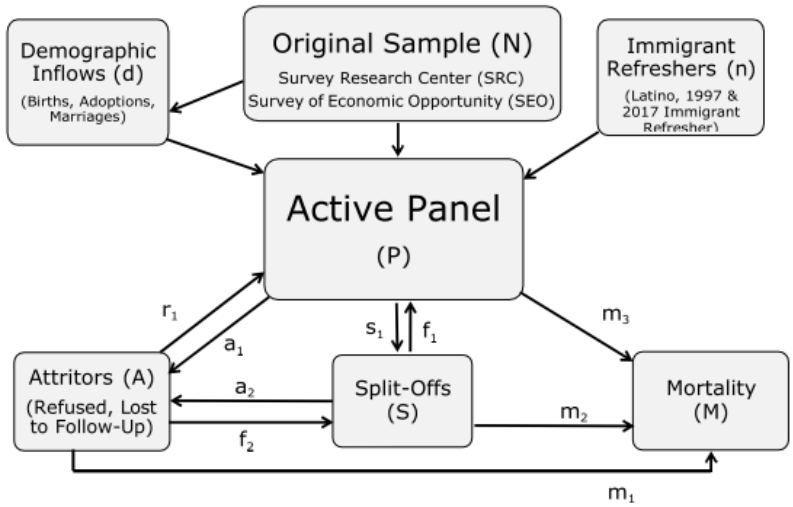
When it comes to gender, the 2007-09 recession hit young, less-skilled men hardest [e.g., Hoynes et al. (2012), Rothstein (2017), Sironi (2017)]. Atherwood and Sparks (2019) analysed microdata from the American Community Survey and concluded that men and women holding at least a high school diploma performed equally through the period from 2006 to 2017 in respect to median annual wages, but with men of all education gradients showing a higher level of wages, especially those holding some sort of graduate diploma. The authors also explain that, for both genders, Asians and Hispanics were better off compared to non-Hispanic whites, while blacks were the most negatively impacted, especially black men.

The main differences between women and men in what regards business cycles are the sensitivity and the sectors they generally work. For instance, women tend to be less sensible to cycles, acting much more like an added workforce during recessions, while men act like the discouraged labor force, and generally work on sectors that are related to the cyclicality [Sironi (2017) and Hoynes et al. (2012)]. Moreover, while in the 80s women saw their path in the labor market not so negatively affected, with an even increasing trend, this is viewed as a secular trend which was reversed before 2007; therefore, the last recession hit women hardest, in relation to the two most severe after the WWII [Hoynes et al. (2012) and Elsby et al. (2016)].

Additionally, the literature shows that college graduates (the finest education gradient) are less likely to be affected in times of economic shock [e.g., Sironi (2017), Rothstein (2017), Schwandt and von Wachter (2019), Atherwood and Sparks (2019)]. As a matter of fact, Rothstein (2017) indicated a stagnation or even decline of general real wages in the U.S. from 2010 to 2014, however with college graduates having showed some recovery compared to 2007 levels. Nevertheless, holding a higher education diploma did not mean a smoother path in the labor market, since bachelors saw their value-to-money lower after 2008 (Atherwood and Sparks, 2019).

2.2. Panel data methods for analysing an unbalanced panel

Now, the econometric methods that will be employed to analyse the dataset object of this work are to be considered. First, it is important to point out here the mechanics of the longitudinal dataset (the sampling and following rules) object of this study. As Figure 2 shows, there are basically three main ways to enter the panel, either by making part of the samples, or by some other natural choice of the sampled individuals, i.e., marriage. Once part of the panel, individuals may split-off from the sampled household (s_1), eventually come to death (m_1, m_2, m_3), attrite (a_1, a_2) and, in this case, become part of the panel again by a recontact (r_1) or because of the follow-status rules (f_1, f_2) (PSID, 2021, p. 9).



Source: PSID (2021, p. 10).

FIGURE 2 – Steady state panel design of the PSID.

By using such a configuration, and considering the multigenerational aspect implied by that, the PSID is a very useful source of information about the population living in the U.S., independently of having American or others roots, since the panel also follows

immigrant families; therefore, reassuring its representativeness. Besides, the number of individuals at each wave is indeed large enough for any statistical analysis focussing on the individual as the cross-section unit, starting with more than 18,000 in 1968, and having interviewed a little more than 26,000 in the 2019 wave (PSID, 2021, p. 15).

2.2.1 Balanced panel analysis

Moving forward, this work shall make use of the econometrics employed to panel data analysis, which differs from a cross-section analysis by including, apart from the obvious cross-section units, the time as a factor – in this study's case, the cross-section units are the individuals, and the time are the waves (years) of interviews. Having said that, the following linear model is primarily considered:

$$(1) \qquad y_t = \beta_0 + \boldsymbol{x}_t\boldsymbol{\beta} + c + u_t,$$

where y_t, $\boldsymbol{x}_t$, c and u_t represent the observable random dependent variable, the vector with the observable random independent variables, an unobservable random variable, and the error term, respectively, for the population of interest, while β_0 and $\boldsymbol{\beta}$ are the (vector) parameters of the equation (Wooldridge, 2010, pp. 281-282). As one can note, the only variable that does not have the time index (t) is the unobservable c, which is then assumed to be constant over the periods and has an implicit parameter identical to the unity.

In fact, as Wooldridge (2010, p. 281) explains, the foremost reason for using panel data is to account and solve for the case of omitted variables problem, and that is why the unobserved variable c appears in Equation (1). In other words, by adding c, possible omitted variables that may affect the response variable y_t, but its partial effect is not reasonable to correctly estimate, is accounted and the parameters can then be estimated. Consequently, as c enters additively along with $\boldsymbol{x}_t$, a structural equation can be written as

$$(2) \qquad \mathrm{E}(y_t \mid \boldsymbol{x}_t, c) = \beta_0 + \boldsymbol{x}_t\boldsymbol{\beta} + c,\ t = 1, 2, \ldots, T,$$

and the only interest lies in the vector of parameters $\boldsymbol{\beta}$ (Wooldridge, 2010, p. 282).

However, to consistently estimate the parameters of interest, more should be assumed about the relation of c with any x_{jt} contained in $\boldsymbol{x}_t$ at any period t. For example, if c is assumed not to be correlated with any x_{jt} at any period t, then it is just another factor affecting the dependent variable which does not pose any trouble in estimating the

parameters. But if it is assumed $\text{Cov}(x_j, c) \neq 0$ for some j at any t, putting the unobserved variable into the error term can cause serious issues for the estimation of $\boldsymbol{\beta}$ (Wooldridge, 2010, p. 281).

Considering Equation (1), another assumption that should be accounted to correctly estimate $\boldsymbol{\beta}$ is

(3) $\quad \text{E}(u_t \mid \boldsymbol{x}_t, c) = 0, t = 1, 2, \ldots, T.$

This has at least the implication that $\text{E}(\boldsymbol{x}_t' u_s) = 0$ for all t and $t \neq s$ (Wooldridge, 2010, pp. 283 and 288). What is implied by such a restriction is the strict exogeneity, i.e., all vectors containing the covariates are orthogonal to the vector of the errors. Additionally, if $\text{E}(\boldsymbol{x}_t' c) = 0$, the pooled Ordinary Least Squares (OLS) regression estimator can be applied, but this is a too strong assumption to carry forward – actually, not much likely, and not of interest for this work as will be explained further – and, rejecting that, pooled OLS is no longer unbiased and consistent (Wooldridge, 2010, p. 283).

Nevertheless, $\text{E}(\boldsymbol{x}_t' c) = 0$ is not going to be assumed for the purpose of this work since this assumption does not make sense for the context of the application here, which relates the explanatory variables to a labor outcome.[3] Having that stated, there is at least one type of estimator that can be excluded from the revision, which is the RE. In other words, there will be room for the unobservable random variable c to be correlated to at least one of the explanatory observable random variables of the vector $\boldsymbol{x}$.

Further, the basic unobserved effects model for a random set of individuals is:

(4) $\quad y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, t = 1, 2, \ldots, T, i = 1, 2, \ldots, N.$

In Equation (4), any covariate that is within the row vector $\boldsymbol{x}$ may be discrete or continuous and is allowed to vary through both the indexes i and t, or at least through one of them. The u_{it} are the idiosyncratic errors and those change across i and t. Finally, the c_i are the individual heterogeneities, which only change across the cross-section units, and are not observed. Such as was stated before, this variable is allowed to be correlated

[3] One of the reasons to assume that the unobserved variable might have a relation to the independent observed variables in the context of a labor outcome function, according to many labor economists, is that the former is viewed to some form of natural ability, which relates to the explanatory variables such as education, that is usually used as a covariate to explain wage [e.g., Fitzgerald et al. (1999), Wooldridge (2010, pp. 282; 2012, p. 463)]. Besides, a Hausman test was devised, and the Chi-squared statistics rules out the consistency of the RE estimator.

to at least one of the observed covariates. These assumptions are known as the FE framework (Wooldridge, 2010, pp. 285-286).

In what concerns the main structural difference between the RE and the FE approaches, the first puts the unobservable random variable within the error term, accounting for the implied serial correlation in the error term, $v_{it} = c_i + u_{it}$, where v_{it} are the composite errors (Wooldridge, 2010, p. 292). However, this is not desired for the FE estimation, as there is no assumption of orthogonality between the terms within $\boldsymbol{x}_{it}$ and c_i. Consequently, as will be further detailed, the only way an observed covariate that is not time-varying, i.e., does not have any change through the years (e.g., race, gender, etc.), to be included in the equation is by interacting it with any other time-varying covariate (Wooldridge, 2010, p. 304).

