Estimating the gender wage gap using Stochastic Frontiers: Some modelling issues

PLEASE DO NOT QUOTE THIS PRELIMINARY VERSION

Antonio Alvarez
University of Oviedo

Graziella Bonanno
University of Salerno

Abstract

This study examines the gender wage gap in Italy using the Stochastic Frontier Approach. Using data from the European Union Statistics on Income and Living Conditions (EU-SILC) survey for Italy in 2021, we estimate a wage frontier model. We find evidence of a gender wage gap, with women earning significantly less than men even after controlling for observed characteristics. Our results are twofold. Specifically, we find that women have lower levels of wage efficiency, meaning that they are less able to convert their productivity into earnings. Second, we show that the size of the gender wage gap depends on a modelling choice regarding where to include in the model the dummy for gender.

***Keywords*:** wage stochastic frontiers, gender wage gap, Italy

JEL codes: J31Wage Level and Structure • Wage Differentials

1. Introduction

The issue of the gender pay gap has been a subject of intense research and policy debate over the past few decades. Even if the gender wage gap declined from 1980 onward, it remains a big issue in modern economies (Blau and Kahn, 2017). Defined as the difference in earnings between male and female workers, the gender pay gap has been attributed to various factors, including differences in human capital, occupational segregation, discrimination, and bargaining power. Despite significant progress in narrowing this gap, it remains a persistent challenge in many countries, including Italy (Furno, 2013).[1]

One way to compute wage gaps is by means of estimating an earnings frontier, as proposed by Herzog et al. (1985). A wage frontier allows for the computation of the maximum wage a worker could earn based on their human capital and other labour market variables that influence wage levels. Frontier functions can be estimated using non-parametric techniques, such as Data Envelopment Analysis (DEA) or parametric methods like Stochastic Frontiers. Originally developed within the context of production economics, Stochastic Frontier Analysis (SFA) helps to estimate the difference between current production and potential production (given inputs). The original formulation of the SFA model is due to the pioneering work of Aigner et al. (1977), and Meeusen and van den Broeck (1977). The main characteristic of stochastic frontiers is to decompose the error term into two components: one representing random noise and the other, an asymmetric term that measures the distance to the stochastic frontier and which can therefore be interpreted as a wage gap.

There is a growing literature that uses SFA to estimate earnings frontiers across various contexts, including studies by Polachek and Yoon (1987, 1996), Robinson and Wunnava (1989), Polachek and Robst (1998), and Zhao et al. (2022). Lang (2005) looks at the wage gap between immigrants and nationals, while Jane (2013) investigates the overpayment in the labour market related to the Professional Baseball in Japan.

A more recent strand of literature employs SF models allowing for some variables to explain the inefficiency term. Some examples, in the field of earnings frontiers are the studies by Díaz and Sánchez (2011), Pérez-Villadóniga and Rodríguez-Alvarez (2017) and Bashford-Fernández and Rodríguez-Álvarez (2019). Several variables have been used in the literature to explain the inefficiency term, with gender being the primary focus.

[1] This reduction in the gender wage gap is well documented. For the US, O'Neill and Polachek (1993) show that the female-to-male earnings ratio jumped from 60% in 1980 to almost 72% in 1990. Gunderson (1989) shows a reduction in the gender wage gap for a set of developed countries between 1960 and 1980.

Our objective is to estimate the existence and extent of the gender wage gap in Italy. For that purpose, we estimate a Mincer equation within the framework of a stochastic frontier model, i.e., an earnings frontier. While we do not make any significant methodological contribution, we engage in a thorough discussion of the modelling issues surrounding the specification of wage frontiers to estimate the gender wage gap. Obviously, when the empirical strategy involves estimating a pooled model for both men and women, the gender wage gap depends on the sign and magnitude of the estimated parameter of a gender dummy variable. Several specification alternatives arise regarding the inclusion of the gender dummy. In fact, it can be included in the frontier, in the inefficiency term, or in both. While previous literature has adopted one of these three options, we compare the three of them and show the differences not only in the results obtained but also in the interpretation associated with each option.

An additional feature of our paper that is worth noting is the computation of the gender wage gap in a stochastic frontier framework. Previous literature (e.g., Adamchik and King, 2007; Garcia-Prieto and Gómez-Costilla, 2017) has estimated earnings frontiers using a dummy variable for gender in the inefficiency term and checked the significance of this dummy in order to identify the existence of a gender wage gap, but they have not quantified it.

In summary, we aim to answer two research questions:

1. Is there a gender wage gap in Italy? If so, how large is it?
2. What are the implications of the three modelling options mentioned above?

We use data from the European Union Statistics on Income and Living Conditions (EU-SILC) survey for Italy in 2021. Specifically, we estimate a wage frontier that captures the maximum wage that a worker with a given set of observable characteristics can expect to earn. We then use this frontier to measure the degree of wage inefficiency for each worker, which represents the extent to which a worker's wage deviates from the maximum wage he/she could earn based on observable characteristics, i.e., the gender wage gap.

Our analysis shows that women tend to have lower levels of wage efficiency than men, even after controlling for observable factors. These findings suggest that, while observable factors such as education and experience play a role in determining wages, there are other unobserved factors that contribute to the gender wage gap.

The paper is organized as follows. Section 2 provides a brief introduction to stochastic frontier analysis. In section 3 we review the literature on wage frontiers and the gender pay gap. Section 4

describes the data, while section 5 presents the econometric model. Section 6 shows the results of the analysis. Finally, Section 7 concludes the paper.

2. Stochastic Frontier Analysis

Aigner et al. (1977) and Meeusen and Van den Broeck (1977) independently proposed the estimation of stochastic production frontiers. These models consider that deviations from the production frontier can be decomposed, allowing to separate the random effects, such as climatic events, from the effects of changes in technical efficiency[2]. Since these pathbreaking contributions, stochastic frontiers have been used in contexts different from production functions, such as earnings equations (Herzog et al., 1985) or demand functions (Algieri and Alvarez, 2023).

