Universidad del País Vasco Euskal Herriko Unibertsitatea

EKONOMIA ETA ENPRESA FAKULTATEA
FACULTAD DE ECONOMÍA Y EMPRESA

DEGREE: ECONOMICS AND BUSINESS ADMINISTRATION

2019/2020

YOUTH AND GENDER IN THE LABOR MARKET

Author: Nora de la Fuente Gómez

Director: Javier Fernández-Macho

Bilbao, 24th June, 2020

Table of contents

Table of acronyms

FDD: Final Degree Dessertation

EPA: Active Population Survey/ Encuesta de Población Activa

INE: National Statistics Institute/ Instituto Nacional de Estadística

ILO: International Labour Organization

PwC: Price Waterhouse and Coopers & Lybrand

CSV: Comma-separated values

LRM: Linear Regression Model

MLR: Multiple Linear Regression

OLS: Ordinary Least Squares

CDF: Cumulative Distribution Function

List of figures and tables

YOUTH AND GENDER IN THE LABOR MARKET

ABSTRACT

This is a Final Degree Dissertation (FDD) that examines the gender inequalities in the different possibilities of labor insertion of young people in Spain.

The problem of unemployment is one of the most relevant issues of citizens. In the theoretical part, the literature related to gender and youth in the Spanish labor market is synthesized while in the empirical part, the impact of the chosen factors on the employability situation is examined. For this purpose, different econometric models are used, which are explained and studied in the methodological part of the paper.

To carry out the analysis, the data from the Active Population Survey (EPA) for the last quarter of 2019 provided by the National Statistics Institute (INE) has been considered, from which the most significant variables have been chosen.

1. INTRODUCTION

1.1. Why do this Final Degree Dissertation?

Throughout my university degree, the subjects that I liked the most have always been those related to numbers. I find it interesting to analyze this situation by applying statistics to obtain conclusions related to gender in the work market.

Historically, through our ideas, practices, and relationships, we have built a social system that we denominate patriarchy. This word in its literal sense means "government of the fathers". It is a historical and social construction in which authority is exercised by the male. To change this, women have fought for their rights throughout history, reaching the place they deserve in society.

The struggle for gender equality has made remarkable progress since the 20th century, with significant victories such as the achievement of universal suffrage. However, there are areas in which gender inequality is still evident, as it is in the labor market. In the case of the European Union there is a great difference between countries in this area, with countries such as Sweden and Denmark as the ones that most practice equality, the situation in Spain is better than the general average but below it in some specific aspects, like time management or economic differences.

Therefore, I consider that the situation of women in the labor market today is interesting to be studied at a national level, and I, as a young woman, felt identified with the subject of the project. Even though this is a topic that has been investigated, I believe that it is essential to investigate the updated study of the factors that most affect this, so I thought that this research project would be useful to me and would add value to my knowledge. Furthermore, with my studies finished, I will jump into the labor market next year, so this study is of interest to me.

1.2. Background

a. Youth in the labor market

Unemployment remains the greatest concern of Spaniards.

This can be seen in details such as the high unemployment rate, which is still around 14%, or the under-utilization of the labor force, which would affect 23% of workers, some 5.4 million in the last quarter of 2019, as published in the Active Population Survey conducted by the National Institute of Statistics

(*Encuesta de Población Activa, cuarto trimestre de 2019*, 2020) ("Unemployment and labor underutilization," 2019).

In calculating the number of people affected by this "underutilization of the workforce", the International Labor Organization (ILO) takes into account the unemployed, those with part-time employees who want a full-time job and those who have given up on finding a job but are willing to return to the labor market.

In the report released by Eurostat, the European statistical office shares that in November 2019, the highest young unemployment rates were recorded in Greece (with 32,5% in the third quarter of 2019), followed by Spain with 32,1% and ahead of Italy (28,6%) (*Euro area unemployment at 7.5%*, 2019).

The youth unemployment rate is the number of people aged 15 to 24 unemployed as a percentage of the labor force of the same age. It should only include people under 25 years old. Specifically, citizens between 16 and 24 years old, according to the Eurostat platform.

Young people are entering the labor market later and later, either because of long periods of study or because of the difficulty of finding their first job. Therefore, "the number of young people who have not yet worked by the time they reach the age of 30 is beginning to be relevant, even if it is a minority" ("Tasa de paro juvenil en España," 2018). Still, the youth unemployment rate includes young people between 16 and 24 years old.

According to the report published by the National Institute of Statistics for the last quarter of 2019, employment has fallen among children under 25, with 32,100 fewer employed people between 16-19 and 46,600 fewer employed people between 20-24 years old (*Encuesta de Población Activa, cuarto trimestre de 2019*, 2020, pp. 2).

b. Gender perspective

To focus on inequalities between the sexes it is essential to make some specifications. The word "sex" refers to the biological, physical, anatomical, and physiological characteristics of human beings, which define them as men and women and it is determined by the nature with which one is born, while "gender" refers to the social values and opportunities that society assigns to individuals in a differentiated way as corresponding to each sex.

It should be added that, traditionally, certain professions have been associated with the male and the female genders. These patterns are exceedingly difficult to be changed, as they are very deeply introduced in society, and make gender inequality grow. In fact, there are more women in studies and professions with lower salaries (*Coste de oportunidad de la brecha de género en el empleo*, 2020, pp. 8).

Also, stereotypes are another type of factors that have a direct impact on the barriers to women's professional career decisions and on the way they participate in the labor market (*Coste de oportunidad de la brecha de género en el empleo*, 2020, pp. 9).

c. Sex discrimination in employment

The unemployment situation affects women more, as it is women who are mostly in part-time employment, and therefore they are the most affected by underemployment. "Spanish women tend to be employed more on a part-time job than men" (*Coste de oportunidad de la brecha de género en el empleo*, 2020, pp. 12).

According to the data released by the National Institute of Statistics, in Spain, there were more than two million women in part-time work compared to 762,000 men in the last quarter of 2019 (*Encuesta de Población Activa, cuarto trimestre de 2019*, 2020, pp. 18).

Furthermore, according to the report made by PwC in 2018, the salary gap between men and women is higher in part-time contracts than in full-time ones. Specifically, the hourly wage difference between the genders with part-time work was 1.8 euros, that is, women were paid 14.9% less per hour than men, at part-time (*Coste de oportunidad de la brecha de género en el empleo*, 2020, pp. 23).