Now, considering Equation (3) and the assumption of possible orthogonality between the unobserved and observed terms, the idea of a consistent estimation of the vector of parameters $\boldsymbol{\beta}$ according to the FE approach is to transform the equations represented by Equation (4), so the terms that are constant in time, specifically the c_i, are then eliminated (Wooldridge, 2010, p. 302; 2012, pp. 484-485). A popular approach is the within transformation, which consists of averaging the values of the random variables of Equation (4) to further subtract from each period:

$$(5) \qquad \ddot{y}_{it} = y_{it} - \bar{y}_i = (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)\boldsymbol{\beta} + u_{it} - \bar{u}_i = \ddot{\boldsymbol{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it},\ t = 1, 2, \ldots, T,\ i = 1, 2, \ldots, N.$$

This approach results in the "time-demeaned" equation [e.g., Wooldridge (2010, p. 302), Croissant and Millo (2018, pp. 2-3)]. The $\bar{y}_i$, $\bar{\boldsymbol{x}}_i$, and $\bar{u}_i$ mean the averaged values across the time for the dependent variable, the explanatory variables, and the idiosyncratic errors, respectively, for each cross-section i. Implied is also the elimination of any constant variable that is not the c_i, a drawback previously mentioned.

Other important reason to not let time-constant observable random variables without interaction in Equation (4) – assuming time-demeaning – is the rank condition assumption. In other words, to estimate $\boldsymbol{\beta}$ consistently, apart from assuming strict exogeneity, as in Equation (3) but in the context of the time-demeaned variables, the FE estimator behaves well asymptotically if $\text{rank}(\sum_{t=1}^{T} \text{E}(\ddot{\boldsymbol{x}}_{it}'\ddot{\boldsymbol{x}}_{it})) = K$, where K is the number of covariates within $\ddot{\boldsymbol{x}}_{it}$ for any period t (Wooldridge, 2010, p. 304). This is so because

any constant in Equation (4), after the time-demeaning approach, would be zero for any period and any cross-section unit, which makes the K x K matrix $E(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i)$ not full-rank (Wooldridge, 2010, p. 304) – being $\ddot{\mathbf{X}}_i$ a T x K matrix containing the time-demeaned variables for each cross-section i along the time.

So, considering both assumptions for the consistency of the FE estimator, pooled OLS can be applied in the FE approach, and the FE estimator turns out to be as follows:

$$(6) \qquad \hat{\boldsymbol{\beta}}_{\mathrm{FE}} = \left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)^{-1}\left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i\right) = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{\boldsymbol{x}}_{it}'\ddot{\boldsymbol{x}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{\boldsymbol{x}}_{it}'\ddot{y}_{it}\right),$$

which is also called the within estimator since it uses the time variation within each cross-section (Wooldridge, 2010, p. 304). Letting the unobserved random variable c_i to covary with the observables turns out the between estimator, which uses only variation across the cross-section units, inconsistent (Wooldridge, 2010, p. 304). In fact, Generalized Least Squares can also be applied to a FE approach, however, this is generally done under the failure in some asymptotic assumptions, which will not be discussed here.[4] Nonetheless, the asymptotic analysis here is as $N \to \infty$, while T is held fixed.

2.2.2 Attrition analysis

Equation (6) is utterly true, without any modification, and under assumptions about the rank condition and strict exogeneity previously made, for a balanced panel dataset. That is, considering a random draw from the population, all cross-section units are observed exactly T times. In other words, considering a panel made up of N individuals – such as the PSID – to effectively estimate $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$, none of those can drop out. This is a somewhat strong, and possibly unrealistic, condition for any longitudinal household survey to be carried out – if this is to use the same units for a long period (Zabel, 1998, p. 502). In fact, as was demonstrated in Figure 2, attrition in the context of the PSID is expected.

So, considering a balanced panel is not feasible for the dataset to be used, there should be made additional assumptions concerning the consistency of the FE estimator for an unbalanced panel. Firstly, $t = 1$ is the first period of which every cross-section unit is observed, and $t = T$ is the last possible period to be in the panel (Wooldridge, 2010, p. 837). Now, for any random draw from the population of interest – in the case of the PSID,

[4] See Wooldridge (2010), chapter 10, for a more detailed discussion about this topic.

the population residing in the U.S. –, s_{it} is the selection indicator which is identical to 1 when the observed random variables ($\boldsymbol{x}_{it}, y_{it}$) are available, and 0 otherwise (Wooldridge, 2010, p. 837). In short, the FE estimator changes in the following way: $\hat{\boldsymbol{\beta}} = N^{-1} s_{it} \hat{\boldsymbol{\beta}}_{\text{FE}}$, where,

(7) $\ddot{\boldsymbol{x}}_{it} \equiv \boldsymbol{x}_{it} - T^{-1} \sum_{r=1}^{T} s_{ir} x_{ir}$,

(8) $\ddot{\boldsymbol{y}}_{it} \equiv \boldsymbol{y}_{it} - T^{-1} \sum_{r=1}^{T} s_{ir} y_{ir}$,

and $T_i \equiv \sum_{i=1}^{T} s_{it}$, with T_i being the number of periods observed for each cross-section i, so the within transformation is applied only for the available periods (Wooldridge, 2010, p. 829).

Additionally, the strict exogeneity assumption and the rank condition now are conditional on the selection indicator. In fact, to be consistent on unbalanced panels, now FE should have $\text{E}(s_{it} \boldsymbol{x}_{it}' u_{it}) = 0$ for all t (Wooldridge, 2010, p. 829). More specifically, for the case of strict exogeneity, consider

(9) $\text{E}(u_{it} \mid s_{it}, \boldsymbol{x}_{it}, c_i) = 0, t = 1, 2, \ldots, T,$

but the selection indicator is not necessary whenever attrition is completely random. What Equation (9) rules out is some sort of non-randomness in selection, i.e., whenever selection is partially correlated to the idiosyncratic errors (Wooldridge, 2010, pp. 829-830). This, along with the rank condition and the assumption of an unbiased asymptotic variance, makes FE on unbalanced panels unbiased and consistent, and the test statistics can be undertaken (Wooldridge, 2010, p. 829).

On the other hand, if the selection that makes the panel unbalanced is indeed correlated to the errors, the FE approach to estimate the parameters as shown before does not produce reliable results. Fitzgerald et al. (1999) and Zabel (1998) presented different methods to tackle such possibility, both using data from the PSID and in the context of a response variable related to labor outcomes. In fact, it is plausible to suspect that attrition may be correlated to a labor outcome, for example, wage or unemployment. In this case, any estimation using the methods demonstrated until now would result in a biased and/or unrealistic inference about the population from which the sample was drawn.

Therefore, it is necessary to understand whether attrition might pose a hazard for the statistics or not. This can be accomplished by several ways. For example, Fitzgerald et al. (1999, pp. 145-147) proposed, among many methods, an "inversion test", which is essentially the effects of future attrition on the first period outcome variables. By this method, it is possible to test for the differences between non-attritors and the full sample. Wooldridge (2010, pp. 832-833) also discusses two other tests whose mechanics employ either a future selection indicator, $s_{i,\,t+1}$, or a variable counting the number of additional periods that a cross-section unit stays in the panel, $r_{i,\,t+1}$. For both, considering FE estimation, T should be higher than two, and a t-test for the significance of $s_{i,\,t+1}$ or $r_{i,\,t+1}$ is undertaken.

Now, if attrition is indeed related to the idiosyncratic errors according to the testing, a correction must be implemented. As before, there is not only one way to work out a solution here. So, amongst all the methods discussed in chapter 19 of Wooldridge (2010, pp. 837-845), the focus here is on the use of IPW with the estimates of the weights to be constructed sequentially. For that case, the objective function to be estimated is as follows:

$$(10) \qquad \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\frac{s_{it}}{\hat{p}_{it}}\right) q_t(\boldsymbol{w}_{it}, \boldsymbol{\theta}),$$

where $\boldsymbol{w}_{it} \equiv (y_{it}, \boldsymbol{x}_{it})$, $\boldsymbol{\theta}$ are now the parameters to be estimated, $q_t(\boldsymbol{w}_{it}, \boldsymbol{\theta})$ is the objective function at each period – for least squares, this is just the squared residual function –, s_{it} is the usual selection indicator, and $\hat{p}_{it}$ is an estimated weight for each time t and cross-section unit i (Wooldridge, 2010, pp. 840-841).

According to Wooldridge (2010, p. 842), these estimated weights are nothing more than the products of some fitted probabilities, worked out for every additional wave an individual stays in the panel, such as:

$$(11) \qquad p_{it}(\boldsymbol{\delta}_t^{\mathrm{o}}) \equiv \pi_{i2}\left(\boldsymbol{\gamma}_2^{\mathrm{o}}\right) \cdots \pi_{it}\left(\boldsymbol{\gamma}_t^{\mathrm{o}}\right),\ t = 2, \ldots, T.$$

In Equation (11), $p_{it}(\boldsymbol{\delta}_t^{\mathrm{o}})$ is the product of the probabilities fitted by the probit models $\pi_{it}\left(\boldsymbol{\gamma}_t^{\mathrm{o}}\right)$, for each period t and cross-section unit i. $\boldsymbol{\delta}_t^{\mathrm{o}}$ is then the set of the "true" parameters

of each probit model estimated until period t (Wooldridge, 2010, p. 841), and, as a consequence, $\boldsymbol{\gamma}_t^{\text{o}}$ are the "true" parameters for each probit model at time t.[5]

The probit models are represented as

$$(12) \qquad \pi_{it}\left(\boldsymbol{\gamma}_t^{\text{o}}\right) \equiv \text{P}(s_{it} = 1 \mid \boldsymbol{z}_{it\text{-}1}, s_{i,t\text{-}1} = 1),$$

where s_{it} is the usual selection indicator, and $\boldsymbol{z}_{it\text{-}1}$ is a vector of lagged observable random variables that might account for attrition (Wooldridge, 2010, p. 842). As one can reason, Equation (12) is the likelihood of a person being present at wave t in the panel, conditional on a set of variables that might be good predictors of attrition, and on being present at wave t-1. Following that, is straightforward that at $t = 1$ there is no need to estimate this likelihood, as $s_{i1} \equiv 1$ for every cross-section unit.[6]

