An Association Model for Bivariate Data with Application to the Analysis of University Students' Success.

Journal:	*Journal of Applied Statistics*
Manuscript ID:	CJAS-2014-0370.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	14-Aug-2014
Complete List of Authors:	Enea, Marco; University of Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche Attanasio, Massimo; University of Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche
Keywords:	bivariate ordinal response, Dale's model, maximum penalized likelihood estimation, models for association, students' performance
2010 Mathematics Subject Classification:	62P25, 62G05

SCHOLARONE™
Manuscripts

To appear in the *Journal of Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–12

An Association Model for Bivariate Data with Application to the Analysis of University Students' Success.

Marco Enea[ab] and Massimo Attanasio[a*]

[a] *Dipartimento di Scienze Economiche, Aziendali e Statistiche, University of Palermo, Palermo, Italy*; [b] *London Metropolitan University, London, UK*

(*Received 00 Month 20XX; accepted 00 Month 20XX*)

The academic success of students is a priority for all universities. We analyze the students' success at university by considering their performance in terms of both "qualitative performance", measured by their mean grade, and "quantitative performance", measured by university credits accumulated. The data **comes** from **an Italian University** and **concerns** a cohort of students enrolled at the Faculty of Economics. To jointly model both the marginal relationships and the association structure with covariates, we fit a bivariate ordered logistic model by penalized maximum likelihood estimation. The penalty term we use allows us to smooth the association structure and enlarge the range of possible parameterizations beyond that provided by the usual Dale model. The advantages of our approach are also in terms of parsimony and parameter interpretation, while preserving the goodness of fit.

Keywords: bivariate ordinal response; Dale's model; maximum penalized likelihood estimation; models for association; students' performance

1. Introduction

During the last decades, the universities of many European countries have undergone several structural reforms aiming at creating a common higher **educational** framework **as well,** at making the European universities more attractive for international students and scholars. One of the most serious problems of the Italian system, before the **introduction** of the **2001 Bologna process**, was the low rate of success and the presence of long survivor students, defined as those students who stay much **longer than** the legal duration of the degree course. **In spite of the introduction of the 2001 Bologna structure**, the problems concerning the **long survivors students** and the drop-outs are still present with **just a slight decline in** incidence. **The Italian** statistical literature for the analysis of students careers is vast and it covers both statistical modeling and indicators. Most of the studies consider retrospective cohort settings, which can be divided into two groups. The first one includes: multilevel models [16], **more or less complex transition models** [20, 23], logit and bivariate logit [9, 24], generalized linear models with **random** effects [14] and multistate models in the presence of competing risks [1, 10, 11, 17]. The latter group **often** focused on the construction of some composite indicators related to the students' careers [4, 25] for the prediction of their success [2].

In this paper the student career is analyzed by two response variables: the first one **is grades and is "qualitative", the second one is credits and is "quantitative"**.

*Corresponding author. Email: marco.enea@unipa.it.

URL: http://mc.manuscriptcentral.com/cjas

For Peer Review Only

According to us, those two variables **can summarize** the "University success", which can be analyzed through a set of covariates considered potential determinants of the success. This is done by jointly modelling the **above** two outcomes through a bivariate ordered logistic model (BOLM) [5, 12], in order to estimate the association structure between the two outcomes and the determinants of **such** association.

The data consist of a student cohort enrolled in a 3-year Economics Degree (EcD) in the academic year 2007-08. The statistical unit is the student **career. The** accumulation of credits and mean grade are measured at the end of the fourth year of **University attendance, allowing one year graduation** delay.

The usual approach to model two categorical responses is the bivariate logistic model [19], which provides better estimates comparing to two univariate models, or **one can use** the Dale's [5] model for ordered responses which can **also** be estimated using the former modelling framework **as well** [18].

In this paper, we use a penalized approach [8], potentially able to smooth both the marginal and the association parameters, across the response categories, as a smoothed alternative to the classical Row/Column (RC) Dale's parameterization. Our approach allows **us** to overcome some computational difficulties present in BOLMs, enlarges the range of the usual RC parameterization in a parsimonious way, while providing advantages in terms of parameter interpretation and goodness of fit.

Finally, the aim of this paper is twofold: to analyze in a novel way the University performance with a bivariate response model and, at the same time, to suggest some refinements of the estimation procedure introducing ad hoc penalty, which allows us to obtain smooth estimates of the association structure. From the applied point of view, the aims of the analysis are:

(1) to find the factors associated with the academic success with respect to grade and credits, marginally;
(2) to detect the determinants of the association between the two responses, in order to describe four student profiles: low credits and low grades, low credits and high grades, high credits and low grades and, high credits and high grades.

The latter point is crucial and shows the advantage of the BOLM on other models. In fact, classical univariate modeling approaches could provide insight only on the former aim, ignoring the association structure and the information thereof.

In the following Section, after describing the construction of the variables, **we present** the data set in terms of cross-classification of the responses and description of the covariates. Section 3 deals with the BOLM and how the penalty term is used for the analysis. The analysis is reported in Section 4. Section 5 includes the results and the discussion.

2. The EcD data set

A retrospective cohort study was conducted including all the freshmen, a total of 627 students, enrolled in the 3-year Economics Degree Course (EcD) **at the University of Palermo**, in the year 2007-2008. As already **stated** the two response variables are **grades and credits**.

The Italian University mean grade assessment scale ranges from 18 to 30 cum laude, while **60 credits are assigned annually to each student**. The grade variable is - strictly speaking - measured on an ordinal scale, even if it is quantitative. In order to analyze the association structure we need to categorize the grade variable. **According to us, a suitable way to categorize such variable is following the European**

For Peer Review Only

Credit Transfer System (ECTS), since it is based on percentiles. The percentiles are employed in the Erasmus System [6] and they are considered one of the best ways to compare grades among different schools/educational systems. Following this approach, we convert the grades into the E-A ordinal scale, that is, from E (sufficient) to A (excellent). Erasmus aims at classifying students on the basis of the quality of their performance, summarized by their mean grade. Its structure consists of five subgroups of successful students, of which the top 10% is awarded an A-grade, the next 25% a B-grade, the following 30% a C-grade, the following 25% a D-grade and the bottom 10% an E-grade. Actually this categorization will eventually allow comparisons with other Italian and European (after suitable adjustments) Universities, but this is not considered in the present paper. Figure 1 shows the empirical cumulative distribution function (ECDF) of the means of the students' original grades and the corresponding percentiles. We called *GRADE* the variable resulting from such categorization.

