

ORIGINAL ARTICLE

Open Access



# An analysis of Italian university students' performance through segmented regression models: gender differences in STEM courses

Andrea Priulla, Nicoletta D'Angelo\*  and Massimo Attanasio

\*Correspondence:

[nicoletta.dangelo@unipa.it](mailto:nicoletta.dangelo@unipa.it)

Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy

## Abstract

This paper investigates gender differences in university performances in Science, Technology, Engineering and Mathematics (STEM) courses in Italy, proposing a novel application through the segmented regression models. The analysis concerns freshmen students enrolled at a 3-year STEM degree in Italian universities in the last decade, with a focus on the relationship between the number of university credits earned during the first year (a good predictor of the regularity of the career) and the probability of getting the bachelor degree within 4 years. Data is provided by the Italian Ministry of University and Research (MIUR). Our analysis confirms that first-year performance is strongly correlated to obtaining a degree within 4 years. Furthermore, our findings show that gender differences vary among STEM courses, in accordance with the care-oriented and technical-oriented dichotomy. Males outperform females in mathematics, physics, chemistry and computer science, while females are slightly better than males in biology. In engineering, female performance seems to follow the male stream. Finally, accounting for other important covariates regarding students, we point out the importance of high school background and students' demographic characteristics.

**Keywords:** Higher education, Segmented regression, Students' performance, Gender differences, STEM

## Introduction

In the last years, studies on students' university experiences have been increasingly common (Salanova et al. 2010; Mega et al. 2014; Freeman et al. 2014). In Italy, several reasons have fostered this interest: first, the Bologna and the Bergen and Lisboa processes; second, the reform of the Italian University system in 2001 introducing the two-step degree program with a bachelor degree and a master degree; and, lastly, the central government's new funding system based on the regularity of students' careers.

Moreover, Italian university students, as in most of the other western countries (Mostafa 2019), are not particularly likely to enroll at Science, Technology, Engineering and Mathematics (STEM) courses. What is more, STEM courses have higher overall dropout rates than the other courses and there are often few women enrolled (Attanasio et al. 2018). Indeed, in Italy, since 1989, more females enroll at Italian universities than

men, but females are still underrepresented in almost all the STEM fields and overrepresented in nursing, the humanities, and the law schools (De Vita and Giancola 2017). On average across OECD countries in 2017, females made up 30% of new entrants to bachelor level in STEM fields, and 77% of new entrants to health and humanities at the bachelor's level (OECD 2019).

This paper aims at investigating the differences among STEM courses in terms of female and male performances. To measure performance (university success), we focus on the accumulation of CUs (university credits) during the first year, since it represents an important moment in the students' path at university.

As highlighted by Barone and Assirelli (2020), individuals tend to favor degree courses that increase their chances of succeeding, with females more likely to follow some fields of study and the males others. Therefore, this paper aims at providing an insight into the performance of freshmen university students in Italy. Our hypothesis is that gender differences can also vary among the different STEM courses: we are mindful of the recent work of Barone et al. (2019) that has claimed the presence of a care-technical divide within STEM courses.

The literature review on gender differences in STEM is in "[Literature review on gender differences in STEM courses](#)" section, which is essentially an international glance with some references to the Italian context. In "[Data, variables, and methods](#)" section the data, the variables and the methods are introduced, for Italian longitudinal university data with a focus on STEM courses. "[Analysis of university students' performance](#)" section provides a data analysis. In particular, "[Exploratory analysis](#)" subsection deals with an exploratory analysis of the data, "[Modeling strategy](#)" subsection outlines the modeling strategy based on the segmented regression models, and "[Results](#)" subsection contains the results of the analysis. Finally, in the conclusions we try to connect our findings with the theory reported in "[Literature review on gender differences in STEM courses](#)" section.

### **Literature review on gender differences in STEM courses**

We briefly describe some papers on the gender gap in STEM, firstly concerning the high school and then the university, with some references to the international literature and some specific references to the Italian experience. These papers generally focus on student performance in math tests, since mathematics can be seen as an "indicator" of STEM ability, and it can, as such, be used as a proxy for future university success.

First, it should be noted that the definition of STEM can differ from country to country (Fan and Ritz 2014). For example, medicine, structural engineering and sports science are not always included in definitions. Core STEM subjects typically include: mathematics; chemistry; computer science; biology; physics; architecture; and, general, civil, electrical, electronics, communications, mechanical, and chemical Engineering (UK-Parliament 2020). In our analyses, we will consider the above mentioned core definition excluding architecture.

As to students' performance in high school, many authors analyze gender differences in STEM. There is an extensive literature addressing the underperformance of males from the first schooling years. Females tend to do better than males in reading test scores, in grades completion and repetition at school, in the likelihood to choose academic educational programs in upper secondary school, in tertiary education attendance, and in bachelor graduation rates (Legewie and DiPrete 2012).

While, from a psychological point of view, it has been often hypothesized that females have an innate predisposition to prefer educational paths in the humanistic and caring disciplines, a wide range of theories have been discussed over the years to explain these differences in educational choices. For instance, (Sherman 1980) discussed that family, as well as school environment and teachers' attitudes, have a strong influence in directing males and females to develop different attitudes towards certain subjects and skills, affecting their educational choices.

An interesting explanation of this phenomenon comes from (Barone et al. 2019), who highlights the absence of accurate high school information on the long-term job opportunities related to specific degree courses. Here we are thinking of the economic rewards or career opening, leading students to base their choice only on their preferred subjects or "dream" occupations, which are often gender-stereotyped. These differences in degree course choice reflect those present in worldwide culture. The authors also note another divide beyond the humanistic-scientific cleft, namely the care-technical divide. "Consequently, females are not underrepresented in all the STEM courses, but mostly in the more technical ones, such as engineering and computer science, while they are numerous in biology or health-care professions, fields historically related to the traditional female stereotype. In fact, some fields of study prepare students for care-jobs, while others can address students to a care-job like teaching as a second-best option, such as some scientific fields like mathematics and biology." (Barone et al. 2019). It would be misleading to consider all the STEM courses to be masculinized, because biology is indeed more care-oriented than, for example, computer science. In this respect, we will consider STEM courses separately to better examine the gender divide.

The stereotyped divide between male and female fields and occupations is obviously mirrored in university choices. Although females usually perform better at university than men, female students may face more serious difficulties in STEM, leading them to switch to a non-STEM program in the second year and, with regards to some scientific courses, to quit their university career altogether (Attanasio et al. 2018). A possible explanation of this phenomenon is given by Hall and Sandler (1982), who defined STEM courses, especially engineering ones, as a "chilly climate" for female students, saying that faculty express higher expectations for male students, or lead females to feel their ambitions are not as important as the ambition of their male colleagues. It is worth noting that other theories exist, such as the rational choice theory, which argues that individuals tend to prefer educational options that enhance their chances of success (Barone and Assirelli 2020). This less well-known theory conceptualizes gender differentiation as an outcome of both socialization processes and rational choice factors (Gabay-Egozi et al. 2015). According to this theory, students who are more career-oriented display a lower propensity to enroll at care-oriented courses. This issue, together with females preferring soft fields because they give less importance to career prospects, lead to few females choosing a more technical career path.

From a sociological point of view, an interesting theory is given by Correll (2001). The author states that gender differences in mathematics do not seem to be responsible for the large differences between the numbers of males and females enrolling at fields requiring a higher level of mathematical competence. She argues that cultural beliefs about gender and mathematics affect the choices of males and females towards educational paths leading to STEM careers in a different way. Indeed, the author claims that some individuals

probably come to personally believe that males are better at math, though females have been shown to be less likely than males to hold stereotypical views about mathematics. “Therefore, if a girl believes that males are better at math, she might view mathematical competence does not match her female gender identity, leading her to doubt her mathematical ability, and consequently to decrease her interest in careers requiring high levels of mathematical competence. However, it is only necessary that individuals perceive that others hold these gendered beliefs with respect to mathematics to lead to biased self-assessments of their ability and reduce their performance.” (Correll 2001). The main conclusions of her work showed that, since males tend to overestimate their mathematical competence relative to females, males are also more likely to pursue activities that will lead to STEM careers. This is an interesting explanation, even if it necessarily made with reference to the US.

