UNIVERSITÀ DEGLI STUDI DI PALERMO
dSEAS dipartimento scienze economiche aziendali e statistiche
department of economics business and statistics
Working Papers

ISSN 2611-0172, volume IV, 2020

Estimating the number of changepoints in segmented regression models: comparative study and application

Nicoletta D'Angelo · Andrea Priulla

Abstract This paper deals with the problem of selecting the number of changepoints in segmented regression models. The aim is to review selection criteria, namely information criteria and hypothesis testing, and to propose a novel application in the context of students' careers in higher education. The performance of the selection criteria is assessed through simulation studies. Furthermore, we investigate the relationship between University students' performances and one of their main determinants, finding out that this relationship is actually broken-line.

Keywords Information criteria · Hypothesis testing · Segmented regression · Changepoints · Higher education

Riassunto

Questo articolo affronta il problema della selezione del numero di punti di svolta nei modelli di regressione segmentata. L'obiettivo è quello di offrire una panoramica dei metodi di selezione, in particolare criteri i di informazione e verifica di ipotesi, e proporre una nuova applicazione nel contesto della carriere degli studenti nell'istruzione superiore. La performance dei criteri di selezione è valutata attraverso studi di simulazione. Inoltre, indaghiamo la relazione fra il successo degli studenti universitari e uno dei suoi principali predittori, scoprendo che questa relazione è effettivamente segmentata.

Parole chiave *Criteri d'informazione - Test d'ipotesi - Regressione segmentata - Punti di svolta - Istruzione superiore*

Nicoletta D'Angelo, Andrea Priulla
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy
E-mail: nicoletta.dangelo@unipa.it, andrea.priulla@unipa.it

1 Introduction

Segmented line regression is 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). Segmented or broken-line models are regression models where the relationships between the response and one or more explanatory variables are piecewise linear, namely represented by two ore more straight lines connected at unknown values. These values are commonly referred to as *changepoints* or *breakpoints*. The main advantage of these models lays in the interpretability of the results, while also achieving a good trade-off with flexibility, typically achieved by non-parametric approaches.

This paper deals with the problem of estimating the number of changepoints in segmented regression models, widely discussed by many authors as Lerman (1980) and Kim et al. (2000).

Therefore, the first aim of this work is to study and compare the performance of different criteria in selecting the right number of changepoints, via simulations. As far as the information based criteria we consider the Akaike Information Criterion, the Bayesian Information Criterion and the generalized Bayesian Information Criterion, while as regards hypothesis testing, we consider the Davies' test and the Score test.

In order to explore the applicability of the considered framework, an original application is proposed, dealing with the academic performance of university students in Italy. Academic student performance is a crucial issue for university policy-makers, today more than ever (Adelfio et al., 2014).

In the last years, studies concerning the students' careers at University have been an important topic in the academic statistical literature, following the introduction of the so-called 3+2 reform. In Italy, as well as in all the other western countries, there has been a poor propensity to enroll in one of the several scientific degree courses, that are characterized by higher dropout rates with respect to the other courses and by a lower proportion of the female component. However, despite their actual minority registered in some specific courses, such as mathematics, computer science and engineering, female students show a better performance with respect to their male colleagues, in terms of dropout and graduation rates within 4 years.

Previous studies showed that performance is a complex phenomenon, often characterized by students personal inclinations and their subjective choices during the path to graduation. Several recent papers deals with students performance and its determinants. This can be broadly classified into two main directions, the first one accounting for students social and demographic characteristics (Cheesman et al., 2006; Birch and Miller, 2006; Grilli et al., 2013), and the second one accounting for the students previous performance (Tattersall et al., 2006; Van Bragt et al., 2011; Horn et al., 2011). In this paper we focus on the latter. Indeed, Adelfio et al. (2014), proposing a new measure for students' performance, showed no significant effects of the social and demographic variables. Attanasio et al. (2013) highlighted the crucial role of the credits earned at the end of the first year as a good and simple predictor of the success, in a retrospective study. Following these results, the second aim of this paper is to analyse the relationship between University students' performance and their main determinants, assuming broken-line effects.

The analysis focuses on freshmen enrolled in scientific courses at Universities in Italy in 2014, assuming the number of credits earned during the first year as one of the main determinants of their success.

The aim of this paper is to provide an overview of the segmented regression model, together with a review of the main tools useful for estimating the number of changepoints. Furthermore, the main novelty of this work lays in the original application of the segmented regression model for studying the academic students' careers.

All the analyses are performed using the `segmented` package (Muggeo, 2008) of the software R (R Core Team, 2019) and all the codes of the simulations and analyses carried out throughout the paper are available from the authors.

The structure of the paper is as follows. Section 2 introduces the segmented regression model and Section 3 reviews criteria suitable for model selection in this context. Section 4 presents simulations to study the performance of the given criteria and Section 5 proposes an original analysis of the students' careers in higher education in Italy. The paper ends with conclusions in Section 6.

2 Background on the segmented regression models

The segmented linear regression is expressed as

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

where g is the link function, x_i is a broken-line covariate and z_i is a covariate whose relationship with the response variable is not broken-line. We denote by K_0 the true number of changepoints and by ψ_k the K_0 locations of the changes in the relationship, that we shall call from now on *changepoints*. These are selected among all the possible values of the range of x. The term $(x_i - \psi_k)_+$ is defined as $I(x_j > \psi_k)$ that is $(x_i - \psi_k)I(x_i > \psi_k)$. The coefficient θ represents the non broken-line effect of z_i, β represents the effect for $x_i < \psi_1$, while δ_k is the vector of the differences in the effects. Finally, ϵ_i is a generic error term. Throughout the paper we only consider models with Gaussian iid errors $\epsilon_i \sim N(0, \sigma^2)$.

The basic statistical problem that we deal with in this paper is the identification of the number of changepoints K_0. The estimation of their locations, that is the vector of ψ_k, and the broken-line effects, represented by β and the vector δ, may also be of interest.