Gender inequality in the labor market continues to persist and it can be seen reflected in different aspects. Women in the EU earn on average almost 15% less per hour than men. Spain is in an intermediate situation with 13.9% ("Brecha salarial de género en Europa," 2020).

Women have represented different revolutions throughout the 20th century and during the first years of the 21st century. Throughout this time, in most parts of the world, they have obtained the right to vote, as well as access to education and employment.

However, real equality is far from achieving it, as inequality and discrimination gaps still exist around the world.

In the academic year 2016-2017, women in university were 55% compared to 45% of men, and in the case of graduated students, the number of women increased to 60% ("Mujeres graduadas en educación superior," 2017).

According to the data of the National Institute of Statistics, in the rate of university education or higher, prepared by PwC, in 2018 the proportion of women between 25 and 64 years old was higher than that of men (*Higher proportion of women than men with a high education level*, 2018).

But this situation is not reflected in the labor market. As mentioned before, women still have lower occupation rates and higher unemployment rates and part-time jobs.

Based on the data of the National Institute of Statistics on employment rates based on education levels, in 2018, the gender gap was 12.1 points, which was higher among the basic education groups. This data shows the importance of education in the participation in the labor market. But it is not enough. Spain has one of the lowest female employment rates in the European Union, below 60% for the population aged 20-64 (*Estadísticas de empleo*, 2019).

Apart from that, motherhood affects women very much, while for men it does not. According to a study carried out by Foretica, a group of 66 big companies and corporate social responsibility professionals, maternity penalizes working careers for 70.6% of women. The work stoppage suffered by women's careers during the last weeks of pregnancy and the recovery and breastfeeding periods has a direct impact on their careers, their working conditions, and their salaries ("La maternidad penaliza la carrera profesional para el 70,6% de las mujeres," 2019).

In the same way, women continue to spend more time on family and domestic tasks such as childcare and housecleaning and consequently have less availability for paid work. The need to combine work with these activities also affects the type of work schedules (full versus part-time) (*Coste de oportunidad de la brecha de género en el empleo*, 2020, pp. 9).

The Foretica report states that there is a double segregation within the business system. On the one hand, the horizontal segregation, which refers to the concentration of women and men in different areas of activity, positioning women in low-wage and low-skill occupations. This generates a higher precariousness in women's employment, which is related to maternity and is reflected in the number of women who work

part-time, which currently stands at two out of every three part-time jobs. On the other hand, vertical segregation gives rise to the well-known glass ceiling, a metaphor used to represent the invisible barriers that women face when trying to reach positions of higher responsibility.

Where equity fails most is in the distribution of power, in which equality is analyzed in decision-making and which continues to be mainly in the hands of men. Concerning the presence of women in listed companies, "there are 268 female directors throughout the continuous market, representing 20.3% of the total of 1,320 members on the Boards of Directors" (Atrevia, 2018, pp. 6).

"Women are underrepresented in the labor market in participation and occupation issues. On equal terms, women have fewer opportunities." By the end of 2019, the male unemployment rate was over 12%, while 16.3% of the female workforce was unemployed ("Tasa de paro EPA," 2019).

Factors such as the wage gap, under which women are paid less than men for doing the same work and having the same responsibility, are important elements that make it difficult to achieve equality.

According to provisional data for 2017, "the gender gap in hourly wages increases with age, from a value of 8.5 in the 25-34 age group to a value of 22.5 in the 55-64 age group" ("Brecha salarial de género en salarios por hora," 2017).

In turn, according to the report prepared by the PwC consulting firm, age is one of the variables with a strong influence on the wage gap. According to the results for the year 2017, the wage difference increases the older the workers are. This result shows a smaller difference between the young population (*Coste de oportunidad de la brecha de género en el empleo*, 2020, pp. 22).

2. WORK OBJECTIVES

The main objective of this work is to examine whether gender inequality exists among young people when they enter the labor market in Spain. In particular, it is intended to analyze the causes or reasons that determine the employment situation of an individual, from the gender perspective, and referring to young aged ones.

When taking the groups of young people aged 16-24 as a reference, the number of individuals with a job was very small. For this reason, it has been decided to extend the range to young people up to 34 years of age. In this way, it has been possible to include more variables to consider if they affect when having a job or not having it.

The fundamental questions that are sought to be resolved are: does gender inequality exist among young people when they enter the labor market? What are the factors that most affect this?

The main objectives of the project are the following:

- ✓ Setting the socioeconomic problem and the related variables which are the object of the study.
- ✓ Obtaining the data from available statistical sources.
- ✓ Describing the relevant variables with the help of graphical and statistical tools.
- ✓ Specifying an appropriate model and estimating it with the available data.

3. MODEL

To carry out the analysis of the relationship between youth and gender in the labor market the Active Population Survey has been used, which is obtained from the INE (National Institute of Statistics) online source. "This is a continuous, quarterly research aimed at families, whose main purpose is to obtain data on the labor force and its various categories (employed, unemployed), as well as on the population outside the labor market (inactive)." (National Institute of Statistics, 2020)

The database used to carry out the work has been elaborated from the microdata of the mentioned Active Population Survey. The original data file was in CSV format, so it has been transformed to be able to be read. This process has been done through excel by importing the data from the CSV file.

After that, a restriction has been done filtering the data by age criterion, since the age of interest is the one of the individuals between 16 and 34 years old. Besides, the explanatory variables that will be mentioned later have been chosen. Apart from that, the data has been organized in such a way that the statistical software would not be confronted with blank cells. When it was over, the file was ready to be read by the software.

For an appropriate data processing and econometric model estimation, the spreadsheet file was imported into the statistical software Gretl. Gretl is an econometric software designed for the statistical analysis and estimation of econometric models. Since this has been the program used during the semester in the econometrics subject, it has been the one selected.

Once inside Gretl, the last step was to create the dummy variables of the qualitative variables used. To include qualitative variables in an econometric model, it is necessary to create dummy variables. These are variables that take two values: 1 if the observation presents the characteristic in question and 0 otherwise. This process is done within Gretl, editing the attributes, and treating these variables as discrete so that dummy variables could be added to the selected discrete ones. This step is necessary to be able to estimate the models.

