3. Data and Methods

In this chapter, the research method is discussed, followed by a discussion of the source of the data to be used in this study is presented. The chapter also specified the model and the data analysis techniques that are used.

3.1 Data

This empirical analysis is based on SOEP data. To obtain relevant quantitative variables the platform SOEP was used in the present study. Collected and processed by Deutsches Institut für Wirtschaftsforschung, a German Institute for Economic Research, SOEP database provides various socioeconomic variables for evaluation including household composition, earnings, satisfaction indication. The database also capture data on other related variables like education, attitude and personalities, family and social networks, level of satisfaction and occupational biographies (Goebel, et al., 2019). SOEP data is available in both cross-sectional and longitudinal databases. It started with a cross-sectional data structure but in 2012 polled data format, SOEPlong format, was introduced to reduce complexity and improve accessibility. The format harmonized different variables (Goebel, et al., 2019).

The database was built by a survey on private households from East and West German starting shortly after German reunification in 1984 and continued year after year. The database contains data of almost 30,000 individuals and about 15,0000 household. Over the years the data captured has been enhance by improving the mode of data collection and by improving the sample representation. When the survey was started the sample size included Sample A and Sample B, representing the population of private household and guest households respectively (Wagner, Frick, & Schupp, 2007). In 1990 with the enlargement of the German territory a new sample was introduced, Sample C, a longitudinal microdata, that allowed for the analysis from one regime to another. With the high wave of immigrants from

1985 to early 1990s the survey sample was increased to consider these immigrants. Therefore, in 1994/1995 a new sample E which involved about of 20,000 households was added and by 2000 Sample F with about 40,000 households was added to the study (Wagner, Frick, & Schupp, 2007). Sample G was created in 2002 to represent high-income households which also lead to the introduction of wealth measures. Over the years, there survey has also introduced additional representative samples which have assisted in adding observations and providing a tool for analysis (Wagner, Frick, & Schupp, 2007).

The data on family and individual social relationship covers different stages of the human life cycle including marriage, divorce, friendship and death. The income social security and taxes also covers information on different form of financial information like debts, assets, inheritance, taxes and pension. The income variable is used to gain information of all sources of household income throughout the year and the information is harmonized periodically to give annual income of the previous calendar year (Grabka, 2020). To check internal consistency, SOEP survey focused on both individual and household incomes (Goebel, et al., 2019). In the most recent years, it is captured under the wages and salary with variables like the current gross labor income and the current gross net labor income for the primary and secondary job (SOEP Group, 2020). SOEP data on work and employment consist of individual information with regards to different engagement aspects like first job, part-time engagements, work hours and attitude to work. Data like fitness, drug addiction, doctor visitation, insurance plan and sports are covered by the survey on health care. The data set also comprises of the demographic and population characteristics. Information like the sex, date, place and history of births, and the family relationship of each interviewer are captured.

Using datasets from the SOEP database, the research will evaluate the Gender Gap Pay, with regards to the manner in which each gender respond to these data sets. The SOEP

database was effective for the study because it offered comprehensive data which is up to date because of the annual updates. The data spans through the periods researched 2005-2009 and 2014- 2018. Other strengths of this data include the data format and the long duration of the studies or the surveys. The data is also fairly accessible, and this is important for the users. SOEP (n.d), however, emphasizes on the need to improve the user-friendliness of SOEP data, as well as its documentation.

3.2 Methodology

This paper seeks to examine the determinants of change in the gender wage gap across two time periods, between 2005-2009 and 2014-2018. More specifically, the ~~research~~ seeks to understand whether the gender pay gap has significantly changed over the study periods. Using regression and decomposition to analyses data from data from the German Socio-Economic Panel Study (SOEP), the ~~research~~ seeks to understand whether the gender pay gap has significantly changed over the study periods and what determinant have resulted to this change. ~~Some of~~ these determinants include include age, wage, and occupation, which form the independent variable evaluated against gender wage pay, the dependent variable.

The study also investigates the gender wage gap in variable occupations by first investigating the statistical significance of the relationship between the variable 'earning' and the gender pay gap. The model is based on the enterprise characteristics influencing earnings of individual's characteristics, analyzing the gap in hourly earnings. The variable 'earnings' focuses on the wages of the females and the males. The relevant wage gaps were calculated by obtaining the differences between the male and the female wages in each category to give the possibility to confirm or to desagree with the fact of wage discrimination, stating women have lower incomes, lower incomes and less advantageous terms of employment and position than men.

3.3 Equation model

~~The data from the SOEP is analyzed in two stages~~: a regression analysis and a decomposition analysis in the equations as. Before decomposition, it is important to determine if the differences in rewarding human capital factors of two groups is statistically significant. For this research, the Mincer wage equation was used to create an econometric model for the pay rate which includes all explanatory variables.

$$\ln w_i = \beta_0 + \beta_1 AGE_i + \beta_2 EDU_i + \beta_3 EXP_i^2 + \sum_{i=5}^{n} \beta_1 X_i + \varepsilon_i \qquad (2)$$

Where the variables are, i - individual, w - wage, $ßo$ is constant and $ßk$ are paramenters covering observed characteristics. In the calculation we can assume, that experience is exogenous ε

The equation above evaluate pay rate with factors like age, education, experience and gender. As we have information about wages of women, men, and their individual characteristics, this will be done by a running a regression with wage gap as the dependent variable, and age, gender, occupation, income, and family status as the predictors/ independent. If the difference pay rate between the two groups, male and female. is statistically significant then it is decomposed.