2.1 Estimation

For estimation purposes, we consider a reparametrization of the segmented model that has the advantage of an efficient estimating approach via the algorithm discussed in Muggeo (2003,

2008), fitting iteratively the generalized linear model:

$$g(E[Y|x_i]) = \beta_1 x_i + \sum_k \delta_k \tilde{U}_{ik} + \sum_k \gamma_k \tilde{V}_{ik}^{-}, \tag{2}$$

where $\tilde{U}_{ik} = (x_i - \tilde{\psi}_k)_+$, $\tilde{V}_{ik}^{-} = -I(x_i > \tilde{\psi}_k)$. The parameters β and δ_k are the same of Equation (1), while the γ are the working coefficients useful for the estimation procedure. At each step the working model in Equation (2) is fitted and new estimates of the changepoints are obtained via:

$$\hat{\psi}_k = \tilde{\psi}_k + \frac{\hat{\gamma}_k}{\hat{\delta}_k}$$

iterating the process up to convergence.

2.2 Example with a single changepoint

Let's consider, as an example, a model with a single breakpoint ψ_1:

$$E\left[Y|x_i\right] = \beta_0 + \beta_1 x_i + \delta_1 \left(x_i - \psi_1\right)_+$$

This specification is particularly appealing, as it also allow to visually assess in a single graph the segmented relationship between the response and the only one variable included whose effect is assumed to be broken-line. For instance the left panel of Figure 1 displays a segmented relation between a segmented covariate, that takes $n = 100$ equispaced values ranging from 0 to 1, and the response variable given by

$$y_i = 2 + 15x_i - 8(x_i - 0.2) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 0.3).$$

The right panel of Figure 1 displays the output of the model fitted through the `segmented` package, that estimates the only changepoint equal to $\hat{\psi}_1 = 0.2036$. The other parameters are equal to $\hat{\beta}_0 = 2.014$, $\hat{\beta}_1 = 14.839$ and $\hat{\delta}_1 = -8.220$. In particular $\hat{\beta}_1$ represents the left slope in Figure 1, being the effect of x when x_i is less than the estimated changepoint $\hat{\psi}_1$, that is when $x_i < 0.2036$. In order to obtain the right slope, that is the effect of x when $x_i > 0.2036$, we have to sum $\hat{\beta}$ and $\hat{\delta}_1$, obtaining $\hat{\beta}_1 + \hat{\delta}_1 = 14.839 - 8.220 = 6.6189$, being of course lower that the left slope.

3 Selecting the number of changepoints

In a more general context, with multiple changepoints as in Equation (1), we need to select only 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 $\hat{\delta}_k \approx 0$. Therefore, the basic problem, that is selecting the number of significant changepoints in the

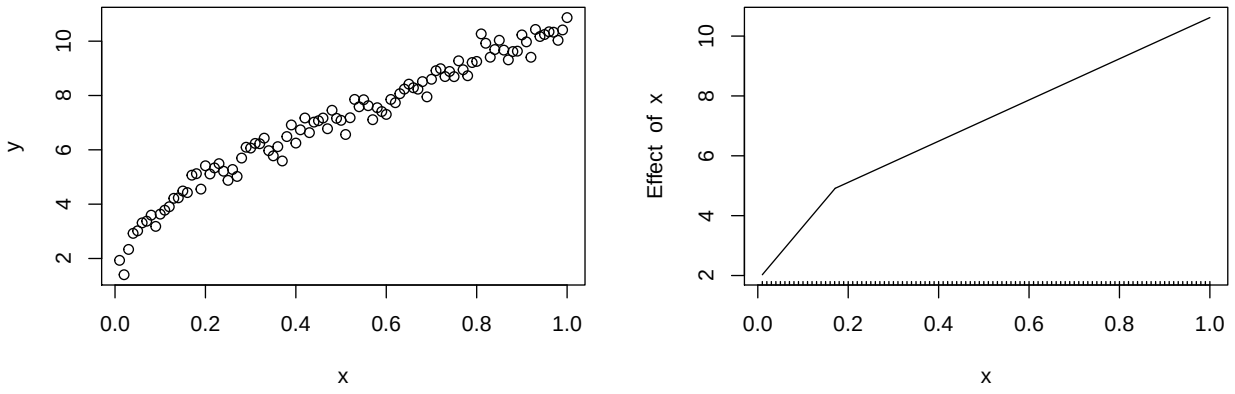

Fig. 1: Simulated segmented relationship

model (1), means selecting the significant variables among $V_1, \ldots, V_k$, from model (2), where K^* is the number of estimated changepoints. The fitted 'optimal' model will have $\hat{K} \leq K^*$ changepoints selected by any criterion. It is important to notice that these models are not nested, so likelihood ratio tests for model selection cannot be used.

Furthermore, the usual statistics can not be used to verify the existence of a changepoint since this parameter is present only under the alternative hypothesis, leading to a non-linear problem because the regularity conditions of the log-likelihood are not satisfied.

Basically, we will need to select the $\hat{\psi}_1, \ldots, \hat{\psi}_{\hat{K}}$ among the $\hat{\psi}_1, \ldots, \hat{\psi}_{K^*}$ via any selection criterion. The changepoints $\hat{\psi}_1, \ldots, \hat{\psi}_{\hat{K}}$ will be a subset of the estimates $\hat{\psi}_1, \ldots, \hat{\psi}_{K^*}$, since one or more changepoints are not included due to the deletion of one or more variables V_k by means of the given selection criterion. Therefore, it should be noticed that, while $\hat{\psi}_1, \ldots, \hat{\psi}_{K^*}$ are the estimates maximising the likelihood with K^* changepoints, there is no guarantee that the subset $\hat{\psi}_1, \ldots, \hat{\psi}_{\hat{K}}$ constitutes also the best estimate for the number of changepoints. Much of the literature is concerned with the problem of determining the 'best' subset of independent variables, and Hocking (1976) summarizes various selection criteria, reviewed below. These can be classified into two major approaches, namely information criteria and hypothesis testing.

3.1 Information criteria

The first information criterion to be considered is the well-known Akaike Information Criterion (Akaike, 1974)

$$AIC = -2 \log L + 2p$$

where L represents the likelihood function and p stands for the actual model dimension quantified by the number of estimated parameters, including $\hat{\beta}$, the $\hat{\delta}$ and $\hat{\psi}$ vectors in the segmented regression models, that is $p = 1 + 2\hat{K}$.

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.