Now, the key assumption for this method to work is as

$$(13) \qquad \text{P}(s_{it} = 1 \mid \boldsymbol{v}_{i1}, \ldots, \boldsymbol{v}_{iT}, s_{i,\,t\text{-}1} = 1) = \text{P}(s_{it} = 1 \mid \boldsymbol{z}_{it\text{-}1}, s_{i,\,t\text{-}1} = 1),$$

for all $t \geq 2$, where $\boldsymbol{v}_{it} \equiv (\boldsymbol{w}_{it}, \boldsymbol{z}_{it\text{-}1})$, and it allows for attrition to be strongly correlated to past outcomes on the dependent (y) and independent ($\boldsymbol{x}$) observable random variables of interest (Wooldridge, 2010, pp. 842-843). This model can be called either by selection on observables or sequential missing at random, as it is based on the missing at random (MAR) assumption [e.g., Fitzgerald et al. (1999), Hoonhout and Ridder (2019), Wooldridge (2010, pp. 840-841)]. Under the assumption from Equation (13), and because $s_{i1} \equiv 1$, it is easy to arrive to Equation (11) by considering the following:

$$(14) \qquad p_{it}^{\text{o}} \equiv \text{P}(s_{it} = 1 \mid \boldsymbol{v}_i) = \text{P}(s_{it} = 1 \mid \boldsymbol{z}_{it\text{-}1}, s_{i,t\text{-}1} = 1)\cdots\text{P}(s_{i2} = 1 \mid \boldsymbol{z}_{i1}).$$

Additionally, Wooldridge (2010, p. 823) demonstrates why applying IPW on the missing data problem on the dependent variable using just the observed outcomes can recover the population mean of any function $\boldsymbol{w}_{it}$. This is achieved by taking iterated expectations as

[5] "True" parameters are called like this here to distinguish the true values of the parameters in an M-estimation context from other candidates.

[6] The theoretical details of estimating a probit model will not be covered here, as the main goal is to demonstrate ways to tackle attrition in panel datasets. However, all the necessary underlying methods for doing such can be approached in chapters 15 and 19 of Wooldridge (2010).

$$
\begin{aligned}
(15) \quad \mathrm{E}\left[s_{it}\mathrm{g}(\boldsymbol{w}_{it})/p^{\mathrm{o}}_{it}\right] &= \mathrm{E}\{\mathrm{E}\left[s_{it}\mathrm{g}(\boldsymbol{w}_{it})/p^{\mathrm{o}}_{it} \mid \boldsymbol{v}_i\right]\} \\
&= \mathrm{E}\left\{\mathrm{E}\,(s_{it} \mid \boldsymbol{v}_i)\mathrm{g}(\boldsymbol{w}_{it})/p^{\mathrm{o}}_{it}\right\} \\
&= \mathrm{E}\left\{\mathrm{P}\,(s_{it} = 1 \mid \boldsymbol{v}_i)\mathrm{g}(\boldsymbol{w}_{it})/p^{\mathrm{o}}_{it}\right\} \\
&= \mathrm{E}\left\{p^{\mathrm{o}}_{it}\,\mathrm{g}(\boldsymbol{w}_{it})/p^{\mathrm{o}}_{it}\right\} \\
&= \mathrm{E}[\mathrm{g}(\boldsymbol{w}_{it})].
\end{aligned}
$$

Moreover, as it is also stated by Wooldridge (2010, pp. 500-502 and 843-844), estimating the probabilities by Maximum Likelihood, instead of taking some known values, is more efficient, considering that the conditional distributions of all s_{it}, conditional on $\boldsymbol{v}_i$, are fully specified, and the estimation in the first step accounted.[7] In effect, estimated weights imply that standard deviations should be corrected for it, as opposed to known weights. This can be done, for example, by a simulation method like bootstrapping (Wooldridge, 2010, pp. 438-442).

[7] For a better understanding of why this is true, see chapter 13 of Wooldridge (2010).

3. ANALYSING THE EFFECTS OF THE GREAT RECESSION ON HOURLY WAGES IN THE U.S.

After the revision of the literature concerning the main subjects of this research, this chapter touches the application of the methods priorly discussed on data retrieved from the PSID. As the main goal here is to identify the effects that the Great Recession inflicted on hourly wages of the American workers, the waves selected to be analysed are waves 35 (2007) to 41 (2019) – seven in total, as from a wave to another there is a difference of 2 years. Firstly, a more detailed revision about the characteristics of the PSID data to be worked will be given. After that, the application of the methods worked out so far will be carried forward.

3.1. Pre-processing data from the PSID

The PSID was originally designed in 1968, at the Michigan University's Survey Research Center (SRC), to fulfil the interest of continuing a national assessment on poverty in the U.S. and forming a representative sample of the American society at that time to be interviewed in a yearly basis. This initial study had a sample of 1,872 low-income households (an over-sample) and a nationally representative sample of 2,930 families, the SRC sample. By now, 41 waves of interviews were completed, covering information about more than 82,000 people and as many as seven generations within sample families represented (PSID, 2021, p. 8).

The individuals who are of interest in this work are those whose employment statuses may have changed between a wave to another, but they all answered positive worked hours at each wave. This is so because a wage offer analysis is not of interest, in which case, the prolonged unemployment should be also accounted to not incur into a self-selection problem of this order. Besides, only non-institutionalized individuals, from the SRC sample, aged 18 or more, and being either the head of a household or the head's spouse/partner were included in the dataset object of this work.

Additionally, another restriction to the selection was imposed, following Fitzgerald et al. (1999) and Wooldridge (2010, p. 837): individuals who attrite at any of the waves after 2007 and come back to the panel at any point in time were considered only until the last wave before attrition. Also, no additional entries were allowed, i.e., individuals entering

the panel between 2009 and 2019 were not included in the analysis. By doing such, the methods for correcting for attrition, if it is an issue for the analysis, are consistent.[8]

Complimentary to that, some adjustments on a few variables had to be undertaken to get more reliable values. Actually, three of the explanatory covariates that are key to a wage function, according to the standard Mincerian earnings equation [e.g., Acemoglu (2002, p. 17), Fournier and Koske (2012, p. 9)], are not much reliable in the PSID: age, education, and working experience. This is because of either a wrong entry (in the case of age, it is just impossible for an individual to have a lower value at time t+1 after time t), or because the variable is not updated frequently by the PSID team – a kind of background section, which is only observed in just a few situations, such as a new entry, and for all other years this value is just brought forward.

In the case of the age variable, the procedure to adjust is straightforward: the minimum value for each individual was taken to be the first; after that, 2 years were added subsequently at each wave. By configuring like this, all individuals had their ages linked to the number of years between waves. Yet, for the other two variables, something more refined had to be done.

Taking the working experience first, there are at least three possible variables in the PSID to take into account whenever this covariate is of interest, two regarding years of experience since 18 years old, and other related to job tenure. This last was not considered since its values do not imply a sequence, which can cause some confusion about its true representation (Brown and Light, 1992). For example, if someone who had worked 7 years for a company, and has now 7 years of experience for another company, but is dismissed and hired again by the former company, the next year this variable will be 8, however, it does not make it clear for which company; thus, this may be a source of bias since the impact of the job tenure might be different from someone that is indeed 8 years straight working for a company.

Nonetheless, the main issue for the two left is that the update only happens when an individual enters the survey, re-enters, or when they change status (head/spouse/partner).

[8] Moreover, as the PSID has a structure that follows the participants across the waves, and only adds new ones by either a recontact or under a refreshment in its base, it is not expected that additional entries move the results towards any direction significantly (Fitzgerald et al., 1999, p. 142).

This is, suppose an individual entered the survey in 2007, and some analyst wants to observe experience for this individual after 2 waves since 2007, the value might be the same as 2007, even if this individual had worked between the waves. Another possibility is to observe a value at the initial wave of the analysis that is actually a value related to any other previously wave, so the true value was first appointed in a year outside the range of the chosen waves. In fact, both situations should happen if no care was taken to assess the reliability of the documented values of the variables.

Having stated the drawbacks, the way to proceed to the adjustments follows closely the one that was devised by Blau and Kahn (2017) – however, no logit or probit model was implemented for this stage of the work, the focus is solely on the support variables and the heuristics adopted by the authors. So, using data from the 1985 wave[9] to the 2019 wave, for all individuals that were present in the 2007 (base year of the analysis), the mechanics of the adjustment is as follows: for the first wave of every individual (less than or equal to 2007), the biggest value between the two PSID variables of experience since 18 years old (full-time or a more generic) was assigned; for further waves, experience was calculated based on one of the statements (for one to be true, the others should be false, a kind of "if else" statement):

- If the difference between two adjacent waves is 1 (year), and at least one of the values for the variables of hours worked and weeks worked is positive, then 1 year of working experience was accounted.
- If the difference between two adjacent waves is 1, and at least one of the variables of hours worked and weeks worked are either not observed or null, no experience was accounted, i.e., 0.
- If the difference between two adjacent waves is 2, the year is 2003 or higher, the variable of weeks worked in the gap year[10] is positive, the variable of whether employed in gap year indicates work, and at least one of the values for the variables of hours worked and weeks worked is positive, then 2 years of working experience was accounted.

[9] According to Blau and Kahn (2017, p. 856), in 1985 the PSID asked for all respondents for an update in the background section's variables, including the two regarding working experience since 18 years old.

[10] Gap year is the year which is not covered by the PSID starting on the 1999 wave. In this case, a wave was only undertaken at each 2 years.

