Kristian McInroy Bergene

The Gender Pay Gap in Europe

An empirical study of the differences in earnings between genders

Graduate thesis in Economics
Supervisor: Hildegunn E. Stokke
Trondheim, June 2016

Norwegian University of Science and Technology
Faculty of Social Sciences and Technology Management
Department of Economics

NTNU
Norwegian University of
Science and Technology

Preface

This thesis marks the end of my time at NTNU and is the culmination of many rewarding hours spent studying the field of economics. My days here in Trondheim have been an absolute joy. I am grateful for the excellent guidance from Hildegunn E. Stokke and the numerous hours my dad has spent proofreading, as well as the boys at 6b for their LaTeX knowledge. I would also like to extend my thanks to all the legendary guys who have stopped by Pareto FK for all the great moments on (and off) the pitch!

One coincidence while concluding this work was the news of Muirfield maintaining its ban on women members. My grandfather was for many years a member of the golf club, and I certainly find it slightly ironic that I am now finishing up a thesis on the topic of male/female discrimination as this news emerged from Scotland given my (though distant) association with the club!

Kristian McInroy Bergene
Trondheim, June 1, 2016.

Abstract

This thesis uses panel data for several European countries to examine factors that affect the gender pay gap, and to examine how the gender pay gap evolves with age. The analysis finds a strong correlation between age and the gender pay gap. The differences in pay between men and woman is found to be small for workers in their 20's, before a steep increase in the male wage premium is observed for workers in their 30's and a subsequent stabilisation.

Further analysis find the percentage of the population in the age group 25-29 to have a strong negative effect on the overall gender pay gap across Europe. This result is robust to different model specifications. I also find indications that the pay gap is positively correlated with the GDP per capita. Additionally, the effect of female educational attainment is found to have a negative effect on the pay gap.

Contents

List of Tables

1 Introduction

The gender pay gap is an important topic on the European political agenda. Article 157 in the Lisbon Treaties states:

Each Member State shall ensure that the principle of equal pay for male and female workers for equal work or work of equal value is applied.

The reduction of the gender pay gap has been an objective of the European employment strategy since 1999, with increasing policy efforts since then. Despite this long-standing legislation on equal pay, women in Europe earn less than men.

The differences in earnings between men and women is a complex issue with origins that go beyond the legal framework. Vladimir Spidla[1] states that *women and men are legally equal, but they are not economically equal*[2].

Narrowing the gender pay gap is not just a morally important issue of equality. A report from the McKinsey Global Institute in 2015 concludes that greater equality of pay, hours worked and access to full time jobs between genders in the workforce could increase the global GDP by between $12 and $28 trillion by 2025[3].

This thesis attempts to point to some underlying factors that contribute to the differences in earnings between men and women in Europe. By including different explanatory variables I hope to uncover trends and point to factors that contribute to the gender pay gap. I will also take a closer look at how the gender pay gap evolves with age. In summary, the main focus of this thesis can be expressed as:

> *What factors contribute to the gender pay gap across Europe, and how do the differences in earnings between men and women develop during working life?*

1.1 General information

The gender pay gap is a phrase used repeatedly in this thesis, as well as variations such as the wage gap, the pay gap, the earnings gap and the abbreviation GPG. All these phrases relate to the differences in pay between men and women, given by the male wage

[1] Member of the European Commission responsible for employment, social affairs and equal opportunities

[2] This quote is found in the foreword of Emerek et al. (2006).

[3] The full report can be found at http://www.mckinsey.com/global-themes/employment-and-growth/how-advancing-womens-equality-can-add-12-trillion-to-global-growth

premium. That is, a gender pay gap of fifteen percent indicates that men earn on average fifteen percent more than women.

1.2 Structure

The structure of the thesis is as follows:

Chapter 2 gives an overview of previous research and findings on the subject of the pay gap.

Chapter 3 provides descriptive statistics on the gender pay gap in Europe, as well as information regarding the data used in the empirical analysis.

Chapter 4 presents the econometric framework used in the empirical analysis and considers the potential issues that may arise.

Chapter 5 gives the results of the empirical analysis, and closing remarks are offered in *chapter 6*.

2 Existing Literature and Previous Research

The gender pay gap is extensively researched in economic literature and the issue is looked at from several different angles. In this chapter I present an overview of some of the existing literature focusing on the gender pay gap over time, causes of the pay gap and how the differences in earnings between men and women evolve through working life.

2.1 The gender pay gap in a historical perspective

The level of male and female earnings has converged significantly when we consider the situation today compared to the immediate post World War II era. Blau & Kahn (2000) provide a good summary of the development of the GPG in the United States from the mid 50's to the early 00's. In the mid 20th century, female participation in the work force was mostly concentrated in low-paying jobs. Blau & Kahn (2000) point to a change in the late 70's, with more women "branching out" to what had previously been considered predominantly male professions[4]. This led to a considerable convergence, with the female earnings ratio[5] jumping from about 60% to about 75% to that of male earnings by the early 90's. In Blau and Kahn's studies, the convergence decelerated in the 1990's and the differences in earnings have been relatively stable since. In Europe historical observations are similar, as reported by Emerek et al. (2006).

2.2 Explanations for the gender pay gap

When discussing the reasons for the differences in earnings between men and women, economists usually distinguish between the explained and the unexplained pay gap. The explained GPG is the part of the gap that can be explained by observable differences between working men and women. The unexplained part is what remains of the gap after the observable differences have been controlled for[6]. The unexplained GPG may be considered to be discrimination between men and women.

Altonji & Blank (1999), Blau & Kahn (2000) and Blau & Kahn (2016) provide a thorough overview of recent research and the factors that are found to contribute to the gender pay gap.

[4] A different study by Flyer & Rosen (1994) found that almost half of female college graduates in 1960 went on to become teachers, compared to less than 10% in 1990.

[5] Female earnings ratio is female earnings expressed as a share of male earnings.

[6] The unexplained part of the GPG is also referred to as the adjusted gender pay gap.

In human capital theory, differences in pay are traditionally explained by characteristics such as age, education and experience. In recent years economically advanced countries (and many developing countries) observe higher education levels for women than for men[7] and so it is difficult to point to educational differences when explaining the pay gap. Blau & Kahn (2007) report that educational attainment level should reduce the GPG by 6.7% in their sample, meaning that the level of educational attainment actually contributes to the unexplained part of the pay gap, rather than the explained part.

Work experience is generally considered a factor with strong explanatory power on the gender pay gap. Mincer & Polacheck (1974) highlight the importance of the traditional division of gender roles within families to explain why women accumulate less work experience than men. Becker (1985) builds on this and discusses whether longer hours of home-related work for women lead to decreased efforts in their working career, reducing productivity and thus also the amounts they earn. Blau & Kahn (2007) find that experience in the labour force explains 10.5% of the pay gap in their sample.

Relating to the effect family life has on female wages is the effect of childbirth. In the past, starting a family would often lead to the mother leaving employment for a significant time period or even leaving the workforce alltogether. In recent times this is not observed as much, and the effect of childbirth on wages is also considered to be tied to policies around parental leave. Indeed, a study by Waldfogel (1998) finds that the negative effect childbirth has on women's wages is significantly reduced amongst mothers who have job-protected maternity leave.

In addition to work experience, the gender differences in some occupations and industries form a big part of the explained wage difference. Although occupational segregation has steadily decreased since the 1970's some differences between men and women still remain. Blue-collar jobs are still dominated by men, and men continue to be overrepresented in managerial positions. Blau & Kahn (2007) find that occupational and industry categories explain 27.4% and 21.9% of the GPG respectively. Women are also often concentrated in lower-paying industries and firms in occupations where both men and women are equally represented (Blau & Kahn (1997), Groshen (1991), Bayard et al. (1999)). The gender balance by occupation and industry certainly explains part of the pay gap between men and women, but at the same time it raises the question of whether these unequal distributions are caused by gender differences in qualifications or whether women face labour market discrimination.

[7] See section 3.3 in Blau & Kahn (2016).

After controlling for worker characteristics, economists find that the gender pay gap remains. Blau & Kahn (2007) find that 41% of the wage gap cannot be explained when controlling for observed differences, which gives a female wage ratio of 91% compared to men in their sample. Additionally, Bayard et al. (1999) find almost half of the gender pay gap to be unexplained in their matched employer-employee dataset[8] covering virtually all industries and occupations across the USA.

Another way to analyse homogeneous workers is offered by Wood et al. (1993) who focus on graduates from University of Michigan Law School between 1972 and 1975, and find a male wage premium of 13% fifteen years after graduation when controlling for observable characteristics.

Finally, Albrecht et al. (2004) use data from the Netherlands and show that the pay gap is due to differences in returns to labour market characteristics rather than the differences in the characteristics themselves.

2.3 The gender pay gap during working life

A separate part of the literature on the differences in earnings between men and women focuses on how the GPG evolves during working life. Stokke (2016) applies matched employer-employee register data for Norway and identifies a non-existant pay gap on entry to the labour market for equal workers, but also observes a rapid increase over the first 10-15 years of work and a subsequent stabilisation of the gap. When observable factors are controlled for, this indicates lower returns to worker characteristics for women with lower returns to work experience as the dominant factor. Stokke (2016) also notes that an increased level of education decreases the gender discrimination in the labour market.

The difference in wage development between men and women is often referred to as the early-career effect, and is found in other studies including Manning & Swaffield (2008), who uses the British Household Panel Suvery (BHPS) and finds a sizeable gender pay gap 10 years after entry to the labour market. Bertrand et al. (2009) focus on MBA graduates between 1990 and 2006 from a top US business school, and find that even though labour incomes are nearly identical for men and women from the outset of their working careers, the male earnings advantage reaches almost 60 log points at ten to 16

[8] Matched employer-employee datasets are used to analyse workers who work in the same firm and/or same profession to provide comparability between individuals.

years after graduating.

Additionally, Blau & Kahn (2000) give an overview the wage gap within different age cohorts in 1978, 1988 and 1998, again showing evidence of the early-career effect, even with reduction of the gender pay gap within each cohort over time.

In contrast to these findings, Kunze (2005) analyses West-German workers with apprenticeship training and finds a large wage gap on entry to the labour market. In her study, the gap is stable during the early years of the career.

3 Data Presentation and Descriptive Statistics

Introduction

In this chapter I present the data used in this thesis. The purpose of the data is to present statistics on the gender pay gap across Europe and examine the factors that contribute to the gender pay gap. To do this, I present some descriptive statistics on the subject, then explain in more detail how the data for my empirical analysis is built up.

All the data is collected from Eurostat[9]. Eurostat's task is "to provide the European Union with statistics at European level that enable comparisons between countries and regions"[10]. It should be noted that the data is gathered from different databases in Eurostat's archives. I will provide a more detailed explanation during this chapter.

The data used in the empirical analysis is divided into two different sets of data. The shorter dataset is used to analyse the developments of the GPG over working life, and the longer dataset is used in the main analysis.

Additionally, descriptive statistics are based on data from the period 2007-2013. Data from this period includes values on the gender pay gap disaggregated by sector, industry, contract type and age group.

3.1 Descriptive statistics

In this section I present some descriptive statistics to give a general overview of the gender pay gap across Europe, as well as across different sectors and age groups. The descriptive statistics create a base for what I show later in my empirical analysis.

3.1.1 Data information

The data described here is exclusively collected from Eurostat's database on the GPG within it's statistics on the labour market across Europe[11]. The data includes observations on the pay gap in 30 European countries over the time period 2007-2013, with observations based on hourly wages and includes both full-time and part-time employees.

[9] Eurostat is the statistical office for the European Union.

[10] http://ec.europa.eu/eurostat/about/overview gives a further description of Eurostat's role and responsibility within the European Union.

[11] http://ec.europa.eu/eurostat/cache/metadata/en/earn_grgpg2_esms.htm provides a detailed description of data collection and calculation methodology.