In the framework of panel data, a general stochastic production frontier model can be given by:

$$y_{it} = \beta' x_{it} + v_{it} - u_{it} \quad (1)$$

where subscript i indexes individuals and subscript t indexes time , y_i represents output produced by firm i at time t (in our case, wage), x_{it} is a vector of inputs (in our case, characteristics of the worker), β is a vector of unknown parameters to be estimated, v_{it} is a symmetric random disturbance which captures the effect of statistical noise, whereas u_{it} is a non-negative stochastic term that is assumed to be independent from v and to capture distance from the stochastic frontier. When u=0, the observation lies on the technological frontier and is therefore efficient. When u>0, the observation is below the frontier, indicating that it is technically inefficient, i.e., in our case indicates the existence of a wage gap (the observed wage is different form the wage given by the frontier after taking into account all the x variables included in the wage frontier).

Since we are interested in finding which variables explain the efficiency of the workers, i.e., which variables are behind the fact that some workers do not achieve the potential wage they could obtain given their characteristics, we estimate several models that modify equation (1) by allowing the inefficiency term u to be a function of some exogenous variables z. The general form of this type of models is:

$$y_{it} = \beta' x_{it} + v_{it} - u_{it}(z_{it}) \quad (2)$$

[2] See Alvarez and Arias (2014) for a survey on stochastic frontier modelling.

There are two possible alternative specifications of u(z), depending on the way that the variables z affect the distribution of u. In particular, they can affect the mean or the variance of the distribution of u.

Battese and Coelli (1995) (hereafter referred to as BC95) is the most popular model among practitioners in order to allow technical inefficiency to be a function of some exogenous variables. They allow the variables z to explain the mean of the pre-truncated distribution of u. The inefficiency term can be expressed in the following way:

$$u_{it} = z_{it}\delta + w_{it} \tag{3}$$

where z are the explanatory variables associated with technical inefficiency and w_{it} is defined by the truncation of a normal distribution with zero mean and variance σ^2, such that the point of truncation is $-z_{it}\delta$.

The other alternative is to model the variance of u. Reifschneider and Stevenson (1991) was the first paper to incorporate heteroskedasticity in the stochastic frontier model. Caudill et al. (1995) (from now on referred to as CFG95) assumed that u exhibits multiplicative heteroskedasticity, a choice that we will use in this paper. In particular, the CFG95 model suggests an exponential function:

$$u_{it} \sim N^{+}(0, \sigma_{it}^2), \quad \sigma_{it} = exp(\delta z_{it}) \tag{4}$$

where the + sign indicates truncation of the distribution at zero.

Modelling the variance of the one-sided error term is very important since the presence of heteroskedasticity in u will yield biased estimates of both the frontier parameters and the efficiency scores. This result differs markedly from the typical effect of heteroskedasticity in the two-sided error term v, which causes the variances of the parameter estimates to be biased. For this reason, the heteroskedastic model (CFG95) will be our preferred specification.

The parameters of the stochastic frontier and the model for the technical inefficiency effects in equation (2) are estimated simultaneously by maximum likelihood. If the dependent variable is measured in logs, the technical efficiency (TE) of unit i in period t can be calculated as:

$$TE_{it} = e^{-u_{it}} \tag{5}$$

Given that u is non-negative, the formula in (5) ensures that the TE index is bounded between 0 and 1.

3. Wage stochastic frontiers and gender

The gender pay gap has been a long-standing challenge in many countries, and it has been attributed to various factors. Early studies primarily focused on differences in observable characteristics, such as education, work experience, and occupation, which explained only a small portion of the wage gap (Blau and Kahn, 2007). However, more recent studies have highlighted also the importance of unobservable factors, such as discrimination, bargaining power, and preferences (Goldin, 2014). The literature on the gender wage gap is too large to be appropriately summarized in this section. We will concentrate on those papers that have used stochastic frontiers to estimate the existence of the gender wage gap, focusing on the main modelling issues.[3]

There are two empirical strategies to estimate the gender wage gap in the framework of stochastic frontiers. Some papers estimate separate frontiers by gender (e.g., Croppenshedt and Meschi, 2000; Oglobin and Brock, 2005; Watson, 2014), while others estimate a single frontier, identifying men and women through a gender dummy (e.g., Adamchik and King, 2007; Garcia-Prieto and Gómez-Costilla, 2017).

An interesting modelling issue of the papers that include a dummy for gender refers to where in the model they choose to introduce it. For example, Adamchik and King (2007), and Garcia-Prieto and Gómez-Costilla (2017) introduce the dummy just on the frontier (i.e., in the deterministic part of the equation). In this way they allow for the possibility that the frontier of men and women can be different. In fact, they find a positive and significant sign for the dummy variable of males, indicating that the frontier of women is below that of men. On the other hand, Díaz and Sánchez (2011), Pérez-Villadóniga and Rodríguez-Álvarez (2017b) and Bashford-Fernández and Rodríguez-Álvarez (2019) choose to introduce the gender dummy just in the inefficiency model. This modelling option implies that men and women share the same frontier, but the model allows for differences in the ability to achieve this potential wage, i.e., the distance to the frontier of men and women can be different, indicating that they face different barriers to achieve the frontier. Finally, Pérez-Villadóniga and Rodríguez-Álvarez (2017a) introduce the gender dummy both on the frontier and in the inefficiency term, allowing for a double role of gender: as a determinant of the frontier but also as a determinant of the distance to the frontier. A new methodology is proposed by Perez-Villadóniga et al. (2025) examining the wage gap among workers holding managerial positions in Spain through a Latent Class-Stochastic Frontier model to analyse the wage gap between male and female managers.

[3] Bashford-Fernández and Rodríguez-Álvarez (2019) perform an exhaustive review on wage stochastic frontiers and gender.

Interestingly, the only paper that justifies its modelling option is Bashford-Fernández and Rodríguez-Álvarez (2018). They claim that: "The aspect of gender was not incorporated into our estimation of the Mincer equation frontier because there are no a priori reasons to expect that gender may influence productivity, meaning that both men and women should be able to reach the same wage frontier given their human capital". While this may be true, the opposite may as well happen; that is, the wage frontier for men and women may be different due to some reasons, the most cited one being discrimination against women. Therefore, we think that this is an empirical issue and therefore, the more flexible model, i.e., that with the gender dummy both on the frontier and in the inefficiency term, is preferred. The estimation will reveal if the estimated coefficients of these dummies are significant or not.