With the quantitative and qualitative data obtained in the Active Population Survey of the National Statistics Institute during the last quarter of 2019, the dependent and explanatory variables described below have been chosen.

3.1. Variables

a. Dependent variables

There are different ways to carry out an analysis that allows us to find out which factors influence the employment situation (being employed or not).

Given the purpose of the work, previously exposed, it is opportune to use a variable that shows the employment situation of the individual. For this purpose, a binary dependent variable has been taken as a reference, which has been called “employment situation”.

In the variables exposed in the data of the National Institute of Statistics, two variables that could be useful for this were found.

Specifically, on the one hand, the variable "TSIDAC" was found, which refers to the situation of self-perceived activity in which the individual was in the week before the reference week, and on the other hand, the variable "TRAREM", which tells whether or not the individual has carried out paid work during that week.

The two variables are binary: the first is divided into the groups "working" and "looking for a job" and the second, instead, into "yes" and "no".

For this study the dependent variable that considers whether the individual has performed paid work during the week before the reference week ("TRAREM") will be used, since explanatory variables related to the other variable, that is, the search for work, are going to be used. The variable that will be used takes the values 1 if the individual did perform paid work and 2 if he or she did not. It has been transformed into a dummy variable, that is, it takes the value 1 if the paid work has been done and 0 if not.

This variable will show the probability of being employed that an individual has depending on the explanatory variables used and, in this way, it will indicate how each one of them influences.

b. Explanatory variables

Based on both the survey conducted by the INE and the aspects commented on at the presentation of the case study, a series of variables considered appropriate to examine and on which the empirical model will be focused have been chosen. These variables will help measure the variation in the employment situation of the individuals, each of them affecting in a greater or lesser degree.

For the analysis, the following variables have been taken: age, number of hours per week dedicated to work, age at which the highest level of education was achieved, the average number of months the individual has been looking for a job, sex of the individuals, level of education, nationality, the type of work schedule, the marital status and if the individual has been looking for a job in the 4 previous weeks.

The first variable, AGE, collects the years of the individuals. It has been decided to focus on all those people aged between 16 and 34, since this allows analyzing the impact of more variables. It is divided into four groups: age 16 is for individuals aged 16-19; age 20 is for those aged 20-24; age 25 for those aged 25-29 and age 30 for those aged 30-34. This is the age range considered relevant at the time of studying young people in the employment situation in Spain and this variable is going to be considered as quantitative.

The next one, HOURS, refers to the number of hours that the individual regularly dedicates to work per week. This is a quantitative variable and it is considered relevant since it is closely related to the employment situation of the individual. This variable was expressed in HHMM format, so it has been converted into hours with a decimal point, with the help of excel.

The following variable, AGEST, represents the age at which the individual reached his or her maximum level of studies. It may be interesting to deduce when each individual incorporates himself or herself into the labor market. It is also a quantitative variable.

The last quantitative variable, LENGTH, indicates the time in months that the individual has been looking for work. This variable has been transformed into a quantitative one, taking the average number of months the individual has spent looking for it.

The next variable, SEX, is one of the most important to consider given the topic of the project. It is a binary variable that refers to the nature of the individual. On the platform used to carry out the analysis, this variable is marked as "discrete" so that the program itself can then transform it into the corresponding dummy variables. Its study will help to know if there is a differentiation between the genders.

Continuing with the discrete variables, the EDUCATION variable indicates the level of education of the individuals. It is divided into 5 groups, being the following: 1 ("P2"=primary education); 2 ("S1"=first stage of secondary education); 3 ("SG"=second stage of secondary education, general orientation); 4 ("SP"=second stage of secondary education, professional orientation); 5 ("SU"=higher education). The variables "illiterate" and "incomplete primary education" have been eliminated since they did not seem to be relevant. The platform will give the respective dummy variables for each one of them. The education variable will indicate the importance of education when it comes to being employed or not.

Another variable, NATIONALITY, indicates the country to which the person belongs. It is made up of 2 groups: 1 for Spanish individuals or those with dual nationality and 2 for foreigners. As they are also discrete, the platform will make the dummy variables.

The other important variable is the type of work schedule of the individual. It is a qualitative variable and takes the value 1 if the workday is complete, 2 if it is partial, and 3 if the individual is not working. This variable will indicate the difference between the two cases when it comes to having a job. The platform will give the corresponding dummy variables.

Regarding the MS variable, which refers to the marital status of the individual, it has been divided into only two groups. At first, it was classified into four groups (single, married, widowed, and separated or divorced). To make the analysis simpler and clearer, in this case, the last two have been added within the single group, leaving the variable in two categories: 1 if the individual is single, widowed or separated and 2 if the person is married.

The last variable, SEARCH, indicates if the individual has been looking for a job in the last 4 weeks. Through this variable, the involvement of the individual when looking for a job will be considered. It takes the value 1 if the answer is yes, 2 if it is not and 3 it seems to refer to those individuals that are working. With this variable, certain problems can be deduced when estimating the model, since the third group refers to those who “actually work” and it seems contradictory when analyzing the dependent variable of the employment situation of the individuals (if they are working or not). For this reason, it has been decided to include the third group in the second one, that is, “not looking for a job”.

Table 1: Information about the explanatory variables

Explanatory variables	Type	Definition
AGE	Quantitative	Age range of the individuals
HOURS	Quantitative	Number of hours per week regularly spent on the work
AGEST	Quantitative	Age at higher level of studies
LENGTH	Quantitative	Average number of months looking for a job
SEX	Qualitative	2 dummy variables: Man (M) → 1 Yes ; 0 No Woman (W) → 1 Yes ; 0 No
EDUC	Qualitative	5 dummy variables: Primary (P2) → 1 Yes ; 0 No Secondary, first stage (S1) → 1 Yes ; 0 No Secondary, second stage, general (SG) → 1 Yes ; 0 No Secondary, second stage, professional (SP) → 1 Yes ; 0 No Higher education (SU) → 1 Yes ; 0 No
NAT	Qualitative	2 dummy variables: Spanish or double nat (SP) → 1 Yes ; 0 No Foreigner (FG) → 1 Yes ; 0 No
WORKSCH	Qualitative	3 dummy variables: Complete (CP) → 1 Yes ; 0 No Partial (PT) → 1 Yes ; 0 No Not working (NT) → 1 Yes ; 0 No
MS	Qualitative	2 dummy variables: Single, widowed or separated → 1 Yes ; 0 No Married → 1 Yes ; 0 No
SEARCH	Qualitative	3 dummy variables: Searching → 1 Yes ; 0 No Not searching → 1 Yes ; 0 No

Source: *Own elaboration*

3.2. Methodology

To begin the analysis, it is necessary to select the variables to be included in the model, as some of the previously selected ones may not be relevant.