The methodology used in this database is based on the calculations of the four-yearly Structure of Earnings Survey (SES) carried out by Eurostat in 2002, 2006 and 2010. The reported values in the years between the SES are based on national sources with the same coverage as the SES. Eurostat reports that the common definitions and concepts are agreed by the countries included in the database, creating a solid base for examining the GPG over time.

The values presented on the gender pay gap are referred to as the unadjusted gender pay gap. Eurostat gives the following explanation: *As an unadjusted indicator, the GPG gives an overall picture of the differences between men and women in terms of pay and measures a concept which is broader than the concept of equal pay for equal work. A part of the earnings difference can be explained by individual characteristics of employed men and women and by sectoral and occupational gender segregations*[12]*.*

The unadjusted gender pay gap is determined by direct observation in the countries without taking into account factors that are considered correlated with wage differences between men and women.

As referenced in chapter 2, many studies look at the adjusted pay gap to examine how much of the gap can be explained by gender-specific factors. However data on the adjusted pay gap between men and women are not available for the nations used in this study, and are not presented in the descriptive statistics in this chapter.

Eurostat data does not cover all sectors in each country. Only enterprises with 10 or more employees are included in the data, and self-employed workers are excluded. The data only covers what Eurostat calls economic sections B to S, excluding O[13]. Agricultural workers are also not included. These exclusions are not necessarily a problem, as wage-setting in these sectors are often special cases[14].

[12] See section 3.4 at http://ec.europa.eu/eurostat/cache/metadata/en/earn_grgpg2_esms.htm

[13] A complete overview of the different sections can be found in the appendix.

[14] For example, it's generally hard to determine an hourly wage for agricultural workers, as well as the self-employed.

The values presented are given as the male wage premium. That is, a gender pay gap of 15% indicates that men earn on average 15% more than women. The observed values can be expressed by the following relationship

$$GPG = \frac{W_M - W_F}{W_M}\% \tag{1}$$

where

GPG is the Gender Pay Gap

W_M is the Gross Hourly Male Average Earnings

W_F is the Gross Hourly Female Average Earnings

3.1.2 The gender pay gap across Europe

Table 1: The Gender Pay Gap Across Europe

Country	Mean GPG	Std. Dev.	Min.	Max.	Obs
Austria	24.143	0.900	23	25.5	7
Czech Republic	23.457	1.876	21.6	26.2	7
Germany	22.386	0.418	21.6	22.8	7
Slovakia	21.114	1.378	19.6	23.6	7
United Kingdom	20.171	0.808	19.1	21.4	7
Finland	19.929	0.730	18.7	20.8	7
Iceland	19.700	2.235	17.7	24.0	7
Greece	19.500	3.905	15	22	3
Switzerland	18.500	0.632	17.8	19.3	6
Netherlands	17.900	1.150	16	19.3	7
Hungary	17.857	1.193	16.3	20.1	7
Cyprus	17.786	2.238	15.8	22	7
Spain	17.657	1.361	16.1	19.3	7
Denmark	16.714	0.587	15.9	17.7	7
Sweden	16.100	0.924	15.2	17.8	7
Norway	16.029	0.610	15.1	17.0	7
Lithuania	15.986	4.340	11.9	22.6	7
France	15.871	0.867	15.1	17.3	7
Ireland	13.750	1.995	11.7	17.3	6
Latvia	13.686	1.135	11.8	15.5	7
Bulgaria	13.129	0.858	12.1	14.7	7
Portugal	11.586	2.346	8.5	14.8	7
Belgium	10.086	0.146	9.8	10.2	7
Romania	9.571	1.700	7.4	12.5	7
Luxembourg	9.100	0.632	8.6	10.2	7
Poland	8.157	3.711	4.5	14.9	7
Malta	7.100	1.319	5.1	9.2	7
Italy	5.800	0.885	4.9	7.3	7
Croatia	4.850	2.092	2.9	7.4	4
Slovenia	2.443	1.976	-0.9	5.0	7
Total	**15.053**	**5.820**	**-0.9**	**26.2**	**201**

Table 1 gives a general overview of the gender pay gap across Europe (2007-2013). It shows significant differences between the countries. Some observations are missing for some countries, and some countries show quite a large change in the gender pay gap over the relatively short time period. The overall average GPG is close to 15%, with 40% of the countries lying within three percentage points and 60% within five percentage points of this value. Looking at the extreme values we find Austria, The Czech Republic, Germany, Slovakia and the United Kingdom as the only countries with an average above

20% for the time period, whereas Slovenia, Croatia and Italy each report a mean gender pay gap under 6%. Slovenia has the lowest of all these countries with just 2.44%, and even reported a negative gender pay gap of -0.9% in 2009, indicating an income inequality slightly in favour of females that year. These low GPG results seem surprising, but one reason could be low reported activity rates for women in these countries. In fact Italy reports an activity rate of only 55% for females in the age group 25-64 over the time period, and even less in the late 90's and early 00's. This will be addressed further later in this chapter and in 5.

We should note too that some standard deviations are quite large resulting from changes over time. Therefore we should be careful when drawing conclusions from these data without explaining these changes.

3.1.3 The gender pay gap in different age groups

The table below demonstrates how the gender pay gap changes with age.

Table 2: The Gender Pay Gap in Age Groups

Age Group	**Mean GPG**	**Std. Dev.**	**Min.**	**Max.**	**Obs**[15]
< 25	4.071	4.864	-10.6	19.2	168
$25 - 34$	7.814	5.445	-5.8	22.6	168
$35 - 44$	16.152	6.549	2.3	32.7	168
$45 - 54$	17.374	7.536	1.0	38.2	168
$55 - 64$	15.882	9.620	-9.8	38.2	168

Table 2 shows a general tendency for the gender pay gap to increase with age, then flatten out and even decrease slightly for the oldest age group. The gender pay gap doesn't vary much between the three oldest groups, but there are a few countries that report less of a gap for the oldest group (Croatia, Malta, Romania and Slovenia[16]). All of these countries have reported a negative gender pay gap at some point during the time period for this age group, with Slovenia reporting the largest negative value of -9.8 in 2009. Due to these fluctuations, the standard deviations are quite large.

[15] As we can see from the observations, we do not have values for all the countries. An overview of the missing observations is found in the appendix.

[16] The activity rate for females in the oldest age group is likely to be low in these countries.

This table also evidences the early-career effect discussed in chapter 2. The wage gap is small upon entry to the labour market, then increases over the next 20 years before stabilising.

3.1.4 The gender pay gap within different contract types

As a general pattern across Europe, there are substantially more women than men in part time employment. 32.2% of women aged 15-64 in employment in the European Union were on part-time contracts, compared to 8.8% of men in 2014[17]. In the table below I show the differences in the observed gender pay gap for the two contract types.

Table 3: The Gender Pay Gap Part Time vs. Full Time

Contract	Mean GPG	Std. Dev.	Min.	Max.	Obs.[18]
Part time	10.162	10.596	−16.6	35.6	146
Full time	13.188	5.935	−1.2	22.2	146

The average gender pay gap is higher for full time employed than part time employed. However we should note that there is a large standard deviation across the part time values where observations range from -16.6 to 35.6. We cannot conclude that the mean gender pay gap for part time workers is significantly different from zero.

The mean gender pay gap values reported here are both lower than the overall mean GPG reported earlier. This could be due to the fact that observations are missing from some of the countries where the gender pay gap is relatively large. Austria, for example, reported the highest mean GPG, but contributed to this table with only one observation in 2010.

If the part-time salary is lower than the full-time salary, and if females are overrepresented in part-time work, then the gender pay gap will increase when these two contract types are presented as an aggregate sum[19]. Also, if part-time workers are found more in the public sector (where earnings are generally lower than in the private) then the aggregate pay gap will increase when more women are employed here.

[17] Figures quoted from Eurostat's Statistics Explained section, and the article on employment statistics. The article is found at http://ec.europa.eu/eurostat/statistics-explained/index.php/Employment_statistics

[18] An overview of the missing observations are found in the appendix

[19] However this will likely not be observed as much when considering hourly wages.

3.1.5 The gender pay gap in the public and private sector

The majority of countries in Europe have a lower pay gap within the public sector than the private sector. This is shown in the table 4.

Table 4: The Gender Pay Gap Public vs. Private Sector

Economic Control	Mean GPG	Std. Dev.	Min.	Max.	Obs[20]
Public	12.401	6.861	−3.6	24.4	173
Private	17.792	5.512	2.8	27.8	173

The gender pay gap is on average about five percentage points lower in the public sector than in the private sector. However this is not observed in all countries. Bulgaria, Croatia and Romania report the opposite, with Latvia and Hungary reporting a change from a bigger gap in the private sector to a bigger gap in the public sector over the period. In the Netherlands, Finland and Sweden there is no significant difference between the public and private sectors.

There is also a tendency for more women to be employed in the public sector than in the private sector [21]. This may affect the gender pay gap, as earnings in the public sector are generally lower than in the private sector, especially later in working life.

[20] Overview of missing observations in appendix.

[21] (Dupuy et al. (2009) reports this in section 3.

3.1.6 The gender pay gap in different industries

To show which industries the gender pay gap is most significant in, I include table 5 which gives numbers on the gender pay gap in different sectors of the economy.

Table 5: The Gender Pay Gap Part in Different Industries

Sector	**Mean GPG**	**Std. Dev.**	**Min.**	**Max.**	**Obs.**
Finance	30.096	7.880	7.0	49	170
Retail	21.466	5.895	9.5	34.9	175
Health and Social Work	19.989	7.990	4.3	41.5	160
Science and Technology	19.664	8.941	−17	37.3	165
Information and Communication	19.950	6.257	5.3	35.5	171
Arts, Recreation and Entertainment	18.473	10.229	2.4	60.9	157
Real estate	16.963	9.850	−16.8	44.8	160
Accommodation and Food Service	13.831	5.176	0.4	29.7	165
Education	12.294	6.936	−4.7	36	173
Administrative and Support Service	9.415	10.652	−32.3	29	163

The financial sector stands out. With a mean gender pay gap of 30% it quite clearly "outperforms" other industries. There are some large standard deviations, mainly due to observed differences between countries.

3.2 Data for the empirical analysis

The dataset used in my empirical analysis includes more obervations than the data I have discussed so far. The data covers a time period stretching from 1997-2013, and includes observations from 24 different countries across Europe[22]. The variables included in the dataset are gathered from different databases in Eurostat. Because we follow the same countries over time, the data is set up as panel data with the gender pay gap as the dependent variable and several explanatory variables. In this section I go into further detail on each variable.

3.2.1 Dependent variable

The dependent variable is the variable we wish to explain, in this case the unadjusted gender pay gap. The observations from 2007 onwards are gathered from the same database as the figures used to present table 1 in the descriptive statistics. A different database in Eurostat is used to gather the data on pre-2007 observations. This database consists of values on the gender pay gap produced by national reports for each country. In general the coverage described earlier also applies here, but there are some exceptions. The methodology used for data collection change for some countries over the time period, and this causes some methodological breaks in the data series. An extensive overview of the deviations from the main methodology in the different countries is presented in the appendix and discussed further in chapter 4. The values presented here are aggregated values for the economy as a whole, and have not been disaggregated by sector and age groups, as was done for the 2007-2013 data.

To illustrate how the gender pay gap has developed over time, I have included figure 1 showing the differences in four countries (Estonia, Italy, Norway and Romania).

[22] A list of the included countries can be found in the appendix.