Regression is performed separately on men ($\ln y_i^m$) and women ($\ln y_i^f$) with the equations below:

$$\ln y_i^m = \beta_0^m + \sum_{k=1}^{k} x_{ki}^m \beta_k^m + u_i^m \qquad (3)$$

$$\ln y_i^f = \beta_0^f + \sum_{k=1}^{k} x_{ki}^f \beta_k^f + u_i^f \qquad (4)$$

This model was used to investigate the statistical significance of the relationships between the dependent and the independent variables The variables in the equation are *yi* is the natural log of hourly earnings for observation I, *xki* are explanatory variables for job and enterprise characteristics influencing earnings of individuals I, *ßo* is constant and *ßk* are paramenters covering observed characteristics, *ui* is a disturbance term for observation, M is for male and F is for female.

Decomposition methods, have been widely use to study group differentials. The decomposition method developed by Blinder and Oaxaca was chosen for this study to investigate the variables surrounding pay gaps between male and female. Its use was significant for this research because it allow in the breakdown group difference with regards to the study variables (Sinning, Hahn, & Bauer, 2008). In this research the analysis is based on the "twofold" Blinder□Oaxaca decomposition (Yun, 2004). With F for female and M for male, using the twofold decomposition, the mean difference for the dependent variable between the two groups can be computed with the equation below:

$$\overline{Y}^{M} - \overline{Y}^{F} = \underbrace{\left[\left[\overline{F\left(X_{M}'\beta^{*}\right)} - \overline{F\left(X_{F}'\beta^{*}\right)}\right]\right]}_{Explained} + \underbrace{\left[\left[\overline{F\left(X_{F}'\left(\beta_{M} - \beta^{*}\right)\right)} - \overline{F\left(X^{F\prime}\left(\beta^{*} - \beta_{F}\right)\right)}\right]\right]}_{Unexplained} \quad (5)$$

Y, *X* and *β* standing for N *X 1*, *N X K* and *K X 1* vector of dependent variable outcomes, N *X 1* This equation is based on the assumption that the dependent variable, independent variables and coefficients respectively.

The first part of the equation forms the present a situation in which women characteristic were rewarded like men, while the second part focuses on remuneration effect, where the pay difference between the two groups arise from discrimination. The problem with this approach has been on the validity of functional assumption. There are characteristic that are obvious for men but might never be observed on female. Therefore, by making this assumption the data from this decomposition can be misleading (Djurdjevic, & Radyakin, 2007).

To an answer the question on how gender pay gap has developed in Germany between the two observation periods, a *t* test in carried out. Paired *t* tests or paired sample *t* test is used when we are interested in the difference between two variables for the same subject. Often the two variables are separated by time. In this case we use a *t* test to compare the means of the wage gaps between 2005-2009 and 2014- 2018. It is used to establish if there is a difference in the gender pay gap between the two time points. To calculate the gender pay gap, measure of what women pay relative to men, the average women wage for the duration will be divided my men's wage. The ratio will then be used in the Paired sample t test. The dependent variable in this case will be the gender pay gap. The paired t-test will determine if the mean difference between the gender pay gap for 2005-2009 and for 2014-2018 is statistically significant different to zero. This is to determine if the means of the wage gaps are significantly different between these two periods. As the values of the wage gaps are compared between the two periods, the intermediate or contingent variable is gender. The samples used in this case are not independent, they are just same samples taken at different times. To answer the question on how the gender pay gap situation has changed in Germany comparing to the certain period selected, correlational test should be assessed. Person test providing statistical relationship between two continuous variables can be applied in this

case. Within this test coefficient is an important variable, where +1 is positive relationship and -1 is the negative relationship.

References

Djurdjevic, D., & Radyakin, S. (2007). Decomposition of the gender wage gap using matching: an application for Switzerland. *Swiss Journal of Economics and Statistics, 143*(4), 365-396.

Leythienne, D., & Ronkowski, P. (2018). *A decomposition of the unadjusted gender pay gap using Structure of Earnings Survey data*. Publications Office of the European Union.

Grabka, M. M. (2020). *SOEP-core v35-codebook for the $ pequiv file 1984-2018: CNEF variables with extended income information for the SOEP* (No. 772). SOEP Survey Papers.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., et al. (2019). The German Socio-Economic Panel (SOEP). *Journal of Economics and Statistics , 239*(2), 345–360.

Sinning, M., Hahn, M., & Bauer, T. K. (2008). The Blinder–Oaxaca decomposition for nonlinear regression models. *The Stata Journal*, *8*(4), 480-492.

SOEP Group. (2020). SOEP-Core v35 – PGEN: Person-Related Status and Generated Variables. *SOEP Survey Papers*.

Spiess, M. / Kroh, M. (2007): Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2006), DIW Data Documentation (forthcoming). Berlin.

Wagner, G. G., Frick, J. R., & Schupp, J. (2007). The German socio-economic panel study (SOEP): Scope, evolution and enhancements. . *Schmollers Jahrbuch - Journal of Applied Social Science Studies, 127*(1), 139-169.

Yun, M. S. (2004). Decomposing differences in the first moment. *Economics Letters*, 82(2), 275–280.