Another information criterion is the Bayesian Information Criterion (Schwarz et al., 1978)

$$BIC = -2\log L + p\log(n)$$

that includes both a penalty for the number of estimated parameters p (like the AIC) and for the logarithm of the number of observations n. We may refer directly to the Gaussian special case

$$BIC = n\log\hat{\sigma}^2 + p\log(n)$$

where $\hat{\sigma}^2$ is the error variance, defined as $\frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$, which is a biased estimator for the true variance. In terms of the residual sum of squares (RSS) the definition is $BIC = n\log(RSS/n) + p\log(n)$. When picking from several models, the one with the lowest BIC is preferred. The BIC is an increasing function of the error variance $\hat{\sigma}^2$ and an increasing function of p. That is, unexplained variation in the dependent variable and the number of explanatory variables increases the value of BIC. Hence, lower BIC implies either fewer explanatory variables, better fit, or both. The BIC generally penalizes parameters more strongly than the AIC, though it depends on n and p.

For a typical linear regression model, it is well understood that the traditional best subset selection method with the Bayesian Information Criterion can identify the true model consistently (Shao, 1997; Shi and Tsai, 2002). With a fixed predictor dimension, (Wang et al., 2009) have demonstrated that the tuning parameters for high dimensional model selection procedures selected by a BIC type criterion can identify the true model consistently, and similar results are further extended to the situation with a diverging number of parameters for both unpenalized and penalized estimators. Therefore, the definition of the generalized BIC based on Gaussian distributed iid errors is

$$gBIC = \log(\hat{\sigma}^2) + p\frac{\log(n)}{n}C_n$$

where C_n is a known constant (e.g. 1, $\sqrt{n}$, $\log n$, $\log\log n$).

The definition reduces to $gBIC = -2\log L + p\log(n)C_n$ in the case of non Gaussian errors (which we will also refer to when dealing with Binomial response).

In general, the larger C_n, the more parsimonious the selected model. Note the gBIC reduces to the usual BIC when $C_n = 1$. The same considerations for the BIC hold in this case, that is, when choosing from several models, the one with the lowest gBIC is the one to be preferred.

3.2 Hypothesis testing

Testing for the existence of a changepoint means that we are dealing with the following system of hypothesis:

$$\begin{cases} \mathcal{H}_0: & \delta_k = 0 \\ \mathcal{H}_1: & \delta_k \neq 0 \end{cases}$$

Evaluating the existence of a changepoint is actually a non-regular problem, due to the fact that ψ_k is present only under the alternative $\mathcal{H}_1$. This problem makes usual statistical tests, such as the Wald or the likelihood ratio test, useless, because of the lack of a reference null distribution, even asymptotically. Therefore, we review below two tests used to evaluate the presence of a changepoint.

The first test proposed is the Davies' Test (Davies, 1977), which is an asymptotic test useful for dealing with hypothesis testing with a nuisance parameter present only under the alternative. Assuming fixed and known changepoints, the procedure computes K 'naive' test statistics for the difference-in-slope δ_1, seeks the lowest value and corresponding naive p-value (according to the alternative hypothesis), and then corrects the selected (minimum) p-value by means of the K values of the test statistic.
Considering the case of multiple changepoints $\psi_1 < \psi_2 < \cdots < \psi_k$ and relevant K test statistics, Davies defined an upper bound for the p-value given by

$$\text{p-value} \approx \Phi(-M) + V \exp\left(-M^2/2\right)(8\pi)^{-1/2}$$

where

- $\Phi(\cdot)$ = Standard Normal distribution
- $M = \max\left[S(\Psi_k)\right]_k$ is the maximum of the K test statistics
- $V = \Sigma_k(|S(\Psi_k) - S(\Psi_{k-1})|)$ is the total variation of $\{S(\psi_k)\}_k$.

Although the Davies' test is useful to test for the existence of a changepoint, it is not considered to be ideal to individuate the number of changepoints or their location. Indeed, the alternative hypothesis $\mathcal{H}_1$ actually states the existence of at least one additional changepoint, that is $K_0 > k$ when $\delta_k \neq 0$.

The second one is a pseudo-score test proposed by Muggeo (2016), which is based on an adjustment of the score statistic. This approach requires quantities only from the null fit and thus it has the advantage that it is not necessary to estimate the nuisance parameter under the alternative. The proposed statistic has the form:

$$s_0 = \frac{\bar{\varphi}^T(I_n - A)y}{\sigma\{\bar{\varphi}^T(I_n - A)y\}^{\frac{1}{2}}}$$

where,

– $(I_n - A)y$ is the residual vector under $\mathcal{H}_0$, with I_n the identity matrix, A the hat matrix and y the observed response vector;

– $\bar{\varphi} = \{\bar{\varphi}_1, ..., \bar{\varphi}_n\}^T$ is the vector of the means of the nuisance parameter ψ_k averaged over the range $\{\mathcal{L}, \mathcal{U}\}$, i.e. $\bar{\varphi} = K^{-1} \sum_{k=1}^{K} \varphi(x_i, \psi_k), i = 1, ..., n$, which doesn't depend on ψ_k, so the score can be computed even under $\mathcal{H}_0 : \delta_k = 0$ when ψ_k is not defined.

3.2.1 Proposed sequential hypothesis testing

An alternative approach to select the number of changepoints relies on sequential hypothesis testing procedure. Typically, this consists in 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. Depending on rejection or not of the null hypothesis, the procedure can test for the next hypothesis system by increasing the number of changepoints specified in $\mathcal{H}_0$ or decreasing the one postulated under $\mathcal{H}_1$, respectively (Kim et al., 2000).

In this paper we propose a different sequential procedure, in order to identify the correct number of changepoints resorting to the pseudo-score test or Davies' test.

Contrary to the procedure of Kim et al. (2000), our proposal has the advantage of not being limited to test for a maximum number of additional changepoints fixed a priori. Indeed, the previously explained procedure makes testing for more than two additional changepoints with the pseudo-score unfeasible. Our proposal overcomes this problem by making it possible to test for any number of additional changepoints thanks to the sequential procedure.

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. The p-value for each hypothesis can be obtained via the Davies or the pseudo-score test. Furthermore, we control for over-rejection of the null hypotheses at the overall level α, employing the Bonferroni correction comparing each p-value with α/K_{max}. Of course, setting the Bonferroni correction to α/K_{max} means putting ourselves in the most conservative setting.