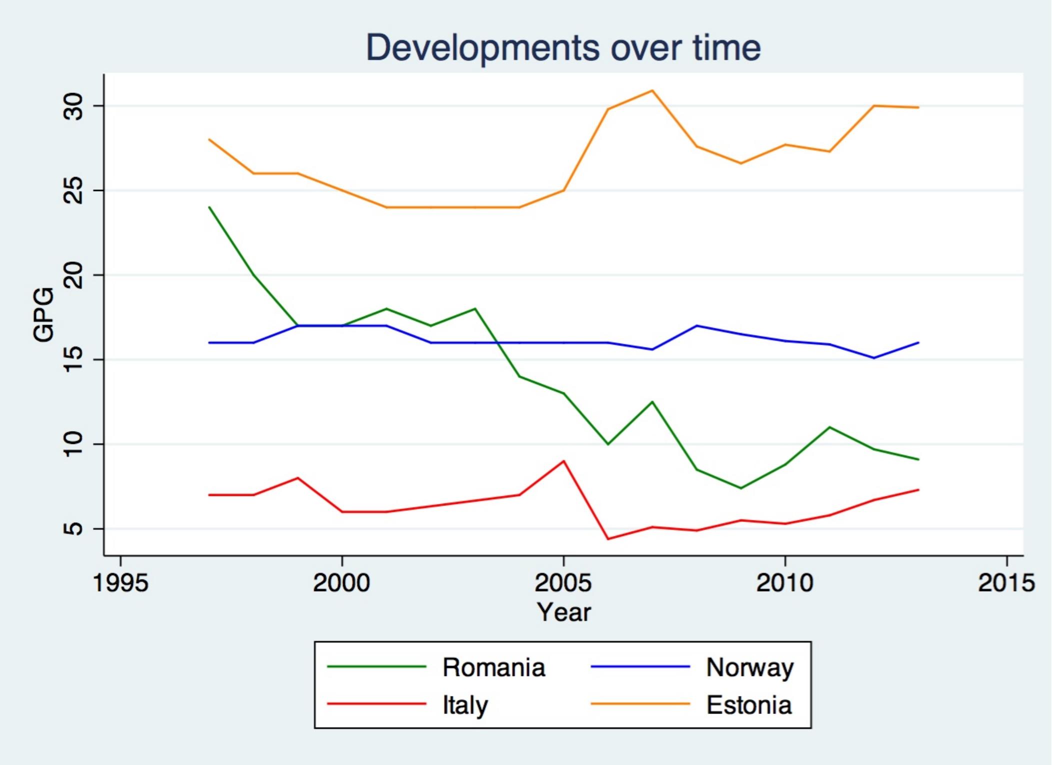

Figure 1: Developments of the Gender Pay Gap Over Time

These four countries were chosen to highlight some of the differences across the countries included in my sample. Norway and Italy are examples of countries that reported a stable gender pay gap over the time period, although their respective levels are quite different. Romania report a significant drop in the gender pay gap and Estonia have seen a slight increase over the time period.

From the graph we also see some shocks around 2006 and 2007. This could be a result of the changes in data collection methodology between the two databases used in the empirical analysis. This is something I address in chapter 4.

Table 6 illustrates the developments in the mean gender pay gap aggregated across all countries in the sample. The table shows the GPG at five year intervals in 1997, 2002, 2007 and 2012 respectively.

Table 6: The Gender Pay Gap Development

Year	Mean GPG
1997	17.913
2002	17.810
2007	17.263
2012	16.021

This shows a total change of about two percentage points over the period. There is some variation within these observations with an extreme low at 14.917% in 2005, and an extreme high at 17.92% in 1998. However the overall trend shows a slight decline in the unadjusted gender pay gap over the period.

3.2.2 Explanatory variables

In my main model and subsequent robustness checks I include a number of explanatory variables to examine the factors I expect to contribute to the gender pay gap. What follows is a description of each of these explanatory variables.

3.2.3 Rate of young individuals in population

I have included this variable to take the demographic of the population into account. From table 2 in the descriptive statistics and previous literature regarding the early-career effect, we would expect a lower gender pay gap when there is a higher proportion of people in the 25-29 age group.

This variable is calculated as individuals in the 25-29 age group as percentage of the whole population[23]. The values are gathered from Eurostats database on population statistics[24].

[23] The figures on individuals in the 25-29 age groups as percentage of the employed population was not available in Eurostat.

[24] http://ec.europa.eu/eurostat/cache/metadata/en/demo_pop_esms.htm gives a thorough description of the data compilation.

3.2.4 GDP per capita

To study whether there is a correlation between the monetary wealth of a country and the gender pay gap, I have included the GDP per capita. By including this, I hope to determine whether growth in the economy leads to an increase or a decrease in the gender pay gap[25].

The values are gathered from Eurostat's overview of national accounts[26]. The values from Eurostat are reported in Euro at current prices. I have transformed the values to constant prices using a GDP deflator provided by the world bank[27]. The base year is chosen as 2010.

To interpret the results of the GDP per capita, the variable has been transformed to a logarithmic variable. We need to take this into account when we interpret the results later.

3.2.5 Females with higher education

I have defined this variable as the percentage of females in the population who have completed tertiary education[28]. A higher level of education is generally thought of as one of the main factors associated with a higher income. By including this variable we can examine whether an increased number of females with higher education reduces the gender pay gap, as one might expect.

The values are calculated as annual averages of Eurostat's quarterly Labour Force Survey (EU-LFS) data[29].

3.2.6 Employment rate of older females

Because of the observations that the gender pay gap increases with age, I have included this variable to see if a higher rate of females in employment the age group 55-64 years

[25] It could be argued that GDP per capita is an insufficient parameter to explain the wealth of the inhabitants of the country as it does not take factors like income distribution into account. However, for the purpose of this thesis, it can be a useful tool.

[26] Further information on the data compilation in the overview of national accounts is given at http://ec.europa.eu/eurostat/cache/metadata/en/nama_esms.htm

[27] The GDP deflator is found at http://data.worldbank.org/indicator/NY.GDP.DEFL.ZS

[28] Eurostat defines tertiary education as level 5-8, which includes short-cycle tertiary education, bachelor's or equivalent level, master's or equivalent level, doctoral or equivalent level

[29] Further information regarding the data is found at
http://ec.europa.eu/eurostat/cache/metadata/en/edat1_esms.htm

old does in fact cause the gender pay gap to increase. These values are gathered from the EU-LFS and are defined as the percentage of the total female population in the age group 55-64 years who are employed.

3.2.7 Fertility rate

Relating to the discussion regarding childbirth causing discontinuous working lives resulting in less work experience, I include the variable $fertrate$. This gives the fertility rate in countri i at time t. The fertility rate is given as the mean number of children born alive to a woman during her lifetime. A positive relationship between the gender pay gap and fertility rates would be in line with the discussion mentioned above. However the effect of this variable may be different across countries. If having children means leaving work completely in some of the included countries, we may get different results.

The values are gathered from Eurostat's section on population in the database considering fertility[30].

3.2.8 Age of mother when first child is born

I include the variable *agebirth* to further examine the effect childbirth has on female wages. The variable gives us the average age of mothers in each country when their first child is born. It is not necessarily clear which effect this variable will give.

The values are gathered from Eurostat's section on population in the database considering fertility.

3.2.9 Activity rate

The next variable I include is *actrate*. The variable is defined as the percentage of females within the population in the 25-64 age group who are economically active[31]. The inclusion of this variable relates to the figures shown in table 1 and the countries with the lowest reported pay gaps. As mentioned, the activity rates for females in these countries may be quite low. This could mean that women in these countries choose not to work over working for a relatively low wage.

[30] http://ec.europa.eu/eurostat/cache/metadata/en/dem0_fer_esms.htm provides further statistical presentation and methodology behind the reported values.

[31] Economically active individuals are defined as individuals in employment.

The values are gathered from Eurostat's database on employment[32].

3.2.10 Rate of unemployment, young individuals

I have included the variable *unemp_young* to examine the younger demographic more closely. The variable gives the unemployment rate in the 25-29 age group. Eurostat defines unemployed individuals as those who are without work but actively seeking it. The values are part of the EU-LFS series in Eurostat.

It has been well documented that the credit crunch had a big effect on unemployment rates of young people in several countries in Europe, most notably Spain. As the descriptive statistics and previous research indicate, the gender pay gap is significantly lower for younger people and we might expect the overall gender pay gap to increase if the rate of unemployment amongst young people increases. This variable is included to test whether we observe this, or if any other effect can be found.

3.2.11 Females in part time work

To account for the amount of females engaged in part time employment I include the variable fem_part. This variable is calculated as female part time workers as a percentage of the total number of females in employment. It is included to examine whether we get the same results as indicated by the descriptive statistics, i.e. that the gender pay gap is smaller for part time workers than for full time workers. Additionally, as mentioned in chapter 2, literature on the gender pay gap often points to women's decisions to choose more flexible working hours due to family responsibilites relating to childbirth etc. The values are gathered from the EU-LFS series in Eurostat.

3.2.12 Additional age groups

To further take the demographic of the population into account, I include three additional variables for age groups. These three variables give us the number of individuals within the respective age group as a percentage of the whole population.

[32] http://ec.europa.eu/eurostat/cache/metadata/en/lfsa_esms.htm gives further information on data coverage and data collection.

The three age groups included are 30-39, 40-49 and 50-64. The values are collected from Eurostat's database on population statistics[33].

3.2.13 Table of explanatory variables

Table 7: Statistics on Explanatory Variables

Variable	Mean Value	Std. Dev	Min.	Max.	Obs
gpg	16.190	6.603	-0.9	30.9	377
$age25\_29$	6.994	0.795	5.4	9.1	377
log_gdp	10.017	0.646	8.504	11.345	377
fem_edu	22.643	8.647	6.1	40.1	377
emp_oldfem	37.320	14.246	9.6	70.3	377
$fertrate$	1.557	0.238	1.13	2.06	375
$agebirth$	27.436	1.673	23.3	30.8	351
$actrate$	69.785	7.971	46	84.6	377
$unemp\_young$	9.528	4.874	2	33.3	377
fem_part	26.370	16.410	5.1	77.2	377
$age30\_39$	14.664	1.289	11.8	17.5	377
$age40\_49$	14.433	0.947	12.5	17.0	377
$age50\_64$	18.120	1.549	13.4	21.7	377

Table 8: Description of Explanatory Variables

Variable	Explanation
gpg	The unadjusted gender pay gap
$age25\_29$	Rate of population between 25-29 years old.
log_gdp	The natural logarithm of the GDP per capita in constant prices.
fem_edu	The rate of females with tertiary education.
emp_oldfem	The rate of employment for females between 55-64 years old.
$fertrate$	The fertility rate.
$agebirth$	The average age of mothers when first child is born.
$actrate$	The activity rate of females in the 25-64 age group.
$unemp\_young$	The rate of unemployment in the 25-29 age group.
fem_part	The rate of females in part time work.
$age30\_39$	Rate of population in the 30-39 age group.
$age40\_49$	Rate of population in the 40-49 age group.
$age50\_64$	Rate of population in the 50-64 age group.

[33] http://ec.europa.eu/eurostat/cache/metadata/en/demo_pop_esms.htm gives a thorough description of the data compilation.

3.3 Data for further age analysis

In light of the statistics presented on the gender pay gap in different age groups in 2 and the early-career effect, I take a closer look at the way the gender pay gap developes during working life. As presented in section 2.3 and touched upon earlier in this chapter, several studies have been made in this area.

To support my analysis, I have included a different dataset based on Eurostat's Structure of Earnings Survey in 2002, 2006 and 2010. The data contains observations from 13 different countries[34]. These countries were the only ones to report values for all three years.

Although the data is gathered from three different years, the methodological framework for each year is reported by Eurostat as being more or less the same. This gives the data good comparability over time[35]. Data coverage is subject to the same limitations that were described in section 3.1.1. The data is reported as being representative for the population, and the data has been revised until considered fit for publication.

The dataset is built up of observations of hourly wages for males and females in five different stages of working life, separated into five age groups: 20-29, 30-39, 40-49, 50-59 and 60-69. We have one observation for females, and one observation for males in each age group in each of the three years, giving a total of 10 observations per country each year. This gives us a total of 390 observations.