4. Data and variables

Our analysis is based on a sample of 6,347 individuals from the European Union Statistics on Income and Living Conditions (EU-SILC) for Italy in 2021. This survey provides a wide range of details concerning the labour market and individual characteristics of the respondents.[4] Specifically, our analysis focuses on individuals who have permanent, full-time contracts, and excludes those employed in the agricultural and public sectors. In line with previous studies (e.g., Zveglich et al., 2019; Díaz and Sánchez, 2011) we further restrict the sample to individuals aged between 25 and 65 years old.

We estimate a Mincer equation (Mincer, 1974) with the dependent variable being the logarithm of the annual in-work income.[5] This is calculated using respondents' self-reported earnings from their main job in the previous year.

As explanatory variables, we account for the number of hours worked during the year, since our dependent variable is not wage per hour. *Worked hours* represents the number of hours that the individuals declare to work in their main job. As suggested by economic theory, we also include a set of human capital proxies, such as education and work experience. *Work experience* is the number of years spent in paid work. As for education, EU-SILC contains information on the highest level of education attained according to the International Standard Classification of Education (ISCED) level

[4] The guidelines and the description of EU-SILC target variables (2021 version), can be found in https://circabc.europa.eu/sd/a/f8853fb3-58b3-43ce-b4c6-a81fe68f2e50/Methodological%20guidelines%202021%20operation%20v4%2009.12.2020.pdf).

[5] The in-work income includes both cash and non-cash income from work. The tax at source and the social insurance contributions are deducted. Data on wage are in national currency (euro).

successfully completed. While some papers convert this discrete variable into a continuous one by assigning years to each schooling category (e.g., Pérez-Villadóniga and Rodríguez-Álvarez, 2017a), we opt for binary variables. Specifically, we include four binary variables capturing the attainment of primary (*D_edu_primary)*, lower secondary (*D_edu_lowsec)*, upper secondary (*D_edu_upsec)*, vocational and training education (*D_edu_training)*, with tertiary or post-tertiary education serving as the reference category.

In order to control for the type of job, we include a set of dummies that capture some categories of the International Standard Classification of Occupations (ISCO). In detail, *D_manager* is equal to 1 if the respondent is a manager, 0 otherwise; *D_professionals* is equal to 1 if respondent is a professional. *D_admin* is equal to 1 for clerical support workers. *D_service_workers* is equal to 1 for employees in service and sales areas. *D_craft_workers* is equal to 1 if the respondent works in craft and related trades. The category of plant and machine operators and assemblers is the control group.

Another relevant aspect of the data refers to the sector of activity (NACE Rev. 2). We include 3 dummies: *D_manufacturing*, *D_trade* and *D_construction*. The reference group is *Services*.

Other control variables included in the estimation are marital status, as well as regional dummies. *D_married* takes value 1 if the individual has ever been married; *Single* is the control group. Four dummy variables, *D_North-East*, *D_Centre*, *D_South* and *D_Islands,* account for the macro-regions of residence according with the NUTS 1 classification. *North_West* is the omitted group. Moreover, we add *D_access_Internet*, which is equal to 1 if the respondent declares to have an Internet connection at home for personal use.

Finally, we include a dummy variable for our main variable of interest, gender. *D_Female* is equal to 1 if the respondent is female, and 0 for males. As explained in the introduction, we will include it not only in the frontier but also in the inefficiency term when the model allows for that possibility.

Table 1 provides some descriptive statistics of all the variables used in the empirical analysis. Out of the 6,347 observations, 57.5% are men compared to 42.5% women.

If we look at the difference in annual earnings across genders, females and males earn on average 21,116 and 24,342 Euros, respectively. Therefore, female wage is 86.7% of male wage, a figure rather high compared to similar countries. In fact, Italy is a country with low gender wage gap. Blau and Kahn (1996) show that for a sample of European countries plus the United States and Australia, Italy is the country with the highest mean percentile ranking of women in the male wage distribution.[6]

[6] The mean percentile ranking of women in the male wage distribution.

With regards to the main worker characteristics, the maximum of worked hours per week are lower for females than for males (about 36 and 40, respectively). Similar picture results from the mean values of *Work_experience* reported in Table 1, indeed, while males have experienced on average about 21 years of paid work, females only 19 years.

As for education, a tiny share of the sample declares to have primary or less than primary education (1.3%). While 20% of the sample gets a lower secondary education level, almost half (46.4%) of the individuals declares that they have obtained an upper secondary level qualification. Individuals that declare to have a vocational education account for about two percentage points. The residual part of the sample (30%) is made up of individuals who have obtained a bachelor or a post-graduate degree.

Referring to the type of job, the sample is composed as follows: 1.8% managers, 49% professionals, 16.8% clerical support workers, 12.4% employees in service and sales areas, 11.4% craft workers, and 8.7% machine operators, and assemblers.

As the sector of activity is concerned, 53.5% of the individuals work in the service sector, with manufacturing making up 29.7% of the sample, while trade and construction account for just 10.9% and 5.9%, respectively.

Marital status is one of the most significant variables in explaining the gender pay gap. Blau and Kahn (1992) show that the gap is substantially larger among married workers. For example, the female-to-male earnings ratio in the US (1985-1988) was 59.4% for married workers compared to 95.5% for single workers (notably, this ratio was 1.02 in Germany). Our sample consists of married, separated and single individuals in the following proportions: 54.4%, 7.7% and 37.8%, respectively. Following Herzog et al. (1985), we classify married and separated individuals into the same category, 'ever-married', which accounts for 62.2% of the sample.

As for the geographical composition of the sample, 26.6% of respondents live in the North-East of the country, 23.4% live in the North-West of Italy (control group), 27% live in the central regions, and 16.8% and 6.2% of respondents live in Southern Italy and the Islands, respectively.