Some variables can be perceived as linearly dependent on other explanatory variables. These variables cannot be included together in the model, as this would create multicollinearity problems.

Given that the variable WORSKSCH, the one that examines the work time of the individual, is composed of a third group named "not working", seems to create problems when examining the dependent variable that analyzes just that. That is why it has been decided to remove it from the analysis.

After having selected the variables that will be used for the study, the model from which the impact of gender on young people in employment will be obtained, is going to be defined.

In the fourth quarter of 2019, the sample in Spain had 163,152 individuals. By restricting the model based on age criteria, since only young individuals between 16 and 34 years old are the ones to be analyzed, the model leaves a current sample of n=28,341 individuals.

In this case, a binary dependent variable model will be used. The model is used to explain situations in which the variable of relevance (in this case the employment situation) is binary, that is, it can only take two values (being employed or not).

Being Y_i=1 if the individual is employed

Y_i=0 if it is not

During the university studies, the model used has been the Linear Regression Model (LRM). The application of this model may not be suitable for the explanation of problems in which the expected response takes only two values; in this case, being employed or not. That is why certain models that are useful for solving problems with a binary dependent variable will now be considered.

To introduce the different models that can be useful for the study, several examples are going to be made.

By having a model,

$$Y= \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u_i$$

Recognizing the unbiasedness assumption, where the error term u_i has conditional mean zero, $E(u|x_1,\dots,x_k)=0$, this model can be expressed as,

$$E(y|X)= \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

The basic point of the explanation of the linear regression model with binary dependent variable is that this model considers the probabilities of an event happening, meaning, it helps to know how each factor affects the probability of success of the projects.

To make the estimation, the simplest procedure is to apply the **Linear Probability Model** (LPM).. This model is a variation of the Linear Regression Model focused on the explanation of qualitative events and it can be interpreted in probabilistic terms. The key point is that the expected value of *y* is always equal to 1, expressed in another way: P(y=1|X)=E(y|X), its probability of success, which in our case means being employed. In this way, the model says that the probability of success is a linear function of the variables *x,* which can be expressed as follows:

$$P(Y = 1/X_2,X_3,\dots X_k)= \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k$$

The common interpretation of β_i can no longer be used, since changes in *y* are not continuous with changes in X. The regression coefficient β_i represents now the variation in the probability of Y=1 when X_i varies by one unit, keeping all other explanatory variables constant:

$$\Delta P(y=1|x)= \beta_i \Delta x_i$$

When using binary independent variables, the coefficient obtained "measures the difference that is predicted for the probability concerning the group base".

Also, since the probabilities have to sum 1, it can be said that P(y=0|X)=1-P(y=1|X), which is also a linear function of the X variables. Linear Probability Model (LPM) is called like this because of the linearity of the β parameters. For its estimation, the Ordinary Least Squares (OLS) method is used, as a common model of Multiple Linear Regression (MLR).

Disadvantages of the LPM

Although the coefficients obtained from the Linear Probability Model have useful interpretations, the model has some disadvantages that make it weak compared to other models.

The first weakness is that, with certain combinations of the independent variables, it is possible to have results where the probability is negative or greater than one. These results create an interpretation problem and require a change in the way the model is understood and estimated.

The second weakness of the model is that probability cannot be considered to be related linearly to the independent variables for all values. Certain effects on the dependent variable vary in magnitude as the independent variable increases and the LPM is not able to capture those changes.

Apart from this, the variance of the perturbation is heterocedastic, which means that the size of the error term differs across values of the independent variables. To satisfy the regression assumptions and be able to trust the results, the residuals should have a constant variance, and this does not happen.

Descriptive analysis of the variables of relevance

Before developing the model, the correlations between the employment situation and the sex and age variables are shown below, based on the research carried out by María Romero Meléndez (*LAS GESTORAS DE EXPLOTACIONES AGRARIAS: análisis de sus características mediante modelos MLP y PROBIT*, 2013), as it is interesting to examine the relationships with the variables that appear to be the most significant. Although the model will include more variables, only the ones that seem to be the most relevant are shown here.

As it has already been mentioned, the endogenous variable "EMPLOYMENT SITUATION" takes the values 1 (if the person is employed) and 0 (if not). That is why the interpretation of the signs obtained will be the higher or lesser probability for the person to be employed.

EMPLOYMENT SITUATION AND SEX:

Table 2: LPM two variables

Modelo 1: estimaciones MCO utilizando las 28341 observaciones 1-28341
Variable dependiente: DESitu_1

Variable	*Coeficiente*	*Desv. típica*	*Estadístico t*	*valor p*	
const	0,45772	0,00412775	110,8885	<0,00001	***
Woman	-0,0552483	0,0058738	-9,4059	<0,00001	***

Media de la var. dependiente = 0,430436
Desviación típica de la var. dependiente. = 0,495146
Suma de cuadrados de los residuos = 6926,48
Desviación típica de los residuos = 0,494384
R^2 = 0,00311215
R^2 corregido = 0,00307698
Grados de libertad = 28339
Log-verosimilitud = -20248,5
Criterio de información de Akaike = 40501
Criterio de información Bayesiano de Schwarz = 40517,5
Criterio de Hannan-Quinn = 40506,3

Dependent variable: Employment situation

As can be seen in the results, using the 28,341 observations, it is less likely for a woman to be employed than a man, since the coefficient that corresponds to this variable is negative. Specifically, a woman is 5.52% less likely to be employed than a man.