- If the difference between two adjacent waves is 2, the year is 2003 or higher, the variable of weeks worked in the gap year is positive, the variable of whether employed in gap year indicates work, and at least one of the variables of hours worked and weeks worked are either not observed or null, then 1 year of working experience was accounted.
- If the difference between two adjacent waves is 2, the year is 2003 or higher, the variable of weeks worked in the gap year is not observed or null or the variable of whether employed in gap year does not indicate work, and at least one of the values for the variables of hours worked and weeks worked is positive, then 1 year of working experience was accounted.
- In any other case, especially those of attrition between waves (in this case, attrition is allowed since this could have happened before 2007), the result follows a routine which tries to cover all years of start and end of a job according to the observed in the PSID variables related to that (e.g., year of start of job1, end year of job2, etc.), then, the result is an approximate integer depending on the range covered by these variables, with minimum value 0 and maximum value being the difference between waves.[11]

After calculating the working experience between each wave, the routine to calculate the accumulated working experience, i.e., the cumulative sum of the experiences worked out as above, is as following: whenever an updated value of one of the two PSID variables of years of experience since 18 years old is higher than the cumulative sum of the calculated experience, this value is accounted and all values before that wave are adjusted; if it does not happen, the accumulated experience is just the cumulative sum of experience (taken the adjusted values as before or not). This step of accounting for updates in the PSID variables tries to approximate more to the real values, which is of utterly interest.

As a final note, an adjustment on the values of the PSID variables of years of experience since 18 years old which had values that were not possible, considering the age of the individual, was also implemented priorly to the routine just described. In these

[11] All variables' names that were used as support to calculate the experience are in Table VI in the appendix

cases, the values were truncated to the integer that, summing up to 18, resulted in the individual's age for that wave.

Now, considering the variable of education in the PSID, this also makes part of the background section, however, a less expensive method was undertaken here. In short, there is a PSID variable that accounts for the year of the last degree earned, which was used as a support to change, whenever needed, the value of the PSID variable for a more accurate one. The routine implemented is as follows:

- If the first wave observed has no observed value for education, then the closest observed value for any of the following waves is taken.
- Considering waves higher than the first, if education is not observed or the value is lower than for the last wave, the last wave's value is taken.
- Otherwise, education is just like the PSID variable.

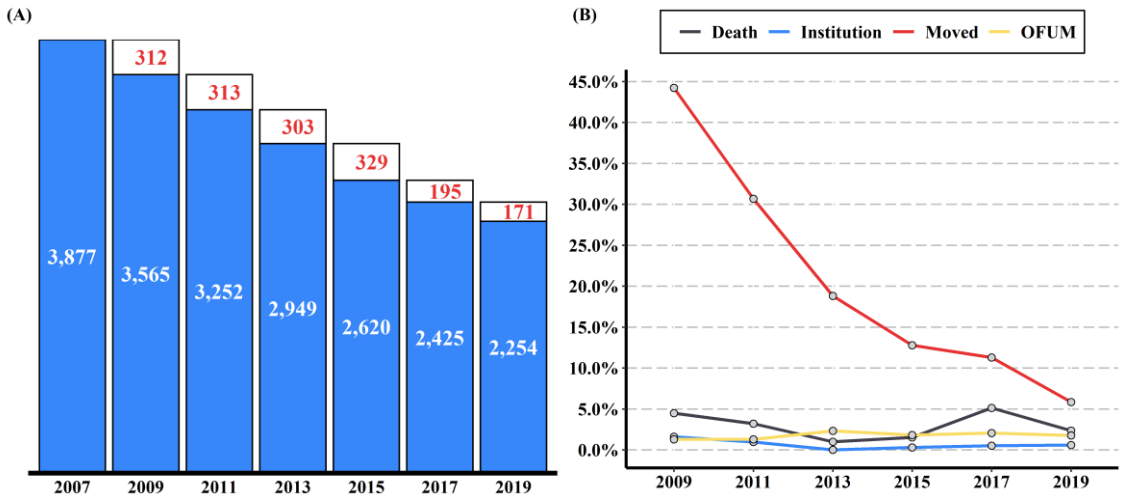
In fact, the methods worked out does not guarantee an error-free variable. Actually, except for the age variable, which is straightforward to see why is plausible the routine applied, the new values for education and working experience are based on a somewhat more realistic measure than the ones that are included in the PSID files, but with possible measurement errors. Anyhow, this possible issue is not going to be accounted here; therefore, the values resulted from the routines are taken as the best possible approximation, bearing in mind the variables available as support.

As a final comment on the selection of the set of individuals, all who had missing values for any of the explanatory variables chosen to estimate the wage function were excluded. This was done since this kind of selection is related to the MAR assumption.[12]

Finally, Figure 3 depicts the number of participants at each wave according to the restrictions described above. Taking Figure 3.A, it is easy to see that the proportion of attritors through the years from 2009 to 2015 is in an increasing pattern, while for the last two years it slows down. Figure 3.B states the known reasons for attrition, at least from the point of view of being a respondent (which means, in terms of the PSID, being at the family unit at the time of the interview, and, by the purpose of this research, being either

[12] In summary, in this context, MAR consists on the assumption that selection may be correlated with the explanatory variables $\boldsymbol{x}$, but not with the error term u, i.e., $E(u \mid \boldsymbol{x}, s) = E(u \mid \boldsymbol{x}) = 0$, where s is the selection indicator (Wooldridge, 2010, p. 795).

a household's head or the spouse/partner): i) death; ii) an individual that become institutionalized; iii) an individual that moved out from the family unit between adjacent waves; and, iv) an individual who lost the status in the family of head or spouse/partner (OFUM). These reasons in 2009 accounted for more than 50% of the attrition, but, for some reason, especially for the "movers", this proportion lost the track across the waves.



Source: Panel Study of Income Dynamics, public use dataset. (2021).

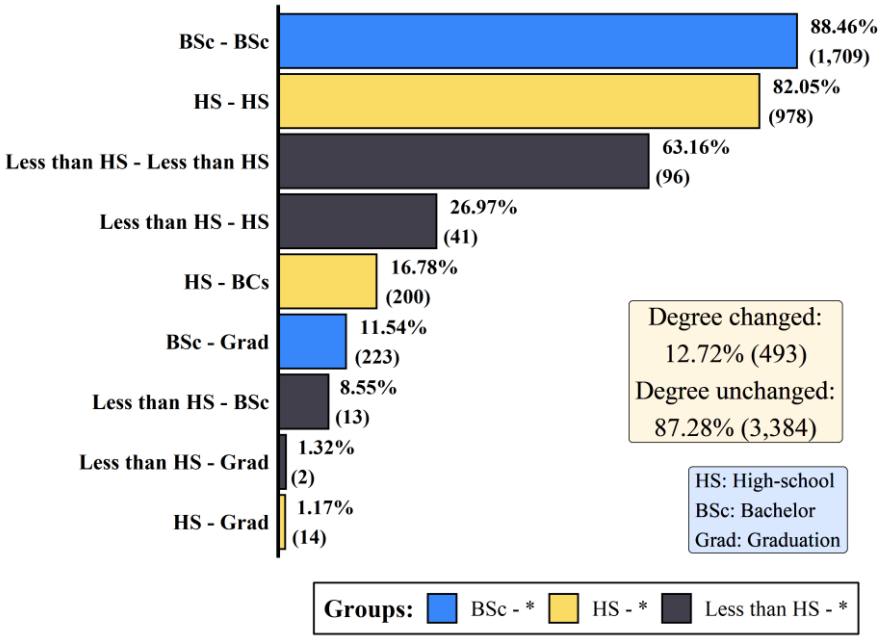
FIGURE 3 – Number of participants and attritors at each wave (A) and known reasons for attrition (B).

Moreover, the dataset to be analysed is composed majority by men (median of 54.4%), whites (median of 93.3%), non-unionized workers (median of 87.8%), bachelors for at least most of the time (median of 51%), people living outside the Southern region of the U.S. (median of 68%), married people (median of 75.5%), homeowners (median of 75.1%), and mortgage payers (median of 63.8%). Also, the median labor income is US$56,925.00, with standard deviation of US$111,733.30 (values were corrected for inflation by using 2021 as the base year)[13].[14]

Complimentarily, Figure 4 and Figure 5 depict the different education paths that individuals took across the waves, and the hourly wage's distribution for each year, excluding the outlier values, respectively. The former compares the education degree from the 2007 wave to the last wave the individual stayed in the panel (graduates never change their status, so no need to appear in the bars). Considering a more traditional life

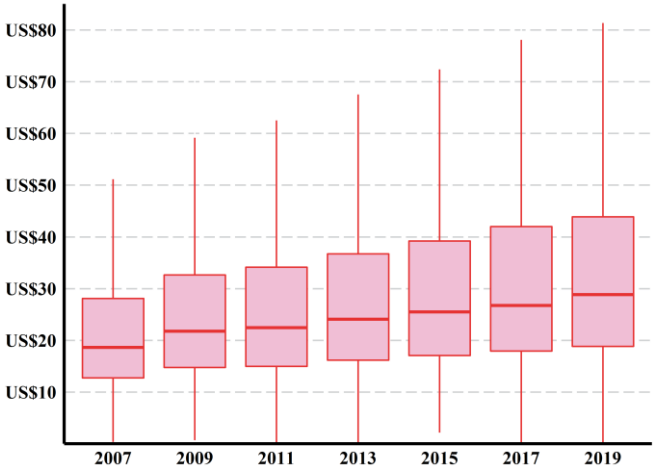
[13] The CPI Inflation Calculator from the U.S. Bureau of Labor Statistics was used here (BLS, 2021).
[14] Tables VII, VIII and IX in the Appendix display a more detailed information about the variables.

trajectory, mostly related to the post-war generations (Sironi, 2017), it is expected that the great majority of the individuals do not change the education degree after in the labor market. However, one still can see examples of great change across the waves. Also, for the hourly wages, the period ranging from the recession (2009 wave) to right after it (2011) seems slightly stagnated and is clear the augmented dispersions after 2007.[15]



Source: Panel Study of Income Dynamics, public use dataset. (2021).