Table 1. Descriptive Statistics of the sample

Variables	All sample			Males			Females		
Wage (In-work income)	6,347	22970.69	13103	3,649	24341.78	15109.96	2,698	21116.32	9443.56
Worked hours	6,347	38.284	6.225	3,649	39.822	5.204	2,698	36.203	6.857
D_Female	6,347	0.425	0.494						
Work experience	6,347	20.390	10.534	3,649	21.134	10.628	2,698	19.384	10.321
D_edu_primary	6,347	0.013	0.113	3,649	0.016	0.126	2,698	0.009	0.092
D_edu_lowsec	6,347	0.203	0.403	3,649	0.255	0.436	2,698	0.134	0.341
D_edu_upsec	6,347	0.464	0.499	3,649	0.478	0.500	2,698	0.445	0.497
D_edu_training	6,347	0.021	0.144	3,649	0.018	0.132	2,698	0.026	0.158
D_manager	6,347	0.018	0.133	3,649	0.023	0.151	2,698	0.011	0.105
D_professional	6,347	0.490	0.500	3,649	0.428	0.495	2,698	0.573	0.495
D_admin	6,347	0.168	0.374	3,649	0.145	0.352	2,698	0.198	0.399
D_service_worker	6,347	0.124	0.330	3,649	0.103	0.304	2,698	0.152	0.359
D_craft_workers	6,347	0.114	0.318	3,649	0.179	0.383	2,698	0.026	0.160
D_manufacturing	6,347	0.297	0.457	3,649	0.387	0.487	2,698	0.174	0.379
D_trade	6,347	0.109	0.312	3,649	0.106	0.308	2,698	0.113	0.316
D_construction	6,347	0.059	0.236	3,649	0.095	0.294	2,698	0.011	0.103
D_married	6,347	0.622	0.485	3,649	0.621	0.485	2,698	0.622	0.485
D_rural_areas	6,347	0.209	0.406	3,649	0.215	0.411	2,698	0.200	0.400
D_access_internet	6,347	0.919	0.274	3,649	0.913	0.282	2,698	0.926	0.262
D_NorthEast	6,347	0.266	0.442	3,649	0.267	0.442	2,698	0.266	0.442
D_Centre	6,347	0.270	0.444	3,649	0.265	0.442	2,698	0.275	0.447
D_South	6,347	0.168	0.374	3,649	0.174	0.379	2,698	0.159	0.365
D_Islands	6,347	0.062	0.241	3,649	0.059	0.236	2,698	0.066	0.248

Source: own elaborations on data from EU-SILC for Italy.

5. Empirical strategy

5.1 Econometric model

Our objective is to measure the differences between the potential income of each worker, given their socio-economic characteristics, and the income actually received. We estimate a Mincer equation within the framework of a stochastic frontier. Our more general model is the following:

$$lnW = \mathrm{f}(\mathrm{X}, \mathrm{G}) + \mathrm{v} - \mathrm{u}(\mathrm{G}) \qquad (6)$$

where the dependent variable is the natural logarithm of the annual wage, X are a set of explanatory variables (worked hours, education, experience) along with control variables for sector, occupation, region and some personal characteristics (place of residence, marital status, access to Internet). G is a dummy variable for gender (1 if female). The error term consists of two components: v which is a random variable accounting for noise, and u, which is a one-sided random variable that accounts for inefficiency (the distance to the frontier).

An important aspect of our specification is that the dependent variable is annual wage instead of the most typical choice, wage per hour (e.g. Garcia-Prieto and Gómez-Costilla, 2017). The main reason for this decision is that in the model where the dependent variable is log of wage per hour, if we undo the log and take the log of number of hours to the right-hand side of the equation, we can clearly see that we are imposing a coefficient of 1 to the log of hours worked. Obviously, this restriction is probably not supported by the data. Blau and Kahn (1996) share this argument, but they also estimate their earnings functions using as dependent variable both the natural logarithm of wage and the natural logarithm of hourly wage.

By including the gender dummy in the frontier, we allow for different frontiers for men and women. If the estimated coefficient of G on the frontier is negative, it will indicate that women can achieve a lower potential salary than men with the same observed characteristics. However, we also believe that women may face more challenges than men in reaching their potential (frontier) salary and, therefore, we introduce gender into the inefficiency model, making the distribution of u a function of the dummy G. Specifically, we estimate the heteroskedastic model proposed by Caudill, Ford and Gropper (1995), where u_i is distributed as Half-Normal, $u_i \sim N^+(0, \sigma_{ui}^2)$. In our case, the variance of the pre-truncated distribution of u is a function of gender, expressed as:

$$\sigma_{ui}^2 = \exp(\delta_0 + \delta_1 G_i) \tag{7}$$

As stated in the Introduction, we will also estimate two restricted models. First, we will consider that the effect of gender takes place only on the frontier and therefore the gender dummy will be introduced only as a shifter of the frontier. In this model, we will assume that the inefficiency term u follows a half-normal distribution which is the same for men and women. Second, we will introduce the gender dummy only in the inefficiency term, using the same specification as in (7).

5.2 Estimation of the gender wage gap

The GWG is the difference in (expected) wage due exclusively to gender. That is

$$E[W|\mathrm{X}, G = 0] - E[W|\mathrm{X}, G = 1] \tag{8}$$

Since our dependent variable is in logs, we start by taking the expectation of lnW in equation (6). Since the expected value of the deterministic part is itself and the expect value of v is zero, we have:

$$\mathrm{E}[lnW|X, G] = \mathrm{f}(X, G) - E[u(\mathrm{G})|X, G] = f(X, G) - E[u(G)|G] \tag{9}$$

The expectation of the frontier is unit-specific. Therefore, to compute the differences in wage due solely to gender we evaluate it at the mean of the data:

$$E[lnW|X = \overline{X}, \mathrm{G}] = f(\overline{X}, G) - E[u(G)|G]] \quad (10)$$

In order to compute the gender wage gap, we evaluate this expression separately for men and women:

- Men: $E[lnW|X = \overline{X}, G = 0] = f(\overline{X}, G = 0) - E[u(G)|G = 0]$ (11)
- Women: $E[lnW|X = \overline{X}, G = 1] = f(\overline{X}, G = 1) - E[u(G)|G = 1]$ (12)