In Italy, several papers look at how males do better than females in math tests, and some explanations have been suggested to try to explain this gap. These studies are mainly based on INVALSI tests which are administered to students through the schooling years. For instance, the gender gap in math test scores in Italy, which is one of the countries with the largest differential between males and females, shows that females systematically underperform in relation to males, even after controlling for a set of individual and family background characteristics. These results show how the average gender gap increases with children’s age and becomes larger among top-performing children. Therefore, females’ underperformance in mathematics could explain the tendency for females to follow non-scientific careers (Contini et al. 2017).

For the university level, some papers have highlighted the importance of differentiating the analysis of students’ performance with respect to enrolment courses. For instance, (Cheryan et al. 2009) examine the determinants of participation in computer science courses, showing that the interest in computer science is influenced by the exposure to environments associated with computer scientists. The conclusions drawn in their paper is that changing stereotypical computer science environments could inspire a new interest in pursuing for this specific degree choice. Eccles (2007) analyzes why females continue to be underrepresented in the physical sciences and engineering in universities and colleges. Her analysis suggests that the main explanation for gender differences in the physical sciences and engineering occupations is the difference placed on different types of occupations by males and females. Looking at Italian high school students, (Barone and Assirelli 2020) highlight the key role of curricular track choices, stating that this single factor mediates the most gender differences in access to engineering, and information and communication technologies courses at university. “This is because curricular choices in high school are heavily segregated along gender lines and curricular track displays a strong influence on field of study choices”. In biological sciences courses, classified by Barone et al. (2019) as care-oriented, little attention has been paid to the performance of females in comparison with males or perceptions of stereotype threat (Lauer et al. 2013). In particular, (Simon 2010) studied gender differences in knowledge and attitude towards biotechnology. His studies follow those suggested by Correll (2001), in which, more knowledge in biotechnology decreased students’ probability of being pessimistic about science. But for females more knowledge in biotechnology actually led to a greater probability of pessimism. Biology courses are considered an exception among STEM fields, since they are female-dominated. In fact, (Eddy et al. 2014) states that: “Often,

gender differences are assumed to be present only in fields where males outnumber females and where there is a strong emphasis on math, but we are seeing it in undergraduate biology classrooms that do not focus on math - where females make up about 60 percent of the class - indicating that this could potentially be a much more systemic problem. It's likely this is not unique to physics or biology, but rather true of most undergraduate classrooms."

Finally, some recent studies deal with gender gap at university in Italy, with an insight relating to STEM courses: "Females have more success in terms of bachelor graduation in geology, biology, biotechnology, and statistics while they seem to suffer in all the remaining STEM courses, especially in mathematics" (Enea and Attanasio 2020).

## Data, variables, and methods

### Data

The data comes from the ANS (Anagrafe Nazionale Studenti), which is the database of Italian university students. Each freshman enrolled at an Italian university represents a statistical unit/record, which can be divided into two main parts: the first regarding high school background, and the second, divided into  $k$  parts (each one representing an academic year), which contains variables on their university career. In this way, we can analyze students' performance longitudinally, taking in their whole university trajectory.

To study students' performance through their ability to get the degree within 4 years, cohorts of students are analyzed in 4-year time intervals. Also, this allows for a follow up looking at their progress from enrolment to the completion of the bachelor degree (or dropout). We will choose the 2014 cohort, the most recent available cohort. This will allow us to cover a period long enough to observe the completion of the degree.

Students enrolled at an online university are excluded from the study, because they behave differently in terms of degree rates. They obtain their bachelor degree more rapidly than students enrolled at a non-online university. We include students enrolled at both private and public universities. That distinction is not important since STEM courses provided by private universities in Italy are limited and, therefore, not comparable to those of public universities. Also, we do not exclude dropout students from our analysis, since the bachelor completion rate would, were we to ignore them, be overestimated. Moreover, high school grades in scientific subjects could be useful for understanding university performance, but we do not have this type of information in our data. Previous high school variables and other personal characteristics are available, and they are named "admission covariates" due to their availability at the moment of enrolment at university. In addition, we do not know whether students are enrolled part-time. Finally, a limit of this study is that we do not have any information on family socio-economic background.

### Variables

Our analysis aims at modeling the probability of getting the bachelor degree within 4 years (i.e. the response variable), with respect to the number of CUs earned at the end of the first year (CU), plus a set of covariates. Those are:

- CU: *university credits*, which ranges from 0 to 60 (which are the annual credits);
- gender;

- *age*: *age at enrolment*, dichotomized in  $\leq 19$  and  $> 19$ . We chose this dichotomization since a student with a regular path enrolls at least at the age of 19;
- *macro-region*: *macro-region of enrolment*, categorized in North, Center, South, and Islands;
- *HSdiploma*: *high school diploma*, categorized in Classical “liceo”, Scientific “liceo”, Technical institute, Vocational institute, Other “liceo”, and Abroad/Other. The first two “licei” are the traditional preparatory high schools for university;
- *HSmark*: *high school final mark*, which ranges from 60 to 101, where 101 identifies “100 cum laude”;
- *degree course*: which identifies the 3-year STEM degree of enrolment. Those are classified into two main groups:
  - *Care-oriented* courses: biology, biotechnology, and mathematics
  - *Technical-oriented* courses: chemistry, computer science, engineering, natural sciences (which includes both geology and environmental sciences), physics, and statistics.

Mathematics is included in the care-oriented group, following (Barone et al. 2019), since most of the students enroll at this course with the aim of taking up a teaching career.

## Methods

We analyze the gender-specific students’ performance in STEM courses through the application of the *segmented regression models*. Generally speaking, segmented regression models allow us to obtain a more synthetic representation and better interpretation of the students’ progress at university through the changepoints, both analytically and graphically, when compared to other standard methods widely used in the literature. All the analyses are performed using the R `segmented` package (Muggeo and et al 2008), and the codes of the analyses carried out throughout the paper are available from the authors.

### *Background on the segmented regression models*

Segmented or broken-line models are regression models where the relationships between the response and one or more explanatory variables are piecewise linear and, as such, represented by two or more straight lines connected at unknown points. These models are a common tool in many fields, including epidemiology, occupational medicine, toxicology and ecology, where usually it is of interest to assess threshold values where the effect of the covariate changes (Ulm 1991; Betts et al. 2007). The main advantage of this approach is the easy interpretation given by two components: the changepoint (or the changepoints) and the slopes.

These models represent a good trade-off between flexibility and computational burden, like the usual non-parametric approaches. Recent papers deal with applications of segmented regression models in higher education (Li et al. 2019), but to the best of our knowledge, this paper represents the first application of segmented regression models applied to predict university success.

The segmented linear regression is expressed as

$$g(E[Y|x_i, z_i]) = \alpha + z_i^T \theta + \beta x_i + \sum_{k=1}^{K_0} \delta_k (x_{i,k} - \psi_k)_+ \quad (1)$$

where  $g$  is the link function,  $x_i$  is the broken-line covariate and  $z_i$  is a covariate vector whose relationship with the response variable is a non broken-line. We denote by  $K_0$  the true number of changepoints and by  $\psi_k$  the  $K_0$  locations of the changepoints in the observed phenomenon. These  $K_0$  are selected among all the possible values in the range of  $x$ . The term  $(x_i - \psi_k)_+$  is defined as  $\sum_i I(x_i > \psi_k)$  that is  $(x_i - \psi_k)I(x_i > \psi_k)$ . The parameter estimates  $\theta$  represent the non broken-line effects of  $z_i$ ,  $\beta$  represents the effect for  $x_i < \psi_1$ , while  $\delta$  is the vector of the differences in the effects.

The parameters to be estimated are: the number of changepoints  $K_0$ ; their locations  $\psi_k$ ; and the broken-line effects, represented by  $\beta$  and  $\delta$ . For the estimation procedure, we refer to Muggeo (2003). Typically, we would need to select the significant changepoints by removing the spurious ones. Indeed, whether the generic  $\hat{\psi}_k$  is not significant, the corresponding covariate  $V_k$  should be a noise variable, as it would be  $\hat{\delta}_k \approx 0$ . The fitted 'optimal' model will have  $\hat{K} \leq K_0$  changepoints selected by any criterion. Indeed, literature is concerned with the problem of determining the 'best' subset of independent variables, conducted with two major approaches, namely information criteria and hypothesis testing (Hocking 1976).