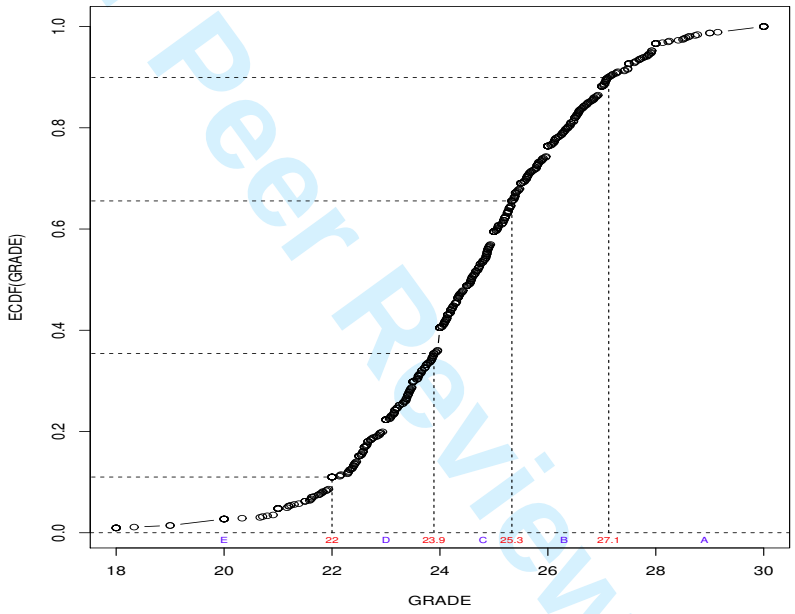


Figure 1. Empirical cumulative distribution function of the means of the students' grades, according to the ECTS percentiles, for the EcD data set.

The second response variable, *CREDITS*, accounts for the quantitative performance of the student, measured by the CFU, the Italian acronym for credits. The natural classification is given by 60 CFU required per year. We also considered 30 as an extra cut-point, since it is the minimum threshold required to apply for a scholarship or some other benefits such as accommodations, according to **our** University regulations. Summarizing, the categorization is: 1 (<30), 2 ($\geq 30, <60$), 3 ($\geq 60, <120$), 4($\geq 120, <180$), 5 (≥ 180). Table 2 shows the cross-classification of the students according to the two responses, along with the corresponding log-global odds ratios (log-GORs, defined in the next Section) corrected by the usual 0.5 adjustment factor.

The variables $CREDITS$ and $GRADE$ seem to be positively associated, although the presence of a negative log-GORs for $GRADE =$ "B" and $CREDITS =$ "1".

The covariate coding is reported in Table 2, along with the observed percentages for

URL: http://mc.manuscriptcentral.com/cjas

Table 1. Cross-classification of 627 students according **to** responses GRADES and CREDITS for the EcD data set. Within brackets: log-global odds ratios.

	CREDITS				
GRADE	1	2	3	4	5
E	41	15	11	2	0
	(1.80)	(2.06)	(2.47)	(3.49)	
D	23	32	68	25	5
	(0.47)	(0.73)	(1.31)	(2.49)	
C	39	25	69	26	30
	(0.32)	(0.51)	(1.12)	(1.60)	
B	21	21	39	31	41
	(-0.55)	(0.01)	(0.93)	(1.65)	
A	21	3	7	3	29

the **categorical** covariates or the medians for the continuous ones.

Table 2. Covariate coding and percentages/medians for the EcD dataset.

Variable	Level/Range	%/Median	Description
Age	≤ 19	85.0	Age at enrollment
	20	8.0	
	> 20	7.0	
Sex	female	50.0	Sex
	male	50.0	
HSgrade	$0 - 40$	20	High school final grade scaled at the minimum mark (60)
HStype	Other	7.9	High school type
	Classical	11.9	
	Scientific	28.3	
	Technical	51.3	
	Vocational	0.6	
Income	$0 - 175.6$	28.6	Annual family income (in thousands)
	NA	5.0	Missing data

The first cut-off of variable Age is 19, the expected age at enrolling for "regular" students. The second cut-off, 20, is also considered in order to distinguish students that are one-year late **in enrolment**. Variable HSgrade is quantitative and ranges onto a $60 - 100$ scoring scale but, for **easier interpretation**, it has been scaled in the interval $0 - 40$.

3. Estimating a BOLM by penalized maximum likelihood

Let $A_1 \times A_2$ be a two-way table cross-classifying two ordered responses A_1 and A_2, respectively with D_1 and D_2 categories, $\boldsymbol{\pi}$ be the underlying vector of cell probabilities in lexicographic order. Define the row and column marginal cumulative probabilities as

$$\mu_{r.} = P(A_1 \leq r) = \sum_{i \leq r} \pi_{i.}, \quad \mu_{.c} = P(A_2 \leq c) = \sum_{j \leq c} \pi_{.j},$$

URL: http://mc.manuscriptcentral.com/cjas

For Peer Review Only

and the first quadrant cumulative probabilities (that is the upper-left quadrant) as

$$\mu_{rc} = P(A_1 \leq r,\, A_2 \leq c) = \sum_{i \leq r} \sum_{j \leq c} \pi_{ij},$$

with $r = 1, \ldots, D_1$, $c = 1, \ldots, D_2$. By difference we obtain

$$\begin{aligned} P(A_1 \leq r,\, A_2 > c) &= \mu_{r.} - \mu_{rc}, \\ P(A_1 > r,\, A_2 \leq c) &= \mu_{.c} - \mu_{rc}, \\ P(A_1 > r,\, A_2 > c) &= 1 - \mu_{r.} - \mu_{.c} + \mu_{rc}. \end{aligned}$$

By choosing the cumulative odds as ordinal risk measures, and the logit function as link function, we obtain the row and column *global logits*, defined as

$$\log \phi_{1r} = \text{logit}[P(A_1 \leq r)] = \log(\mu_{r.}) - \log(1 - \mu_{r.}), \tag{1}$$

$$\log \phi_{2c} = \text{logit}[P(A_2 \leq c)] = \log(\mu_{.c}) - \log(1 - \mu_{.c}), \tag{2}$$