For simplicity, we outline the algorithm when the maximum number of changepoints is $K_{max} = 3$, restricting the analyses to a contained number of changepoints.

The procedure is as follows:

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

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

via the Score or Davies' test. If it is not rejected then $\hat{K} = 0$ and the procedure stops at this step. Otherwise, we proceed with the algorithm;

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

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

   If it is not rejected then $\hat{K} = 1$ and the procedure stops. Otherwise, we proceed fitting the following model;

3. Fit a segmented model with $\hat{K} = 3$ and test

$$\begin{cases} \mathcal{H}_0: & \delta_3 = 0 \quad (K_0 = 2) \\ \mathcal{H}_1: & \delta_3 \neq 0 \quad (K_0 > 3) \end{cases}$$

   If the null hypothesis is not rejected then $\hat{K} = 2$. Otherwise, $\hat{K} = 3$.

It is important to highlight that, using the Davies test, even if based on the rejection of the last test the number of changepoints selected is equal to 3 (or more in general K_{max}), the actual number could be larger, as at each step we are actually testing for (at least) one additional changepoint.

4 Simulation studies

In order to study the performance of the previously introduced criteria to select the right number of changepoints, simulation studies are run.

We simulated from four different scenarios, generating and then fitting models with different true values of the number of changepoints, namely $K_0 = 0, 1, 2, 3$. We consider three different sample sizes $n = 100, 250, 500$ and we first include only one segmented covariate, taking equispaced values ranging from 0 to 1. The segmented models used for the simulations are reported in Table 1, considering iid Gaussian errors with standard deviation equal to $\sigma = 0.3$. As for the hypothesis testing, we fix $\alpha = 0.05$ and as far as the penalization of the gBIC we have chosen $C_n = \log\log n$. The estimated number of changepoints is obtained fitting segmented models using `segmented` library (Muggeo, 2003, 2008) over 500 simulations.

Table 2 reports the results in terms of percentage of correctly selected number of changepoints for each criterion. For each K_0 we fit a set of four models with $\hat{K} = 0, 1, 2, 3$. With regards to information based criteria, we select the 'best' model by choosing the one with lowest value of the given information criterion. As for the hypothesis testing, we choose the best model by applying the procedure proposed in Section 3.2.1. Conditional frequencies are reported in the rows and the interpretation of the results is as follows. For instance, simulating a model with $K_0 = 0$ and $n = 100$, the AIC picks the right number of changepoints 60% of the times, while the remaining percentage is spread among all the other possibilities. Therefore, a criterion that

Table 1: Linear segmented regression models fitted for the simulations aimed at testing the performance of the criteria for selecting the number of changepoints

K_0	model
0	$y_i = 2 + 15x_i + \epsilon_i$
1	$y_i = 2 + 15x_i - 8(x_i - 0.2)_+ + \epsilon_i$
2	$y_i = 2 + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + \epsilon_i$
3	$y_i = 2 + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + 10(x_i - 0.75)_+ + \epsilon_i$

Table 2: Percentages of correctly selected number of changepoints by each criterion (based on 500 runs and three different sample sizes n=100, 250, 500) - Gaussian reponse variable

	$n = 100$				$n = 250$				$n = 500$			
AIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.600**	0.192	0.130	0.078	**0.612**	0.164	0.142	0.082	**0.556**	0.198	0.120	0.126
1	0.000	**0.656**	0.214	0.130	0.00	**0.590**	0.256	0.154	0.000	**0.572**	0.254	0.174
2	0.000	0.002	**0.712**	0.286	0.000	0.000	**0.656**	0.344	0.000	0.000	**0.656**	0.344
3	0.000	0.000	0.010	**0.990**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
BIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.966**	0.030	0.004	0.000	**0.994**	0.006	0.000	0.000	**0.992**	0.008	0.000	0.000
1	0.000	**0.970**	0.022	0.008	0.000	**0.982**	0.016	0.002	0.000	**0.996**	0.004	0.000
2	0.000	0.020	**0.950**	0.030	0.000	0.000	**0.986**	0.014	0.000	0.000	**0.996**	0.004
3	0.000	0.000	0.096	**0.904**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
gBIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.996**	0.004	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000
1	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000
2	0.000	0.072	**0.926**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000
3	0.000	0.000	0.270	**0.730**	0.000	0.000	0.008	**0.992**	0.000	0.000	0.000	**1.000**
Davies	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.982**	0.014	0.004	0.000	**0.994**	0.006	0.000	0.000	**0.992**	0.008	0.000	0.000
1	0.000	**0.994**	0.004	0.002	0.000	**0.986**	0.006	0.008	0.000	**0.994**	0.002	0.004
2	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000
3	0.000	0.000	**0.682**	0.318	0.000	0.000	0.112	**0.888**	0.000	0.000	0.000	**1.000**
Score	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.986**	0.014	0.000	0.000	**0.976**	0.024	0.000	0.000	**0.992**	0.008	0.000	0.000
1	0.000	**0.996**	0.004	0.000	0.000	**0.986**	0.014	0.000	0.000	**0.998**	0.002	0.000
2	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.999**	0.001
3	0.012	0.048	0.286	**0.654**	0.000	0.000	0.016	**0.984**	0.000	0.000	0.000	**1.000**

Table 3: Linear segmented regression models fitted for the simulations aimed at testing the performance of the reviews criteria for selecting the number of changepoints, with an additional variable whose effect is non broken-line

K_0	model
0	$y_i = 2 + 4z_i + 15x_i + \epsilon_i$
1	$y_i = 2 + 4z_i + 15x_i - 8(x_i - 0.2)_+ + \epsilon_i$
2	$y_i = 2 + 4z_i + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + \epsilon_i$
3	$y_i = 2 + 4z_i + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + 10(x_i - 0.75)_+ + \epsilon_i$

perfectly selects the right number of changepoints should report values equal to 1 in the main diagonal of the table, and zeros in all the other entries.