We refer to D'Angelo and Priulla (2020) for a complete description of both the problem of estimating the number of changepoints and the criteria adopted. In that paper, a modified version of the usual procedure for the selection of the number of changepoints is proposed. This version is based on sequential hypothesis testing, and its "validity" is assessed through simulations, proving that the proposal correctly identifies the true number of changepoints and, in particular, it outperforms all the considered information-based criteria competitors in the binomial case. Therefore, the procedure reported below will be performed throughout the analyses in this paper.

#### ***Sequential hypothesis testing procedure for the choice of $K_0$***

An approach for the selection of the number of changepoints is proposed in Kim et al. (2000), relying on sequential hypothesis testing procedure. It consists of performing different hypothesis tests starting from  $\mathcal{H}_0 : K_0 = 0$  vs.  $\mathcal{H}_1 : K_0 = K_{max}$ , where  $K_{max}$  is fixed a priori. If the null hypothesis is rejected, the procedure tests for the next hypothesis by increasing the number of changepoints specified in  $\mathcal{H}_0$  or by decreasing the one postulated under  $\mathcal{H}_1$ . D'Angelo and Priulla (2020) propose a different sequential procedure, to identify the correct number of changepoints through the pseudo-score test or through the Davies' test.

Starting from  $\mathcal{H}_0 : K_0 = 0$  vs  $\mathcal{H}_1 : K_0 = 1$ , and depending on the tests' results, the procedure ends testing at most  $\mathcal{H}_0 : K_0 = K_{max} - 1$  vs  $\mathcal{H}_1 : K_0 = K_{max}$ , and selecting up to  $K_{max}$  changepoints. Furthermore, we control for the over-rejection of the null hypotheses at the overall level  $\alpha$ , employing the Bonferroni correction comparing each p-value with  $\alpha/K_{max}$ . Of course, setting the Bonferroni correction to  $\alpha/K_{max}$  is conservative.

As compared to the procedure in Kim et al. (2000), the proposal of (D'Angelo and Priulla 2020) has the advantage of not being limited to test for a maximum number of additional a priori fixed changepoints. Indeed, the proposal of (Kim et al. 2000) makes testing for more than two additional changepoints with the pseudo-score unfeasible, because the current implementation of the pseudo-score test in R does not allow for testing for  $\mathcal{H}_0 : K_0 = K$  vs  $\mathcal{H}_1 : K_0 = K + 3$ . D'Angelo and Priulla (2020) overcome this problem accommodating for any number of additional changepoints through the sequential procedure, outlined below.

**Steps of the procedure** For  $K_{max} = 2$  the procedure is as follows:

- 1 Fit a segmented model to the data, with  $\hat{K} = 1$  and test

$$\begin{cases} \mathcal{H}_0 : \delta_1 = 0 & (K_0 = 0) \\ \mathcal{H}_1 : \delta_1 \neq 0 & (K_0 > 1) \end{cases}$$

via the Score or Davies' test. If  $\mathcal{H}_0$  is not rejected then  $\hat{K} = 0$  and the procedure stops at this step. Otherwise, go to the next step.

- 2 Fit a segmented model with  $\hat{K} = 2$  and test

$$\begin{cases} \mathcal{H}_0 : \delta_2 = 0 & (K_0 = 1) \\ \mathcal{H}_1 : \delta_2 \neq 0 & (K_0 > 2) \end{cases}$$

If  $\mathcal{H}_0$  is not rejected then  $\hat{K} = 1$ , otherwise,  $\hat{K} = 2$ .

In practice, the iterative procedure with the Davies' test always stops as it gets  $\hat{K} = 2$ , even if the actual number can be larger. This is because each step tests for at least an additional changepoint.

## Analysis of university students' performance

### Exploratory analysis

This section is divided into three parts: the first concerns the enrolment, the second the first year, and the third the relationship between high school completed and university performance.

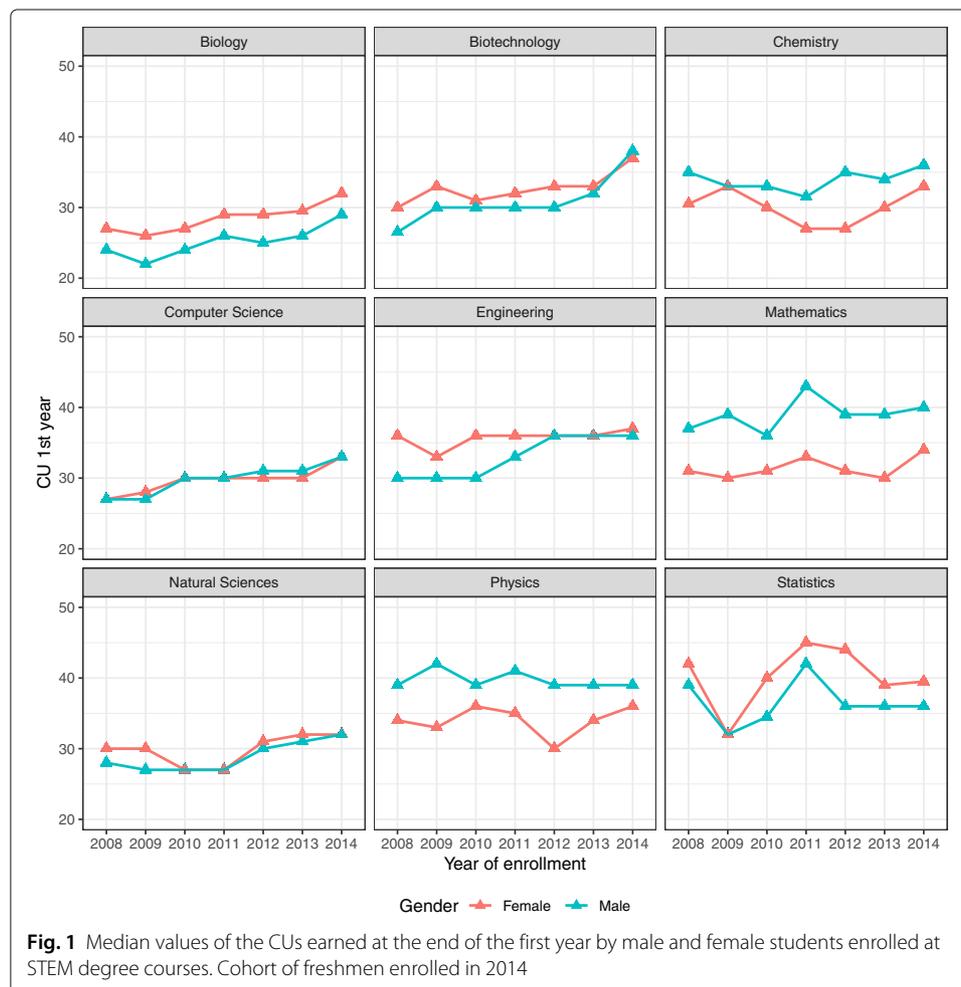
First, we investigate the female enrolment rate in the last decade in STEM courses in Italy. Female students enrolled for the first time in STEM courses, and the female enrolment rate for the 2008 and 2014 cohorts are reported in Table 1. The 2008 cohort is also included to get a temporal comparison. At first glance, it is evident that female students prefer to enroll at care courses, such as biology or biotechnology, with percentages ranging from 64 to 74%. Conversely, female students have a lower interest towards technical-oriented courses, with percentages between 12 and 25%. Looking at the differences between the two cohorts, the overall number of enrolled students has increased in 6 years, regardless of gender. Although the total number of enrolled female students remains almost stable, the percentage of females enrolled decreased by more than 2%.

**Table 1** Female students enrolled for the first time in STEM courses, and female enrolment rate of the 2008 and 2014 cohorts

	Cohort					
	2008			2014		
Degree course	F	%	Total	F	%	Total
<b>Biology</b>	4785	73.6	<b>6505</b>	3945	72.4	<b>5446</b>
<b>Biotechnology</b>	2044	64.0	<b>3194</b>	1966	65.3	<b>3013</b>
<b>Chemistry</b>	1042	48.2	<b>2163</b>	1046	44.8	<b>2334</b>
<b>Computer Science</b>	388	14.7	<b>2646</b>	544	13.6	<b>3992</b>
<b>Engineering</b>	5505	22.5	<b>24,506</b>	6864	24.5	<b>28,027</b>
<b>Mathematics</b>	1330	61.8	<b>2152</b>	838	56.2	<b>1490</b>
<b>Natural Sciences</b>	1123	48.5	<b>2315</b>	1165	49.6	<b>2349</b>
<b>Physics</b>	610	34.1	<b>1787</b>	664	31.7	<b>2092</b>
<b>Statistics</b>	326	46.6	<b>700</b>	330	41.5	<b>796</b>
<b>Total</b>	<b>17,153</b>	<b>37.3</b>	<b>45,968</b>	<b>17,362</b>	<b>35.0</b>	<b>49,539</b>

Furthermore, course-specific differences can be identified with respect to gender composition. In particular, the percentages of female students decreased by more than 5% in mathematics and statistics from 2008 to 2014. On the contrary, an increase is recorded in biotechnology, natural sciences and engineering. Engineering reports a significant increase in the total number of females enrolled, together with computer science. The most striking increase is recorded in computer science, with 40%, followed by engineering courses, with almost 25% more female students.