FIGURE 4 – Changes in education degree along the waves.



Source: Panel Study of Income Dynamics, public use dataset. (2021).

FIGURE 5 – Hourly wage's distribution, with outliers excluded, at each wave.

[15] Labor income in the PSID is related to the year previous to the interview.

3.2. Panel data methods applied on the PSID dataset

Now, the methods discussed in chapter 2 will be applied to the dataset worked out as described in the last section. All applications (and the routines previously explained) were undertaken using the statistical software R. In fact, to get in touch to the raw PSID data, there is a package for R, "psidR", that exports data from the individual and family files of the PSID, for all the desired waves, samples, and variables to be worked out in the R environment by using the functions "build.panel" and "getNamesPSID" (Oswald, 2021).

As it was already uncovered, the main interest of this research lies in assessing the effects of the Great Recession on the American workers' hourly wages, controlled for some sociodemographic characteristics, and with another goal of understanding whether being part of one of the generations that were the core of the labor market for the period chosen – i.e., Boomers (born between 1946 - 1964), Gen X (born between 1965 - 1980), and Millennials (born between 1981 - 1996) – brought about any possible negative or positive impact.[16] So, beyond this categorical variable about the generations, other covariates that entered the equation as explanatory factors were: education (categorical variable with four possible values)[17]; experience and a quadratic of experience; a union dummy; a South dummy (whether the individual lived in the Southern region of the U.S. or not); a female dummy; a black race dummy; a marriage dummy; dummies for working either on the finance, insurance, and real estate (FIRE) industries or on the construction and manufacturing (CM) industries; a categorical for four different periods – right before the recession, during the recession, right after the recession, and the other recovery periods until 2019; and the number of children in the family unit. The PSID variables used here are all displayed in Table VI in the appendix.

The first analysis here is to assess a possible correlation between attrition and the dependent variable, which is the natural logarithmic of the hourly wage, using all variables described above as explanatory. As there are some of these that are time-invariant (e.g., female, race), and FE is assumed for this estimation, some interactions were worked out, namely: i) the categorical variable for periods with the generations'

[16] The start and end years of each generation follows the stated by Beresford Research. (2020).

[17] Instead of working with the integer values of years of education, which is somewhat inconsistent due to the way the PSID documents this variable, the following categories were devised: Less than High School, for individuals with less than 12 years of education; High School, for those with exactly 12 years; Bachelor, for those with any value between 13 and 16 years; and, Graduate, for those with exactly 17 years.

variable, the dummies FIRE and CM, separately, the dummy for black race, and the dummy for being female; ii) number of children with the dummy for being female, the dummy for being married (here, marriage is a constant through the years for all individuals), and the dummies for being female and married; iii) education with the generations' variable, the dummy for being female, and the dummy for black race.[18]

Following the methods for testing for attrition according to Wooldridge (2010), and discussed in chapter 2, Table I displays some results for the FE estimations on the unbalanced panel considering two models, which, besides the set of explanatory variables already discussed and their interactions, have the following differences: an additional variable representing future selection ($s_{i,t+1}$) was included in Model 1, i.e., whether the individual participated in wave t+1 or not (attrition); while in Model 2 an additional variable representing the number of waves after wave t ($r_{i,t+1}$) each individual has in the panel was included. To estimate both, the R package "plm" and the function "plm" along with the chosen arguments "within", for the model parameter (i.e., this is the same as the within estimator worked out in chapter 2), and "individual", for the effect parameter (i.e., this is linked to the type of effect c_i has for the estimation), were applied on the PSID dataset (Croissant and Millo, 2008).

TABLE I

COMPARISON BETWEEN MODELS WITH DIFFERENT ATTRITION VARIABLES

Model 1	Log(Hourly Wage)	Model 2	Log(Hourly Wage)
$s_{i,t+1}$	-0.020 (0.013)	$r_{i,t+1}$	0.159 (0.039) ***
All explanatory variables	…	All explanatory variables	…
Observations	20,942	Observations	20,942
Within R^2	0.145	Within R^2	0.145
F-Statistic	61.250 *** (df = 47; 17,018)	F-Statistic	61.608 *** (df = 47; 17,018)

Source: Panel Study of Income Dynamics, public use dataset. (2021).
Note 1: *** p-value < 0.01.
Note 2: Standard deviation in parenthesis.

In fact, as the different coefficients shown in Table I do not agree to each other in terms of significance, no conclusive interpretation can be made so far. Anyhow, another possible test covered in chapter 2 is the "inversion test". Table II shows some results of

[18] Although some of the interactions occurred between dummy variables, not all dummies are time-invariant, then some variability through the periods is seen and the rank assumption is valid.

this test applied to the dataset. In short, the test compared the 2007 characteristics of non-attritors with the full sample for that year. The comparison here is done by interacting a dummy variable for being always in the panel with the explanatory variables (except the categorical for the periods, which is not included here since it does not make sense for a static analysis). Here, two of the coefficients of interest related to the generations are significant for at least 90% of confidence when interacted with the dummy for being always in; however, this dummy is not significant at all. Additionally, by applying an F-statistic considering a restricted model which drops the interactions and the dummy for being always in the panel, it is not rejected the hypothesis of null coefficients for these variables, at a 10% significance level.

TABLE II

INVERSION TEST

	Log(Hourly Wage)	
Intercept	2.25 (0.09)	***
Always In	-0.11 (0.12)	
Less than High School : Always In	-0.07 (0.10)	
Graduation : Always In	-0.14 (0.06)	**
Bachelor : Always In	-0.02 (0.04)	
Experience : Always In	0.01 (0.01)	
Experience^2 : Always In	-0.00 (0.00)	
Union : Always In	-0.05 (0.06)	
South : Always In	-0.04 (0.04)	
Millennials : Always In	0.15 (0.08)	*
Boomers : Always In	-0.15 (0.07)	**
FIRE : Always In	-0.02 (0.07)	
CM : Always In	-0.02 (0.05)	
Black : Always In	0.12 (0.08)	
Children : Always In	0.00 (0.02)	
Female : Always In	-0.04 (0.04)	
Married : Always In	-0.02 (0.04)	
Explanatory variables w/ no interactions	…	
Observations	3,877	
R^2 | Adjusted R^2	0.278 | 0.272	
Residual Std. Error	0.577 (df = 3,845)	
F-Statistic	47.812 *** (df = 31; 3,845)	
F-Statistic (interactions and Always In)	7.629 (df = 16; 3,861)	

Source: Panel Study of Income Dynamics, public use dataset. (2021).
Note 1: * p-value < 0.1; ** p-value < 0.05; *** p-value < 0.01.
Note 2: The term ":" represents interaction.
Note 3: Standard deviation in parenthesis.
Note 4: The rationale for the last F-test follows from chapter 7 of Wooldridge (2012, pp. 240-248). The function "anova" from the package "stats" was used here.

Now, considering the tests undertaken, and the fact that attrition may be a potential source of bias for the estimation of the wage function (perhaps weak, but still significant for at least the consistency of some covariates), the IPW approach is to be used. Table III indicates the differences between attritors and non-attritors for some variables that may indicate a way to build a model that accounts for selection/attrition, considering only the initial wave, where the dataset was complete:

TABLE III

DIFFERENCES BETWEEN ATTRITORS AND NON-ATTRITORS FOR SELECTED FEATURES

Features	Non-attritors	Attritors	
Millennials (prop.)	15%	15%	
Gen X (prop.)	47%	47%	
Boomers (prop.)	38%	38%	
Less than High School (prop.)	3%	5%	***
High School (prop.)	28%	34%	***
Bachelor (prop.)	51%	48%	**
Graduation (prop.)	17%	13%	***
Employed (prop.)	96%	96%	
Median wage	47,342.00	44,545.00	
Std. dev. wage	75,560.00	115,989.00	***
Hours worked	2,079	2,099	
Income/mortgage – 20% (prop.)	5%	7%	**
Income/mortgage – 40% (prop.)	11%	11%	
Income/mortgage – 60% (prop.)	23%	19%	*
Income/mortgage – 80% (prop.)	18%	16%	
Income/mortgage – 100% (prop.)	43%	46%	
Rent (prop.)	27%	28%	
FIRE (prop.)	9%	7%	
CM (prop.)	18%	24%	***
Black (prop.)	7%	6%	
South (prop.)	31%	33%	
Married (prop.)	75%	72%	**
Might move (prop.)	37%	38%	
Person interview (prop.)	1%	2%	***
Homeowner w/ mortgage (prop.)	63%	62%	
Female (prop.)	47%	42%	***
Time of interview	86.79	87.48	
Weeks out of work	16	15	
Weeks unemployed	1	1	
Weeks worked	32	32	
Union (prop.)	13%	10%	*

Source: Panel Study of Income Dynamics, public use dataset. (2021).
Note: * p-value < 0.1; ** p-value < 0.05; *** p-value < 0.01.

In effect, Table III shows the significance of the differences between the parameters for both groups. The tests were done using the package "stats" for R language. For example, in the case of the median wage, a Mann-Whitney test was performed using the function "wilcox.test"; to test the differences between standard deviations of wages, the "var.test" function was implemented for the F test; for proportions and continuous variables, functions "prop.test" and "t.test" were used, respectively.[19]

In short, the main differences between attritors and non-attritors are related to: i) education, and especially concerning high education diplomas; ii) the spread of the distribution of the wages; iii) have worked in the CM industries; iv) have been married; v) have been interviewed by a person; vi) being a woman; and, vii) have been part of any labor union. Apart from the already discussed variables, other features were analysed here, such as the time an interview lasted and whether it was undertaken personally or by telephone, and the likelihood of moving to another place to live by the next wave, which may be good proxies for future attrition (Zabel, 1998). Also, an index which relates the proportion of the mortgage to be paid and the annual family income was devised; however, not much likely to influence attrition.