It appears evident that the AIC overestimates the number of changepoints more frequently with respect to the other considered criteria. This is a reasonable result since it is well known that the AIC tends to overestimate the number of parameters. Also, this is the reason why at a glance it could seem to correctly pick the number of changepoints when $K_0 = 3$, as the percentage of correctly identification approaches to 1. This is due to the fact that we have not considered alternative hypotheses with more than 3 changepoints, that would likely be selected. The BIC and gBIC seem to behave better, as well as the Davies' and pseudo-score test. An exception is represented by the case in which $K_0 = 3$ and $n = 100$, since the Davies' test underestimates the number of changepoints on average. Overall, we notice that the gBIC outperforms its competitors in all the considered scenarios, especially as n increases. Other simulations studies, omitted for brevity, show that a sample size larger than $n = 500$ leads to the same results.

Then, we perform other simulations considering the same models in Table 1, with an additional non broken-line variable z_i, whose effect is set equal to $\theta = 4$. The models considered are reported in Table 3. We consider both a continuous variable $Z \sim Beta(\alpha_1 = 1, \alpha_2 = 2)$ and a dichotomous variable $Z \sim Bernoulli(\pi = 0.5)$. These additional results are reported in Tables 9 and 10 of the Appendix A, respectively. Overall, we do not identify any relevant differences in the results when a non broken-line variable is added to the linear predictor, especially as n increases.

Finally, in Table 5 we report the results of fitting Logit models, whose linear predictors are reported in Table 4. Noticing that the overall performance of the considered criteria is worse if compared to the Gaussian case, we chose to increase the sample size to $n = 2500, 5000, 10000$ for the Binomial case. Contrary to the Gaussian case, the performance of the BIC and gBIC is poor as K_0 increases. This could be due to the 'wrong' choice of the penalization C_n as well as to the excessively increased sample size n leading to over-penalization. Overall, we can conclude that when fitting Logit models with large sample sizes the proposed sequential procedure based on hypothesis testing performs better in the identification of the right number of changepoints, and therefore decisions on the final model to be chosen should be based on these results.

Table 4: Linear segmented regression models fitted for the simulations aimed at testing the performance of the reviews criteria for selecting the number of changepoints - Binomial case

K_0	linear predictor
0	$x_i\beta = -1 + 11x_i$
1	$x_i\beta = -1 + 11x_i - 20(x_i - 0.2)_+$
2	$x_i\beta = -1 + 11x_i - 20(x_i - 0.2)_+ + 25(x_i - 0.5)_+$
3	$x_i\beta = -1 + 11x_i - 20(x_i - 0.2)_+ + 25(x_i - 0.5)_+ - 14(x_i - 0.8)_+$

Table 5: Percentages of correctly selected number of changepoints by each criterion (based on 500 runs and three different sample sizes n=2500, 5000, 10000) - Binomial reponse variable

	$n = 2500$				$n = 5000$				$n = 10000$			
AIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.610**	0.236	0.114	0.040	**0.640**	0.198	0.110	0.052	**0.568**	0.192	0.178	0.062
1	0.000	**0.642**	0.226	0.132	0.000	**0.606**	0.248	0.146	0.000	**0.598**	0.218	0.184
2	0.000	0.034	**0.596**	0.370	0.000	0.014	**0.648**	0.338	0.000	0.010	**0.644**	0.346
3	0.000	0.000	0.036	**0.964**	0.000	0.000	0.002	**0.998**	0.000	0.000	0.000	**1.000**
BIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.998**	0.002	0.000	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000
1	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000
2	0.000	0.036	**0.870**	0.094	0.000	0.016	**0.876**	0.108	0.000	0.012	**0.878**	0.110
3	0.000	0.000	**0.550**	0.450	0.000	0.000	0.186	**0.814**	0.000	0.000	0.006	**0.994**
gBIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000
1	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000
2	0.000	0.036	**0.870**	0.094	0.000	0.020	**0.872**	0.108	0.000	0.014	**0.876**	0.110
3	0.000	0.000	**0.774**	0.226	0.000	0.000	**0.552**	0.448	0.000	0.000	0.176	**0.824**
Davies	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.998**	0.002	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.990**	0.010	0.000	0.000
1	0.000	**0.998**	0.002	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.996**	0.004	0.000
2	0.000	0.000	**0.866**	0.134	0.000	0.000	**0.872**	0.128	0.000	0.000	**0.874**	0.126
3	0.000	0.000	0.460	**0.540**	0.000	0.000	0.060	**0.940**	0.000	0.000	0.000	**1.000**
Score	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.988**	0.012	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.990**	0.010	0.000	0.000
1	0.000	**0.996**	0.004	0.000	0.000	**0.988**	0.012	0.000	0.000	**0.992**	0.008	0.000
2	0.000	0.000	**0.862**	0.138	0.000	0.000	**0.862**	0.138	0.000	0.000	**0.870**	0.130
3	0.000	0.000	0.352	**0.648**	0.000	0.000	0.064	**0.936**	0.000	0.000	0.000	**1.000**

5 Analysis of the University students' performance

The analysed data come from the ANS (Anagrafe Nazionale Studenti). These data are individual and longitudinal, indeed each record represents a student, and information about their social-demographic characteristics and their University career are available as covariates. In this paper, we consider only the freshmen enrolled at a scientific degree course in the academic year 2014-15. The aim is to analyse the University success, defined as obtaining the bachelor degree up to four years from the first enrolment, considering the following variables:

- *bachelor degree up to four years*, which takes value 1 if the student has graduated up to four years from the first University enrolment, and 0 otherwise.
- *number of credits (CFUs) earned at the end of the first year*, which is considered as a good predictor of the University success. The natural classification is given by 60 CFUs required per year, making its values ranging from 0 to 60.
- *gender*, because of the importance of accounting for the relevant differences present in the scientific degree courses between male and female students;

As we aim at studying the students' success, modelling the probability of obtaining the bachelor degree, the appropriate model is the Logistic Regression Model.