These observations are used to create several interaction variables combining gender and age groups to examine how the gender pay gap evolves over working life. We may expect to find some similar results to those presented in the descriptive statistics earlier in this chapter.

[34] The countries included can be found in the appendix

[35] More information regarding the metadata of this dataset can be found at http://ec.europa.eu/eurostat/cache/metadata/en/earn_ses_main_esms.htm

Table 9: Variable Description

Variables	Description
$logwage$	The natural logarithm of the mean hourly wage
$MALE$	Gender dummy which takes the value 1 if subject is male, 0 if female
$twenties$	Age dummy which takes the value 1 if subject is in their 20's
$thirties$	Age dummy which takes the value 1 if subject is in their 30's
$forties$	Age dummy which takes the value 1 if subject is in their 40's
$fifties$	Age dummy which takes the value 1 if subject is in their 50's
$maletwenties$	Interaction term combining $MALE$ and $twenties$
$malethirties$	Interaction term combining $MALE$ and $thirties$
$maleforties$	Interaction term combining $MALE$ and $forties$
$malefifties$	Interaction term combining $MALE$ and $fifties$

4 Econometric Framework and Challenges

4.1 Introduction

In this chapter I explain the empirical specification used in chapter 5 and the reasoning behind my specification. I examine the potential challenges that may arise, and look at possible corrections for these.

I start by addressing the econometric challenges present when using panel data and go on to explain the theoretical framework behind the main model, then I present the regressions estimated in chapter 5. Thereafter, I go through the structure used for extra analysis on the gender pay gap in different age groups, before discussing some of the issues that can arise from data problems and possible actions to address these issues.

4.2 Econometric challenges

To determine whether a causal relationship exists between two variables, the term *ceteris paribus* is central (Wooldridge (2002)). *Ceteris paribus* is a latin term meaning "all else equal". If, when all other factors are held constant, a change in the explanatory variable causes a change the dependent variable, we say that we have a causal relationship between the two variables.

The econometric model I explain here is designed to show causal relationships between the dependent and explanatory variables. The expected value of the dependent variable in a ceteris paribus analysis can be expressed as

$$E(y|x, Z) \tag{2}$$

where y is the dependent variable, x is the explanatory variable of interest, and Z represents a row vector of control variables[36].

The data is organised as panel data and so we have multiple observations of the same units over time. This gives us two dimensions of observations, unit and time, indicated by subscripts i and t respectively. In this case the units refer to countries, and time refers to the year.

[36] My analysis does not include control variables specifically, but this illustration is still relevant to present the mechanics of the analysis.

The panel data structure gives us the opportunity to follow developments in the gender pay gap in different countries over time. Organising the data in this way, however, does have some drawbacks.

To explain how the model is estimated, and to illustrate some of the issues we must consider, I will start by presenting the following simple regression

$$y_{it} = \beta_0 + \beta_1 X_{it} + u_{it} \tag{3}$$

where

y_{it} is the gender pay gap

β_0 is the constant term

X_{it} is a row vector of explanatory variables and associated coefficient vector β_1 [37]

u_{it} is the stochastic error term

Under certain assumptions this regression could be estimated by the Ordinary Least Squares (OLS) method and give us consistent and unbiased results (Woolridge (2009)). The OLS method minimizes the sum of squared errors in the data to give a linear representation of the relationship between the dependent and explanatory variables. To accept this, the following assumptions would have to hold:

1. The model is linear in its parameters.

2. Random sampling.

3. No perfect collinearity, i.e. none of the explanatory variables are constant, and there is no exact linear relationship between the explanatory variables.

4. Exogenous explanatory variables. We can illustrate this as $E(u_{it}|X_{it}) = 0$. Specifically, this means we assume that the explanatory variables are not correlated with the error term.

[37] $X_{it} = [X_{1it}, X_{2it}, ..., X_{Jit}]$ is a row vector with dimension $(1xJ)$, and $\beta_1 = \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1J} \end{bmatrix}$ is the associated column vector with dimension $(Jx1)$

For a ceteris paribus analysis to reveal causal relationships it is critical that assumption 4 holds. If assumption 4 does not hold, we have what we call endogenous explanatory variables. This could result from one of four main problems in the analysis: *Measurement error*, *omitted variable bias*, *simultaneity bias* and *misspecification of the model* (Woolridge (2009)). I will now present each of these problems to identify which are relevant in this case and explain what can be done to remove the issues they cause.

4.2.1 Measurement error

A measurement error occurs when the observed value of a variable deviates from its actual value. We usually consider two types of measurement error, random and systematic. Random measurement errors are generally not a problem so long as the errors are relatively few and small. Systematic measurement errors in explanatory variables are more serious and can cause bias towards zero in our dependent variable. This is illustrated by Woolridge (2009) on page 318-322.

In the analysis that follows, the data is gathered from Eurostat with some of the data reported from statistical offices of different countries. From the metadata included in the Eurostat databases, measurement errors in explanatory variables are expected not to be a serious problem.

4.2.2 Omitted variable bias

The problem of omitted variable bias occurs when a variable with explanatory power that is correlated with one or more other explanatory variables is not included in the model. The omitted variable is, in part, captured by the error component. Furthermore, because the omitted variable is correlated with one or more of the included explanatory variables, these included explanatory variables are correlated with the error term, thus breaking assumption 4 and causing biased estimates. The problem can arise from lack of data, or ignorance.

To be certain of avoiding omitted variable bias we would, in theory, have to include all variables with explanatory power on the gender pay gap. In reality this is difficult to achieve and so I propose a different way of resolving the issue.

First we look at the error term u_{it}. Because we are using panel data we must consider the error term to be composite (Woolridge (2009), page 457). The two components of the error term are:

1. The idiosyncratic component ϵ_{it}. This component captures unobserved factors that vary between units and over time which affect the dependent variable.

2. The individual component α_i. This component captures unobserved heterogeneity, i.e. factors that vary between units but are assumed to be constant over time. These time-invariant factors can, for example, be cultural differences between the countries that have an effect on the gender pay gap.

The error term can now be expressed as

$$u_{it} = \alpha_i + \epsilon_{it} \tag{4}$$

and for assumption 4 to hold we must demand that

$$E(\epsilon_{it}|X_{it}) = 0 \tag{5}$$

$$E(\alpha_i|X_{it}) = 0 \tag{6}$$

Considering assumption 4 we must require that the explanatory variables are uncorrelated with each component of the error term for OLS to give consistent and unbiased results. Equation (6) tells us that even though the idiosyncratic error term is not correlated with the explanatory variables, we could still have a problem of omitted variable bias. When equation (6) does not hold we get $E(\alpha_i|X_{it}) \neq 0$ and a problem of heterogeneity. This is a result of unobserved time-invariant factors between countries, and is perhaps the central problem to consider in this analysis.

To address the issue of omitted variable bias I use the Fixed Effects method for estimation. More details are provided later in this chapter.

Three further assumptions should be made regarding the two error components for the estimation to give "BLUE"[38] estimates Woolridge (2009):

1. No autocorrelation between the idiosyncratic component and homoscedasticity (constant variance): $corr(\epsilon_{it}, \epsilon_{js}|X_{it}) = 0$ and $var(\epsilon_{it}|X_{it}) = \sigma^2_\epsilon$

2. No autocorrelation between the individual component and homoscedasticity: $corr(\alpha_i, \alpha_j|X_{it}) = 0$ and $var(\alpha_i|X_{it}) = \sigma^2_\alpha)$

[38] BLUE is short for Best Linear Unbiased Estimator

3. The components of the error term are not correlated: $corr(\epsilon_{it}, \alpha_j) = 0$ for all i, t and j.

These assumptions need to hold for our regression to give us BLUE, but they do not need to hold to achieve consistent and unbiased estimates. If the assumptions do not hold this will invalidate the standard errors, and thus statistical inference theory will not give precise results (Verbeek (2012)).

It would be difficult to argue that autocorrelation and heteroskedasticity are not present in the error term in my analysis. To compensate for this I utilise cluster-robust standard errors[39] at country level. This ensures more reliable standard errors and allows me to use classic inference theory.

4.2.3 Simultaneity bias

Simultaneity bias occurs when one of the explanatory variables is determined simultaneously with the dependent variable. This will generally cause the dependent variable to be correlated with the error term, which in turn causes bias and inconsistency in the estimates (Woolridge (2009)). This is also known as reverse causality (Verbeek (2012), page 146-147).

For example, if we assume that the gender pay gap in a country affects the rate of females who complete tertiary education, then using this variable to explain the gender pay gap would cause a problem of simultaneity.

There could also be an unobserved variable that affects both the gender pay gap and the rate of females completing tertiary education. In this case we would have a problem of endogeneity. However if this unobserved variable is time-invariant the Fixed Effects method removes the issue, as we will see later in this chapter.

4.2.4 Misspecification of the model

Misspecification of the model is a special case of the omitted variable problem, and is caused by omitting a variable which can be expressed as a linear function of the included explanatory variables in the model. This can happen when the linear regression does not describe the true relationship between the dependent and the explanatory variables.

[39] Further information on clustered standard errors in Woolridge (2009) page 495, Verbeek (2012) page 389-390

For example, if the true relationship between the gender pay gap and the effect of further education is given by a positive, but decreasing relationship, the model will be misspecified if the squared term of the effect of education is not included. This is something that is not controlled for in my analysis.

4.3 Additional problems

4.3.1 Explanatory variables are correlated with the idiosyncratic error component

Measurement error, omitted variable bias, simultaneity bias and misspecification of the model would lead to explanatory variables correlating with the error term. This can be dealt with straightforwardly using the Fixed Effects method if the variables are correlated with the individual component of the error term.

However if there is a correlation between the explanatory variables and the idiosyncratic component, the issue becomes more complicated. We could still resolve the issue using either the Instrument Variable (IV) method or the two-stages least squares (2SLS) method (Woolridge (2009)).

Unfortunately data restrictions limit my ability to implement either of these methods successfully, and they will not be discussed further in this chapter.

4.3.2 Self-selection

By definition, the gender pay gap is only observed for employed individuals. If the sample of individuals we observe (employed) differs systematically from those not observed (unemployed) because of attributes and characteristics, we will have sample selection bias. As mentioned earlier, some countries show a lower rate of females in employment, which may affect the estimation of the gender pay gap.

Selection into employment is usually handled by considering variables such as marital status, number of children and other household members' income (Beblo et al. (2003), Albrecht et al. (2004)). In particular, the presence and age of children is a strong predictor for the choice of engaging in employment for females.

(Dupuy et al., 2009) have attempted to correct for self-selection based on data from Eurostat, and conclude that the restricted set of variables available causes the modelling of selection into employment to be a tedious exercise. Additionally, they note that their way of modelling the selection “does not seem to play a significant role in the age group 22-55”.

Due to the difficulty in modelling I do not be taking self-selection into account in my analysis.

4.4 Econometric Framework

As explained in section 4.2.2 the problem of heterogeneity is difficult to ignore when we have a panel data structure involving different countries. A simple OLS regression would consider all variations in the data (i.e. the variation between countries and the variation over time) but would not control for variables that are unique to a country, such as cultural differences that affect the gender pay gap. To correct this I implement a different strategy to my regression - the Fixed Effects method.

4.4.1 The Fixed Effects model

By transforming the model to deviations from individual means we are able to isolate variations within each country when estimating our model. This eliminates the individual component of the error term, and thus time-invariant unobserved differences between countries are transformed away from the model, removing the problem of heterogeneity.