Second, we examine the performance of students during their first year at university by looking at the number of CUs (university credits) earned. The median values of CUs earned at the end of the first year for male and female students are shown separately in Fig. 1, with the cohorts of freshmen enrolled in the academic years from 2008 to 2014. It appears clear that student performance varies among the different degree courses and with respect to gender. Indeed, males outperform females in mathematics. On the one hand, as expected, female students show a better performance in the more care-oriented courses, such as biology and biotechnology. On the other hand, male students show better results in the more technical-oriented courses, such as physics and chemistry. Despite being considered as one of the most “masculinized” courses, engineering does not show a significant gender gap in students’ performance, with female students



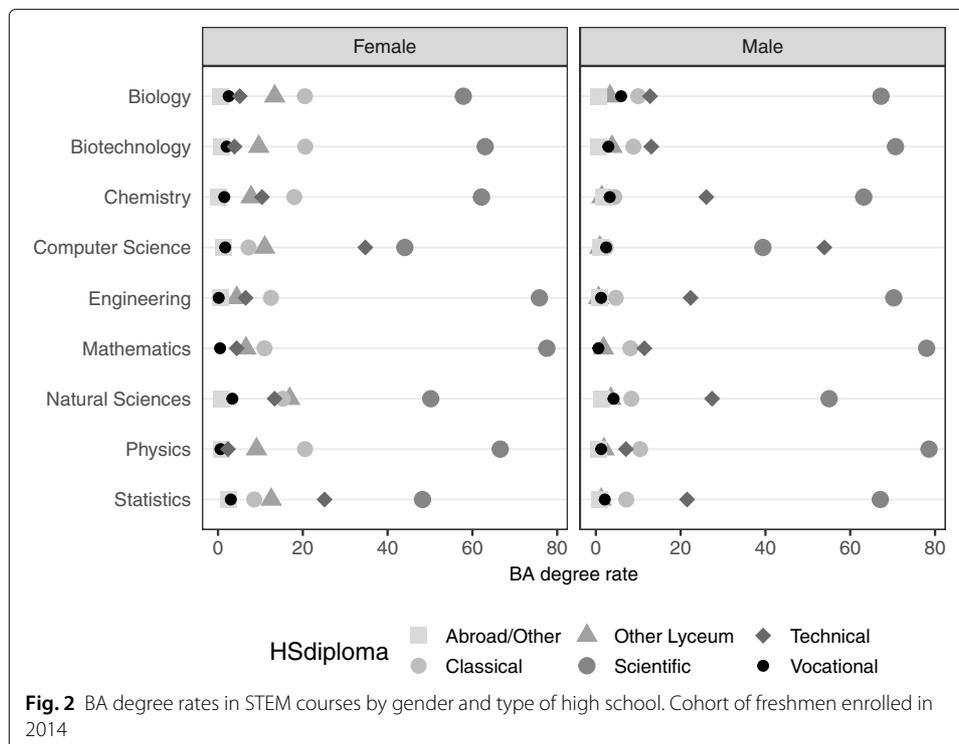
**Fig. 1** Median values of the CUs earned at the end of the first year by male and female students enrolled at STEM degree courses. Cohort of freshmen enrolled in 2014

enjoying slightly better results compared to their male colleagues. Furthermore, even after a slight improvement in the most recent years, natural and computer sciences are the courses where students exhibit the greatest difficulties, and there are no significant gender differences in the performance.

Third, we investigate the relationship between the high school the student attended and university performance by computing the BA degree rates for male and female students, separately. The rate is computed as the number of students who obtained the degree within 4 years over the total of enrolled students of the corresponding cohort. These are shown in Fig. 2. Here there are some differences among degree courses by school type. There seems to be a clear separation between students from a scientific “liceo” and others. This is especially so when compared to students who completed their education at a vocational school. Gender differences come up clearly too. In fact, female students with a classical “liceo” diploma perform better than their male counterparts, while male students who completed their education at a technical school outperform females with the same educational background. A possible explanation for this result can be addressed by the fact that females from a classical “liceo” are likely to be more involved than males from the same background, in facing the challenge of enrolling at a STEM degree course. Students, meanwhile, from a scientific “liceo” seem to perform better in each scenario. Only in computer science courses do male students from technical schools achieve better results than those from a scientific “liceo”. Students who completed their education abroad perform the worst, regardless of their gender and degree course.

**Modeling strategy**

Previous works have suggested a strong relationship between the number of CUs earned during the first year and the probability of getting the BA degree (Attanasio et al. 2013).



In our application,  $\pi$  is the probability of obtaining the BA degree within 4 years from the first university enrolment, that represents the probability of success. The coefficients  $\alpha$  and  $\lambda$  are the intercept and the slope of CU.

To fit (1) we proceed in the following way:

1. We first fit the model (2):

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \lambda\text{CU}_i \tag{2}$$

that accounts for only the covariate CU, to assess its effect on the probability of success.

2. Secondly, to investigate whether this relationship can be considered as segmented, i.e. whether there exist some thresholds in the CUs after which a significant change in the probability of success is recorded, we fit a segmented logistic regression model of the form:

$$\begin{aligned} \log\left(\frac{\pi}{1-\pi}\right) = & \alpha + \lambda_1\text{CU}_{male,i} + \lambda_2\text{CU}_{female,i} + \theta_1\text{gender}_i \\ & + \sum_{j=1}^J (\beta_j\text{CU}_{j,i} + \sum_{k=1}^{K_j} \delta_{j,k}(\text{CU}_{j,i} - \psi_{j,k})_+) \end{aligned} \tag{3}$$

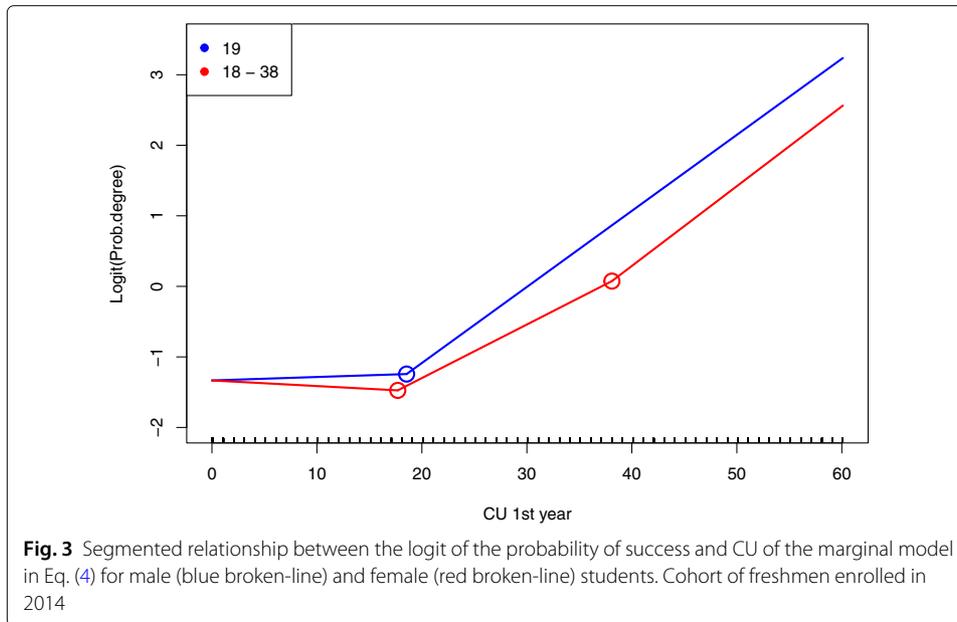
where  $z_i$  is just the variable gender, and  $x_i$  is CU. We include gender as the first and only non-segmented variable in the segmented model because we want to first assess the significance of this variable.