$r = 1, \ldots, D_1 - 1$, $c = 1, \ldots, D_2 - 1$. By choosing the cross-product of quadrant probabilities as ordinal association measure, and the natural logarithm function as link function, the *global log-odds ratio* (or *log-global odds ratio* or log-GOR) is defined as:

$$\log \psi_{rc} = \log \frac{P(A_1 \leq r,\, A_2 \leq c) P(A_1 > r,\, A_2 > c)}{P(A_1 \leq r,\, A_2 > c) P(A_1 > r,\, A_2 \leq c)} = \log \frac{\mu_{rc}(1 - \mu_{r.} - \mu_{.c} + \mu_{rc})}{(\mu_{r.} - \mu_{rc})(\mu_{.c} - \mu_{rc})}. \tag{3}$$

Given the three parameters $\mu_{r.}$, $\mu_{.c}$, and ψ_{rc}, the corresponding cumulative joint probabilities can be obtained through the following inversion formula:

$$\mu_{rc} = \begin{cases} \frac{1}{2}(\psi_{rc} - 1)^{-1}(a_{rc} - \sqrt{a_{rc}^2 + b_{rc}}) & \text{if } \psi \neq 1, \\ \mu_{r.}\mu_{.c} & \text{if } \psi = 1, \end{cases} \tag{4}$$

where $\boldsymbol{a_{rc}} = 1 + (\mu_{r.} + \mu_{.c})(\psi_{rc} - 1)$ and $\boldsymbol{b_{rc}} = -4\psi_{rc}(\psi_{rc} - 1)\mu_{r.}\mu_{.c}$. If the cumulative probabilities $\mu_{r.}$ and $\mu_{.c}$ satisfy the constraints $\mu_{r.} < \mu_{r+1,.}$ for $r = 1, \ldots, D_1 - 1$, and $\mu_{.c} < \mu_{.,c+1}$ for $c = 1, \ldots, D_2 - 1$, and the global odds ratios are not dependent on the categories, that is $\psi_{rc} = \psi$, then (4) is a Plackett [22] distribution. Consider the following proportional odds BOLM:

$$\begin{cases} \log(\phi_{1r}(\boldsymbol{x}_i)) = \beta_{10r} + \boldsymbol{\beta}_1' \boldsymbol{x}_i, \\ \log(\phi_{2c}(\boldsymbol{x}_i)) = \beta_{20c} + \boldsymbol{\beta}_2' \boldsymbol{x}_i, \\ \log(\psi_{rc}(\boldsymbol{x}_i)) = \beta_{30rc} + \boldsymbol{\beta}_3' \boldsymbol{x}_i, \end{cases} \tag{5}$$

with $r = 1, ..., D_1 - 1$, $c = 1, ..., D_2 - 1$, and where ϕ_{kr}, $k = 1, 2$ are global odds, ψ_{rc} are global odds ratios (GORs). Parameters β are unknown and $\boldsymbol{x}_i$ is the covariate vector of length p for the ith unit, with $i = 1, ..., m$, where m is the observed number of response configurations. Notice that (5) is a system of $D_1 \times D_2 - 1$ equations, in which the covariates are supposed to have a proportional effect on the categories of the responses. Dale [5] already proposed a model similar to (5), in which $\log(\psi_{rc}(\boldsymbol{x}_i)) = \alpha + \gamma_r + \delta_c + \sigma_{rc} + \boldsymbol{\beta}_3' \boldsymbol{x}_i$, that is, the log-global odds ratio structure, is modelled by an RC-type structure with

For Peer Review Only

base, row, column, and interaction effects. For the estimation of such model, the following uniqueness constraints are **chosen**: $\gamma_{D_1-1} = \delta_{D_1-1} = 0$ and $\sigma_{r,D_2-1} = \sigma_{D_1-1,c} = 0$, $r = 1, ..., D_1 - 1$, $c = 1, ..., D_2 - 1$. The Dale model does not require marginal scores for the responses and it is also invariant under any monotonic transformation of the marginal responses. Further, since the model is based on global odds ratios, collapsing adjacent row or column categories does not produce any effect in parameter interpretation. That is in contrast with the RC Goodman [13] **model,** which uses local cross-ratios. On the other hand, Glonek and McCullagh [12] write the *multivariate logistic model*:

$$\boldsymbol{C}' \log(\boldsymbol{L}\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}, \tag{6}$$

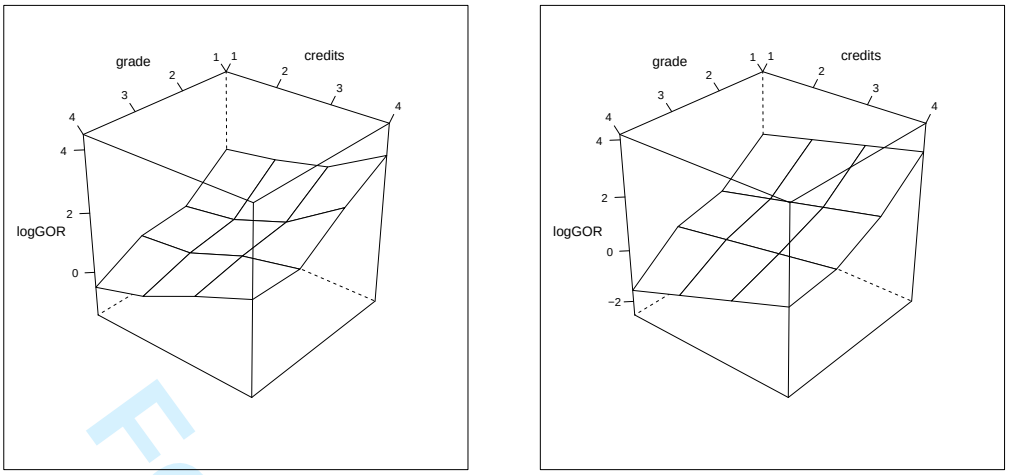
where $\boldsymbol{C}$ is a contrast matrix, $\boldsymbol{L}$ is a matrix of 0's and 1's such that $\boldsymbol{L}\boldsymbol{\pi} = \boldsymbol{\mu}$, $\boldsymbol{\eta} = \boldsymbol{C}' \log(\boldsymbol{L}\boldsymbol{\pi})$ is the parameter vector of interest, and $\boldsymbol{X}$ is the $m \times p$ design matrix. For the bivariate case, the components of $\boldsymbol{C}' \log(\boldsymbol{L}\boldsymbol{\pi})$ are symbolically denoted by $\boldsymbol{\eta} = (\eta_\varnothing, \boldsymbol{\eta}'_{A_1}, \boldsymbol{\eta}'_{A_2}, \boldsymbol{\eta}'_{A_1A_2})'$, where $\eta_\varnothing = \log(\sum \boldsymbol{\pi}) = 0$ is the null contrast and the remaining vectors have elements specified by (1), (2), and (3), respectively. **We replace (5) with (6) because it is more computationally convenient and this** can be generalized with respect to the number and the type of responses.