Logistic regression models are typically employed when the aim is to study the probability of a given event, depending on a set of covariates. In the context of regression models, the response variable is a dichotomous variable Y, taking only 0-1 values, and the model matrix X can accommodate for continuous as well as categorical variables. The variable Y is distributed as a Bernoulli with parameter depending on the realizations of the covariates:

$$E[Y|X = x] = P\{Y = 1|X = x\} = \pi.$$

First, we fit a logistic regression model

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} \tag{3}$$

that does not assume any segmented relationship between the covariates and the response variable ($\hat{K} = 0$). Here $p(x_i)$ is the probability of obtaining the bachelor degree up to four years from the first University enrolment, that represents the probability of success. x_1 and x_2 are the variables `CFU` and `gender`, respectively. In detail, the baseline is set as `females` for `gender` and 0 for `CFU`. The summary of the coefficients of the estimated model (3) is reported in Table 6. $\hat{\beta}_2$ is the coefficient of `gender` that, being negative, indicates that the probability of success is lower for male students with respect to their female colleagues. This is a reasonable result, since several previous studies have already highlighted that female students perform better than male students. $\hat{\beta}_1$ represents the change in the logit of the probability of success for each additional

Table 6: Coefficients of the model (3)

	Estimate	Std. Error	z value	Pr(> |z|)
α	-2.207	0.032	-67.9	<2e-16 ***
β_1	0.071	0.001	83.4	<2e-16 ***
β_2	-0.551	0.030	-18.3	<2e-16 ***

CFU. The estimated value of $\hat{\beta}_1 = 0.07$ proves, as one would expect, that earning more CFUs during the first year actually increases the probability of getting the degree up to four years.

Given that the number of CFUs earned during the first year is significant in explaining the probability of getting the degree, we aim at testing if this relationship can actually be assumed to be broken-line. This means that the number of credits earned during the first year would modify the probability of students' success differently after a given threshold of earned CFUs. We estimate models of the form:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \theta \texttt{gender}_i + \beta \texttt{CFU}_i + \sum_{k=1}^{K} \delta_k (\texttt{CFU}_i - \psi_k)_+ \tag{4}$$

with $\hat{K} = 1, 2, 3$.

In Table 7 we report the information criteria values and the p-values for each step of the sequential hypothesis testing outlined in Section 3.2.1, of the models (4). As previously discussed, we should pick the model with the lowest information criteria value. As for the sequential hypothesis testing, we should select the number of changepoints for which the corresponding test is no longer significant, that is our application correspond unequivocally to $\hat{K} = 2$. A segmented relationship seems to be more appropriate with respect to a classical linear relationship, given that the model without changepoints is never selected as the best one by any of the information based criteria.

Following the results of the simulation studies, that have shown that the proposed sequential procedure outperforms its information based criteria competitors in Logit segmented models, we chose the model with $\hat{K} = 2$, that is the one selected by the sequential procedure based on both the Davies' and Score tests.

The summary of the coefficients of the selected model is reported in Table 8 and the broken-line relationship between the logit of the probability of success and the number of credits earned at the end of the first year is displayed in Figure 2.

In this model, $\hat{\theta}$ is the parameter of the non broken-lined variable `gender`. A negative value of this estimate proves again that the probability of success is higher for females. Then, all the other estimated parameters concern the segmented variable `CFU`. In particular, $\hat{\beta}$ is the effect of this covariate when $x_i < \hat{\psi}_1$, that is, when the number of earned CFUs is less than 24. $\hat{\delta}_1$ and $\hat{\delta}_2$ are the changes in the slope when $\hat{\psi}_1 < x_i < \hat{\psi}_2$ and $x_i > \hat{\psi}_2$, respectively. Positive values of these parameters indicate an increase in the logit of the probability of success. In detail, a larger value of $\hat{\delta}_1$ highlights a more abrupt change in correspondence of $\hat{\psi}_1$.

Table 7: Information criteria of the fitted models (4) and p-values of each step of the sequential procedure

	$\hat{K}$			
Criterion	0	1	2	3
AIC	27288.98	26446.08	26424.7	**26424.55**
BIC	27313.72	26487.31	**26482.42**	26498.76
gBIC	27354.51	**26555.28**	26577.59	26621.11
Davies' test	0.00	0.00	**0.72**	1.00
Score test	0.00	0.00	**0.15**	0.79

Table 8: Coefficients of chosen model with $\hat{K} = 2$

	Estimate	Std. Error	z value	Pr(> |z|)
α	-1.606	0.036	-44.5	< 2e-16 ***
θ	-0.643	0.031	-20.6	< 2e-16 ***
β	0.022	0.003	7.6	3.28e-14 ***
δ_1	0.074	0.006	13.0	NA
δ_2	0.045	0.008	5.5	NA
ψ_1	24.445	1.052	-	-
ψ_2	43.000	1.685	-	-

In summary, the probability of getting the degree up to four years is overall higher for female students. It has also been confirmed that the number of CFUs earned during the first year influences the given probability of success. Furthermore, we have proven the presence of two thresholds in the number of CFUs after which the probability of success significantly increases. The most important threshold is represented by the first changepoint that is estimated to be $\hat{\psi}_1 = 24$ CFUs. Then, the students that perform the best are the ones that earned more than $\hat{\psi}_2 = 43$ CFUs.

6 Conclusions

This paper presents an overview of the segmented regression model, together with a review of the main tools useful for estimating the number of changepoints, whose performance is tested through simulation studies. Furthermore, an original application of the logistic segmented regression model for studying the academic students' careers has been provided.

First, we have reviewed some of the most useful tools in selecting the right number of changepoints in segmented regression models context and we have compared their performance through simulations studies. Among the different criteria examined, simulations have revealed that the generalized Bayesian Information Criterion performs better when considering a Gaussian distributed response variable. As for the proposed sequential procedure based on hypothesis