EMPLOYMENT SITUATION AND AGE:

Table 3: LPM two variables

Modelo 2: estimaciones MCO utilizando las 28341 observaciones 1-28341
Variable dependiente: DESitu_1

Variable	*Coeficiente*	*Desv. típica*	*Estadístico t*	*valor p*	
const	-0,604555	0,0114475	-52,8111	<0,00001	***
AGE	0,0455089	0,000490442	92,7917	<0,00001	***

Media de la var. dependiente = 0,430436
Desviación típica de la var. dependiente. = 0,495146
Suma de cuadrados de los residuos = 5328,99
Desviación típica de los residuos = 0,433641
R^2 = 0,23303
R^2 corregido = 0,233003
Grados de libertad = 28339
Log-verosimilitud = -16533,1
Criterio de información de Akaike = 33070,3
Criterio de información Bayesiano de Schwarz = 33086,8
Criterio de Hannan-Quinn = 33075,6

Dependent variable: Employment situation

In this case, the probability of being employed is higher as age increases. As mentioned above, the age range is between 16 and 34 years old, so it is within these ages that the possibility of being employed increases the older the individual is, by 4.55%, specifically.

4. EMPIRICAL RESULTS

An OLS regression has been performed. The dependent variable is employment situation (it takes two values, 1 if the individual is employed and 0 if it is not) and all the explanatory variables have been included (having discarded the previously mentioned variable):

Despite the possible problems mentioned above, several conclusions can be drawn from this model. In this case, the sign of the coefficients is more relevant.

Table 4: LPM Results

Modelo 3: estimaciones MCO utilizando las 28341 observaciones 1-28341
Variable dependiente: DESitu_1

Variable	*Coeficiente*	*Desv. tipica*	*Estadistico t*	*valor p*	
const	-0,0560341	0,0173962	-3,2210	0,00128	***
AGE	0,00936877	0,000426313	21,9763	<0,00001	***
AGEST	0,00124753	0,000707582	1,7631	0,07789	*
LENGTH	0,00237734	0,000186989	12,7138	<0,00001	***
HOURS	0,0193177	0,000111614	173,0767	<0,00001	***
Man	-0,000296569	0,00326788	-0,0908	0,92769	
Mar	-0,0561531	0,00566195	-9,9176	<0,00001	***
For	0,0277129	0,00572116	4,8439	<0,00001	***
Searching	-0,199166	0,00605381	-32,8994	<0,00001	***
P2	-0,0471942	0,0106075	-4,4491	<0,00001	***
S1	-0,042516	0,00635264	-6,6927	<0,00001	***
SG	-0,0507272	0,00587903	-8,6285	<0,00001	***
SP	-0,00413971	0,00621924	-0,6656	0,50565	

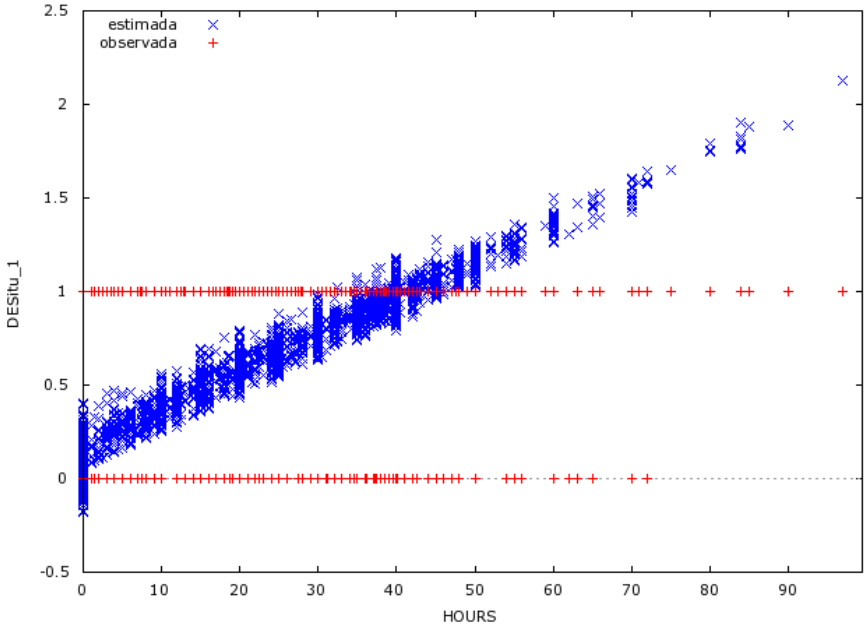
Media de la var. dependiente = 0,430436
Desviación típica de la var. dependiente. = 0,495146
Suma de cuadrados de los residuos = 2043,9
Desviación típica de los residuos = 0,268609
R^2 = 0,705834
R^2 corregido = 0,70571
Estadístico F (12, 28328) = 5664,29 (valor p < 0,00001)
Log-verosimilitud = -2953,49
Criterio de información de Akaike = 5932,97
Criterio de información Bayesiano de Schwarz = 6040,25
Criterio de Hannan-Quinn = 5967,49

Dependent variable: Employment situation

As can be seen here, broadly speaking, it can be said that the probability of being employed increases when the AGE, AGEST, LENGTH and HOURS quantitative variables increase and when the individual is foreigner. In contrast, this probability decreases when the individual is a man, is married, is looking for a job, and with an education lower than the superior education. All this must be interpreted *ceteris paribus*, meaning, keeping all the other variables constant.

Like it has been commented before, the LPM model presents several limitations. As it can be seen in figure 1, the linear specification can produce probability predictions that are less than 0 and greater than 1, which is not appropriate since the probability must be limited to between 0 and 1. In this case, the example has been taken against the exogenous variable HOURS.

Figure 1: Employment situation against Hours (LPM)



That is why it is more appropriate to use a Probit or Logit model. Despite being more complicated when estimating and presenting the results and interpreting them, they are more useful for carrying out this analysis as they better model the probability of success as a function of the explanatory variables. In this case, the estimated values of the probabilities are between 0 and 1, solving the problem previously mentioned in the Linear Probability Model.

In the LPM is modeled that the probability of Y=1 is linear:

$$Pr(Y = 1|X) = \beta_0 + \beta_1 X$$

What we want is this:

i. $Pr(Y=1|X)$ to be increasing in X for $\beta_1 > 0$, and
ii. $0 \leq Pr(Y=1|X) \leq 1$ for all X

This requires using a nonlinearity functional form for the probability.