So, after that step, and following what is written in chapter 2 to undertake the correction for attrition according to the IPW model, the probit models for each year after the initial wave were estimated (the covariates included and all the results for the 6 models can be checked in Table V in the appendix).[20]

Finally, with the results of the sequential weights worked out according to the chosen probit models, applying the IPW to the objective function is straightforward. Table IV depicts the main results of the estimated wage function, with and without the weights.[21]. Notice that, because the weights were in fact estimated in a previous step, the standard deviations of the weighted estimation are not the correct ones. So, a non-parametric bootstrap routine was implemented to estimate the correct standard deviations, which are reported in Table IV.[22] This was needed since the weights are not known beforehand, and by doing that, an extra variation is introduced in the final weighted estimates.

[19] See R Core Team (2020) for a better understanding of the functions implemented.

[20] In order to that, the "glm" function of the "stats" package for R was used (R Core Team, 2020).

[21] To estimate the weighted model, the parameter "weights" of the "plm" function was supplied with the estimated weights (Croissant and Millo, 2008).

[22] This routine considered 1,000 estimations using a subsample of 2,000 individuals from the first wave.

TABLE IV

ESTIMATED COEFFICIENTS FOR THE UNWEIGHTED AND WEIGHTED FUNCTIONS

	Unweighted		Weighted (IPW)[a]	
	Log(Hourly wage)		Log(Hourly wage)	
Less than High School	-0.07 (0.10)		-0.07 (0.13)	
Bachelor	-0.06 (0.05)		-0.07 (0.06)	
Graduation	-0.04 (0.06)		-0.04 (0.09)	
Experience	0.08 (0.00)	***	0.08 (0.00)	***
Experience^2	-0.00 (0.00)	***	-0.00 (0.00)	***
Union	0.07 (0.02)	***	0.07 (0.03)	***
South	-0.05 (0.02)	*	-0.05 (0.03)	
Recession (2008)	0.05 (0.02)	***	**0.05 (0.02)**	*
Early-recovery (2010)	-0.02 (0.02)		-0.02 (0.03)	
Recovery (2012-2018)	-0.04 (0.02)	*	**-0.05 (0.03)**	
FIRE	0.02 (0.03)		0.01 (0.04)	
CM	0.04 (0.02)	**	**0.04 (0.03)**	
Children	0.03 (0.01)	**	**0.02 (0.02)**	
Recession : Millennials	0.07 (0.03)	***	**0.07 (0.04)**	*
Early-recovery : Millennials	0.09 (0.03)	***	**0.09 (0.04)**	**
Recovery : Millennials	0.11 (0.03)	***	0.10 (0.04)	***
Recovery : Boomers	0.06 (0.02)	***	**0.07 (0.03)**	**
Recovery : FIRE	0.07 (0.03)	**	**0.08 (0.04)**	*
Recovery : CM	0.05 (0.02)	***	**0.06 (0.03)**	**
Early-recovery : Black	-0.07 (0.04)	*	**-0.08 (0.05)**	
Recovery : Black	-0.09 (0.03)	***	**-0.09 (0.04)**	**
Recovery : Female	-0.03 (0.02)	*	**-0.03 (0.02)**	
Female : Children	-0.04 (0.02)	**	**-0.04 (0.03)**	
Less than High School : Millennials	0.27 (0.13)	**	**0.30 (0.18)**	
Graduation : Millennials	0.25 (0.08)	***	**0.26 (0.12)**	**
Graduation : Black	0.27 (0.15)	*	**0.34 (0.23)**	
Bachelor : Black	0.30 (0.10)	***	**0.34 (0.15)**	**
Female : Children : Married	0.04 (0.02)	*	**0.04 (0.03)**	
Other interactions[b]	…		…	
Observations	20,942		20,942	
Within R^2	0.145		0.144	
F-Statistic	62.523 *** (df = 46; 17,019)		32.316[c] *** (df = 46; 17,019)	

Source: Panel Study of Income Dynamics, public use dataset. (2021).

(a) Bootstrapped standard deviations reported.

(b) The p-values for the coefficients of the remained interactions are all above 10%, for both estimations, resulting in not significantly different from zero effects.

(c) The F-statistics for the weighted estimation was calculated using a bootstrapped covariance matrix, supplying it to the "pwaldtest" function of the "plm" package (Croissant and Millo, 2008).

Note 1: * p-value < 0.1; ** p-value < 0.05; *** p-value < 0.01.

Note 2: The term ":" represents interaction.

Note 3: Standard deviation in parenthesis.

Note 4: Highlighted coefficients in the IPW estimation means a different significance power when compared to the unweighted.

4. INTERPRETATION OF THE RESULTS

This chapter will analyse the results related to both Table III and Table IV. So, regarding the former, in summary, considering only the initial wave where all individuals in the dataset were part of, an attritor is less likely to be a woman, to participate in labor unions, to get a high education degree (BSc or higher), to be married, to undertake the PSID interview by telephone, and more likely to work on the construction and manufacturing industries. Additionally, compared to the non-attritors group, the distribution of the 2007 wave' wages of the attritors is more skewed, which may indicate that, although the median wage was not significantly different between the groups, attritors might be also those whose wages were on the top of the distribution, something similarly stated by Fitzgerald et al. (1999, p. 142).

When it comes to the estimation of the objective function (Table IV), the IPW approach chosen for correcting the possible attrition issue yields only a few different interpretations of the estimated coefficients, compared to the unweighted. Mostly, these differences are in respect to a somewhat lower power of significance of the estimated coefficients for the weighted model, since the standard deviations should be corrected for the fact that the weights were estimated beforehand.

Additionally, the probit models worked out as in Table V in the appendix seem to lose the explanatory power regarding the last two waves, as per the Likelihood Ratio (LR) test performed, which may translate into somewhat meaningless weights for the case of attrition, as the covariates that predict it may not be as good as expected for all waves. Nonetheless, the percent correctly predicted specification test was also performed and show exactly the opposite, having in fact shown satisfactory rates for all waves.

Anyhow, in relation to the semi-elasticities shown in Table IV, the coefficients estimated for the education levels are the most intriguing. (Both models consider having a high school diploma as the baseline of the comparison.) In effect, none of the estimated coefficients for the levels left was significantly different from zero. Further, all of them would actually decrease the wage, something that does not follow the literature reviewed in what concerns higher education diplomas – in general, it has been stated by many authors that more education indeed yields a higher pay. These effects may be related to some sort of erroneous values for the education variable as already discussed, even after

the correction routine implemented, or a possible correlation to the idiosyncratic errors, something that is not covered in this work.

Nevertheless, higher education diplomas' categories from the education variable, when interacted with the dummy for black race, show estimated coefficients which are significant at a 10% significance level and yield the highest impact on hourly wage, for the unweighted estimation. For example, black people holding a bachelor's or graduation degree are 24% and 23% better paid than whites with a high school diploma, as this last demographic group is the baseline.[23] However, the IPW estimation only considers significantly different from zero blacks holding the highest education gradient diploma, with a higher effect on hourly wages of 27% faced to the baseline group. These effects on blacks are greatly downsized after 2012 (2013 wave on the PSID), by 9% for both estimations.

Now, regarding the generations, considering Gen X being the baseline, Millennials have shown better and most often significant results. For example, when interacted to the education variable, higher graduated (graduate level) Millennials accrue about 21% and 22% more dollars than someone from Gen X holding a high school diploma, respectively, by the unweighted and the weighted estimations. However, what is not expected at all, and again against the economic theory, is that someone from the Millennials generation not holding any diploma is better off by about 20% than someone from Gen X with a high school diploma, according to the unweighted estimation. It is not actually expected that a less educated individual earns a higher wage per hour.

Also, Millennials are significantly better off than Gen X individuals in what respect to the periods analysed. For both estimations, the recession year covered by the survey (2008) shows a higher pay for Millennials in about 12%, when compared to Gen X individuals from the pre-recession period. This difference slows down for the 2011 wave to 7%, also for both. Through the recovery years, according to unweighted estimation, Millennials kept the positive difference registered right after the recession, yet slowing just a bit more to 5% according to the IPW estimation. Boomers, in contrast, earn about

[23] All results that consider coefficients of the interactions should be interpreted as the coefficient of the interaction plus the coefficient of the time-variant variable. In this case, considering the coefficient of the interaction between bachelor and black, for Model 1, one has to work out the following: 0.30 + (-0.06), the first being the coefficient of the interaction and the last being the coefficient for bachelor, according to Table IV.

2% more according to both estimations, when compared to the baseline group, in the recovery years.

By any means, the coefficients of the periods' variables show somewhat expected directions. This is, compared to the pre-recession period, the hourly wages for the recessionary year may show some rigidity and is significantly above by about 5%. This pattern seems to be reversed in the early-recovery period, but with just a small impact and not significantly different from zero. Nevertheless, for the unweighted estimation, the recovery years follow the decreasing path and even puts the wages down by about 4%, compared to the pre-recession year analysed, almost throwing away the gain coming from the rigidity seen in the recession period.

In more general terms, both estimations reflect the importance to gain years of labor experience and be associated to a labor union. Considering having an additional year of work, an employee living in the U.S. can expect to have their hourly wage raised by about 15%, when compared to a non-unionized worker. In contrast, as per the unweighted estimation, living in the Southern region of the U.S. and working in one of the CM industries would decrease an employee's hourly wage by about 1% (5% - 4%); something that cannot be interpreted using the same coefficients from the IPW estimation, since these are not significantly different from zero.

As a matter of fact, blacks can be seen as the most negatively affected workers when assessing the years after 2012 (recovery coefficients). For both estimations, compared to white Americans of the pre-recession period, blacks working in the U.S. have been earning, on average, about 9% less on hourly wages. This was indeed expected, as the economic literature presented in chapter 2 showed that this ethnic group is constantly "losing" dollars after a big economic downturn.