The conditional expectation of the frontier is straightforward but the expected value of a half-normal, needs some elaboration. We take the following expression from Kumbhakar and Lovell (2000) (p. 77)

$$\mathrm{E[u(G)|G]} = \sigma_u(\mathrm{G})\sqrt{\frac{2}{\pi}} = \sqrt{\exp\left(\delta_0 + \delta_1 G\right)}\sqrt{\frac{2}{\pi}} \quad (13)$$

This expression can be evaluated for G=1 and G=0, yielding an estimate of average inefficiency different for males and females. Therefore, we can now compute the GWG (in logs) as the expected (log) wage for men and women at the mean of the data:

- Men: $E[lnW|X = \overline{X}, G = 0] = f(\overline{X}, G = 0) - \sqrt{\exp\left(\delta_0\right)}\sqrt{\frac{2}{\pi}}$ (14)
- Women: $E[lnW|X = \overline{X}, G = 1] = f(\overline{X}, G = 1) - \sqrt{\exp\left(\delta_0 + \delta_1 G\right)}\sqrt{\frac{2}{\pi}}$ (15)

So, the Gender Wage Gap (in logs) is:

$$E[lnW|X = \overline{X}, G = 0] - E[lnW|X = \overline{X}, G = 1] \quad (16)$$

At this point, it only rests to move from wages in logs to wages in levels. Since $\exp(E[W]) = GM(W)$, i.e., the geometric mean of W, we can compute the GWG

$$GWG = GM(W|G = 0) - GM(W|G = 1) \quad (17)$$

6. Econometric results

Table 2 reports the estimated models on data of Italian individuals from EU-SILC in 2021. We present two estimations, one with the full sample and another one with those individuals that reported to work 40 hours. The estimation of this second model serves to avoid the problems associated with the specification on number of hours worked. However, before commenting the results related to the estimated parameters associated to regressors, we first look at the models' diagnostics. In the last block of Table 2, we report the λ parameter, which indicates the importance of the inefficiency effects, strongly supports the use of stochastic frontiers instead of the standard

ordinary least square method. [7] Finally, Model CFG95 with *D_female* on both frontier and variance of *u* is the specification to be preferred, as it is documented by the lowest value of the Akaike Information Criterion (AIC) statistics (Burnham and Anderson, 2004).[8]

In order to investigate the role of gender, we estimate three models: the first one is the standard specification of Aigner, Lovell and Smith (1977) (hereafter, ALS77), where the dummy *D_Female* enters just in the frontier. The two other options correspond to the CFG95 model, in which we first consider *D_Female* as a determinant of the variance of the inefficiency component, and then, we include the gender dummy in both the frontier and the variance of u.

Starting with the results of the models that use the full sample, the three specifications used yield very similar results. The estimates of the parameters of all the control variables are significant and have the expected signs. The estimated parameters of both worked hours and experience are positive, and the coefficients are significant across all model options.

As regard the effects of education, we find unsurprisingly results. Indeed, the estimated coefficients for the dummies capturing the ISCED of individuals are significant and negative. This is in line with the previous literature (Oglobin and Brock, 2005) highlighting higher wages for individuals with tertiary education with respect to lower levels of education.

Interesting findings emerge when we look at the estimated coefficients for the types of job. Even if always significant, these effects are mixed in terms of sign. Indeed, we estimated positive coefficients for managers, professionals and clerical support workers, and negative coefficients for workers in services and craft. This means that, while the first three categories achieve higher levels of wage than machine operators (which is the control group), workers in services and craft register lower wages than the control group.

As for the sector of activity is concerned, workers in Manufacturing receive higher wages than those employing in Services (the estimated parameter is always significant and positive). The opposite happens for Trade and Construction, for which we estimate negative coefficients.

In addition, while we estimate positive coefficients for both *D_married* and *D_Internet*, we find a negative effect of living in a rural area. Finally, not surprisingly, the levels of wage achieved in the North-West of Italy are higher than in other regions.

[7] λ is equal to $\frac{\sigma_u}{\sigma_v}$, where the zero value of this parameter indicates that deviations from the frontier are only due to random error, while values greater than 1 indicate that the distance from the frontier is mostly due to inefficiency.

[8] AIC is equal to [2*k-2*Log-likelihood], where k is the number of all estimated parameters.

6.1 The role of gender

We now turn to our topic of interest, the effect of gender. As it is customary in this literature, the estimated parameters for the dummy variable *D_Female* are negative and significant when used as an explanatory variable in the frontier. This is line with a mainstream literature, which is well consolidated (MacPherson and Hirsch, 1995; Adamchik and King, 2007).

However, when *D_Female* is allowed to influence the variance of *u*, the impact is positive, meaning that females appear to be more inefficient than males. This is an interesting result since the interpretation of a larger estimated variance of *u* for females means that women are further away from their own frontier than men are from theirs. This may be an indication that women face more difficulties than men in achieving the frontier.

In Tables 3.a and 3.b, we present the estimation of the gender wage gap for the three models considered, for both full sample and sample of individuals that have worked 40 hours, as derived in the sub-section 5.2. The results indicate significant differences among the three models suggesting that the placement of the gender dummy variable in the model is an important modelling choice, and this holds for both sets of estimations. In detail, for the full sample of individuals, we find that the GWG calculated with the CFG95 specifications is lower than the GWG obtained from the ALS77 model. Conversely, in the second set of estimations (when only individuals worked 40 hours per week are involved) the GWG calculated for the full CFG95 model is very high (3407 euro), and in the other two cases the GWGs are lower than this.