To better analyze gender differences, we accommodate two instrumental covariates into the equation:  $\text{CU}_{male}$  and  $\text{CU}_{female}$ . The baseline profile is: { 0 for  $\text{CU}_{male}$  and  $\text{CU}_{female}$  } and { female for gender}. The covariate gender is indexed by  $j$ , which corresponds to two different segmented relationships, and  $K_{male}$  and  $K_{female}$  are the gender changepoints to be estimated. We will call this model the marginal model.

The segmented regression estimation procedure works plugging in  $\hat{K}_j = 1, 2$  for  $j = \{\text{male}, \text{female}\}$ , separately. In this way, we compare 5 models, given by the combination of the two  $\hat{K}_j$  plus the null model with  $\hat{K} = 0$ . Then, we apply the sequential hypothesis testing procedure outlined in “[Sequential hypothesis testing procedure for the choice of  \$K\_0\$](#) ” section to select the “best” number of changepoints. Basically, the fitted segmented models with  $K$  changepoints are compared to the models with  $K-1$  changepoints. The model selected overall the courses provides  $\hat{K}_{male} = 1$  and  $\hat{K}_{female} = 2$ . In Fig. 3 the broken-line relationship between the logit of the probability of success and CUs is displayed. The first changepoints are not distant, and after them, the two lines are roughly parallel.

3. Model (3) can be further specified, including the admission covariates, obtaining Eq. 4.

$$\begin{aligned} \log\left(\frac{\pi}{1-\pi}\right) = & \alpha + \lambda_1\text{CU}_{male,i} + \lambda_2\text{CU}_{female,i} + \theta_1\text{gender}_i + \theta_2\text{macro-region}_i \\ & + \theta_3\text{HSdiploma}_i + \theta_4\text{HSmark}_i + \theta_5\text{age}_i + \theta_6\text{degreecourse}_i \\ & + \theta_7\text{HSdiploma}_i * \text{gender}_i \\ & + \sum_{j=1}^J \left( \beta_j\text{CU}_{j,i} + \sum_{k=1}^{K_j} \delta_{j,k}(\text{CU}_{j,i} - \psi_{j,k})_+ \right). \end{aligned} \tag{4}$$



where  $z_i$  now contains all the admission covariates, and  $x_i$  is CU. We are aware that there is a strong relationship between the covariate CU and the admission covariates, as CUs are determined at the end of the first year, but the inclusion of  $z_i$  leads to a more than 20% improvement in the fitting, due to the prolonged effect of  $z_i$  on the probability of getting the degree.

The baseline profile for the admission covariates is: {0 for CU}, {female for gender}, {Islands for macro-region}, {60 for HSmark}, {Biology for degreecourse}, {Other "liceo" for HSdiploma}, and { $\leq 19$  for age},

As before, the model selected provides  $\hat{K}_{male} = 1$  and  $\hat{K}_{female} = 2$ . The summary of the parameter estimates, the constant effects ( $\theta$ ) and the broken-line effects ( $\psi$  and  $\delta$ ), is reported in Tables 2 and 3.

Before analyzing the results, it is important to stress that the estimated parameters of the chosen model cannot be considered in the usual "inferential" way, since the dataset is a population. Nevertheless, the usual statistical procedures of model selection and estimation are used to better understand the relationship among variables.

4. Finally, since our interest lies in analyzing the relationship between gender and STEM, we proceed with the estimation of a stratified model. In fact, to avoid inserting several dummies, given by the couples {degreecourse $_l$ \*CU $_j$  with  $l = 1, \dots, 9$ ;  $j = male, female$ }, we fit  $L = 9$  course-specific segmented regression models, as in Eq. 4, as follows:

$$\log\left(\frac{\pi_l}{1 - \pi_l}\right) = \alpha_l + \lambda_{1,l}CU_{male,i} + \lambda_{2,l}CU_{female,i} + \theta_{1,l}gender_i + \theta_{2,l}macro-region_i + \theta_{3,l}HSdiploma_i + \theta_{4,l}HSmark_i + \theta_{5,l}age_i + \theta_{6,l}HSdiploma_i * gender_i + \sum_{j=1}^J (\beta_{j,l}CU_{j,i} + \sum_{k=1}^{K_j} \delta_{j,l,k} (CU_{j,i} - \psi_{j,l,k})) \quad (5)$$

**Table 2** Parameter estimates  $\theta$ 's of the segmented regression model in Eq. 4

Variable	Category	Estimate	S.E.	P value
<b>Intercept</b>		- 1.96	0.16	0.00
<b>Gender f</b>	<b>Male</b>	- 1.03	0.19	0.00
<b>CU male</b>		- 0.02	0.01	0.03
<b>CU female</b>		- 0.03	0.01	0.01
<b>HSdiploma (Other "liceo")</b>	<b>Scientific</b>	- 0.15	0.06	0.02
	<b>Classical</b>	- 0.17	0.07	0.02
	<b>Technical</b>	- 0.39	0.08	0.00
	<b>Vocational</b>	- 0.45	0.15	0.00
<b>Age (<math>\leq 19</math>)</b>	<b>Abroad/Other</b>	- 0.45	0.21	0.03
	<b>&gt; 19</b>	- 0.53	0.04	0.00
	<b>Biotechnology</b>	- 0.26	0.06	0.00
	<b>Chemistry</b>	0.03	0.06	0.69
<b>Degree course (Biology)</b>	<b>Computer Science</b>	- 0.49	0.06	0.00
	<b>Engineering</b>	- 0.29	0.04	0.00
	<b>Mathematics</b>	- 0.32	0.07	0.00
	<b>Natural Sciences</b>	0.08	0.06	0.20
	<b>Physics</b>	- 0.28	0.07	0.00
	<b>Statistics</b>	0.12	0.10	0.23
<b>Macro-region of enrolment (Islands)</b>	<b>North</b>	0.24	0.04	0.00
	<b>Center</b>	0.07	0.05	0.17
	<b>South</b>	- 0.18	0.05	0.00
<b>HS final mark (60)</b>		0.02	0.00	0.00
	<b>Scientific</b>	0.28	0.14	0.05
<b>HSdiploma x Gender (Female)</b>	<b>Classical</b>	0.33	0.16	0.03
	<b>Technical</b>	0.39	0.15	0.01
	<b>Vocational</b>	0.39	0.21	0.07
	<b>Abroad/other</b>	0.50	0.28	0.08

Baselines are in brackets. Cohort of freshmen enrolled in 2014

**Results**

We now present the interpretation of the results of the segmented model in Eq. 4. Looking at Table 2, we notice that the parameter estimates of CU for both male and female students are very close. Then, looking at Table 3, the other estimated parameters  $\hat{\beta}_m, \hat{\beta}_f$  and  $\hat{\delta}_{m,1}, \hat{\delta}_{f,1}, \hat{\delta}_{f,2}$  concern the segmented variable CU. Male students show  $\hat{K}_m = 1$ , which is located at  $\psi_{m,1} = 18.85$ . For female students, we have that  $\hat{K}_f = 2$ , which are located at  $\psi_{f,1} = 15.14$  and  $\psi_{f,2} = 29.22$ . In practice, when the students earn less than around 20 CUs the probability of success does not change, regardless of gender. As shown in Fig. 3, after 20 CUs, the male line is always above the female one, with a slight difference till 30 CUs and afterwards, the two lines run in parallel.