Besides the Dale model, we use a penalty term [8] to constrain the baseline association structure, defined by β_{30rc}, that follows a polynomial surface. This enlarges the range of possible parameterizations of the association structure of (5). Under multinomial sampling with frequencies $\boldsymbol{y}_i \sim M(n_i, \boldsymbol{\pi}_i)$, the fitting or estimation of (5) is performed by penalized maximum likelihood estimation for which the kernel of the penalized log-likelihood is $l_P(\boldsymbol{\beta}) = \sum_{i=1}^{m} \boldsymbol{y}'_i \log(\boldsymbol{\pi}_i) - \frac{1}{2}\tau(\boldsymbol{\beta})$. As a specification for $\tau(\boldsymbol{\beta})$ we use:

$$\begin{aligned}\tau(\boldsymbol{\beta}) = \lambda_1 \sum_{r=s_1+1}^{D_1-1} (\Delta^{s_1}\beta_{10r})^2 + \lambda_2 \sum_{c=s_2+1}^{D_2-1} (\Delta^{s_2}\beta_{20c})^2 \\ + \left[\lambda_3 \sum_{r=s_3+1}^{D_1-1} \sum_{c=1}^{D_2-1} (\Delta^{s_3}\beta_{30rc})^2 + \lambda_4 \sum_{r=1}^{D_1-1} \sum_{c=s_4+1}^{D_2-1} (\Delta^{s_4}\beta_{30rc})^2\right],\end{aligned} \tag{7}$$

where Δ^a is the difference operator of order a, and the λ's are the smoothing (or tuning) parameters. As λ_h increases ($h = 1, ..., 4$), the term can be used to fit nonparametric models where the effects are determined by a polynomial of degree $s_h - 1$, with equally-spaced and integer scores. In particular, when $h = 1, 2$, the estimated marginal intercepts will lie onto a polynomial curve; when $h = 3, 4$, the estimated association intercepts will lie onto a polynomial surface. By varying λ's, (7) allows to fit models with scores "chosen by data". Penalty (7) is the sum of single **terms particularly** known in literature **because they are used** in the P-spline **context. They were introduced by** Eliers and Marx [7]. For example, for the univariate case, Tutz and Scholz [26] use just the first term of (7) to penalize the likelihood of cumulative logistic models while Tutz [27] employs (7) for additive cumulative logistic models. The choice of smoothing parameters and polynomial degrees can be made, for example, on the grounds of the values minimizing the AIC. In addition, (7) can also be used to fit models with row or column effects, to reproduce proportional odds models, or for parameter space regularization. The latter option may be applied **whether** Fisher scoring fails the estimation of (5). **This may occur in** presence of zeros for some configurations of the responses. Then, one can try to estimate (5) by fixing small λ values to stabilize the estimates, as an alternative, or in addition, to the usual *step length* reduction used in line-search iterative algorithms. **Using the latter solution** for this kind of **problem** may produce estimates of the

For Peer Review Only



(a) Oserved log-GORs. (b) Fitted log-GORs.

Figure 2. Observed and fitted (for the baseline) log-GORs, for the EcD data set.

association structure which are too "irregular". In addition, by a comparison on a literature dataset, Enea [9] showed that the use of (7) may be a valid alternative to the Dale's parameterization, in terms of parsimony, parameter interpretation and goodness of fit. **In order to select the model, Tutz and Scholz [26] suggest the use of the classical χ^2 approximation to the asymptotic distribution of LR_P, the penalized likelihood ratio statistic, when fixing small tuning parameters. That result was investigated via simulation by Enea [8].**

4. Results of the EcD data analysis

Figure 2 (a) shows the observed association structure, measured by the log-GORs. Such association structure is modelled using both the Dale's parametrization and the BOLM with the penalty term (7). In both cases, the model and the covariates selection is crucial as well as the association structure. We followed the selection strategy recommended by Lapp *et al.* [18], **by choosing first** the model for the association and then **by proceeding** to the **covariates selection**. **In this paper**, the choice of the association model is carried out on the grounds of the AIC, whereas the selection of covariates is performed by a backward procedure based on the penalized deviance LR_P and the AIC.

Selection of the covariates is performed for a linear model by considering a backward elimination in the following subsets:

$$\begin{aligned}
\mathcal{P}_1 &= \{\text{HStype}_{[123]}, \text{HSgrade}_{[123]}, \text{Sex}_{[123]}, \text{Income}_{[123]}, \text{Age}_{[123]}\},\\
\mathcal{P}_2 &= \{\text{HStype}_{[123]}, \text{HSgrade}_{[123]}, \text{Sex}_{[123]}, \text{Income}_{[123]}\},\\
\mathcal{P}_3 &= \{\text{HStype}_{[123]}, \text{HSgrade}_{[123]}, \text{Sex}_{[123]}\},\\
\mathcal{P}_4 &= \{\text{HStype}_{[123]}, \text{HSgrade}_{[123]}\},\\
\mathcal{P}_5 &= \{\text{HStype}_{[123]}, \text{HSgrade}_{[12]}\},
\end{aligned}$$

where the index of $\text{HStype}_{[123]}$ indicates the presence of HStype in all the 3 equations of model (5), $\text{HStype}_{[12]}$ indicates that it is present in the first and the second equation, and

For Peer Review Only

so on. Let $NUPOM(\mathcal{P}_j)$, $j = 1, ..., 5$, be a proportional-odds model with non-uniform association, that is model (5) fitted on set $\mathcal{P}_j$. Let $NUPOM_{s_3 s_4}(\mathcal{P}_j)$ be the same model taking into account that the association surface is expressed by row and column polynomials of degree $s_3 - 1$ and $s_4 - 1$, respectively. Furthermore, a proportional odds model with uniform association will be indicated by $UPOM(\mathcal{P}_j)$. A $NUPOM(\mathcal{P}_1)$ is the most complex model we have considered. The goodness of fit of the selected model is assessed with respect to the $NUPOM(\mathcal{P}_1)$ by using the LR_P statistic.

The selection of the association structure **as well as** the covariates and the goodness-of-fit test **are** reported in Table 3.