Finally, when the gender variable is put in the equation, according solely to the unweighted estimation, single women, after 2012, and with no descendants have been earning, on average, about 97% of a man's hourly wage from the pre-recession period. Something that could worsen by giving a birth, but that could be reverted with a marriage. Here, again, this interpretation cannot be made by analysing the same coefficients of the IPW estimation, since these are statistically insignificant.

5. CONCLUSIONS

This work investigated, amongst many aspects, the effects that the economic recession triggered by the financial crisis that occurred in the end of 2007 imposed on the hourly wages of the American workers. By taking a well-known household longitudinal survey, the PSID, as an object of investigation, a few conclusions could be extracted, both in terms of the methods that should be implemented to better analyse such a dataset, and the actual impacts for the population of interest.

Firstly, the PSID is indeed a very complex dataset that covers a tantamount of aspects of a family unit. For example, for the last wave released, in 2019, there are more than 5,000 variables just in the family file (PSID, 2021, p. 27). Besides, as the goal of the PSID is to follow the same families, or sampled individuals, through the years, a lot can be stated according to this longitudinal information about each individual's path. However, as an expected drawback coming from a long held micro panel, attrition might happen between waves. This can complicate the data analyses and disturb the consistency of the econometric methods.

Also, the methods that were brought up to test whether attrition is correlated to the error term did not agree to each other in full, which turns out the assumption of self-selection bias a little blurred and not much conclusive. By any means, some features of the individuals were tested to devise a better idea of what can predict the likelihood of an individual to attrite after any wave, following the works of Fitzgerald et al. (1999) and Zabel (1998).

In what concerns the answer to the main question that is the driver of this research, the effects of the Great Recession on the hourly wages showed, in general, that the nominal wage's rigidity may have played a crucial role during the recessionary period in favour of the American workers, with those even earning more than the year before the crisis. However, this effect disappeared through the recovery years, lowering even more after the year right after the end of the recession (2010). Blacks and women can be seen as the biggest losers – as expected, but the gender was only a statistically significant factor for the unweighted estimation –, however education for the former and marriage for the last can buffer the negative impacts.

Surprisingly, the Millennials generation has been earning more dollars than the other two analysed here. This is not an obvious result, as this is the cohort with the youngest workers. The economic literature presented in chapter 2 shows explicitly that youngsters are generally the most negatively affected in times of recessions, when compared to more senior co-workers. A possible explanation in terms of the consistency of the results here is that the main negative impact described and mostly analysed is unemployment, a labor outcome of no direct interest for this work. Further, as explain Blair and Deming (2020), the labor market since 2007 has been attracting more skilled professionals, which means that more educated people or, in recent years, with a somewhat greater ease to use informatic systems are better candidates. Even though young people are generally less skilled, in tacit terms, they are more prone to use new software and work with computers, a positive characteristic then.

Furthermore, the results in Table IV follow in some respects the economic literature. The coefficients for experience have both a positive and a correction, negative but lower than the first, impact – something generally expected when building a labor income function. Also, the minorities' groups – blacks and women, in this case – were the most negatively impacted, as already mentioned. However, the education variable, with its four levels, resulted in somewhat inconsistent estimated coefficients since the directions of the higher gradients (bachelors and graduates) are exactly the opposite from the expected. Perhaps the original values retrieved from the PSID are not very well documented in this regard – what possibly makes the estimation suffering from a measurement error bias –, or this could be due to some form of correlation between the variable and the idiosyncratic errors. Both hypotheses are likely and should be investigated in future research.

Finally, the presence of attrition in the PSID is a well-known issue and any econometric analysis using data from it must take into account such behaviour. As was denoted here, attritors might have different attitudes and labor outcomes. Nonetheless, when correcting for the possible bias due to self-selection coming from the effect of attrition on the estimates, the IPW model's estimated coefficients were not much different from an unweighted estimation on the same unbalanced panel. In effect, the greater differences are related to the gender variable introduced, which was actually detected as a differentiator factor between attritors and non-attritors. This can perhaps be a signal of improvement on the estimation when using the IPW approach.

REFERENCES

Acemoglu, D. (2002). Technical change, inequality, and the labor market. *Journal of Economic Literature*, *40*(1), pp. 7–72.

Atherwood, S., and Sparks, C. S. (2019). Early-career trajectories of young workers in the U.S. in the context of the 2008–09 recession: The effect of labor market entry timing. *PLOS ONE*, *14*(3), pp. 1-30.

BEA: Data Tools. (2021). Bureau of Economic Analysis. Retrieved September 17, 2021, from https://apps.bea.gov.

Bell, D. N. F., and Blanchflower, D. G. (2011). Young people and the Great Recession. *Oxford Review of Economic Policy*, *27*(2), pp. 241–267.

Beresford Research. (2020). *Age Range by Generation*. Beresford Research. Retrieved from https://www.beresfordresearch.com/age-range-by-generation/.

Blair, P. Q., and Deming, D. J. (2020). Structural Increases in Demand for Skill after the Great Recession. *AEA Papers and Proceedings*, *110*, pp. 362–365.

Blau, F. D., and Kahn, L. M. (2017). The Gender Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature*, *55*(3), pp. 789–865.

BLS. (2021). *CPI Inflation Calculator*. U.S. Bureau of Labor Statistics. Retrieved from https://www.bls.gov/data/inflation_calculator.htm.

BLS Data. (2021). U.S. Bureau of Labor Statistics. Retrieved September 17, 2021, from https://data.bls.gov/timeseries/LNS14000000.

Brown, J. N., and Light, A. (1992). Interpreting Panel Data on Job Tenure. *Journal of Labor Economics*, *10*(3), pp. 219-257.

Croissant, Y., and Millo, G. (2008). "Panel Data Econometrics in R: The plm Package." *Journal of Statistical Software*, *27*(2), pp. 1-43.

Croissant, Y., and Millo, G. (2018). *Panel Data Econometrics with R* (1st ed.) [E-book]. Wiley.

Dickens, W. T., and Triest, R. K. (2012). Potential Effects of the Great Recession on the U.S. Labor Market. *The B.E. Journal of Macroeconomics*, *12*(3), pp. 1-40.

Elsby, M. W. L., Shin, D., and Solon, G. (2016). Wage Adjustment in the Great Recession and Other Downturns: Evidence from the United States and Great Britain. *Journal of Labor Economics*, *34*(S1), pp. S249–S291.

Erken, H., Grabska, K., van Kempen, M. (2015). Labor Market Adjustments During the Great Recession: An International Comparison. CPB Background Document, pp. 1-39. From: https://www.cpb.nl/en/publications.

Fitzgerald, J., Moffit, R., and Gottschalk, P. (1999). Sample Attrition in Panel Data: The Role of Selection on Observables. *Annales d'Économie et de Statistique*, *55/56*, pp. 129-152.

Fournier, J. M., and Koske, I. (2012). The determinants of earnings inequality. *OECD Journal: Economic Studies*, *2012*(1), pp. 7–36.

Hoonhout, P. and Ridder, G. (2019). Nonignorable attrition in multi-period panels with refreshment samples. *Journal of Business & Economics Statistics*. *37*(3), pp. 377-390.

Hoynes, H., Miller, D. L., and Schaller, J. (2012). Who Suffers During Recessions? *Journal of Economic Perspectives*, *26*(3), pp. 27–48.

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, *6*(60), pp. 1-5.

NBER. (2021). US Business Cycle Expansions and Contractions. Retrieved September 17, 2021, from https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions.

Oswald, F. (2021). psidR: Build Panel Data Sets from PSID Raw Data. R package version 2.1. https://github.com/floswald/psidR.

Panel Study of Income Dynamics, public use dataset. (2021). Produced and distributed by the Institute for Social Research, University of Michigan, Ann Arbor, MI.

PSID. (2021). *PSID Main Interview User Manual: Release 2021*. Institute for Social Research, University of Michigan.

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rothstein, J. (2017). The Great Recession and Its Aftermath: What Role for Structural Changes? *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *3*(3), pp. 22-49.

Schwandt, H., and von Wachter, T. (2019). Unlucky Cohorts: Estimating the Long-Term Effects of Entering the Labor Market in a Recession in Large Cross-Sectional Data Sets. *Journal of Labor Economics*, *37*(S1), S161–S198.

Sironi, M. (2017). Economic Conditions of Young Adults Before and After the Great Recession. *Journal of Family and Economic Issues*, *39*(1), pp. 103–116.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.) [E-book]. The MIT Press.

Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach* (5th ed.) [E-book]. Cengage Learning.

Zabel, J. E. (1998). An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior. *The Journal of Human Resources*, *33*(2), pp. 479-506.

APPENDIX

TABLE V

ESTIMATED COEFFICIENTS FOR THE YEARLY PROBIT MODELS

	Selection (attrition)					
	2009	2011	2013	2015	2017	2019
Intercept	0.47	-0.43	0.92 **	1.42 ***	1.16 **	1.17 **
	(0.38)	(0.15)	(0.39)	(0.43)	(0.48)	(0.49)
Log(Wage)	0.06 *	0.15 ***	0.02	-0.02	0.03	0.03
	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)
College	0.20 ***	0.12 *	0.22 ***	0.07	-0.06	0.04
	(0.06)	(0.07)	(0.07)	(0.07)	(0.09)	(0.09)
Union	-0.01	0.18 *	0.15	0.15	-0.02	0.05
	(0.09)	(0.10)	(0.10)	(0.10)	(0.11)	(0.12)
Female	0.08	0.19 ***	-0.03	0.09	0.07	0.10
	(0.06)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)
Married	0.25 ***	0.11 *	0.13 *	-0.13 *	0.04	-0.15
	(0.07)	(0.07)	(0.07)	(0.08)	(0.09)	(0.10)
Person interview	-0.71 ***	-0.92 ***	-0.68 ***	-0.53 **	0.03	-0.24
	(0.19)	(0.18)	(0.26)	(0.23)	(0.36)	(0.28)
CM	-0.19 ***	-0.02	-0.05	-0.11	-0.11	-0.09
	(0.07)	(0.08)	(0.08)	(0.08)	(0.09)	(0.10)
Obs.	3,877	3,565	3,252	2,949	2,620	2.425
LR test	67.51 ***	69.31 ***	26.55 ***	18.61 **	3.15	7.20
	(df = 7)	(df = 7)	(df = 7)	(df = 7)	(df = 7)	(df = 7)
PCP[(a)]	85.51%	84.40%	83.25%	80.31%	86.24%	86.93%

Source: Panel Study of Income Dynamics, public use dataset. (2021).