**Table 3** Parameter estimates of the  $\psi$ 's and  $\delta$ 's of the segmented regression model

Variable	Parameter	Estimate	S.E.
<b>CU male</b>	$\psi_{m,1}$	18.85	0.72
<b>CU female</b>	$\psi_{f,1}$	15.14	1.66
	$\psi_{f,2}$	29.22	1.66
<b>CU male</b>	$\delta_{m,1}$	0.12	0.01
<b>CU female</b>	$\delta_{f,1}$	0.08	0.02
	$\delta_{f,2}$	0.05	0.01

Cohort of freshmen enrolled in 2014

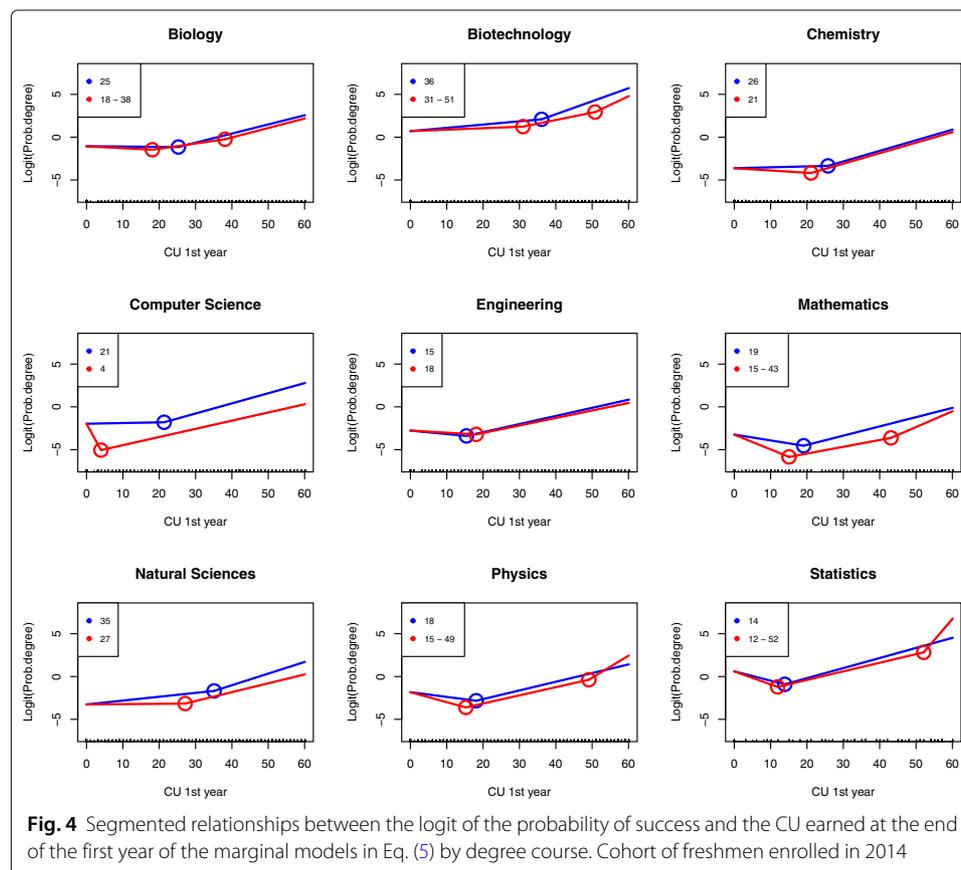
**Table 4** Number of selected changepoints by gender and degree course of enrolment

Test	Gender	Degree course								
		Bio	Biot	Chem	Comp	Eng	Math	Nat	Phy	Stat
Davies	Male	1	1	1	1	1	1	1	1	1
	Female	2	2	1	1	1	2	1	2	2
Score	Male	1	1	1	1	1	1	1	1	1
	Female	2	2	1	1	1	2	1	2	2

Furthermore, looking at the admission covariates in Table 2, the first estimate referred to gender shows a generally better female performance, which is attenuated by the interaction `gender*HSdiploma`. As expected, students coming from a traditional “liceo” have a higher probability of getting the BA degree within 4 years versus those with vocational or technical diploma. Moreover, higher HS final marks lead to an increase in this probability. Students who enroll “late” at university are, unsurprisingly, the ones facing most difficulties in achieving the bachelor degree within 4 years.

In respect of the degree course, statistics students perform slightly better; those enrolled at computer science have substantial difficulties in getting their BA degree; while, natural sciences and chemistry are close to biology. As for the macro-region, students enrolled at southern universities have an overall lower probability of success, followed by island students. Northern students perform the best.

We now look for differences among degree courses by interpreting the results of models fitted as in Eq. 5. Table 4 shows one changepoint for males in all the STEM courses and two changepoints for females in only 5 out of 9 courses. In Fig. 4, the estimated change-



**Fig. 4** Segmented relationships between the logit of the probability of success and the CU earned at the end of the first year of the marginal models in Eq. (5) by degree course. Cohort of freshmen enrolled in 2014

points are displayed, and the parameter estimates of the fitted models are reported in Tables 5, 6, 7, and 8. In detail, we identify some course-specific differences in the analyzed relationship. Only some students display a significant effect before the first estimated changepoint. In computer science, mathematics, and physics this occurs only for female students, with a negative coefficient, meaning that before the threshold the probability of success decreases. The effect of CUs before the estimated changepoint is, on the other hand, significant and positive only for males in engineering, natural sciences, and biotechnology courses. One possible explanation could be that these courses are mostly affected by dropouts. This leads to an overall lower estimated probability of getting the bachelor degree within 4 years when a slight increase occurs with a low number of CUs. The first changepoints are almost always located between 10 and 20 CUs, except for biotechnology and natural sciences, for both males and females. The first female changepoints precede the male ones, but for engineering. In biology, biotechnology, engineering and chemistry, the relationships between the probability of success and the CUs do not show significant gender differences. Other courses, such as computer science and mathematics, highlight significant gender differences in favor of male students.

Tables 5, 6, and 7 report the parameter estimates of the admission covariates, and show some differences. The gender parameter is between  $-1$  and  $0$  in all degree courses but computer science, natural sciences, and statistics. These estimates have to be interpreted

**Table 5** Parameter estimates  $\theta$ 's of the segmented regression models by degree course

Variable	Bio	Biot	Chem
<b>Intercept</b>	− 1.06* (0.35)	0.71 (0.46)	− 3.63* (0.62)
<b>Gender (female)</b>			
<b>Male</b>	− 0.59 (0.45)	− 0.85 (0.58)	0.00 (0.67)
<b>CU male</b>	0.00 (0.02)	0.04* (0.01)	0.01 (0.02)
<b>CU female</b>	− 0.02 (0.02)	0.02 (0.01)	− 0.03 (0.02)
<b>HSdiploma (other "liceo")</b>			
<b>Scientific</b>	− 0.10 (0.11)	− 0.68* (0.18)	− 0.10 (0.29)
<b>Classical</b>	− 0.14 (0.13)	− 0.56* (0.2)	− 0.10 (0.32)
<b>Technical</b>	− 0.35* (0.18)	− 0.72* (0.3)	− 0.36 (0.35)
<b>Vocational</b>	− 0.51* (0.23)	− 0.32 (0.39)	− 0.85 (0.68)
<b>Abroad/other</b>	− 0.57 (0.46)	− 0.29 (0.54)	− 3.13* (1.19)
<b>Age (<math>\leq 19</math>)</b>			
<b>&gt; 19</b>	− 0.41* (0.10)	− 0.30 (0.16)	− 0.74* (0.17)
<b>Macro-region (Islands)</b>			
<b>North</b>	0.36* (0.11)	− 0.23 (0.20)	0.17 (0.20)
<b>Center</b>	0.20 (0.11)	− 0.17 (0.20)	0.09 (0.21)
<b>South</b>	− 0.25* (0.11)	− 0.07 (0.20)	− 0.64* (0.24)
<b>HS final mark (60)</b>	0.01 (0.00)	− 0.01* (0.00)	0.03* (0.01)
<b>HSdiploma x Gender (female)</b>			
<b>Scientific</b>	− 0.14 (0.36)	− 0.24 (0.44)	− 0.40 (0.63)
<b>Classical</b>	− 0.29 (0.41)	− 0.45 (0.49)	− 1.14 (0.72)
<b>Technical</b>	0.30 (0.42)	− 0.20 (0.53)	− 0.65 (0.67)
<b>Vocational</b>	0.45 (0.48)	− 0.54 (0.70)	− 0.21 (0.94)
<b>Abroad/other</b>	0.51 (0.99)	− 0.24 (1.18)	2.66 (1.44)

Baselines are in brackets. Cohort of freshmen enrolled in 2014. Biology, biotechnology and chemistry. The asterisk indicates a corresponding  $p$ -value  $< 0.05$ . Standard errors are in brackets