Table 3. EcD data set: association structure selection (Models Ec1-Ec18), backward selection of the covariates (Models Ec19-Ec22) based and goodness of fit.

Model	Description	# parameters	Vs Model	LR_P	df	p-value	AIC
Association structure selection:							
(1) Polynomial type							
Ec1	$NUPOM(\mathcal{P}_1)$	51	-	-	-	-	3396.7
Ec2	$NUPOM_{33}(\mathcal{P}_1)$	44	Ec1	12.28	7	0.092	3394.9
Ec3	$NUPOM_{23}(\mathcal{P}_1)$	41	Ec1	17.80	10	0.058	3394.5
Ec4	$NUPOM_{13}(\mathcal{P}_1)$	38	Ec1	43.02	13	0.000	3413.7
Ec5	$NUPOM_{32}(\mathcal{P}_1)$	41	Ec1	20.70	10	0.023	3397.4
Ec6	$NUPOM_{31}(\mathcal{P}_1)$	38	Ec1	77.13	13	0.000	3447.8
Ec7	$NUPOM_{22}(\mathcal{P}_1)$	39	Ec1	23.58	12	0.023	3396.3
Ec8	$NUPOM_{21}(\mathcal{P}_1)$	37	Ec1	79.76	14	0.000	3448.4
Ec9	$NUPOM_{12}(\mathcal{P}_1)$	37	Ec1	48.20	14	0.000	3416.9
(2) RC type							
Ec10	$Dale\, RC(\mathcal{P}_1)$	51	-	-	-	-	3395.7
Ec11	$Dale\, R + C(\mathcal{P}_1)$	42	Ec10	14.71	9	0.099	3392.4
Ec12	$Dale\, R(\mathcal{P}_1)$	39	Ec10	78.14	12	0.000	3449.9
Ec13	$Dale\, C(\mathcal{P}_1)$	39	Ec10	39.11	12	0.000	3410.8
(3) Polynomial-RC mixed type							
Ec14	$NUPOM_{R3}(\mathcal{P}_1)$	47	Ec1	9.37	4	0.052	3398.0
Ec15	$NUPOM_{3C}(\mathcal{P}_1)$	47	Ec1	1.38	4	0.847	3390.1
Ec16	$NUPOM_{2C}(\mathcal{P}_1)$	43	Ec1	5.68	8	0.683	3386.4
Ec17	$NUPOM_{R2}(\mathcal{P}_1)$	43	Ec1	20.62	8	0.008	3401.3
(4) Uniform type							
Ec18	$UPOM(\mathcal{P}_1)$	36	Ec1	92.04	15	0.000	3458.7
Covariate selection:							
Ec19	$NUPOM_{2C}(\mathcal{P}_2)$	37	Ec16	6.01	6	0.422	3380.4
Ec20	$NUPOM_{2C}(\mathcal{P}_3)$	34	Ec19	2.53	3	0.470	3376.9
Ec21	$NUPOM_{2C}(\mathcal{P}_4)$	31	Ec20	4.04	3	0.257	3374.9
Ec22	$NUPOM_{2C}(\mathcal{P}_5)$	30	Ec21	5.30	1	0.021	3378.2
Goodness of fit:							
Ec21	-	-	Ec1	18.26	20	0.570	-

The first one is Model Ec1. For this model, a small smoothing value of λ_4=0.0001 has been fixed to stabilize ML estimation. Model comparisons with respect this model imply non-integer degrees of freedom that can be rounded. The choice among Models Ec2-Ec18 **provides the best association structure based on the AIC**. These models involve several NUPOM models of polynomial type (Models Ec2-Ec9) **as well as** the Dale's RC, R+C, R and C models (Ec10-Ec13). **Furthermore, (7) allows us to enlarge** the range of the possible NUPOMs to the hybrid parameterizations R3, 3C, R2 and 2C (Ec14-Ec17). For example, $NUPOM_{R3}$ indicates a model **that** is unconstrained across the rows **but with column effects that follow** a second degree polynomial. **Models**

For Peer Review Only

of the type $R1$ and $1C$ have not been considered since these exactly correspond to the Dale's R and C models.

The models of polynomial type do not seem provide a good fit. Models Ec2 and Ec3 provide a non-significant difference with model Ec1, though the p-values are not so high. The Dale's R+C model Ec11 provides an acceptable fit and it is to prefer to the latter two models according to the AIC. The mixed parameterization provides the best two models, Ec15 and Ec16, **based on** fit and AIC. In particular, model Ec16 provides the best fitted association structure. Such model is linear along $CREDITS$ and unconstrained along $GRADE$. **Figure 2 (b) shows the baseline's student's profile association structure**.

The covariate selection shows that the best model is Model Ec21. **The goodness of fit of such model has been observed by comparing with Model Ec1 (p=0.57)**. The estimates for Model Ec21 are reported in Table 4.

Table 4. Parameter estimates from Model Ec21.

outcomes	parameter	estimate	se	z	p-value
global logits *GRADES*	Intercept[1]	-1.248	0.296	-4.219	0.000
	Intercept[2]	0.307	0.285	1.078	0.281
	Intercept[3]	1.717	0.293	5.856	0.000
	Intercept[4]	3.441	0.322	10.697	0.000
	HStype Classical	-0.767	0.332	-2.312	0.021
	HStype Vocational	0.540	0.920	0.587	0.557
	HStype Scientific	-0.287	0.290	-0.988	0.323
	HStype Technical	0.554	0.277	1.998	0.046
	HSgrade	-0.053	0.006	-8.616	0.000
global logits *CREDITS*	Intercept[1]	-0.238	0.278	-0.857	0.392
	Intercept[2]	0.625	0.278	2.243	0.025
	Intercept[3]	2.122	0.290	7.319	0.000
	Intercept[4]	3.005	0.302	9.958	0.000
	HStype Classical	-0.511	0.322	-1.585	0.113
	HStype Vocational	0.532	0.889	0.598	0.550
	HStype Scientific	-0.826	0.284	-2.907	0.004
	HStype Technical	0.642	0.270	2.381	0.017
	HSgrade	-0.057	0.006	-9.423	0.000
log global odds ratios	Intercept[1][1]	1.776	0.539	3.294	0.001
	Intercept[1][2]	2.245	0.569	3.945	0.000
	Intercept[1][3]	2.713	0.699	3.883	0.000
	Intercept[1][4]	3.183	0.886	3.593	0.000
	Intercept[2][1]	0.363	0.516	0.704	0.482
	Intercept[2][2]	0.840	0.512	1.640	0.101
	Intercept[2][3]	1.316	0.536	2.456	0.014
	Intercept[2][4]	1.794	0.583	3.073	0.002
	Intercept[3][1]	-0.088	0.536	-0.163	0.870
	Intercept[3][2]	0.259	0.527	0.492	0.622
	Intercept[3][3]	0.605	0.538	1.124	0.261
	Intercept[3][4]	0.951	0.569	1.673	0.094
	Intercept[4][1]	-1.555	0.587	-2.646	0.008
	Intercept[4][2]	-0.782	0.575	-1.361	0.174
	Intercept[4][3]	-0.008	0.583	-0.014	0.988
	Intercept[4][4]	0.765	0.611	1.253	0.210
	HStype Classical	-0.243	0.591	-0.412	0.681
	HStype Vocational	-0.250	1.616	-0.155	0.877
	HStype Scientific	-0.135	0.520	-0.259	0.796
	HStype Technical	-1.205	0.496	-2.430	0.015
	HSgrade	0.028	0.012	2.414	0.016