To illustrate the transformation of the model, I start with the following regression

$$y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + \epsilon_{it} \tag{7}$$

This is the same as equation (3). As the individual component of the error term is constant over time it can be expressed as part of the constant term. We can therefore write

$$\beta_0 + \alpha_i = \delta_i \tag{8}$$

and call δ_i our new constant term. Furthermore, we must find the individual means by computing the following transformations

$$\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} \tag{9}$$

$$\overline{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it} \tag{10}$$

$$\overline{\epsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it} \tag{11}$$

$$\overline{\delta}_i = \frac{1}{T}\sum_{t=1}^{T} = \frac{1}{T}T\delta_i = \delta_i \tag{12}$$

This gives us

$$\overline{y}_i = \delta_i + \beta_1\overline{X}_i + \overline{\epsilon}_i \tag{13}$$

By subtracting equation (13) from equation (7) we have performed the within-transformation, and our specification is now given by

$$y_{it} - \overline{y}_i = \beta^{FE}(X_{it} - \overline{X}_i) + \epsilon_{it} - \overline{\epsilon}_i \tag{14}$$

By using this transformed model we only consider variations within each country and eliminate the problem of heterogeneity. We now have an estimator that gives unbiased estimates even when $E(\alpha_i|X) \neq 0$. These time-invariant differences between the countries are likely to be present in my model, which is the reason I have used the Fixed Effects model. When the within-transformation is implemented, the model can be estimated by OLS[40].

One drawback of the Fixed Effects method is that we need variation over time to get precise estimators. Because we now consider less variation than an OLS estimation would,

[40] When the Fixed Effects method is implemented in Stata, the following equation is estimated: $y_{it} - \overline{y}_i + \overline{y} = \beta_0 + \beta_1^{FE}(X_{it} - \overline{X}_i + \overline{X}) + \phi_t + (\epsilon_{it} - \overline{\epsilon}_i + \overline{\alpha}) + \overline{\epsilon}$, where the extra terms are the global means for each variable and $\overline{\alpha} = 0$ (see William Gould (1997)). This explains why the estimated model includes a constant term but my equation does not.

we should expect higher standard errors.

4.4.2 Time dummies

Time Dummies are included to control for yearly effects in my analysis. Generally, panel regressions without these time dummies will fail to control for aggregate variables such as inflation, economic growth and population growth. This would result in these aggregate variables affecting the included variables and we would not get a good estimate of causal relationships.

The inclusion of year dummies is implemented by using the first year of the sample as a base year.[41]

4.4.3 Final model specification

The introduction of time dummies now gives us the final model specification:

$$y^* = \beta^{FE} X_{it}^* + \phi_t + \epsilon^* \tag{15}$$

where

* indicates the within-transformed variables

X_{it} is a vector of explanatory variables presented in table 7 with the corresponding column vector β^{FE} of coefficients

ϕ_t is the time dummies

[41] 1997 is the first year of the sample.

4.5 Econometric specification age analysis

The data used for the further age analysis consists of observations from 2002, 2006 and 2010 in 13 countries as described in section 3.3. Again we have a panel data structure using 13 units with observations over three different years.

Unlike the data used in the main analysis, the build up of this dataset means that we do not have a balanced panel. To control for heterogeneity we have to use a different approach to the one we used for the main dataset. Where heterogeneity was previously corrected for by using the Fixed Effects model, we must now correct for this by using country dummies to accomplish the same. When implementing these dummies we must keep one country as a reference, and in my analysis the United Kingdom will be used as the reference country. The effect of this is virtually the same as the effect of implementing the Fixed Effects method described earlier, and we will only consider variation within each country when we do our analysis.

As they are in the main analysis, year dummies are included to capture the influence of aggregate trends.

The estimation is based on the following regression:

$$logwage_{it} = \beta_0 + \beta_1 MALE + \beta_2 AGE + \beta_3 MALE \times AGE + \phi_t + \gamma_i + u_{it} \tag{16}$$

where

$logwage$ is the hourly wage expressed as a natural logarithm

β_0 is the constant term

$MALE$ is the gender dummy

AGE is a row vector of age groups

$MALE \times AGE$ is a row vector of interaction terms

ϕ_t is the year dummies

γ_i is the country dummies

$\beta_{1,2,3}$ is a column vector of corresponding coefficients

By using a model built up of dummy variables and interaction terms, we are able to get a better understanding of how the gender pay gap changes with age.

The dependent variable in my analysis is the average hourly wage for male and females in the different age groups. The hourly wage is expressed as a natural logarithm.

Inclusion of the gender dummy gives us the male wage premium (or the gender pay gap) in the reference category chosen. In this analysis males in their 60's are chosen as the reference category, and β_1 can be interpreted directly as the GPG for workers in their 60's.

The interaction terms give us the relationship between the various age and gender categories and the reference category. Specifically they give us the difference in the male wage premium in percentage points. To find the male wage premium in a different age class, we can use the following equation

$$gpg_{agegroup_i} = \beta_1 + \beta_{3i} \tag{17}$$

where β_{3i} indicates the age group of interest

4.6 Data issues and possible corrections

As discussed in chapter 3 the data used in this study is gathered from different databases in Eurostat. Here I address data values on the gender pay gap, which is the dependent variable in my analysis.

The gender pay gap data used in my main analysis is gathered from two different databases in Eurostat. One database includes the reported values from 1997 to 2006, the other gives the reported values from 2007 to 2013. Issues may arise due to methodological differences in data collection between the two databases.

As described in chapter 3, the data compiled relating to 2007 to 2013 are based on the SES and follow the guidelines agreed by the countries included in the database. Some deviations are reported from the participating countries, but their effects are minimal. Eurostat reports that ”comparability over time has improved substantially since the GPG has been calculated on the basis of the four-yearly Structure of Earnings Survey”[42].

4.6.1 Data compilation before 2007

Data for the period 1997-2006 present some issues for my analysis. The values from this period are not based on the SES, but rather on national sources. These national sources use different surveys as a basis for calculating the reported values on the GPG, including the European Community Household Panel (ECHP) and the EU Survey on Income and Living Conditions (EU-SILC)[43].

The fact that we have these differences in data collection methodology presents some challenges. The methodological breaks can cause variation in the GPG that our included explanatory variables will capture, giving misleading results. One example is the inclusion of more sections of the economy during the time period. Another example was indicated briefly by figure 1 where we saw a change in reported GPG in the transition period 2006-2007.

Some countries do comment that the changes in reported GPG are a direct result of the changes in methodology. Denmark asserted that the change of source increased the reported gender pay gap by an estimated four percentage points between 2001 and 2002.

[42] See section 16.2 at http://ec.europa.eu/eurostat/cache/metadata/en/earn_grgpg2_esms.htm

[43] Metadata for the time period 1997-2006 is found at http://ec.europa.eu/eurostat/cache/metadata/en/earn_gr_hgpg_es

Similarly the way the gender pay gap data is measured is a cause for concern. Understanding whether the calculation is based on hourly, monthly or annual earnings is essential to interpreting results. For example the variable concerning part-time female workers could give different results depending on this, as the difference between part-time workers' and full-time workers' earnings can be exaggerated when looking at monthly salaries compared to hourly wages. If we assume that part-time workers have a relatively lower monthly salary than full-time workers (compared to the differences in their hourly rate), then the effect of the variable regarding females in part-time work would be overestimated in the analysis if monthly salary data were used.

Looking more closely at the gender pay gap information provided by the countries, we should also be aware that some countries reported changes in the GPG as a result of policy changes rather than methodological breaks. Hungary reported a significant change in the gender pay gap between 2002 and 2003 because of a pay rise in the public sector which includes a lot of women. There are some other examples of this before 2007, but none after 2007. This could cause misleading results as this change in GPG might be picked up by the included explanatory variables.

Eurostat reports that the geographical comparability is limited because of this difference in national data sources, and also that the methodological breaks cause the comparability over time to be less accurate. The geographical comparability issue is not necessarily important as the variation used in my analysis is within-variation implemented by the Fixed Effects model. However comparability over time is essential to get trustworthy results. The unreliable pre-2007 data could give an incorrect estimation of the gender pay gap over time.

Notes regarding data collection and comments pre and post-2007 is from each country is included in the appendix.

4.6.2 Responses to issues

As discussed earlier in this chapter, the Fixed Effect model with time dummies is implemented to address the issue of unobserved heterogeneity and aggregate time-series trends.

Regarding the data collection before 2007 and the data quality issues discussed, it is difficult to implement a direct econometric fix. To determine whether there are problems too large to ignore, I check the robustness of my analysis by performing a new analysis based on better data. For this robustness check I will utilise the more trustworthy data gathered after 2007 to see if there is any indication that any results from my main analysis are due to spurious values from the older data.

Additional robustness checks are performed to see how the main model responds to the inclusion of explanatory variables that are not included in the original specification.

5 Results

In this chapter I present the results from my empirical analysis. The theoretical framework on which the analysis is built is described in chapter 4, and the data used is presented in chapter 3.

The first results I present are the results of the regression based on the dataset modelling the development of the gender pay gap during working life. This gives us an opportunity to look at the early-career effect.

I then present the results of my main model, including the age variable and other explanatory variables to see whether we can find some common trends regarding the gender pay gap across Europe.

Finally I perform some robustness checks relating to the data issues discussed in section 4.6. The checks investigate whether the results of the main model are robust to the inclusion of new explanatory variables and to an alternative time period.

5.1 GPG developments age groups

Table 10 shows the results from the model reported in chapter 4, and the reported values are given by equation (16). The model is based on dummy and interaction variables, with reported values being findings relative to male workers in their 60's.

Table 10: Age Analysis

Dependent Variable	*logwage*
$MALE$	0.195
	(0.0406)***
$twenties$	-0.295
	(0.0406)***
$thirties$	-0.099
	(0.0406)**
$forties$	-0.082
	(0.0406)**
$fifties$	-0.058
	(0.0406)
$maletwenties$	-0.132
	(0.0575)**
$malethirties$	-0.017
	(0.0575)
$maleforties$	0.021
	(0.0575)
$malefifties$	0.002
	(0.0575)
Year Dummies	Yes
Country Dummies	Yes
Observations	390
R^2	0.967
Statistical significance	$^{**}p < 0.05$, $^{***}p < 0.01$

The reported coefficients give us insight into how the hourly wage develops during working life compared to the reference category.

Table 10 shows a coefficient value of 0.195 relating to the gender dummy variable, giving us the male wage premium for the reference category. A value of 0.195 suggests the average hourly earnings for men in their 60's is 19.5% higher than for women in the same age group. This result is highly significant, with a t-value of 4.80.

Dummies for each age group indicate the difference in hourly earnings experienced in each age group compared to the reference category of workers in their 60's. We notice that there is quite a large negative difference for workers in their 20's, but a much smaller difference for workers in their 30's and 40's. There is no significant difference for workers in their 50's.

If we assume that the increase in earnings for workers as they grow older is a result of increased work experience[44] we can use the interaction variables between gender and age to examine whether the returns to work experience are similar for men and women.

The coefficients of interest here are those relating to the interaction variables between age and gender. We can use these coefficients to find the earnings gap in the different age groups. Each coefficient is related to the reference category. We can immediately see that there is no significant difference between the male wage premium for workers in their 60's and the workers in their 30's, 40's or 50's. However there is a big difference when we look at *maletwenties*. Compared to the reference category, there is a huge change in the male wage premium. -0.132 indicates that the gender pay gap for workers in their 20's is 13.2 percentage points lower than for workers in their 60's. Therefore the overall gender pay gap in this age group can be calculated as

$$0.195 - 0.132 = 0.063 \tag{18}$$

This indicates a rather low pay gap near the beginning of the working career with an estimated male wage premium of 6.3% for workers in their 20's. We can see a big jump during the next decade, and a subsequent stabilization at about 19.5%.

This is consistent with the descriptive statistics presented in chapter 3, as well as findings in previous studies on the subject with Stokke (2016), Blau & Kahn (2000) and Bertrand et al. (2009) showing similar results.

Although this is a simplification, it does show some of the signs that were found by Stokke (2016). The higher returns to work experience for men probably contributes to increasing the gender pay gap with age. Factors such as childbirth and establishing families are often considered components in the increase of the gender pay gap during working life. However without data to supplement my analysis we should take caution when attributing observed results to these factors.