In **Probit** regression, the cumulative standard normal distribution function $\Phi(z)$ is used to model the probability that Y=1. The Probit regression model is:

$$E(Y|X) = Pr(Y=1|X) = \Phi(\beta_0 + \beta_1 X)$$

Where Φ (.) is the cumulative normal distribution function and $z = \beta 0 + \beta 1X$ is the z-value or z-index of the Probit model.

In **Logit** regression, the probability of Y=1 is modeled, given X, as the cumulative standard logistic distribution function:

$$Pr(Y=1|X) = F(\beta_0 + \beta_1 X)$$

Where F is the cumulative logit distribution function:

$$F(\beta_0 + \beta_1 X) = \frac{1}{1+e^{-(\beta 0+\beta 1X)}}$$

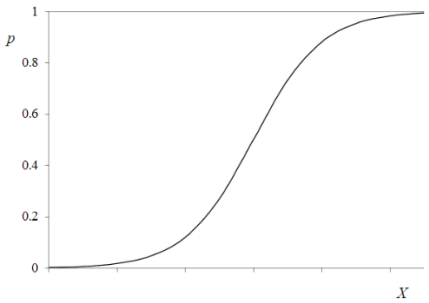
The idea is similar to Probit regression except that a different Cumulative Distribution Function (CDF) is used:

$$F(x)=\frac{1}{1+e^{-x}}$$

Is the CDF of a standard logistically distributed random variable.

Being this a logistic function, since the dependent variable is binary, the relation between the dependent and the explanatory variable will not be linear. This is explained by the fact that the influence that the explanatory variables have on the probability of being Y=1 does not only depend on the value of the coefficients, but also on the value that the explanatory variables take.

The form that the logistic function takes is the following:



The logistic coefficient β_i is calculated by comparing the probability of occurrence of the event with the probability of non-occurrence, so that the estimated coefficients are measures of the changes in the probability ratio, called odds ratio, which is expressed as follows:

$$\frac{Pr\ (ocurrence)}{Pr(non\ ocurrence)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}$$

The interpretation of the sign of the coefficients is the same as in linear models. A β_i positive will increase the probability of the event occurring, a negative one decreases it and a coefficient equal to zero produces no change in the ratio.

Considering the selected explanatory variables, the model to be estimated would be the one below:

$P(y=1|x_i)=F(\beta_0+\beta_1 AGE+\beta_2 AGEST+\beta_3 LENGTH+\beta_4 HOURS+\beta_5 Man+\beta_6 Mar+\beta_7 For+\beta_8 Searching+\beta_9 P2+\beta_{10} S1$

$\beta_{11} SG+\beta_{12} SP)$

This will not be a definitive model since it will be necessary to discard non-relevant variables in the calculation of the estimated probability.

Table 5: Results with Logit Model

Modelo 4: estimaciones Logit utilizando las 28341 observaciones 1-28341
Variable dependiente: DESitu_1

Variable	*Coeficiente*	*Desv. típica*	*Estadístico t*	*valor p*	
const	-4,19711	0,248249	-16,9068	<0,00001	***
AGE	0,10626	0,00584118	18,1915	<0,00001	***
AGEST	0,00127024	0,0103909	0,1222	0,90270	
LENGTH	0,0457788	0,00407179	11,2429	<0,00001	***
HOURS	0,124934	0,00160809	77,6912	<0,00001	***
Man	0,114402	0,0469844	2,4349	0,01490	**
Mar	-0,822497	0,0822216	-10,0034	<0,00001	***
For	0,339207	0,0807508	4,2007	0,00003	***
Searching	-29,5885	58148	-0,0005	0,99959	
P2	-0,524456	0,158754	-3,3036	0,00095	***
S1	-0,431448	0,0907302	-4,7553	<0,00001	***
SG	-0,394438	0,0799548	-4,9333	<0,00001	***
SP	-0,0272269	0,0877712	-0,3102	0,75641	

Media de DESitu_1 = 0,430
Número de casos 'correctamente predichos' = 26179 (92,4%)
f(beta'x) en la media de las variables independientes = 0,029
Pseudo R^2 de McFadden = 0,651007
Log-verosimilitud = -6759,74
Contraste de razón de verosimilitudes: Chi-cuadrado(12) = 25219,1 (valor p 0,000000)
Criterio de información de Akaike (AIC) = 13545,5
Criterio de información Bayesiano de Schwarz (BIC) = 13652,8

Criterio de Hannan-Quinn (HQC) = 13580

Dependent variable: Employment situation

**** indicates significant at the 10 percent level***

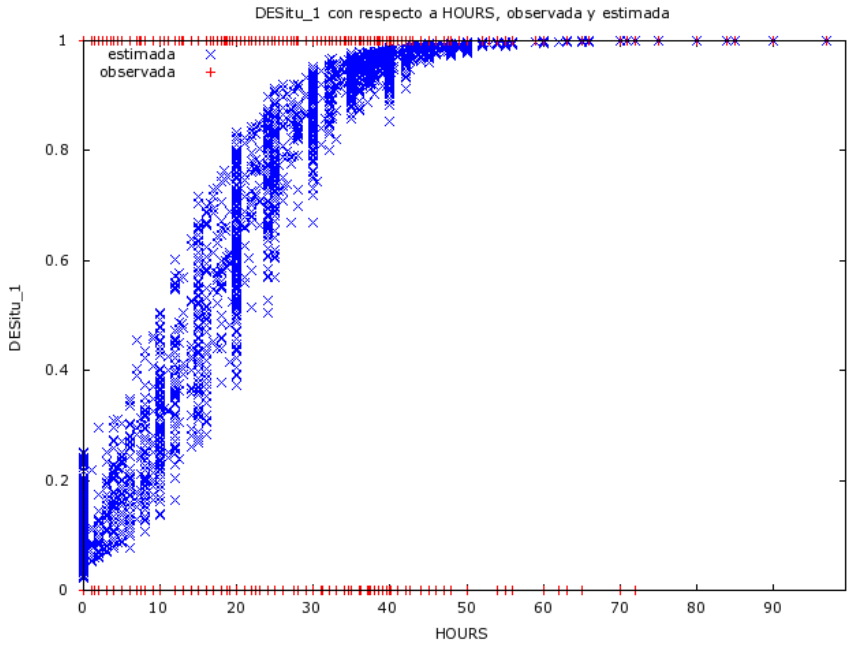
***** indicates significant at the 5 percent level***

****** indicates significant at the 1 percent level***

As can be seen in the results, the sign of the coefficients is the same as in the Linear Probability Model, except for the MAN variable, which is now *positive*. This means that the probability of being employed increases when the individual is a man.