Table 2. Estimation of the wage stochastic frontier models

	Full sample			Only 40 hours per week		
	ALS77	CFG95		ALS77	CFG95	
	D_Female on frontier only	*D_Female* on Var(u) only	*D_Female* on both	*D_Female* on frontier only	*D_Female* on Var(u) only	*D_Female* on both
Worked hours	0.0002***	0.0003***	0.0002***			
	(0.0000)	(0.0000)	(0.0000)			
D_Female	0.0679**		-0.2626***	-0.1162***		-0.0285
	(0.0279)		(0.0685)	(0.0162)		(0.0188)
D_Female* Worked hours	-0.0001***		-0.0001***			
	(0.0000)		(0.0000)			
Work experience	0.0084***	0.0087***	0.0096***	0.0081***	0.0082***	0.0081***
	(0.0005)	(0.0004)	(0.0008)	(0.0005)	(0.0004)	(0.0004)

D_edu_primary	-0.4325***	-0.4166***	-0.4328***	-0.5000***	-0.4887***	-0.4970***
	(0.0348)	(0.0428)	(0.0572)	(0.0726)	(0.0710)	(0.0714)
D_edu_lowsec	-0.3442***	-0.3413***	-0.3284***	-0.3514***	-0.3477***	-0.3556***
	(0.0234)	(0.0244)	(0.0286)	(0.0244)	(0.0234)	(0.0239)
D_edu_upsec	-0.2224***	-0.2155***	-0.2132***	-0.2399***	-0.2317***	-0.2415***
	(0.0020)	(0.0040)	(0.0076)	(0.0057)	(0.0024)	(0.0055)
D_edu_training	-0.1848***	-0.1931***	-0.1370***	-0.1836***	-0.1868***	-0.1908***
	(0.0215)	(0.0178)	(0.0186)	(0.0189)	(0.0104)	(0.0158)
D_manager	0.6219***	0.5865***	0.4409***	0.6112***	0.5874***	0.5930***
	(0.0564)	(0.0559)	(0.0997)	(0.0754)	(0.0680)	(0.0738)
D_professional	0.1814***	0.1602***	0.1773***	0.1689***	0.1520***	0.1645***
	(0.0108)	(0.0126)	(0.0064)	(0.0277)	(0.0283)	(0.0284)
D_admin	0.1004***	0.0757***	0.0741***	0.0700**	0.0467*	0.0671**
	(0.0062)	(0.0085)	(0.0070)	(0.0277)	(0.0280)	(0.0297)
D_service_worker	-0.0071	-0.0279***	-0.0731***	-0.0487**	-0.0696***	-0.0500**
	(0.0057)	(0.0029)	(0.0086)	(0.0236)	(0.0239)	(0.0249)
D_craft_workers	-0.0568***	-0.0603**	-0.0596***	-0.0508**	-0.0520**	-0.0499**
	(0.0215)	(0.0236)	(0.0143)	(0.0207)	(0.0222)	(0.0222)
D_manufacturing	0.0404***	0.0492***	0.0483***	0.0168	0.0216	0.0137
	(0.0093)	(0.0088)	(0.0081)	(0.0114)	(0.0148)	(0.0114)
D_trade	-0.0507***	-0.0511***	-0.0133	-0.0435	-0.0456	-0.0497*
	(0.0176)	(0.0183)	(0.0282)	(0.0267)	(0.0295)	(0.0281)
D_construction	-0.0553***	-0.0368***	-0.0430***	-0.0688***	-0.0549**	-0.0725***
	(0.0120)	(0.0113)	(0.0112)	(0.0216)	(0.0219)	(0.0209)
D_ married	0.1055***	0.0591***	0.0941***	0.0920***	0.0616***	0.0923***
	(0.0164)	(0.0210)	(0.0108)	(0.0122)	(0.0199)	(0.0122)
D_Female*D_married	-0.0621***		-0.0225	-0.0778***		-0.1688***
	(0.0118)		(0.0277)	(0.0206)		(0.0515)
D_rural_areas	-0.0298**	-0.0263*	-0.0142*	-0.0295*	-0.0224*	-0.0295**
	(0.0149)	(0.0137)	(0.0086)	(0.0152)	(0.0127)	(0.0128)
D_access_internet	0.0559***	0.0492***	0.0686***	0.0706***	0.0703***	0.0730***
	(0.0137)	(0.0097)	(0.0211)	(0.0161)	(0.0145)	(0.0137)
D_NorthEast	-0.0146***	-0.0132***	-0.0049***	-0.0056**	-0.0019	-0.0034**
	(0.0024)	(0.0021)	(0.0013)	(0.0022)	(0.0016)	(0.0016)
D_Centre	-0.1010***	-0.0993***	-0.1033***	-0.0713***	-0.0696***	-0.0706***
	(0.0029)	(0.0026)	(0.0018)	(0.0024)	(0.0029)	(0.0022)
D_South	-0.1450***	-0.1347***	-0.1613***	-0.1443***	-0.1358***	-0.1459***
	(0.0090)	(0.0071)	(0.0069)	(0.0036)	(0.0044)	(0.0034)
D_Islands	-0.1278***	-0.1284***	-0.1364***	-0.1460***	-0.1509***	-0.1461***
	(0.0067)	(0.0066)	(0.0051)	(0.0020)	(0.0045)	(0.0046)
Constant	9.7977***	9.6802***	9.7158***	10.3196***	10.2843***	10.3086***
	(0.1054)	(0.1034)	(0.1195)	(0.0266)	(0.0273)	(0.0308)
Var(u)						
Constant	-1.0490***	-1.2724***	-2.0221***	-1.1178***	-1.3442***	-1.2239***

	(0.1297)	(0.1639)	(0.2396)	(0.1462)	(0.1830)	(0.1702)
D_Female		-0.7795***	-8.2338		0.6772***	0.5613***
		(0.1695)	(0.0000)		(0.1210)	(0.1269)
D_Female* Worked hours		0.0007***	-0.0199***			
		(0.0001)	(0.0018)			
D_Female*D_married		-0.3280	0.3358		-0.1927	-0.5805*
		(0.2156)	(5.9843)		(0.2138)	(0.3472)
Observations	6,347	6,347	6,347	3,416	3,416	3,416
N_clust	5	5	5	5	5	5
ll	-3449	-3450	-3987	-1749	-1763	-1731
sigma_u	0.592	.	.	0.572	.	.
sigma_v	0.243	0.249	0.423	0.237	0.243	0.237
lambda	2.436	.	.	2.417	.	.

Source: own elaborations on data from EU-SILC for Italy.

Note(s). Significance levels: *** p<0.01, ** p<0.05, * p<0.1. Clustered standard errors by macro-regions are in parentheses.