**Table 6** Parameter estimates  $\theta$ 's of the segmented regression models by degree course

Variable	Comp	Eng	Math
<b>Intercept</b>	− 1.97 (1.05)	− 2.78* (0.23)	− 3.24* (0.79)
<b>Gender (female)</b>			
<b>Male</b>	− 4.32* (1.09)	− 0.75* (0.30)	− 0.64 (0.95)
<b>CU male</b>	0.01 (0.02)	− 0.04* (0.01)	− 0.07 (0.05)
<b>CU female</b>	− 0.78* (0.39)	− 0.02 (0.01)	− 0.17* (0.05)
<b>HSdiploma (other “liceo”)</b>			
<b>Scientific</b>	− 0.67 (0.39)	− 0.01 (0.13)	0.36 (0.30)
<b>Classical</b>	− 0.64 (0.54)	− 0.09 (0.14)	0.51 (0.37)
<b>Technical</b>	− 1.13* (0.4)	− 0.30 (0.16)	− 0.06 (0.44)
<b>Vocational</b>	− 1.36 (0.75)	− 1.06* (0.53)	− 0.61 (1.01)
<b>Abroad/other</b>	0.89 (1.50)	− 0.13 (0.36)	− 24.93 (573.89)
<b>Age (<math>\leq 19</math>)</b>			
<b>&gt; 19</b>	− 0.49* (0.11)	− 0.58* (0.06)	− 1.32* (0.33)
<b>Macro-region (Islands)</b>			
<b>North</b>	1.40* (0.21)	0.10 (0.06)	0.55 (0.30)
<b>Center</b>	0.93* (0.23)	− 0.07 (0.07)	0.56 (0.31)
<b>South</b>	0.34 (0.22)	− 0.19* (0.07)	− 0.02 (0.32)
<b>HS final mark (60)</b>	0.02* (0.00)	0.02* (0.00)	0.04* (0.01)
<b>HSdiploma x Gender (female)</b>			
<b>Scientific</b>	1.75* (0.60)	0.33 (0.24)	− 0.79 (0.84)
<b>Classical</b>	2.17* (0.78)	0.42 (0.26)	0.17 (0.95)
<b>Technical</b>	1.92* (0.60)	0.53* (0.25)	− 0.04 (0.93)
<b>Vocational</b>	1.90* (0.92)	1.18* (0.58)	0.46 (1.64)
<b>Abroad/other</b>	− 0.38 (1.63)	0.32 (0.45)	1.58 (847.23)

Baselines are in brackets. Cohort of freshmen enrolled in 2014. Computer Science, engineering and mathematics. The asterisk indicates a corresponding  $p$ -value < 0.05. Standard errors are in brackets

together with the interaction effects. In the first three courses, the main effects are compensated for by the interaction effects. The estimates of the covariates CU Male and CU Female range between  $-0.10$  and  $+0.10$ , but those in mathematics, computer science, and statistics have higher negative values. All the other parameters referred to the other admission covariates are usually between  $-1$  and  $+1$ . Some larger negative values are estimated for students from technical and vocational schools. Besides, the most important difference in the interaction is observed in computer science, where the parameters are all positive, save for students who went to high school abroad. This means that being male from a traditional “liceo” or a technical or vocational school leads to a higher probability of success with respect to females. Finally, it is important to note that northern students perform better in almost every degree course, save statistics and biotechnology. In those courses, students enrolled at the islands perform the best.

## Conclusions

STEM and gender has been a recent focus of academic worldwide writing, and quantitative studies on the relationship between the two are essential for better understanding this topic. We restrict our analysis to university comparing success in STEM courses for males and for females and particularly, the relationship between the university career and the first-year performance.

**Table 7** Parameter estimates  $\theta$ 's of the segmented regression models by degree course

Variable	Nat	Phy	Stat
<b>Intercept</b>	− 3.26* (0.56)	− 1.83* (0.69)	0.60 (1.42)
<b>Gender (female)</b>			
<b>Male</b>	− 1.56* (0.55)	− 0.59 (0.74)	− 2.04 (1.63)
<b>CU male</b>	0.04* (0.01)	− 0.06 (0.04)	− 0.11 (0.14)
<b>CU female</b>	0.00 (0.01)	− 0.12* (0.05)	− 0.15 (0.16)
<b>HSdiploma (other “liceo”)</b>			
<b>Scientific</b>	− 0.02 (0.20)	0.07 (0.31)	− 0.79 (0.60)
<b>Classical</b>	0.06 (0.25)	− 0.09 (0.35)	− 0.55 (0.75)
<b>Technical</b>	− 0.22 (0.25)	− 1.24* (0.61)	− 0.73 (0.61)
<b>Vocational</b>	− 0.04 (0.37)	− 0.62 (0.89)	− 0.38 (1.24)
<b>Abroad/other</b>	− 0.64 (0.65)	0.07 (1.09)	2.11 (1.34)
<b>Age (<math>\leq 19</math>)</b>			
<b>&gt;19</b>	− 0.39* (0.13)	− 0.68* (0.2)	− 0.86* (0.27)
<b>Macro-region (Islands)</b>			
<b>North</b>	0.40* (0.18)	0.64* (0.27)	− 1.02 (0.65)
<b>Center</b>	0.19 (0.19)	0.23 (0.28)	− 1.32 (0.69)
<b>South</b>	− 0.13 (0.20)	0.01 (0.30)	− 2.51* (0.73)
<b>HS final mark (60)</b>	0.03* (0.01)	0.01* (0.01)	0.01 (0.01)
<b>HSdiploma x Gender (female)</b>			
<b>Scientific</b>	0.46 (0.44)	− 0.32 (0.61)	1.83 (1.10)
<b>Classical</b>	1.05* (0.53)	0.39 (0.67)	1.24 (1.30)
<b>Technical</b>	0.39 (0.47)	0.48 (0.82)	0.77 (1.12)
<b>Vocational</b>	0.15 (0.62)	0.50 (1.12)	0.91 (1.67)
<b>Abroad/other</b>	0.90 (0.93)	0.16 (1.47)	23.02 (598.91)

Baselines are in brackets. Cohort of freshmen enrolled in 2014. Natural Sciences, physics and statistics. The asterisk indicates a corresponding  $p$ -value  $< 0.05$ . Standard errors are in brackets

Our analysis confirms that first-year performance is strongly correlated to obtaining a degree within 4 years. This relationship often varies between males and females and is in line with Barone's divide between (female) care-oriented and the (male) technical-oriented courses. This divide is consistent save in mathematics, where males outperform females, though mathematics is included by Barone et al. (2019) in the (female) care-oriented group, probably because it was, in the past, a teaching-oriented course. Today a mathematics degree leads to a wider range of careers, with many technical and computer science jobs being taken up by math graduates. Therefore math, we would suggest, be considered both a care and technical degree.

Moreover, it is crucial to stress an upstream pattern in engineering, where it looks that female performance follows the male stream. There is, then, a similar male-female performance in this important area, even if there are still relatively few females. To explain this divide, (Hall and Sandler 1982) suggested the interesting idea of the “chilly climate”, that is, the presence of university environments in which females face greater difficulties in succeeding in some specific STEM careers. In general, our findings show that gender differences vary a lot among STEM courses. These differences are in accordance with the care-oriented and technical-oriented dichotomy (Barone et al. 2019), save, as noted, for mathematics. On the other hand, it is interesting to stress that in Italy the theory of a “chilly climate” at university can be extended to high school. In fact, (Sherman 1980) states that school environment and teachers' decisively shape the attitudes of students, males and females, to certain subjects and skills. This is confirmed by the gender compo-

**Table 8** Parameter estimates of the  $\psi$ 's and  $\delta$ 's of the segmented regression models by courses

Variable	Parameter	Bio	Biot	Chem
CU male	$\psi_{m,1}$	25.27 (2.32)	36.10 (2.53)	25.82 (3.12)
CU female	$\psi_{f,1}$	18.12 (2.36)	30.97 (3.06)	21.10 (2.57)
	$\psi_{f,2}$	38.10 (3.08)	50.81 (2.52)	-
CU male	$\delta_{m,1}$	0.11* (0.02)	0.11* (0.02)	0.11* (0.02)
CU female	$\delta_{f,1}$	0.08* (0.02)	0.07* (0.02)	0.15* (0.03)
	$\delta_{f,2}$	0.05* (0.02)	0.11* (0.05)	-
Variable	Parameter	Comp	Eng	Math
CU male	$\psi_{m,1}$	21.35 (2.70)	15.45 (0.85)	19.10 (3.99)
CU female	$\psi_{f,1}$	3.98 (1.24)	18.10 (1.36)	15.07 (2.23)
	$\psi_{f,2}$	-	-	43.10 (4.17)
CU male	$\delta_{m,1}$	0.11* (0.02)	0.13* (0.01)	0.18* (0.05)
CU female	$\delta_{f,1}$	0.87* (0.39)	0.11* (0.01)	0.25* (0.05)
	$\delta_{f,2}$	-	-	0.10* (0.05)
Variable	Parameter	Nat	Phy	Stat
CU male	$\psi_{m,1}$	35.09 (3.58)	18.10 (3.00)	13.89 (5.73)
CU female	$\psi_{f,1}$	27.12 (2.82)	15.25 (2.90)	11.92 (5.26)
	$\psi_{f,2}$	-	49.10 (3.61)	52.10 (2.54)
CU male	$\delta_{m,1}$	0.09* (0.02)	0.16* (0.04)	0.23 (0.14)
CU female	$\delta_{f,1}$	0.10* (0.02)	0.21* (0.06)	0.25 (0.16)
	$\delta_{f,2}$	-	0.16 (0.10)	0.39 (0.31)

Cohort of freshmen enrolled in 2014

sition of the scientific “liceo” and the technological-technical high schools in Italy, where females represent only, respectively, 43% and 17% of the graduates in 2017. These percentages represent how the gender gap in the scientific-technological ambit is still present even if female participation has increased in the past 50 years.