[44] This might be a simplification, as there are likely other factors.

5.2 Main model

To build up my analysis, I start with a simple regression, then expand it by adding more explanatory variables. In total I will run four different regressions to see how the estimated coefficients react to the inclusion of new variables[45].

The four model specifications are given by equation (15), where X_{it} represents the explanatory variables presented in table 7[46]. The fourth specification gives the result of my main model.

Table 11: Main Model

Dependent variable	*gpg*	*gpg*	*gpg*	*gpg*
$age25\_29$	-1.600	-1.803	-1.532	-1.609
	(0.715)** [a]	(0.619)***	(0.607)**	(0.571)**
log_gdp		4.716	5.008	3.803
		(1.892)**	(2.063)**	(1.581)**
fem_edu			-0.224	-0.237
			(0.134)	(0.127)*
emp_oldfem				0.077
				(0.119)
$cons$	29.820	-15.490	-16.844	-6.243
	(5.298)***	(20.697)	(22.708)	(15.820)
Country Fixed Effects	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Observations	377	377	377	377
Period	1997-2013	1997-2013	1997-2013	1997-2013
R^2 Within[b]	0.1758	0.2150	0.2373	0.2442
Statistical significance	$*p < 0.10$	$**p < 0.05$	$***p < 0.01$	

[a] All reported standard errors are cluster robust and clustered on country level

[b] R^2 gives us an indication of how well our regression approximates the actual data, providing a goodness-of-fit measure. A higher value of R^2 indicates higher explanatory power. With the Fixed Effects method, all variation used is within-variation, and the R^2 reported is calculated from the OLS regression used on the transformed data.

[45] Correlation matrices for all included variables are included in the appendix.

[46] The main model will include four explanatory variables from table 7.

To interpret the coefficients we must first consider the way the variables are specified. With the exception of *log_gdp* all variables are specified as rates and so it is straightforward to interpret these coefficients. An increase in a variable by one percentage point leads to a β percentage point change in the dependent variable, *gpg*.

log_gdp is the only logarithmic variable in the analysis. In this case the coefficient can be interpreted as the percentage point change in the dependent variable given a one unit increase in *log_gdp*. For *log_gdp* to increase by one point the constant GDP per capita would have to double in size.

The first coefficient we address is $age25\_29$. The negative coefficient values in the model specifications are consistent with the descriptive statistics detailed in chapter 3, earlier studies mentioned in chapter 2 and the analysis presented in table 10. The coefficients are stable in all model specifications with values ranging from -1.532 to -1.803. This suggests that an increase in the proportion of population aged 25-29 by one percentage point would decrease the gender pay gap by 1.6 percentage points in the final model specification. All reported coefficients are statistically significant at a 5% level and even at a 1% level in specification 2.

In specification 2 I introduce *log_gdp*. The results show a positive relationship between the GDP per capita in constant prices and the gender pay gap. The coefficients change when introducing more variables to the specification, but considering that the GDP per capita would have to double in size to observe a percentage point change in the GPG equal to the coefficients, the variation is not very significant. The results suggest we can establish a connection between the gender pay gap and the level of the GDP per capita in a country, with an estimated effect in the final model specification of 3.8 percentage points following an increase in GDP per capita in constant prices of 100%. All coefficients are significant at a 5% level.

The interpretation of these results is not obvious as variation in the observed gender pay gap across the countries presented in table 1 did not show a clear pattern of a higher gender pay gap in wealthier countries. One interpretation may be that countries with a higher GDP per capita have higher employment rates than less wealthy countries. If these extra workers are largely employed in low-wage jobs, and if women dominate this group, then we would observe the relationship my model predicts.

fem_edu gives us the relationship between the rate of females in the population completing tertiary education and the gender pay gap. We expect an increase in this variable would cause a reduction in the wage gap as higher education is generally associated with higher wages. The results from the model support this with negative values reported in both specifications. In model specification 3 the coefficient falls just outside the 10% significance level, but the inclusion of emp_oldfem in specification 4 gives the coefficient statistical significance at 10% level with a p-value of 0.075.

The coefficient suggests than an increase by one percentage point in the rate of females completing tertiary education reduces the gender pay gap by 0.237 percentage points in the final model specification. The coefficient is stable in both specification 3 and 4.

In the final model specification I introduce the variable emp_oldfem. This gives us the relationship between the employment rate of females in the 55-64 age group and the gender pay gap. The effect is estimated to be small but positive, with a coefficient of 0.077. However a large standard error causes the variable to fall well below the threshold to give this result any statistical significance. We cannot therefore conclude that this coefficient is different from zero. This is perhaps unsurprising when we consider that the gender pay gap is relatively stable after the early-career effect documented in the age analysis and by Stokke (2016), and the majority of females in the workforce probably fall into age groups in 30's and above.

5.3 Robustness checks

In this section I present four alternative specifications of the main model to check the robustness of the results obtained using the main model.

The first robustness check addresses the data issues described in chapter 4, and the subsequent tests are designed to investigate whether previously omitted variables affect the gender pay gap.

5.3.1 Alternative time period

As discussed in chapter 4 some problems arise from the methodological differences in data collection between countries prior to 2007. For my first robustness check I will use the same model specification as in my main analysis but limit the data to that compiled between 2007-2013. The methodological differences in data collection between the countries is greatly reduced over this period and I consider the data quality to be better. This should decrease the risk of my explanatory variables picking up effects principally due to changes in methodology, and so the coefficients may give more trustworthy estimates of the effect the explanatory variables have on the gender pay gap.

Reducing the time period is not without its complications however, as the within-variation required for the Fixed Effects model to give accurate results will most likely be reduced. This may cause higher standard deviations and lower estimates, resulting in coefficients with a lower significance level than previously presented. However the regression will still indicate whether the results of the main model are trustworthy or not.

Table 12: Robustness Check 1, Alternative Time Period

Dependent variable	*gpg*
$age25\_29$	-1.647
	(0.637)**
log_gdp	1.600
	(3.984)
fem_edu	-0.118
	(0.104)
emp_oldfem	-0.093
	(0.129)
$cons$	18.962
	(36.906)
Country Fixed Effects	Yes
Year Dummies	Yes
Observations	163
Period	2007-2013
R^2 Within	0.2517
Statistical significance	$^{**}p < 0.05$

Straight away we can see the effect the reduction in variation gives us. In general significance levels are lower, with the exception of the coefficient relating to young individuals in the population. This is perhaps as expected, but there are also some positives to consider.

There is little difference between the coefficient values generated from this limited data and those produced by the main model. The only variable to stand out is emp_oldfem, which now suggests a negative relationship between the rate of employment of older females and the gender pay gap. However the estimated coefficient is still not significantly different from zero.

These results indicate that even though some of the estimated coefficients may be caused by variation resulting from changes in data collection methodology, the relationships between the gender pay gap and the explanatory variables predicted by the main model are likely to be present and could provide us with insight into the factors that contribute to the gender pay gap.

Most significantly, the result of the $age25\_29$ coefficient adds weight to the argument that the level of the gender pay gap is highly dependent on the age group we address. The estimated coefficient is still significant at a 5% level even after the reduction in observations.

5.3.2 Additional explanatory variables

The following robustness checks focus on explanatory variables that were excluded from the main model. By including these I examine whether the results from the main model hold up when other variables are included.

The new variables are listed in table 7 and presented in section 3.2.

5.3.3 The effect of childbirth

The first two variables I include are linked to the effect childbirth has on the the working life of females.

To model this I introduce $fertrate$ and $agebirth$ to the main model. These variables are added one at a time. I also perform an additional regression where the only insignificant coefficient from the main model, emp_oldfem, is dropped. This gives us a total of three new specifications.

Table 13: Robustness Check 2, The Effect of Childbirth

Dependent variable	*gpg*	*gpg*	*gpg*
*age*25_29	-1.657	-1.589	-1.541
	(0.606)**	(0.635)**	(0.693)**
log_gdp	4.387	5.634	6.485
	(1.910)**	(2.204)**	(1.983)***
fem_edu	-0.243	-0.270	-0.260
	(0.130)*	(0.140)*	(0.139)*
emp_oldfem	0.067	0.050	
	(0.113)	(0.129)	
fertrate	-3.306	-4.400	-4.493
	(5.276)	(5.578)	(5.522)
agebirth		-0.788	-0.900
		(0.684)	(0.753)
cons	-6.396	5.077	0.613
	(15.430)	(23.462)	(20.727)
Country Fixed Effects	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
Observations	375	327	327
Period	1997-2013	1997-2013	1997-2013
R^2 Within	0.2517	0.2893	0.2870
Statistical significance	$*p < 0.10$	$**p < 0.05$	$***p < 0.01$

We can see that the inclusion of *fertrate* does not yield any significant change to our main model specification. The coefficient itself gives a negative relationship between fertility rate and the pay gap, but with a t-value of -0.63 it is not significantly different from zero. There is a small change in R^2 indicating some added explanatory power when including this variable, however these results suggest there is no reason to interpret the effect of this variable further.

The inclusion of *agebirth* gives a similar conclusion, with a negative but smaller effect on the gender pay gap. The coefficient is certainly more significant than *fertrate* but still well below the threshold to give us any statistical significance. We do observe a bigger increase in R^2, indicating more explanatory power when this variable is included. Once again we should not conclude any significant effects from this result.

The removal of emp_oldfem in the final specification causes a bigger and more significant result in the coefficient concerning GDP per capita. Other than this there is no significant change.

The results presented in table 13 indicate no effect of childbirth on the gender pay gap. The effect is difficult to model with a sample covering several countries, as employment laws may be related to childbirth may be different across the countries in my sample[47].

This robustness check indicate robust results in my main model specification. The inclusion of new variables do not cause a significant change in the interpretation of the existing variables.

[47] One factor that could affect the results are the parental leave policies in different countries. We would expect a country with a liberal policy regarding maternal and paternal leave would report a smaller effect of childbirth than a country with a stricter parental leave regime. A study by (Waldfogel, 1998) finds that the negative effect of childbirth on women's wages is reduced for mother's who have maternity leave.

5.3.4 Work patterns and activity rate

In this test I introduce three new variables relating to activity rates and work patterns. I introduce one variable at a time and again remove *emp_oldfem*, creating four new model specifications.

In the first specification I introduce the variable *actrate*, then *unemp_young* in specification 2, *fem_part* in specification 3 before removing *emp_oldfem* in specification 4.

Table 14: Robustness Check 3, Work Patterns and Activity Rates

Dependent variable	*gpg*	*gpg*	*gpg*	*gpg*
*age*25_29	-1.599	-1.606	-1.697	-1.628
	(0.584)**	(0.584)**	(0.648)**	(0.682)**
log_gdp	3.813	3.713	3.648	4.336
	(1.629)**	(1.750)**	(1.755)**	(1.628)**
fem_edu	-0.243	-0.240	-0.250	-0.251
	(0.131)*	(0.131)*	(0.136)*	(0.140) *
emp_oldfem	0.071	0.063	0.068	
	(0.132)	(0.129)	(0.128)	
actrate	0.019	0.031	0.074	0.118
	(0.181)	(0.178)	(0.182)	(0.163)
unemp_young		-0.026	-0.027	-0.048
		(0.083)	(0.081)	(0.087)
fem_part			-0.077	-0.070
			(0.156)	(0.155)
cons	-7.336	-6.635	-6.295	-14.486
	(22.985)	(23.727)	(23.916)	(23.794)
Country Fixed Effects	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Observations	377	377	377	377
Period	1997-2013	1997-2013	1997-2013	1997-2013
R^2 Within	0.2443	0.2448	0.2479	0.2442
Statistical significance	$^*p < 0.10$	$^{**}p < 0.05$		

Table 14 gives the results of the new model specifications. *actrate* is estimated as being small and positive while $unemp\_young$ and fem_part are estimated as small and negative. However none of the new coefficients are significantly different from zero in any of the model specifications.