Table 3.a Estimation of the gender wage gap: full sample

	ALS77	CFG95	
	D_Female on frontier only	*D_Female* on Var(u) only	*D_Female* on both
$E[lnW\|X=\overline{X},G=0]$	20118.91	19909.22	21024.21
$E[lnW\|X=\overline{X},G=1]$	17132.35	17401.02	18562.30
Gender Wage Gap	**2986.55**	**2508.20**	**2461.91**

Source: own elaborations on data from EU-SILC for Italy.

Table 3.b Estimation of the gender wage gap: individuals worked 40 years

	ALS77	CFG95	
	D_Female on frontier only	*D_Female* on Var(u) only	*D_Female* on both
$E[lnW\|X=\overline{X},G=0]$	20032.23	20012.68	20249.05
$E[lnW\|X=\overline{X},G=1]$	17020.51	17262.53	16842.05
Gender Wage Gap	**3011.73**	**2750.16**	**3406.99**

Source: own elaborations on data from EU-SILC for Italy.

7. Conclusions

The estimation of wage frontiers provides a robust framework for testing the existence of the gender wage gap. The empirical strategy typically involves estimating a pooled model for both men and women, whereby the gender wage gap depends on the sign and magnitude of the estimated parameter for a gender dummy variable. Some specification alternatives arise concerning the inclusion of the gender dummy. Specifically, it can be included into the frontier, the inefficiency term, or both. While previous literature has predominantly adopted one of these three options, we compare all three and show the differences not only in the results obtained but also in the interpretation of the three options.

An additional noteworthy feature of our study is the computation of the gender wage gap within a stochastic frontier framework. Previous research has employed earnings frontiers with a gender dummy to assess the significance of gender disparities; however, they have not explicitly quantified the gap. Our findings reveal that a significant gender wage gap exists to the disadvantage of women, and importantly, that its magnitude varies depending on the chosen modelling approach.

A persistent challenge in studies on the gender wage gap is determining the extent to which it stems from discrimination as opposed to differences in worker or market characteristics. Our study lacks the information to explicitly attribute wage inefficiency to discrimination. In our model, the wage gap is captured as a residual, encompassing the effects of all relevant but unobserved variables. While discrimination is a plausible explanation for this residual, other factors, such as women's self-selection, may also play a role.

References

Adamchik, V., & King, A. (2007). Labor market efficiency in Poland: A stochastic wage frontier analysis. *The International Journal of Business and Finance Research*, 1(2), 41–51.

Aigner, D., Lovell, C. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6(1), 21-37.

Algieri, B., A. Alvarez, (2023), Assessing the Ability of Regions to Attract Foreign Tourists: The Case of Italy, *Tourism Economics*, 29(3), 788–811.

Bashford-Fernández, J. M., Rodríguez-Álvarez, A. (2019) Wage Frontiers in Pre and Post-crisis Spain: Implications for Welfare and Inequality. *Social Indicators Research*, 143(2), 579-608.

Battese, G. E., & Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics*, 20, 325-332.

Blau, F. D., & Kahn, L. M. (1992). The Gender Earnings Gap: Learning from International Comparisons. *The American Economic Review*, 82(2), 533–538.

Blau, F. D., & Kahn, L. M. (1996). Wage Structure and Gender Earnings Differentials: an International Comparison. *Economica*, 3, Supp., S29–S62.

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

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

Caudill, S. B., Ford, J. M., & Gropper, D. M. (1995). Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. *Journal of Business & Economic Statistics*, 13(1), 105-111.

Croppenstedt, A., & Meschi, M. (1997). Assessing Wage Discrimination in Italy. *Centre for International Business Studie*s, London Southbank University, Issue 2, working paper No. 11-98 (pp. 1–25).

Díaz, M.A., & Sánchez, R. (2011). Gender and potential wage in Europe: a stochastic frontier approach. *International Journal of Manpower*, 32(4), 410-425.

Garcia-Prieto, C., & Gómez-Costilla, P. (2017). Gender wage gap and education: A stochastic frontier approach. *International Journal of Manpower*, 38(3), 504–516.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, 104(4), 1091-1119.

Gunderson, Morley. 1989. "Male-Female Wage Differentials and Policy Responses." *Journal Economic Literature*. 27:1, pp. 46–72

Furno, M. (2013). Quantile regression and structural change in the Italian wage equation, *Economic Modelling*, 30, 420-434.

Herzog Jr, H. W., Hofler, R. A., & Schlottmann, A. M. (1985). Life on the frontier: migrant information, earnings and past mobility. *The Review of Economics and Statistics*, 373-382.

Jane, W. J. (2013). Overpayment and reservation salary in the Nippon Professional Baseball League: A stochastic frontier analysis. *Journal of Sports Economics*, 14(6), 563-583.

Kumbhakar, S. C., & Lovell, C. K. (2000). *Stochastic Frontier Analysis*. Cambridge University Press.

Lang, G. (2005). The difference between wages and wage potentials: Earnings disadvantages of immigrants in Germany. *The Journal of Economic Inequality*, 3, 21-42.

MacPherson, D. A., Hirsch, B. T. (1995). Wages and Gender Composition: Why Do Women's Jobs Pay Less? *Journal of Labor Economics*.13:3, 426–71.

Meeusen, W., & van Den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 435-444.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. Columbia University Press, New York, NY.

Ogloblin, C., & Brock, G. (2005). Wage determination in urban Russia: Underpayment and the gender differential. *Economic Systems*, 29(3), 325-343.

O'Neill, June and Solomon Polachek**. (**1993). "Why the Gender Gap in Wages Narrowed in the 1980s." *Journal of Labor Economics,* 11(1), 205–28.

Perez-Villadóniga, M. J., & Rodriguez-Alvarez, A. (2017a). Analysing wage differentials when workers maximize the return to human capital investment. *Applied Economics*, 49(42), 4196–4208.

Perez-Villadóniga, M. J., & Rodriguez-Alvarez, A. (2017b). Comparing the gender gap in gross and base wages. *International Journal of Manpower*, 38(5), 646–660.