(a) This is the percent correctly predicted specification test. The function applied in R was the default "performance_pcp", available on the "performance" package (Lüdecke et al., 2021).

Note 1: * p-value < 0.1; ** p-value < 0.05; *** p-value < 0.01.

Note 2: Standard deviation in parenthesis.

TABLE VI

PSID VARIABLES' CODES

Purpose	Code (wave)[(a)]
Working out new experience and education variables	V11828 (1985); V11829 (1985); V12191 (1985); V12192 (1985); V14169 (1987); V14179 (1987); V14186 (1987); V14247 (1987); V14262 (1987); V14270 (1987); V14277 (1987); V14342 (1987); V14352 (1987); V14359 (1987); V14411 (1987); V14426 (1987); V14434 (1987); V14441 (1987); V22575 (1993); V22928 (1993); V22741 (1993); V23094 (1993); ER7657 (1996); ER7163 (1996); ER10081 (1997); ER10563 (1997); ER12170 (1997); ER12181 (1997); ER17409 (2001); ER17422 (2001); ER17424 (2001); ER17436 (2001); ER17449 (2001); ER17451 (2001); ER17461 (2001); ER17474 (2001); ER17476 (2001); ER17487 (2001); ER17500 (2001); ER17502 (2001); ER17258 (2001); ER17313 (2001); ER17318 (2001); ER17320 (2001); ER17691 (2001); ER17995 (2001); ER17704 (2001); ER17992 (2001); ER18275 (2001); ER17730 (2001); ER18019 (2001); ER18301 (2001); ER17756 (2001); ER18045 (2001); ER18327 (2001); ER17782 (2001); ER18071 (2001); ER18353 (2001); ER17600 (2001); ER17888 (2001); ER17979 (2001); ER18262 (2001); ER17717 (2001); ER18006 (2001); ER18288 (2001); ER17743 (2001); ER18032 (2001); ER18314 (2001); ER17769 (2001); ER18058 (2001); ER18340 (2001); ER17540 (2001); ER17828 (2001); ER18111 (2001); ER17595 (2001); ER17883 (2001); ER18166 (2001); ER17706 (2001); ER17994 (2001); ER18277 (2001); ER17732 (2001); ER18021 (2001); ER18303 (2001); ER17758 (2001); ER18047 (2001); ER18329 (2001); ER17784 (2001); ER18073 (2001); ER18355 (2001); ER17602 (2001); ER17890 (2001); ER12174 (1997); ER12185 (1997); ER77119 (2019); ER77167 (2019); ER77118 (2019); ER77166 (2019); ER18171 (2001); ER72180 (2019); ER72457 (2019); ER72244 (2019); ER72521 (2019); ER72274 (2019); ER72551 (2019); ER72304 (2019); ER72581 (2019); ER72182 (2019); ER72459 (2019); ER72246 (2019); ER72523 (2019); ER72276 (2019); ER72553 (2019); ER72306 (2019); ER72583 (2019); ER18173 (2001); ER11897 (1997); ER11809 (1997); ER11898 (1997); ER11810 (1997).
Variables for the objective function	ER32049 (2019); ER32000 (2019); ER77448 (2019); ER77315 (2019); ER77343 (2019); ER77255 (2019); ER77276 (2019); ER77249 (2019); ER77266 (2019); ER77268 (2019); ER77270 (2019); ER77287 (2019); ER77289 (2019); ER17227 (2001); ER17797 (2001); ER60195 (2015); ER60458 (2015); ER72196 (2019); ER72473 (2019); ER72208 (2019); ER72485 (2019); ER72048 (2019); ER72051 (2019); ER76897 (2019); ER76752 (2019); ER72030 (2019); ER72162 (2019); ER72006 (2019); ER72015 (2019); ER72021 (2019); ER77591 (2019).

Source: Panel Study of Income Dynamics, public use dataset. (2021).

(a) Variables sometimes only exist for a specific range of waves, being superseded by others.

TABLE VII

DESCRIPTIVE STATISTICS OF THE INTEGER AND CONTINUOUS VARIABLES

EXPERIENCE	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
MINIMUM	0.00	2.00	4.00	5.00	7.00	9.00	11.00
1ST QUARTILE	11.00	13.00	15.00	17.00	19.00	21.00	23.00
MEDIAN	19.00	21.00	23.00	24.00	26.00	28.00	30.00
MEAN	19.88	21.77	23.62	25.6	27.48	29.41	31.38
3RD QUARTILE	28.00	30.00	32.00	34.00	36.00	38.00	40.00
MAXIMUM	72.00	74.00	76.00	78.00	80.00	82.00	84.00
HOURLY INCOME	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
MINIMUM	0.35	0.69	0.27	0.22	2.16	0.32	0.29
1ST QUARTILE	12.76	14.77	15	16.17	17.09	17.96	18.83
MEDIAN	18.62	21.75	22.43	24.06	25.49	26.72	28.85
MEAN	25.82	29.03	29.63	32.13	33.03	35.29	37.64
3RD QUARTILE	28.12	32.64	34.24	36.71	39.22	42.00	43.88
MAXIMUM	4,444.44	2,184.49	1,000.00	1,276.60	765.96	1,027.13	888.89
(NO.) CHILDREN	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
MINIMUM	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1ST QUARTILE	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MEDIAN	1.00	1.00	1.00	1.00	0.00	0.00	0.00
MEAN	0.93	0.97	0.98	0.99	0.94	0.9068	0.85
3RD QUARTILE	2.00	2.00	2.00	2.00	2.00	2.00	2.00
MAXIMUM	7.00	9.00	11.00	9.00	8.00	7.00	7.00

TABLE VIII

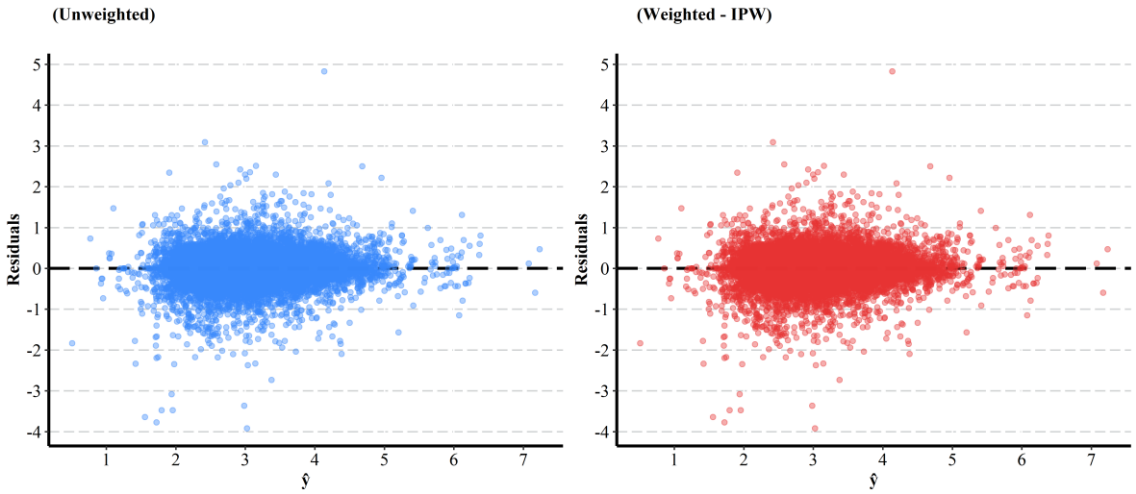
YEARLY COUNTS OF THE BINARY VARIABLES

FEMALE	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
0	2,142	1,957	1,769	1,609	1,411	1,299	1,198
1	1,735	1,608	1,483	1,340	1,209	1,126	1,056
BLACK	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
0	3,620	3,334	3,038	2,756	2,435	2,251	2,092
1	257	231	214	193	185	174	162
UNION	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
0	3,424	3,142	2,842	2,583	2,298	2,130	1,992
1	453	423	410	366	322	295	262
SOUTH	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
0	2,637	2,426	2,230	2,008	1,769	1,622	1,508
1	1,240	1,139	1,022	941	851	803	746
MARRIED	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
0	1,014	897	798	707	642	592	559
1	2,863	2,668	2,454	2,242	1,978	1,833	1,695
FIRE	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
0	3,571	3,294	3,013	2,719	2,408	2,230	2,064
1	306	271	239	230	212	195	190
CM	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
0	3,067	2,855	2,598	2,372	2,095	1,963	1,841
1	810	710	654	577	525	462	413

TABLE IX

YEARLY COUNTS OF THE CATEGORICAL VARIABLES

EDUCATION	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
LESS THAN HIGH SCHOOL	152	74	63	53	48	41	37
HIGH SCHOOL	1,192	933	692	699	606	562	517
BACHELOR	1,932	1,840	1,673	1,513	1,337	1,219	1,127
GRADUATION	601	718	692	684	629	603	573
GENERATION	*2007*	*2009*	*2011*	*2013*	*2015*	*2017*	*2019*
BOOMERS	1,471	1,334	1,209	1,108	983	910	847
GEN X	1,823	1,696	1,560	1,408	1,248	1,146	1,062
MILLENNIALS	583	535	483	433	389	369	345



Source: Panel Study of Income Dynamics, public use dataset. (2021).

FIGURE 6 – Residuals plots for the estimations with and without IPW weights.