Once again, the inclusion of these new variables do little to change my main model. The original coefficients are stable in all new specifications. Additionally, R^2 indicate no significant added explanatory power in any of the new specifications. We conclude that adding these variables does little to improve the main model, and that the main model is robust to the inclusion of the new variables on activity rates and work patterns.

5.3.5 Additional age groups

The final robustness check examines how the inclusion of additional age groups affects the main analysis. From the main analysis and the analysis of the gender pay gap development we find the youngest age group to be the main explanatory factor, and the inclusion of these new age groups will provide insight into whether the pay gap is dependent on the rest of the population demographic.

The new variables are all added at once in the first specification, and again a second specification is estimated without emp_oldfem.

Table 15: Robustness Check 4, Additional Age Groups

Dependent variable	*gpg*	*gpg*
$age25\_29$	-1.692	-1.674
	(0.558)***	(0.558)***
log_gdp	4.373	4.949
	(1.570)**	(1.646)***
fem_edu	-0.184	-0.173
	(0.137)	(0.145)
emp_oldfem	0.038	
	(0.099)	
$age30\_39$	-0.276	-0.333
	(0.359)	(0.410)
$age40\_49$	-0.251	-0.283
	(0.638)	(0.679)
$age50\_64$	-0.760	-0.780
	(0.563)	(0.609)
$cons$	9.124	5.803
	(25.321)	(22.969)
Country Fixed Effects	Yes	Yes
Year dummies	Yes	Yes
Observations	377	377
Period	1997-2013	1997-2013
R^2 Within	0.2686	0.2672
Statistical significance	$^{**}p < 0.05$	$^{***}p < 0.01$

Table 15 shows that the $age25\_29$ and log_gdp coefficients are robust to the inclusion of additional age groups.

None of the estimated coefficients are significantly different from zero, which supports the previous findings on the early-career effect.

We also observe a reduction in the estimated coefficient relating to female education attainment. This reduction leads to insignificant coefficients in both new specifications.

6 Closing Remarks

This thesis uses panel data for several European countries to examine factors that affect the gender pay gap. The data utilised in the empirical research is gathered from Eurostat and used to create two panel datasets as described in chapter 3. The biggest dataset is used in the main model and several robubstness checks, while the smaller dataset is used to examine the development of the gender pay gap with age.

The main model is estimated with the Fixed Effects method and subsequently utilises within-variation to correct for time-invariant heterogeneity in the sample. In the additional model considering developments of the GPG with age country dummies are used to correct for time-invariant heterogeneity. Time dummies are included in both models and subsequent robustness checks to control for aggregate time variables.

The analysis on the gender pay gap development during working life strongly indicates of the early-career effect in the included countries. I find a gender pay gap of 6.3% for workers in their 20's. Further results from the analysis show a significant increase in the gender pay gap during the next decade, and a stabilisation at 19.5% that is broadly constant for the remainder of working life.

The strong link between the gender pay gap and the age group we observe is highlighted further by my main analysis. An increase in the rate of young individuals in the population is found to have a strong negative effect on the gender pay gap in countries across Europe. This effect is robust to the implementation of different model specifications as well as to alternative time periods. This result is consistent with previous studies summarised in chapter 2.

I also find a positive relationship between the GDP per capita in constant prices and the gender pay gap. This indicates a larger observed gender pay gap in wealthier countries. The result is robust to the inclusion of new variables. Additionally an increase in the level of education for females is found to have a small but negative effect on the gender pay gap. This result is significant at a 10% level in the main analysis, but not significant when including additional age variables.

The issue of omitted variable bias has been discussed in this thesis, and one factor I have not been able to include is the effect of employment segregation between men and women in different sectors and industries. Earlier studies indicate this plays a big part in contributing to the gender pay gap, and it is probably relevant when looking at the pay gap across Europe.

Expanding the main model with variables concerning childbirth, activity rates and work patterns and different age groups are not found to give any significant effect on the gender pay gap. Including such variables on a sample of this size does not provide any significant results, and may work better on more individual specific data.

To analyse the gender pay gap across a big sample of countries demands reliable data over time when using the Fixed Effects method. There is some doubt regarding the data compilation before 2007 in this case and the effects this causes on the results of log_gdp and fem_edu. Further analysis based on the more trustworthy data from 2007 and onwards might give more accurate results and could be something to look at for further research.

References

Albrecht, J., Van Vuuren, A., & Vroman, S. (2004). Decomposing the gender wage gap in the netherlands with sample selection adjustments.

Altonji, J. G., & Blank, R. M. (1999). Race and gender in the labor market. *Handbook of labor economics*, *3*, 3143–3259.

Bayard, K., Hellerstein, J., Neumark, D., & Troske, K. (1999). *New evidence on sex segregation and sex differences in wages from matched employee-employer data* (Tech. Rep.). National bureau of economic research.

Beblo, M., Beninger, D., Heinze, A., & Laisney, F. (2003). Methodological issues related to the analysis of gender gaps in employment, earnings and career progression. *Mannheim: European Commission, Employment and Social Affairs DG*.

Becker, G. S. (1985). Human capital, effort, and the sexual division of labor. *Journal of labor economics*, S33–S58.

Bertrand, M., Goldin, C., & Katz, L. F. (2009). *Dynamics of the gender gap for young professionals in the corporate and financial sectors* (Tech. Rep.). National Bureau of Economic Research.

Blau, F. D., & Kahn, L. M. (1997). Swimming upstream: Trends in the gender wage differential in the 1980s. *Journal of labor Economics*, 1–42.

Blau, F. D., & Kahn, L. M. (2000). *Gender differences in pay* (Tech. Rep.). National bureau of economic research.

Blau, F. D., & Kahn, L. M. (2007). The gender pay gap have women gone as far as they can? *The Academy of Management Perspectives*, *21*(1), 7–23.

Blau, F. D., & Kahn, L. M. (2016). The gender wage gap: Extent, trends, and explanations.

Dupuy, A., Fouarge, D., & Buligescu, B. (2009). Development of econometric methods to evaluate the gender pay gap using structure of earnings survey data. *Eurostat Methodologies and Working Papers*.

Emerek, R., Meulders, D., O'Dorchai, S., Beleva, I., Krizkova, A., Maier, F., ... others (2006). The gender pay gap-origins and policy responses.

Flyer, F., & Rosen, S. (1994). *The new economics of teachers and education* (Tech. Rep.). National Bureau of Economic Research.

Groshen, E. L. (1991). The structure of the female/male wage differential: Is it who you are, what you do, or where you work? *Journal of Human Resources*, 457–472.

Kunze, A. (2005). The evolution of the gender wage gap. *Labour Economics*, *12*(1), 73–97.

Manning, A., & Swaffield, J. (2008). The gender gap in early-career wage growth. *The Economic Journal*, *118*(530), 983–1024.

Mincer, J., & Polacheck, S. (1974). Family investments in human capital: Earnings of women. In *Economics of the family: Marriage, children, and human capital* (pp. 397–431). University of Chicago Press.

Stokke, H. E. (2016). The gender wage gap and the early-career effect.

Verbeek, M. (2012). *A guide to modern econometrics, 4th edition.* John Wiley & Sons.

Waldfogel, J. (1998). The family gap for young women in the united states and britain: Can maternity leave make a difference? *Journal of labor economics*, *16*(3), 505–545.

William Gould, S. (1997). How can there be an intercept in the fixed-effects model estimated by xtreg, fe? *Stata FAQ. Available at http://www.stata.com/support/faqs/statistics/intercept-in-fixed-effects-model/*.

Wood, R. G., Corcoran, M. E., & Courant, P. N. (1993). Pay differences among the highly paid: The male-female earnings gap in lawyers' salaries. *Journal of Labor Economics*, 417–441.

Wooldridge, J. M. (2002). Econometric analysis of cross section and panel data.

Woolridge, J. M. (2009). Introductory econometrics: A modern approach 4. ed. *South-Western, Michigan State University*, 378.

A Appendix

A.1 Economic sections in Eurostat

B - Mining and quarrying

C - Manufacturing

D - Electricity, gas, steam and air conditioning supply

E - Water supply; sewerage, waste management and remediation activities

F - Construction

G - Wholesale and retail trade; repair of motor vehicles and motorcycles

H - Transportation and storage

I - Accommodation and food service activities

J - Information and communication

K - Financial and insurance activities

L - Real estate activities

M - Professional, scientific and technical activities

N - Administrative and support service activities

O - Public administration and defence; compulsory social security

P - Education

Q - Human health and social work activities

R - Arts, entertainment and recreation

S - Other service activities

A.2 Countries included

Table 16: List of Countries Included in Datasets

Main Dataset	Age Analysis Dataset
Austria	Bulgaria
Belgium	Czech Republic
Cyprus	Ireland
Czech Republic	Spain
Denmark	Lithuania
Estonia	Hungary
Finland	Netherlands
France	Poland
Germany	Romania
Greece	Slovenia
Hungary	Slovakia
Ireland	Sweden
Italy	United Kingdom
Latvia	
Lithuania	
Luxembourg	
Netherlands	
Norway	
Portugal	
Romania	
Slovenia	
Spain	
Sweden	
United Kingdom	
24 countries	13 countries

A.3 Correlation matrices

A.3.1 Main model

```
             | age25_29  log_gdp  fem_edu emp_ol~m
-------------+------------------------------------
    age25_29 |   1.0000
     log_gdp |  -0.2974   1.0000
     fem_edu |  -0.2973   0.3353   1.0000
  emp_oldfem |  -0.3711   0.1885   0.6354   1.0000
```

A.3.2 Robustness check 1

```
             | age25_29  log_gdp  fem_edu emp_ol~m
-------------+------------------------------------
    age25_29 |   1.0000
     log_gdp |  -0.2282   1.0000
     fem_edu |   0.1314   0.3252   1.0000
  emp_oldfem |  -0.2289   0.1929   0.5791   1.0000
```

A.3.3 Robustness check 2

```
             | age25_29  log_gdp  fem_edu emp_ol~m fertrate agebirth
-------------+------------------------------------------------------
    age25_29 |   1.0000
     log_gdp |  -0.2879   1.0000
     fem_edu |  -0.3506   0.3482   1.0000
  emp_oldfem |  -0.4040   0.1836   0.6164   1.0000
    fertrate |  -0.4190   0.6368   0.5674   0.4629   1.0000
    agebirth |  -0.1806   0.7318   0.3777   0.0977   0.4307   1.0000
```

A.3.4 Robustness check 3

```
             | age25_29  log_gdp  fem_edu emp_ol~m  actrate unemp_~g fem_part
-------------+---------------------------------------------------------------
    age25_29 |   1.0000
     log_gdp |  -0.2974   1.0000
     fem_edu |  -0.2973   0.3353   1.0000
  emp_oldfem |  -0.3711   0.1885   0.6354   1.0000
     actrate |  -0.4076   0.0698   0.5969   0.8213   1.0000
 unemp_young |   0.0614  -0.2994   0.0516  -0.1512  -0.1364   1.0000
    fem_part |  -0.4183   0.6907   0.2375   0.2218   0.1545  -0.3616   1.0000
```

A.3.5 Robustness check 4

```
             | age25_29  log_gdp  fem_edu emp_ol~m age30_39 age40_49 age50_64
-------------+---------------------------------------------------------------
    age25_29 |   1.0000
     log_gdp |  -0.2974   1.0000
     fem_edu |  -0.2973   0.3353   1.0000
  emp_oldfem |  -0.3711   0.1885   0.6354   1.0000
    age30_39 |   0.3328   0.1524  -0.3072  -0.4851   1.0000
    age40_49 |  -0.4226   0.2233  -0.0355  -0.1711  -0.1128   1.0000
    age50_64 |  -0.4368  -0.0574   0.1822   0.3268  -0.3495   0.0576   1.0000
```

A.4 Notes data compilation from countries 1997-2006

The following gives each countries notes on data collection regarding the reported values on the gender pay gap in the time period 1997-2006. The information is gathered from http://ec.europa.eu/eurostat/cache/metadata/en/earn_gr_hgpg_esms.htm

Austria

Since 2003 figures are based on EU-SILC. No data for 2002 are available.