Perez-Villadóniga, M.J., Rodriguez-Alvarez, A. & Roibas, D. (2025): An alternative approach for analysing the gender wage gap in managerial positions: a stochastic latent class model, *Applied Economics*, DOI: 10.1080/00036846.2025.2471037

Polachek, S. W., & Robst, J. (1998). Employee labor market information: comparing direct world of work measures of workers' knowledge to stochastic frontier estimates. *Labour Economics*, 5(2), 231-242.

Polachek, S. W., & Yoon, B. J. (1987). A two-tiered earnings frontier estimation of employer and employee information in the labor market. *The Review of Economics and Statistics*, 296-302.

Polachek, S. W., & Yoon, B. J. (1996). Panel estimates of a two-tiered earnings frontier. *Journal of Applied Econometrics*, 11(2), 169-178.

Reifschneider, D., & Stevenson, R. (1991). Systematic departures from the frontier: a framework for the analysis of firm inefficiency, *Int. Econ. Rev.* 32(3): 715–723.

Robinson, M. D., & Wunnava, P. V. (1989). Measuring direct discrimination in labor markets using a frontier approach: evidence from CPS female earnings data. *Southern Economic Journal*, 212-218.

Slottje, D. J., Hirschberg, J. G., Hayes, K. J., & Scully, G. W. (1994). A new method for detecting individual and group labor market discrimination. *Journal of Econometrics*, 61(1), 43-64.

Watson, D. (2000). UK wage underpayment: implications for the minimum wage. *Applied Economics*, 32(4), 429-440.

Zveglich Jr, J. E., van der Meulen Rodgers, Y., & Laviña, E. A. (2019). Expected work experience and the gender wage gap: A new human capital measure. *Economic Modelling*, 83, 372-383.

Zhao, X., Jiao, Y., & Wu, D. (2022). The impact of Internet use on labor wage distortions: Empirical Evidence from China. *SAGE Open*, 12(2), 21582440221099290.

Appendix

Table A1. Correlation matrix for variables included in the estimated models

	Worked_hours	D_Female	Work_exp	D_edu_primary	D_edu_lowsec	D_edu_upsec	D_edu_training	D_manager	D_professional	D_admin	D_service_worker	D_craft_workers	D_manufacturing	D_trade	D_construction	D_ever_married	D_rural_areas	D_access_internet	D_NorthEast	D_Centre	D_South	D_Islands
Worked_hours	1																					
D_Female	-0.2875	1																				
Work_experience	0.0350	-0.0821	1																			
D_edu_primary	0.0425	-0.0335	0.0170	1																		
D_edu_lowsec	0.0944	-0.1479	0.2159	-0.0578	1																	
D_edu_upsec	0.0493	-0.0329	0.0387	-0.1064	-0.47	1																
D_edu_training	0.0063	0.0267	0.0086	-0.0168	-0.0742	-0.1366	1															
D_manager	0.0720	-0.0451	0.0267	0.0158	-0.0452	-0.0458	-0.0117	1														
D_professional	-0.2411	0.1443	-0.0566	-0.0506	-0.2592	-0.1252	0.0184	-0.1330	1													
D_admin	0.0517	0.0702	0.0258	-0.0178	-0.0719	0.1353	0.0279	-0.0610	-0.4397	1												
D_service_worker	0.0573	0.0730	-0.0494	0.0162	0.0866	0.0584	-0.0187	-0.0511	-0.3684	-0.1689	1											
D_craft_workers	0.1203	-0.2371	0.0430	0.0468	0.2463	-0.0033	-0.0182	-0.0487	-0.3511	-0.161	-0.1349	1										
D_manufacturing	0.1926	-0.2305	0.0284	-0.0163	0.1422	0.0447	-0.0258	-0.0210	-0.1530	-0.036	-0.1386	0.2307	1									
D_trade	0.0648	0.0105	-0.0296	-0.0086	0.0307	0.0634	-0.0126	0.0170	-0.1429	0.0123	0.2660	-0.0171	-0.227	1								
D_construction	0.0891	-0.1770	0.0248	0.0834	0.1048	-0.0078	-0.0044	-0.0042	-0.0821	-0.0397	-0.0642	0.2540	-0.1632	-0.0878	1							
D_ever_married	-0.0328	0.0013	0.3332	0.0174	0.0715	-0.0116	-0.0278	0.0061	0.0116	-0.0034	-0.0327	-0.0106	-0.0038	-0.0057	0.0078	1						
D_rural_areas	0.0216	-0.0179	0.0052	0.0168	0.0334	0.0458	-0.0134	-0.0174	-0.0397	-0.023	-0.0084	0.0564	0.0375	-0.0076	0.0449	-0.0303	1					
D_access_internet	-0.0191	0.0230	-0.0011	-0.0985	-0.1071	0.0183	0.0237	0.0232	0.0900	0.0181	-0.0767	-0.0691	0.0080	-0.0106	-0.0641	0.0479	-0.0016	1				
D_NorthEast	0.0675	-0.0006	0.0553	-0.0090	-0.0265	0.0455	0.0007	0.0223	-0.0113	-0.0093	-0.0094	0.0184	0.0744	0.0216	-0.0157	-0.0324	0.0537	0.0414	1			
D_Centre	0.0218	0.0113	0.0017	0.0060	-0.0106	0.0266	-0.0176	0.0106	-0.0025	-0.0001	0.0246	-0.0222	-0.0253	-0.0129	0.0126	-0.0055	-0.0410	0.0044	-0.3661	1		
D_South	-0.1291	-0.0207	-0.0743	0.0122	0.0467	-0.0520	-0.0043	-0.0199	0.0280	-0.0153	-0.0089	0.0064	-0.0514	-0.0214	0.0264	0.0571	0.0447	-0.0406	-0.2705	-0.2726	1	
D_Islands	-0.0596	0.0139	-0.0352	0.0284	-0.0083	-0.0363	0.0258	-0.0105	0.0211	-0.0089	0.0102	-0.0100	-0.0856	0.0023	0.0237	0.0150	-0.0164	-0.0165	-0.1550	-0.1563	-0.1155	1

Source: own elaborations on data from EU-SILC for Italy.