We first illustrate the **marginal models estimates**. Students coming from classical and scientific high schools seem to provide the best **grades and credits performances. Classical studies students have an estimated odds of receiving low grades which is about 2.15 times lower than the ones coming from "other" high**

For Peer Review Only

schools. Scientific studies students have an estimated odds of taking low credits which is 2.28 times lower. On the contrary, students coming from technical high school are likely to provide the worst grades (GOR= 1.74) and credits (GOR=1.9). HSgrade variable seems to be having about the same significant effect on both the responses. In fact, for both the responses, a 10 marks HSgrade increase implies a 1.7 odds ratio increase of having a positive performance. Of particular interest is that having a the classical high school background is a better predictor for good grades, but a scientific high school background is a better predictor for higher credits.

The association model gives information on the discordant profiles, as high (low) credits and low (high) grades. For instance, technical high school students provide discordant performances. The slope of the association structure in Figure 2 shows how those students are mainly likely to have high grades but low credits. However, they have also chances to have low grades and high credits. Furthermore, the strength of the association between the responses increases proportionally to the higher high school final grade. By considering that the baseline value of HSgrade is 0, we can conclude that students with high values of HSgrade have more chances to have both high grades and credits. Bad high school students are probably bad in terms of grades and credits, even if there are chances that they provide high grades and low credits, as we can see from the significant negative estimate of Intercept[4][1].

5. Conclusions

We investigated on the association between two response variables in order to "better understand" the University career of Italian students in a regression setting. The association model used for these data is very interesting in terms of future applications because it is able to give insight into the relationship between the velocity to get a degree, measured by the credits, and the excellence, measured by the grades. The analysis of this relationship is important to both examine the student's career and his/her job placement after completing the bachelor degree.

Furthermore, we provided a suitable semiparametric alternative to the Dale and the ordinary BOLM models. We used a penalty term of polynomial type for the estimation of a BOLM in order to provide smooth estimates of the association structure, while preserving parsimony and goodness of fit. The penalty term provides a parameterization of the association structure which is easier to interpret, because it is function of the global log-odds ratios. Although the penalty term we inserted is potentially able to smooth both the marginal intercepts and the association structure, we decided to smooth only the latter, because of greater interest. Model selection included a comparison among the proposed semiparametric BOLM, the ordinary BOLM and the Dale model. As result, the semiparametric BOLM provided the best fit on the grounds of the AIC. The choice of the smoothing parameter and the order of the polynomial degrees was based on the AIC as well. The estimated smoothing parameters were found fairly high to provide approximately equally-spaced and integer scores. This result allowed us to approximate the asymptotic behaviour of the penalized likelihood ratio statistic using the classical χ^2 distribution.

The analysis results showed that the classical or scientific high school stu-

For Peer Review Only

dents provide the best performances. The former have better grades while the latter have higher credits. Technical high school students have a lower chance for a good performance. The high-school final grade was found to be a good predictor of students' success. Furthermore, a low family income does not seem to influence students' performances.

The association model showed that the students that have earned a low high school grade, and the technical high school students, are likely to have variable University career. The positive association between the two outcomes increases for the best high-school students.

Finally, the most interesting feature of the model suggested in this paper is its capacity of giving extra information through the association structure, although the procedure is not straightforward. This is due to the fact that model selection involves the association structure, the covariates, and fit. In this respect it is important to balance, case-by-case, these three issues according to the aim and the type of data. Moreover, the potentiality of the method concerns two aspects. On one hand, the applications in which there are two ordered bivariate responses are very common, for instance, in medicine (the HCV and HIV patients are usually monitored by two responses) or in the social sciences (services evaluation often present two or more outcomes; etc.). On the other hand, the refinements suggested can be seen as a first extension of the ordinary BOLM model, which can be eventually developed in other directions.

Authors' contribution

The authors contributed to write this manuscript in the following way. The first author proposed the methodology, developed the R codes and carried out the analyses. He wrote Sections 2-4 and contributed to write Sections 1 and 5. The second author proposed the application, supervised the writing of the article and contributed to write Sections 1 and 5.

Acknowledgements

This paper has been supported by the Italian Ministerial grant PRIN 2008 "Measures, statistical models and indicators for the assessment of the University System", n. 2008WXSLH. We are also very grateful to the two anonymous referees for their constructive comments. A special thank goes to Sandra for the English revision.

References

[1] Ambrogi, F. and Biganzoli, E. and Boracchi, P.: Estimating crude cumulative incidences through multinomial logit regression on discrete cause-specific hazards. Computational Statistics and Data Analysis, **53(7)**, 2767–2779 (2009)