Belgium

Since 2004, data are based on EU-SILC. Data for 2002 and 2003 are not available.

Cyprus

The Gender Pay Gap is calculated on the basis of the average monthly rates of pay extracted from the annual survey of wages and salaries, since 1995. The survey covers all size groups (including the enterprises with 1-9 employees) and collects data for full-time employees in all economic activities of NACE Rev.1.1, including the government sector. The specific survey is conducted on a yearly basis and has October as the reference period. Information is collected for the occupation, gender and the gross earnings and employer's social contribution paid for each employee in the enterprise. An indication is also given concerning the age of the employee. No information is given concerning the educational status and the professional experience of the employees.

According to the specific survey, gross earnings refer to the total gross annual earnings (i.e. normal plus overtime earnings) for actual hours worked, including bonuses paid irregularly during the year.

Czech Republic

Figures are based on median earnings of employees working 30 or more planned hours per week, in enterprises with more than 9 employees. No data are available for 1994 and 1995. The greatest increase of the gender pay gap happened from 1997 to 1998. In these years, the national economy passed through a major depression. Reductions in earnings levels were documented in many groups of employees, especially in the public sector. Also, the earnings of clerks, teachers, shop assistants, etc. fell. The earnings levels of blue-collar workers were not affected as much. Women typically dominate the above-mentioned occupations and also the public sector. In the subsequent years, the situation recovered gradually. The discrepancy between earnings of men and women, in terms of the gender pay gap, had dramatically increased in 1998 (to 25 percentage points). After that, it returned close to the original level (22 percentage points) in 1999. From 2000, the economic situation has been stable and the discrepancy has slowly narrowed.

Denmark

Since 2002, the national structure of earnings survey is used, which covers employees aged 16-64 working 15 or more hours per week in economic activities NACE Rev.1.1 sections C-Q. The weights are based on the number of hours paid and bonuses are excluded. The effect of the change of source on the gender pay gap is estimated to be an increase of 4 percentage points, based on data from 2001. The reason for the change in the gender pay gap between 2001 and 2002 is that the data source was changed. A change in data source also occurred between 1994 and 1995 but it is not possible to explain how much this affected the increase of 4 percentage points over this period.

Estonia

Since 2002, the national survey (covering NACE Rev.1.1. sections A, B, L-O) and the structure of earnings survey (covering sections C-K) have been combined.

Finland

Since 2002 Data, the national structure of earnings data is used to calculate the GPG. Data covers almost all employees despite of their age and working time in all NACE sections. There are some coverage problems especially in micro enterprises and among general managers. Data for 2001 and earlier is based on European Community Household Panel. It is estimated that the structure of earnings data produces around 3 percentage points higher gender pay gap value than the value from the ECHP.

France

The annual labour force survey is used as the source for the gender pay gap for 1994 - 2002. Since 2003, the results are based on the quarterly LFS (Labour Force Survey). The effect of this change of source is an estimated reduction in the gender pay gap of 1 percentage point, following a comparison of data for 2002 from both sources.

Germany

From 1994 to 2001 the gender pay gap was calculated using the European Community Household Panel (ECHP), which is based on converted data of the German Socio-Economic Panel (GSOEP) at the DIW (Deutsches Institut für Wirtschaftsforschung) in Berlin). The known possible drawbacks of household surveys are the rather small sample sizes for employees and the quality of measured earnings and hours worked. Hence from 2002, the gender pay gap is calculated using earnings surveys out of official statistics as far as possible. Since the coverage of earnings surveys in Germany is limited to industry and only a few economic activities out of the service sector, the GSOEP is used to complete the coverage. Three reasons for the differences in 2001 are (a) differences in results for hourly earnings of SES and GSOEP; (b) the ECHP sample was only a sub-sample of the full GSOEP; and (c) the weighting of results of Earnings Survey and GSOEP also uses Mikrozensus distributions, not only the sample distributions of GSOEP, ECHP or structure of earnings surveys. There are no explanations for the change between 1998 and 1999. However the change of source in 2002 is estimated to have increased the gender pay gap by 1 percentage point, from 21

Greece

Since 2003, data are based on EU-SILC. The difference between the results for 2001 and 2003 is attributed to the change in data source.

Hungary

Figures for 2004 got revised. A significant decrease of the gender pay gap took place from 2002 to 2003 (from 16

Ireland

Since 2003 the figures are based on EU-SILC. Data for 2002 are not available.

Italy

Since 2004 data are based on EU-SILC. Data for 2003 are not available. Data for 2002 are available from the European SES 2002, giving a gender pay gap value of 21 per cent. However, this survey is limited in the coverage of economic activities (NACE sections C-K in the private sector) and the results are not comparable to the figures from the ECHP. In a comparison between the ECHP and SES data for 1995, the SES produced a gender pay gap figure which was 14 percentage points above the value from the ECHP.

Latvia

In 2004 the data source has been changed. Since 2004, data are based on hourly earnings of full and part time employees from the Quarterly Survey on Earnings and Employment. The survey covers all NACE sections and all size classes of enterprises. Data for 1998-2003 are based on hourly earnings for employees in the main job from the Survey on Occupations in October of the respective year. This survey covers full-time and part-time employees who had worked full month in October and their wage or salary was not influenced by absence.

Lithuania

The data for 2000-2006 are calculated on the basis on Quarterly Survey on Wages and

Salaries; sole proprietorships are excluded. Only full-time employees are included for 1995 - 1999. Between 1995 and 1996 the minimum wage was increased significantly, which particularly affected women, as a significant proportion of women earned the minimum wage. The change in the gender pay gap between 1998 and 1999 occurred because women's earnings increased more than men's. This followed government increases in earnings for employees in the educational sector, where there is a significant proportion of female employees.

Luxembourg

Data are based on total gross earnings for March of each year, for all employees covered by the social security system, with no age or working time restrictions, including cross-border employees (from neighbouring countries), working in the Grand Duchy of Luxembourg. Officials/employees working for international institutions or bodies established in Luxembourg are excluded. Gross earnings are wages and salaries (including, as appropriate, bonus) before deduction of income tax and wage-related mandatory social security contributions.

Netherlands

Data are based on annual earnings including overtime pay and non-regular payments. The national structure of earnings survey is used.

Norway

Data are based on full-time equivalent monthly salaries, not hourly earnings, using national statistics sources. NACE Rev.1.1 section H is included from 2001 on.

Portugal

Since 2004 the results are based on EU-SILC. The gender pay gap results for 2002 and 2003 have been calculated from a national sub-sample of the ECHP. The difference between the results for 2003 and 2004 is attributed to the change in data source.

Romania

The gender pay gap is expressed as ratio between average monthly gross earnings of women and average monthly gross earnings of men in October. Data source is the Annual survey in enterprises on earnings by occupation groups in October. The survey covers all NACE sections and all size classes. Data refers to employees (full-time and part-time) converted into FTEs.

Slovenia

Employees in public enterprises and employees in private enterprises with more than 2 employees are included.

Spain

Since 2004 GPG figures are based on EU-SILC. For 2002 and 2003 the data are based on earnings data from tax returns and hours worked from the labour force survey. The effect of the change in source after 2001 is estimated to be +3 percentage points. The tax data are from the Agencia Tributaria, which is a census of employees based on the annual income tax returns made by the employers. The population is composed of all employers, enterprises, companies and entities (included the public sector) that pay wages and salaries, except private households with employed persons. This source provides data classified by gender. All employees with any wage payment are included, irrespective of their working time during the year and the age of the employees. The units from Basque Country and Navarra are not included, but it is estimated that this does not have a significant effect on the gender pay gap figure. The decrease in the gender pay gap in 2003 is due to women's annual earnings increasing faster than men's annual earnings.

Sweden

Data are based on full-time equivalent monthly salaries, not hourly earnings, for employees aged 18-64. The figures exclude employees working less than 5 per cent of the full-time hours. The data source is the national structure of earnings survey.

United Kingdom

There is a break in series between 1996 and 1997. Until 1996, the European Community Household Panel (ECHP) was used for calculations. From 1997 onwards, the national panel, transformed into ECHP format, was used. From 2002, the national structure of earnings survey is used. An analysis of the data for 2001 indicated that the national structure of earnings survey produced a gender pay gap estimate which is +2 percentage points higher than the figure based on the national panel source.

A.5 Notes from countries 2007-2013

The following gives each countries notes on data collection regarding the reported values on the gender pay gap in the time period 2007-2013. The information is gathered from http://ec.europa.eu/eurostat/cache/metadata/Annexes/earn_grgpg2_esms_an1.pdf

Cyprus

Mean monthly earnings are used in the calculation of the GPG between the 4-yearly Structure of Earnings Survey (SES).

Czech Republic

Enterprises with 1+ employees are covered by data. The gender pay gap by age, economic control and working time are for NACE sections B to S.

France

The gender pay gap by age, economic control and working time are for NACE sections B to S.

Lithuania

A different method compared to the SES is used to calculate mean hourly earnings from which the gender pay gap is derived.

Slovenia

Data between the 4-yearly SES are estimated on the basis of the annual statistical survey, the Structure of Earnings Statistics, based on existing sources (Statistical Register of Employment and Tax Data) where data on part-time workers are excluded, irregular payments are included and only monthly earnings are available.

Sweden

From 2011, data are based on monthly earnings instead of hourly earnings. The population is aged 18-64 and work at least 5% of full time, excluding overtime hours.

United Kingdom

For 2010, the 2011 ASHE (the Annual Survey of Hours and Earnings) is used instead of the 2010 ASHE which is the basis for the 2010 SES. The 2011 ASHE captures annual earnings for the 12 months that ended on April 5 in reference year. Annual earnings data from the 2010 ASHE are for the period from 6 April 2009 to 5 April 2010. It was not possible to exclude non-regular payments from employees' gross pay. However, payments that relate to a different pay period than the period covered by the survey reference date. The gender pay gap by age, economic control and working time are for NACE sections B to S.

A.6 Tables of missing observations in descriptive statistics

A.6.1 Missing observations in table 2

	2007	2008	2009	2010	2011	2012	2013
Czech Republic	x						
Denmark	x						
Germany	x	x	x		x	x	x
Ireland	x						x
Greece	x	x	x		x	x	x
France	x	x					
Croatia	x	x	x				
Cyprus	x						
Latvia	x	x	x		x	x	x
Luxembourg					x	x	x
Austria	x	x	x		x	x	x
Slovakia	x						
Finland	x						
Sweden	x						
Norway	x						
Switzerland	x						

A.6.2 Missing observations in table 3

	2007	2008	2009	2010	2011	2012	2013
Czech Republic	x	x	x	x	x	x	x
Denmark	x	x					
Germany	x						
Ireland	x						x
Greece	x	x	x		x	x	x
France	x	x	x		x	x	x
Croatia	x	x	x				
Cyprus	x	x	x		x	x	x
Luxembourg					x	x	x
Austria	x	x	x		x	x	x
Slovakia	x						
Slovenia	x	x	x	x	x	x	x
Finland	x	x	x		x		
Sweden	x						
Norway	x	x					
Switzerland	x						

A.6.3 Missing observations in table 4

	2007	2008	2009	2010	2011	2012	2013
Czech Republic	x						
Denmark	x						
Germany	x						
Ireland	x						x
Greece	x	x	x		x	x	x
France	x	x	x		x	x	x
Croatia	x	x	x				
Cyprus	x						
Luxembourg					x	x	x
Malta						x	x
Austria	x	x	x		x	x	x
Slovakia	x						
Finland	x						
Sweden	x						
Norway	x						
Switzerland	x						