Figure 2 clearly shows the form of the non-linear function and that the conditional probabilities always take values between 0 and 1.

Figure 2: Employment situation against Hours (Logit)



In Table 5 the significance of each variable is shown by asterisks: the more asterisks, the greater the significance. Since the variables AGEST and SEARCHING do not appear to be relevant, they will be removed from the model, leaving it as follows:

$$P(y=1|x_i)=F(\beta_0+ \beta_1 AGE+ \beta_2 LENGTH+ \beta_3 HOURS+ \beta_4 Man+ \beta_5 Mar+ \beta_6 For+ \beta_7 P2+ \beta_8 S1+ \beta_9 SG+ \beta_{10} SP)$$

Table 6: Definitive Logit Model

Modelo 5: estimaciones Logit utilizando las 28341 observaciones 1-28341
Variable dependiente: DESitu_1

Variable	*Coeficiente*	*Desv. tipica*	*Estadistico t*	*valor p*	
const	-3,8671	0,13575	-28,4869	<0,00001	***
AGE	0,08178	0,00539733	15,1519	<0,00001	***
LENGTH	-0,00313189	0,00202595	-1,5459	0,12213	
HOURS	0,138026	0,00162426	84,9778	<0,00001	***
Man	0,0521811	0,0456168	1,1439	0,25266	
Mar	-0,739003	0,0800425	-9,2326	<0,00001	***
For	0,265981	0,0769191	3,4579	0,00054	***
P2	-0,637737	0,123296	-5,1724	<0,00001	***
S1	-0,462403	0,0605913	-7,6315	<0,00001	***
SG	-0,308465	0,0643354	-4,7946	<0,00001	***
SP	-0,0978449	0,0774016	-1,2641	0,20619	

Media de DESitu_1 = 0,430
Número de casos 'correctamente predichos' = 26043 (91,9%)
f(beta'x) en la media de las variables independientes = 0,248
Pseudo R^2 de McFadden = 0,622947
Log-verosimilitud = -7303,25
Contraste de razón de verosimilitudes: Chi-cuadrado(10) = 24132,1 (valor p 0,000000)
Criterio de información de Akaike (AIC) = 14628,5
Criterio de información Bayesiano de Schwarz (BIC) = 14719,3

Criterio de Hannan-Quinn (HQC) = 14657,7

Dependent variable: Employment situation

Once the final variables to be included have been chosen, several conclusions can be drawn. In the Logit regression, seven of the ten coefficients of the explanatory variables are statistically significant at 1%, while the other three variables do not show significance. Even so, it has been decided to include them in the model; one of them, SP, because it is dummy and forms part of a qualitative variable, EDUC, and MAN because it is an important factor that has to be taken into account to carry out a correct study.

Regarding the goodness of fit, the R^2 of McFadden stands out. McFadden (McFadden, 1979, p.306) establishes that a pseudo-R squared above 0.2 indicates a good fit and above 0.4 an excellent fit. The McFadden pseudo-R squared in this model is 0.622947, which indicates an excellent model fit. This measure is complemented by a predictive accuracy of 91.9%, which serves to indicate how good the specified model is.

Table 7: Logit Model with slopes in the mean

Modelo 6: estimaciones Logit utilizando las 28341 observaciones 1-28341
Variable dependiente: DESitu_1

Variable	*Coeficiente*	*Desv. típica*	*Estadístico t*	*Pendiente**
const	-3,8671	0,13575	-28,4869	
AGE	0,08178	0,00539733	15,1519	0,0202842
LENGTH	-0,00313189	0,00202595	-1,5459	-0,000776815
HOURS	0,138026	0,00162426	84,9778	0,0342351
Man	0,0521811	0,0456168	1,1439	0,0129427
Mar	-0,739003	0,0800425	-9,2326	-0,183298
For	0,265981	0,0769191	3,4579	0,0659723
P2	-0,637737	0,123296	-5,1724	-0,15818
S1	-0,462403	0,0605913	-7,6315	-0,114692
SG	-0,308465	0,0643354	-4,7946	-0,0765096
SP	-0,0978449	0,0774016	-1,2641	-0,0242688

*Evaluado en la media
Media de DESitu_1 = 0,430
Número de casos 'correctamente predichos' = 26043 (91,9%)
f(beta'x) en la media de las variables independientes = 0,248
Pseudo R^2 de McFadden = 0,622947
Log-verosimilitud = -7303,25
Contraste de razón de verosimilitudes: Chi-cuadrado(10) = 24132,1 (valor p 0,000000)
Criterio de información de Akaike (AIC) = 14628,5
Criterio de información Bayesiano de Schwarz (BIC) = 14719,3

Criterio de Hannan-Quinn (HQC) = 14657,7

Dependent variable: Employment situation

The new Logit model has now been estimated by showing the slopes in the mean, which indicates that the results obtained are interpreted for the "average individual" in the sample. The slopes in the mean are useful to look at the partial effects of the x_i on the probability of success. This tool allows the explanation of the effects of the independent variables on the probability of success.

In other words, the current analysis provides indications of how each variable influences the employment situation.

The on-going analysis allows to have indications of how each variable affects the employment situation of each individual while, as mentioned above, Table 6 shows the most significant variables and Table 7 the marginal effects of every single variable. The following paragraphs are based on the results provided by these two tables.

On the one hand, it is observed that one of the most significant variables is the age of the individuals, which has a positive influence, that is, the older the individual, the greater the probability of being

employed; when age increases by one unit, the probability of being employed increases by 2.03%, *ceteris paribus*. The same happens with the number of hours per week that the individual employs on the job, which also has a positive influence, increasing that probability by 3.42%. In terms of discrimination among the younger population, this can be translated into the fact that, in examining young people from the age of 16 and up, they may be spending more hours on their education than on work, during these early years.

On the other hand, the dummy variable referring to the nationality of the individual indicates that the foreign nationality increases the probability of being employed by 6.60% This result can be explained by the fact that foreigners can come to work in the country with a predetermined contract, so the number of unemployed can be lower in this group.