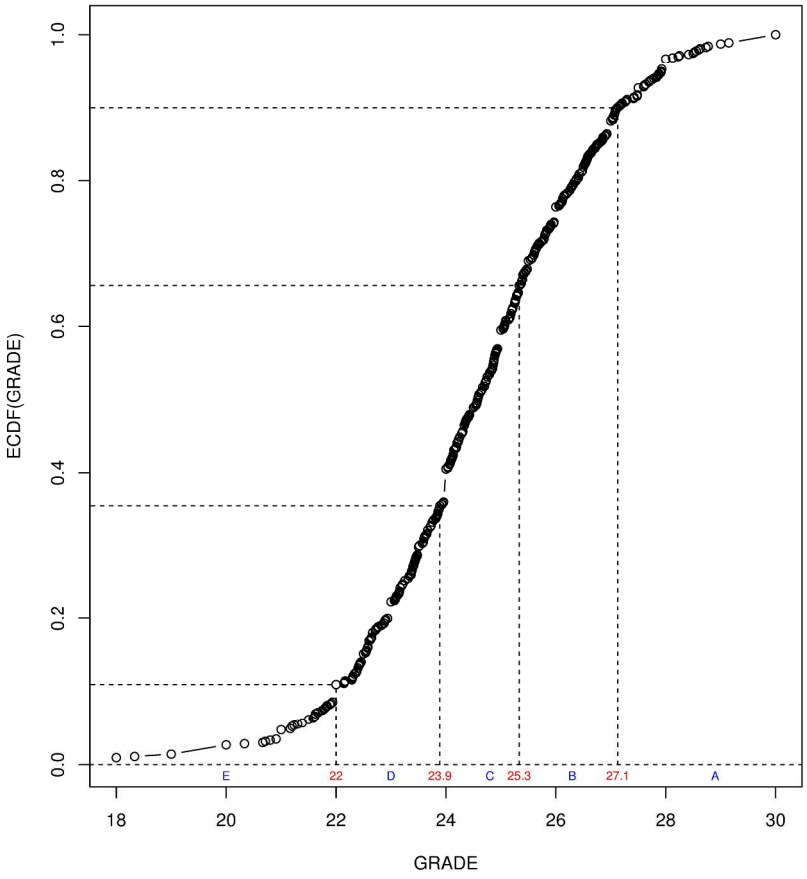
[2] Attanasio, M. and Boscaino, G. and Capursi, V. and Plaia, A.: May the students career performance helpful in predicting an increase in universities income? In P. Giudici, S. Ingrassia, & M. Vichi (Eds.) Statistical Models for Data Analysis. Series in Studies in Classification, Data Analysis, and Knowledge Organization. Springer International Publishing Switzerland, 9–16 (2013)

[3] Bustami, R. and Lesaffre, E. and Molenberghs, G. and Loos, R. and Danckaerts M. and Vlietinck, R.: Modelling bivariate ordinal responses smoothly with examples from ophthalmology and genetics, Statistics in Medicine, **20**, 1825–1842 (2001).