A limit of this paper is the lack of information on students' social and economic background, which is an important covariate for any career. Other useful information might be found using ad hoc surveys in understanding the transition from high school to university, and in determining male and female attitudes and expectations towards STEM studies.

Finally, the novelty of this paper consists in a straightforward representation of the relationship between CUs and the completion of the degree course. This is done via segmented models that allow for the identification of significant change points in CU accumulation during the students' first year at university. That relationship varies between gender and STEM: the probability of getting a degree, conditioning on the CUs at first year, is higher for males in computer science, mathematics, and slightly higher in natural sciences and biotechnology.

#### Acknowledgements

We would like to thank Vito M. R. Muggeo for his fruitful and insightful comments.

#### Authors' contributions

Andrea Priulla and Nicoletta D'Angelo conceived the presented idea, and developed the theoretical formalism. Andrea Priulla designed the computational framework and analyzed the data. The authors contributed equally to the interpretation of results and to the writing of the manuscript. Massimo Attanasio supervised the project. The author(s) read and approved the final manuscript.

#### Funding

This work was supported by Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR), PRIN 2017 'From high school to job placement: micro data life course analysis of university student mobility and its impact on the Italian North-South divide' [grant n. 2017HBTk5P]. P.I. Massimo Attanasio

**Availability of data and materials**

Database MOBYSU.IT [Mobilità degli Studi Universitari in Italia], research protocol MUR - Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenze, Cattolica and Napoli Federico II, Scientific Coordinator Massimo Attanasio (UNIPA), Data Source ANS-MUR/CINECA

**Declarations****Competing interests**

The authors declare that they have no competing interests.

Received: 21 December 2020 Accepted: 26 May 2021

Published online: 05 July 2021

**References**

- Attanasio, M., Boscaino, G., Capursi, V., Plaia, A. (2013). Can the students' career be helpful in predicting an increase in universities income?, In P. Giudici, S. Ingrassia, & M. Vichi (a cura di), *Statistical Models for Data Analysis*: Springer.
- Attanasio, M., Enea, M., Albano, A., Priulla, A. (2018). Analisi delle carriere universitarie nelle lauree scientifiche di base in Italia nell'ultimo decennio. *Induzioni*, 37–66.
- Barone, C., & Assirelli, G. (2020). Gender segregation in higher education: an empirical test of seven explanations. *Higher Education*, 79(1), 55–78.
- Barone, C., Schizzerotto, A., Assirelli, G., Abbiati, G. (2019). Nudging gender desegregation: A field experiment on the causal effect of information barriers on gender inequalities in higher education. *European Societies*, 21(3), 356–377.
- Betts, M.G., Forbes, G.J., Diamond, A.W. (2007). Thresholds in songbird occurrence in relation to landscape structure. *Conservation Biology*, 21(4), 1046–1058.
- Cheryan, S., Plaut, V.C., Davies, P.G., Steele, C.M. (2009). Ambient belonging: how stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology*, 97(6), 1045.
- Contini, D., Di Tommaso, M.L., Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review*, 58, 32–42.
- Correll, S.J. (2001). Gender and the career choice process: The role of biased self-assessments. *American Journal of Sociology*, 106(6), 1691–1730.
- D'Angelo, N., & Priulla, A. (2020). Estimating the number of change-points in segmented regression models: comparative study and application. *DSEAS working papers, IV*. <https://doi.org/10.2139/ssrn.3724965>.
- De Vita, L., & Giancola, O. (2017). Between education and employment: Women's trajectories in stem fields. *Polis*, 31(1), 45–72.
- Eccles, J.S. (2007). Where Are All the Women? Gender Differences in Participation in Physical Science and Engineering. In S.J. Ceci & W.M. Williams (Eds.), *Why aren't more women in science?: Top researchers debate the evidence*. <https://doi.org/10.1037/11546-016>: American Psychological Association.
- Eddy, S.L., Brownell, S.E., Wenderoth, M.P. (2014). Gender Gaps in Achievement and Participation in Multiple Introductory Biology Classrooms. *Cell Biol. Educ*, 13(3), 478. <https://doi.org/10.1187/cbe.13-10-0204>.
- Enea, M., & Attanasio, M. (2020). Gender differences in Italian stem degree courses: a discrete-time competing-risks model. In N.S. Alessio Pollice (a cura di), *Book of short papers - SIS 2020* (pp. 385–390): Pearson.
- Fan, S., & Ritz, J. (2014). International views of stem education. In M.J. de Vries (Ed.), *Proceedings PATT-28 Conference* (pp. 7–14). Orlando.
- Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., Wenderoth, M.P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415.
- Gabay-Egozi, L., Shavit, Y., Yaish, M. (2015). Gender differences in fields of study: the role of significant others and rational choice motivations. *European Sociological Review*, 31(3), 284–297.
- Hall, R.M., & Sandler, B.R. (1982). The classroom climate: A chilly one for women? ERIC.
- Hocking, R.R. (1976). A biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, 32(1), 1–49.
- Kim, H.-J., Fay, M.P., Feuer, E.J., Midthune, D.N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, 19(3), 335–351.
- Lauer, S., Momen, J., Offerdahl, E., Kryjevskaja, M., Christensen, W., Montplaisir, L. (2013). Stereotyped: Investigating gender in introductory science courses. *CBE—Life Sciences Education*, 12(1), 30–38.
- Legewie, J., & DiPrete, T.A. (2012). School context and the gender gap in educational achievement. *American Sociological Review*, 77(3), 463–485.
- Li, K., Zhang, P., Hu, B.Y., Burchinal, M.R., Fan, X., Qin, J. (2019). Testing the 'thresholds' of preschool education quality on child outcomes in China. *Early Childhood Research Quarterly*, 47, 445–456.
- Mega, C., Ronconi, L., De Beni, R. (2014). What makes a good student? how emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology*, 106(1), 121.
- Mostafa, T. (2019). *Why don't more girls choose to pursue a science career? PISA in Focus, No. 93*. Paris: OECD Publishing. <https://doi.org/10.1787/02bd2b68-en>.
- Muggeo, V.M. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19), 3055–3071.
- Muggeo, V.M., & et al (2008). Segmented: an R package to fit regression models with broken-line relationships. *R news*, 8(1), 20–25.
- OECD (2019). *Education at a Glance 2019*, (p. 520). <https://doi.org/10.1787/f8d7880d-en>. <https://www.oecd-ilibrary.org/content/publication/f8d7880d-en>.
- Salanova, M., Schaufeli, W., Martínez, I., Bresó, E. (2010). How obstacles and facilitators predict academic performance: The mediating role of study burnout and engagement. *Anxiety, Stress & Coping*, 23(1), 53–70.

- Sherman, J. (1980). Mathematics, spatial visualization, and related factors: Changes in girls and boys, grades 8–11. *Journal of Educational psychology*, 72(4), 476.
- Simon, R.M. (2010). Gender differences in knowledge and attitude towards biotechnology. *Public Understanding of Science*, 19(6), 642–653.
- UK-Parliament (2020). Science and technology committee—second report: higher education in science, technology, 596 engineering and mathematics (STEM) subjects.HOUSE OF LORDS - Select Committee on Science and Technology - 2nd Report of Session 2012–13.
- Ulm, K. (1991). A statistical method for assessing a threshold in epidemiological studies. *Statistics in Medicine*, 10(3), 341–349.

### **Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

---

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)

---