Regarding marital status, the probability decreases by 18.33% among those who are married compared to those who are single, divorced, or widowed. This difference can be related to work-family conciliation if the term of being married is related to having children and what that entails when it comes to spending more time in the family environment.

Looking at how the variables referring to education influence, the level of education with the highest probability of employment is the higher education (SU), since all the others have a negative influence. The slopes of each of these dummy variables indicate the difference in the probability of being employed compared to the SU base group, keeping the other variables constant. This decrease in probability is reduced as the level of education increases; that is, the decrease in probability is greater at the lower levels (P2) and increases until it reaches the highest levels (SP). This makes sense because of the great consideration given to education around employment, as education is one of the key factors in the labor market associated with better jobs and working conditions, and often with better wages.

Finally, it is observed that it is not significant that the individual is a man or a woman, although the fact of being a man has a positive influence on the probability of being employed. In turn, the variable LENGTH, which refers to the average time in months that the individual has been looking for work, is also not significant in this estimation.

5. CONCLUSIONS

To analyze the Spanish labor market from a gender and youth perspective, several steps have been taken towards this work.

Beginning by presenting the case to be studied, the problem at hand has been analyzed, so that the reason for developing this work could be seen. The relevant Labor Force Survey data obtained from the INE online source was then transferred so that the Gretl statistical software could read it, selecting the variables that would be used later in the study.

Once this has been done, two econometric models have been used to perform the analysis: on the one hand, the Linear Probability Model and the other, the Logit model. Through these two models, the results of the empirical research carried out regarding the determinants of the employment situation have been presented in the preceding pages. Despite not having made use of this model during the Econometrics subject in the university degree, it has been concluded that it is more useful in this case.

Following this, the results obtained have been explained, based on the Logit model, taking into account the application of the techniques that have been studied during the degree.

After this process, we are in a position to answer the question that was raised: does gender inequality exist among young people in Spain, when they enter into de labor market?

Male individuals are more likely to be employed (we see this in the positive coefficient in the slope column of Table 7). Even so, after having eliminated non-significant variables, the variable sex remains as non-significant in the estimation, so it follows that sex is not relevant when examining the employment situation in this period.

Regarding age, we see that the higher the age, the greater the probability of being employed. In turn, the higher the education, the greater the probability of having a job. Taking these two variables into account, we can deduce that young people have a late entry into the labor market due to spending this time on their education, which was already discussed during the introductory part of this paper.

It is also important to take into account the number of hours spent at work, as this may be closely related to the impossibility of reconciling family and work, which tends to affect women more.

As a conclusion, we could say that based on the data presented for the last quarter of 2019, the population with the greatest probability of employment is concentrated among those of foreign nationality, higher education and those who are not married, with the probability increasing with the hours spent at work and the older the individual, keeping all the other variables constant.

Contrasting expectations with results, I was surprised by the fact that the sex of individuals is not relevant when studying the employment situation, since, as has already been stated in the introduction, there is inequality in the labor market between men and women. Despite thinking that I would come up against different results, these outcomes can be explained in terms of how the gender gap increases with age, as has already been mentioned. As we are looking at the young population, 16-34 years old to be precise, the gap may not be accentuated in this range, and therefore that may be the reason why we have obtained these results.

6. BIBLIOGRAPHY

Atrevia. (2018). *Mujeres en los Consejos de las empresas cotizadas*. Retrieved from https://media.iese.edu/research/pdfs/ST-0508.pdf

Brecha salarial de género en Europa. (2020, March 3). Retrieved March 3, 2020, from https://www.europarl.europa.eu/news/es/headlines/society/20200227STO73519/brecha-salarial-de-genero-en-europa-hechos-y-cifras-infografia

Brecha salarial de género en salarios por hora. (2017). Retrieved from https://www.ine.es/ss/Satellite?L=es_ES&c=INESeccion_C&cid=1259925408327&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout¶m1=PYSDetalle¶m3=1259924822888

Coste de oportunidad de la brecha de género en el empleo. (2020, February). Retrieved from https://www.pwc.es/es/publicaciones/diversidad/pwc-closingap-brecha-empleo.pdf

Encuesta de Población Activa, cuarto trimestre de 2019. (2020, January). Retrieved from https://www.ine.es/daco/daco42/daco4211/epa0419.pdf

Estadísticas de empleo. (2019, May). Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Employment_statistics/es

Euro area unemployment at 7.5%. (2019, November). Retrieved from https://ec.europa.eu/eurostat/documents/2995521/10159284/3-09012020-AP-EN.pdf/31cdc9f0-951b-6677-93c7-7646ca6eeb95

F. (2019, March 3). La maternidad penaliza la carrera profesional para el 70,6% de las mujeres. Retrieved from https://foretica.org/la-maternidad-penaliza-la-carrera-profesional-para-el-706-de-las-mujeres/

Higher proportion of women than men with a high education level. (2018). Retrieved from https://ec.europa.eu/eurostat/cache/infographs/womenmen/bloc-2a.html?lang=en

LAS GESTORAS DE EXPLOTACIONES AGRARIAS: análisis de sus características mediante modelos MLP y PROBIT. (2013). Retrieved from https://baobab.uc3m.es/backup_monet_2013_3_22/monnet/IMG/pdf/TFG_-_Maria_Romero_Melendez.pdf

McFadden, D. (1979): "Quantitative Methods for Analysing Travel Behaviour of Individuals: Some Recent Developments" en Hensher, D. y Stopher, P. (Eds.), Behavioural Travel Modelling Editorial Croom Helm (pp. 279-318), Londres.

Mujeres graduadas en educación superior. (2017). Retrieved from https://www.ine.es/ss/Satellite?L=es_ES&c=INESeccion_C&cid=1259925481157&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout¶m1=PYSDetalle¶m3=1259924822888

Tasa de paro EPA. (2019). Retrieved from https://datosmacro.com

Tasa de paro juvenil en España. (2018, December). Retrieved from https://www.tasadeparo.com/tasa-paro-juvenil-espana.html

Unemployment and labor underutilization. (2019). Retrieved from https://ilostat.ilo.org/topics/unemployment-and-labour-underutilization/

Attachments

Excel – Explanatory variables

GRETL