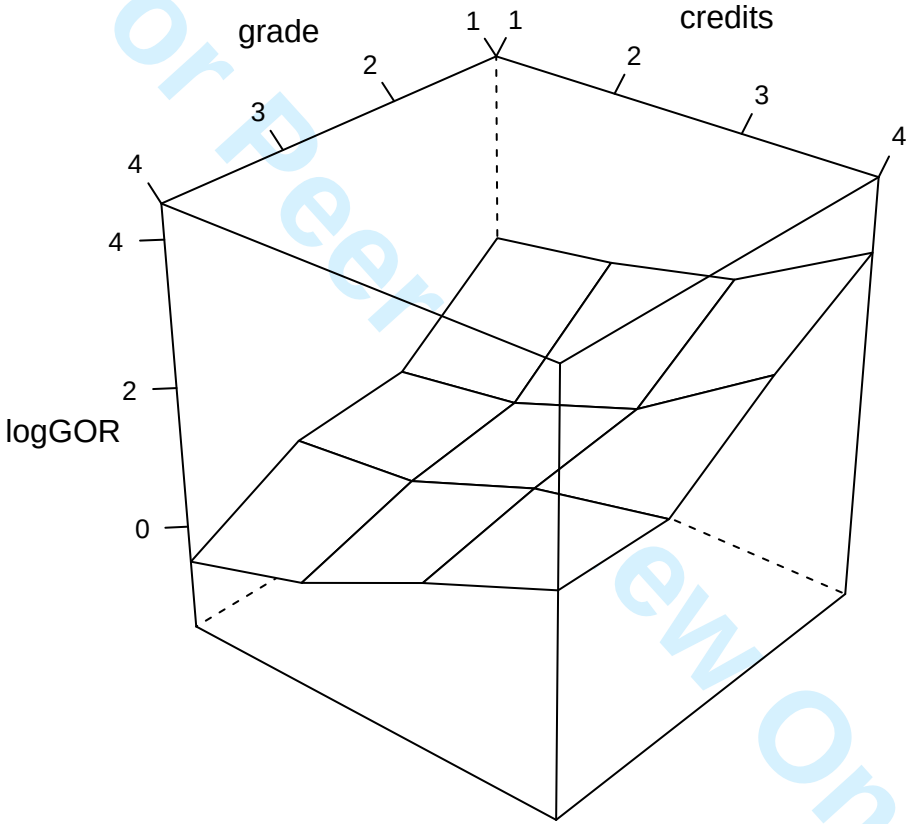
For Peer Review Only

[4] Capursi, V. and Librizzi, L.: La qualitá della didattica: indicatori semplici o composti? In Capursi, V. & Ghellini, G. (ed.) Dottor Divago: Discernere, Valutare e Governare la nuova Universitá. Franco Angeli. ISBN: 978-88-464-9634-8. (2008)
[5] Dale, J. R.: Global Cross-Ratio Models for Bivariate, Discrete, Ordered Responses. Biometrics, **42(4)**, 909–917 (1986)
[6] European Communities: ECTS User's Guide. Bruxelles (2009) ISBN: 978-92-79-09728-7
[7] Eilers, P. H. C. and Marx, B. D.: Flexible smoothing with B-splines and penalties. Statistical Science, **11**, 89–121 (1996)
[8] Enea, M.: Penalized Inference in Multivariate Ordered Logistic Models: Theory and Applications. Ph.D. thesis in statistics and quantitative finance, University of Palermo (2010)
[9] Enea, M.: Penalized Inference in Multivariate Ordered Logistic Models: Theory and Applications. Book of short papers of PHD Theses in Statistics and Applications. Cleup, Padova (2011)
[10] Fasola, S.: Laurea o abbandono: il buongiorno si vede dal mattino? Master thesis, University of Palermo, Palermo (2011)
[11] Foucher, Y. and Giral M. and Soulillou J. P. and Daures J. P.: A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. Statistics in Medicine, **2**, 5381–5393 (2007)
[12] Glonek, G. F. V. and McCullagh, P.: Multivariate logistic models. Journal of the Royal Statistical Society, Series B, **57**, 533–546 (1995)
[13] Goodman, L. A.: Simple models for the analysis of cross-classifications having ordered categories. Journal of the American Statistical Association, **74**(367), 537–552 (1979)
[14] Gori, E. and Montagni, M.: Random effects models for event data. Evaluating effectiveness of universities through the analysis of students careers. Multilevel Modelling Newsletter, **10(1)** (1997), ESRC, England.
[15] Gray, R. J.: Spline-based test in Survival Analysis. Biometrics, **50**, 640–652 (1994)
[16] Grilli, L. and Rampichini, C.: A multilevel multinomial logit model for the analysis of graduates' skills. Statistical Methods and Applications, **16(3)**, 381–393 (2007)
[17] Janssen, J. and Manca, R.: Applied Semi-Markov Processes. Springer. (2006)
[18] Lapp, K. and Molenberghs, G. and Lesaffre, E.: Models for the association between ordinal variables. Computational Statistics and Data Analysis, **28**, 387–411 (1998)
[19] McCullagh, P. and Nelder, J. A.: Generalized Linear Models. Chapman & Hall, London (1989)
[20] Mealli, F. and Pudney, S.: Specification tests for random-effects transition models: An application to a model of the British Youth Training Scheme. Lifetime Data Analysis, **5**, 213–237 (1999)
[21] Molenberghs, G. and Lesaffre, E.: Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. Journal of the American Statistical Association, **89**, 633–644 (1994)
[22] Plackett, R. L.: A class of bivariate distributions. Journal of the American Statistical Association, **60**, 516–522 (1965)
[23] Robinson, R.: Pathways to completion: patterns of progression through a university degree. Higher Education, **47**, 1–20 (2004)
[24] Schizzerotto, A. and Denti, F.: Lost and late - The experience of abandonment and irregularity in the studies in five generations of enrolled at the University of Milan-Bicocca. University Milan-Bicocca, http://nucleo.unimib.it/nucleoHTM/Abbandoni_2004.pdf (2005)
[25] Sulis, I. and Capursi, V.: Building up adjusted indicators of students' evaluation of university courses using generalized item response models. Journal of Applied Statistics, **40(1)**, 88–102 (2013)
[26] Tutz, G. and Scholz, T.: Ordinal regression modelling between proportional odds and non-proportional odds. Technical report, University of Munich, Institute of Statistics (2003)
[27] Tutz, G.: Generalized Semiparametrically Structured Ordinal Models. Biometrics, **59(2)**, 263–273 (2003)

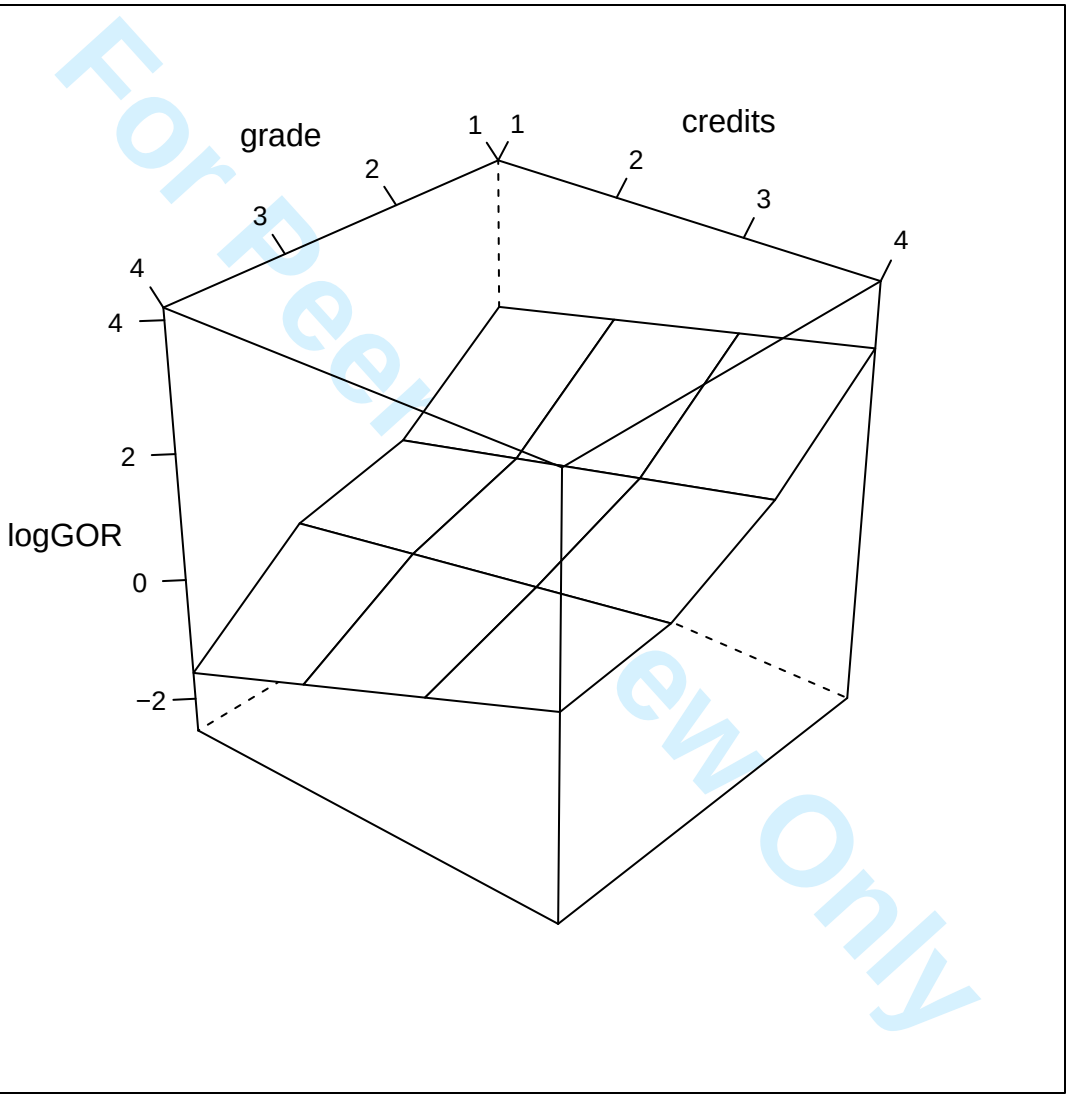
For Peer Review Only



Empirical cumulative distribution function of the means of the students' grades, according to the ECTS percentiles, for the EcD data set.
279x361mm (300 x 300 DPI)



For Peer Review Only



For Peer Review Only