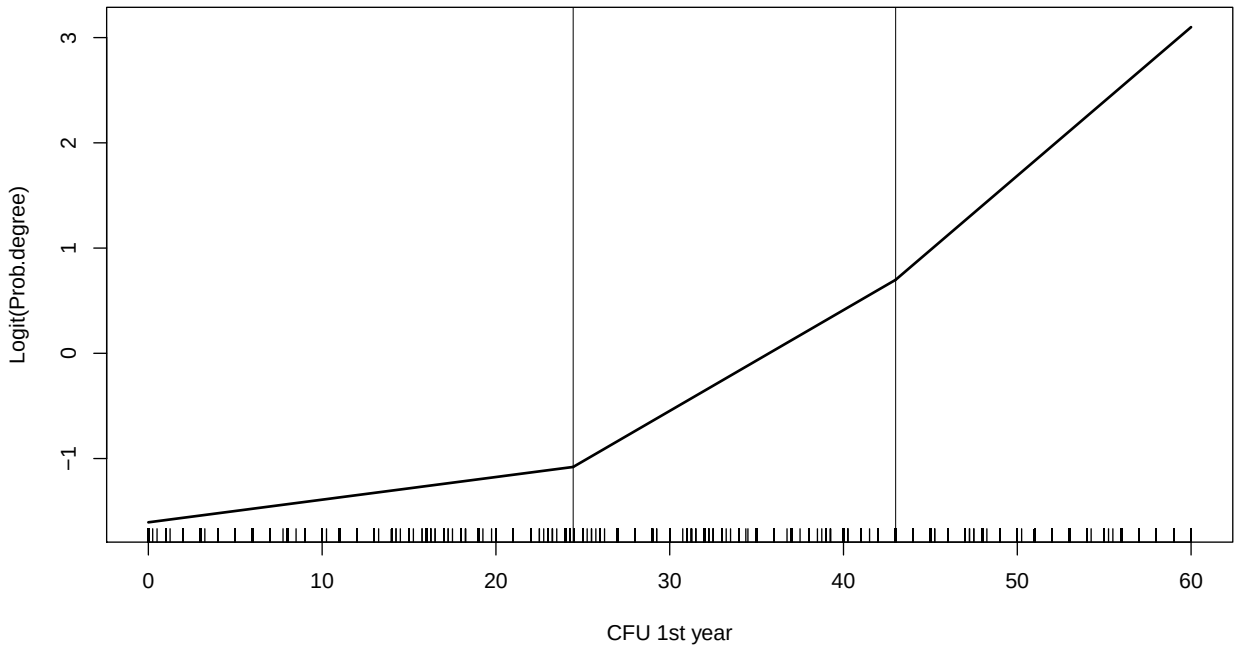

Fig. 2: Segmented relationship between the logit of the probability of success and the segmented variable. The vertical lines are located in correspondence of the estimated changepoints.

testing, this well behaves in the Gaussian case, and outperforms its information based criteria when dealing with Binomial response variables, especially with multiple changepoints.

Thus, these results have some limitations. As far as the information based criteria, we choose the best model based only on the smallest value of the considered criterion, not taking into account the absolute difference among the criteria values. This could be overcomed by employing the Bayes Factor (Kass and Raftery, 1995), that compares the evidences provided by the data in two scientific theories. It is also worth to notice that the simulation studies that we have run are restricted only up to a contained number of changepoints, namely $K_{max} = 3$. Of course our method can be implemented to any fixed K_{max}, but in general is not able to identify more than K_{max} changepoints, as the procedure ends at most to $\mathcal{H}_0 : K_0 = K_{max} - 1$ vs. $\mathcal{H}_1 : K_0 = K_{max}$, and therefore, selecting at most $\hat{K} = K_{max}$.

Secondly, in order to explore the applicability of the considered framework, an original application is presented, dealing with the academic performance of University students in Italy. As previous studies showed that students' performance is a complex phenomenon, often characterized by students' previous performance, we have focused our analysis on freshmen enrolled in scientific courses at Universities in Italy in 2014, assuming that the number of CFUs at the

end of the first year could affect their University success through a broken-line relationship. We have confirmed that the female students generally perform better in the scientific courses, with respect to their male colleagues. This is a relevant result, since the higher proportion of male student enrolled in most of the scientific courses is often misled as a better performance in scientific subjects. Otherwise, it is known that every course is characterized by a proper level of this gender gap in the performance, so it would be of interest to study this relationship in every specific course, both in the scientific and non-scientific fields. Further, we have found out that the effect of the number of credits that students earn during the first year is significant in explaining their University success, and in particular earning more than 24 CFUs significantly increase the overall probability of getting the degree up to four years. In addition, exceeding 43 CFUs further increases this probability.

The topic of gender differences in scientific courses, not explored in detail throughout this paper, is currently of crucial interest in social research and will be investigated in future works. Other determinants of University students' success could be further addressed, both broken-line or not, such as the high school final mark or the region of enrollment, as previous studies have shown significant differences among Italian regions (Attanasio et al., 2018). Furthermore, the phenomenon of University dropout, could be examined in the context of segmented regression models.

Fundings

This work was supported by Ministero dell'Istruzione, dell'Universitá 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-Sud divide' [grant n. 2017HBTK5P]. P.I. Massimo Attanasio

A Additional simulation studies

Table 9: Percentages of correctly selected number of changepoints by each criterion (based on 500 runs and three different sample sizes n=100, 250, 500) - Gaussian response variable and covariate $Z \sim Beta(\alpha_1 = 1, \alpha_2 = 2)$

	$n = 100$				$n = 250$				$n = 500$			
AIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.578**	0.182	0.130	0.110	**0.584**	0.190	0.140	0.086	**0.558**	0.178	0.146	0.118
1	0.000	**0.644**	0.202	0.154	0.000	**0.598**	0.252	0.150	0.000	**0.584**	0.240	0.176
2	0.000	0.000	**0.608**	0.392	0.000	0.000	**0.646**	0.354	0.000	0.000	**0.632**	0.368
3	0.000	0.000	0.008	**0.992**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
BIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.980**	0.018	0.002	0.000	**0.988**	0.012	0.000	0.000	**0.994**	0.006	0.000	0.000
1	0.000	**0.974**	0.026	0.000	0.000	**0.992**	0.008	0.000	0.000	**0.990**	0.008	0.002
2	0.000	0.012	**0.924**	0.064	0.000	0.000	**0.984**	0.016	0.000	0.000	**0.996**	0.004
3	0.000	0.000	0.082	**0.918**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
gBIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**1.000**	0.000	0.000	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000
1	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000
2	0.000	0.086	**0.914**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000
3	0.000	0.000	0.260	**0.740**	0.000	0.000	0.014	**0.986**	0.000	0.000	0.000	**1.000**
Davies	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.994**	0.006	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.996**	0.004	0.000	0.000
1	0.000	**0.996**	0.004	0.000	0.000	**0.998**	0.002	0.000	0.000	**0.988**	0.008	0.004
2	0.000	0.372	**0.624**	0.004	0.000	0.022	**0.976**	0.002	0.000	0.000	**0.998**	0.002
3	0.000	0.000	**0.690**	0.310	0.000	0.000	0.144	**0.856**	0.000	0.000	0.000	**1.000**
Score	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.978**	0.022	0.000	0.000	**0.980**	0.020	0.000	0.000	**0.984**	0.016	0.000	0.000
1	0.000	**0.994**	0.006	0.000	0.000	**0.994**	0.006	0.000	0.000	**0.994**	0.006	0.000
2	0.000	0.322	**0.674**	0.004	0.000	0.008	**0.988**	0.004	0.000	0.000	**0.996**	0.004
3	0.050	0.052	0.280	**0.618**	0.000	0.000	0.008	**0.992**	0.000	0.000	0.000	**1.000**

Table 10: Percentages of correctly selected number of changepoints by each criterion (based on 500 runs and three different sample sizes n=100, 250, 500) - Gaussian response variable and covariate $Z \sim Bernoulli(\pi = 0.5)$

	$n = 100$				$n = 250$				$n = 500$			
AIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.602**	0.186	0.128	0.084	**0.616**	0.162	0.116	0.106	**0.590**	0.200	0.118	0.092
1	0.000	**0.602**	0.232	0.166	0.000	**0.624**	0.236	0.140	0.000	**0.556**	0.296	0.148
2	0.000	0.000	**0.666**	0.334	0.000	0.000	**0.924**	0.076	0.000	0.000	**0.682**	0.318
3	0.000	0.000	0.004	**0.996**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
BIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.970**	0.024	0.006	0.000	**0.996**	0.002	0.002	0.000	**0.998**	0.002	0.000	0.000
1	0.000	**0.972**	0.026	0.002	0.000	**0.988**	0.012	0.000	0.000	**0.990**	0.010	0.000
2	0.000	0.018	**0.926**	0.056	0.000	0.000	**0.998**	0.002	0.000	0.000	**0.998**	0.002
3	0.000	0.000	0.048	**0.952**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
gBIC	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.992**	0.008	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000
1	0.000	**0.992**	0.008	0.000	0.000	**1.000**	0.000	0.000	0.000	**0.998**	0.002	0.000
2	0.000	0.088	**0.908**	0.004	0.000	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000
3	0.000	0.000	0.228	**0.772**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
Davies	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.988**	0.012	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.990**	0.010	0.000	0.000
1	0.000	**0.992**	0.004	0.004	0.000	**0.992**	0.006	0.002	0.000	**0.998**	0.002	0.000
2	0.000	0.394	**0.600**	0.006	0.000	0.002	**0.998**	0.000	0.000	0.000	**0.998**	0.002
3	0.000	0.000	**0.764**	0.236	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
Score	$\hat{K}$				$\hat{K}$				$\hat{K}$			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.968**	0.032	0.000	0.000	**0.986**	0.014	0.000	0.000	**0.984**	0.016	0.000	0.000
1	0.000	**0.996**	0.004	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.992**	0.008	0.000
2	0.000	**0.510**	0.488	0.002	0.000	0.000	**0.998**	0.002	0.000	0.000	**0.992**	0.008
3	0.052	0.082	0.338	**0.528**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**

References

Adelfio, G., Boscaino, G., and Capursi, V. (2014). A new indicator for higher education student performance. Higher Education, 68(5):653–668.

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6):716–723.

Attanasio, M., Boscaino, G., Capursi, V., and Plaia, A. (2013). Can the students' career be helpful in predicting an increase in universities income? In Statistical Models for Data Analysis, pages 9–16. Springer.

Attanasio, M., Enea, M., Albano, A., and Priulla, A. (2018). Analisi delle carriere universitarie nelle lauree scientifiche di base in italia nell'ultimo decennio.

Betts, M. G., Forbes, G. J., and Diamond, A. W. (2007). Thresholds in songbird occurrence in relation to landscape structure. Conservation Biology, 21(4):1046–1058.

Birch, E. R. and Miller, P. W. (2006). Student outcomes at university in australia: A quantile regression approach. Australian Economic Papers, 45(1):1–17.

Cheesman, J., Simpson, N., and Wint, A. G. (2006). Determinants of student performance at university: Reflections from the caribbean. Unpublished Manuscript.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika, 64(2):247–254.

Grilli, L., Rampichini, C., and Varriale, R. (2013). Predicting students' academic performance: a challenging issue in statistical modelling. CLEUP: Cladag 2013 Book of abstracts.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. Biometrics, 32(1):1–49.

Horn, P., Jansen, A., and Yu, D. (2011). Factors explaining the academic success of second-year economics students: An exploratory analysis. South African Journal of Economics, 79(2):202–210.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. Journal of the american statistical association, 90(430):773–795.

Kim, H.-J., Fay, M. P., Feuer, E. J., and Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. Statistics in medicine, 19(3):335–351.

Lerman, P. (1980). Fitting segmented regression models by grid search. Journal of the Royal Statistical Society: Series C (Applied Statistics), 29(1):77–84.

Muggeo, V. (2008). segmented: An r package to fit regression models with broken-line relationships. R NEWS, 8/1:20–25.

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. Statistics in medicine, 22(19):3055–3071.

Muggeo, V. M. (2016). Testing with a nuisance parameter present only under the alternative: a score-based approach with application to segmented modelling. Journal of Statistical Computation and Simulation, 86(15):3059–3067.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schwarz, G. et al. (1978). Estimating the dimension of a model. The annals of statistics, 6(2):461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. Statistica sinica, pages 221–242.

Shi, P. and Tsai, C.-L. (2002). Regression model selection—a residual likelihood approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(2):237–252.

Tattersall, C., Waterink, W., Höppener, P., and Koper, R. (2006). A case study in the measurement of educational efficiency in open and distance learning. Distance Education, 27(3):391–404.

Ulm, K. (1991). A statistical method for assessing a threshold in epidemiological studies. Statistics in medicine, 10(3):341–349.

Van Bragt, C. A., Bakx, A. W., Bergen, T. C., and Croon, M. A. (2011). Looking for students' personal characteristics predicting study outcome. Higher Education, 61(1):59–75.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(3):671–683.