Abstract: The Portuguese Naval Academy (NA) has a significant failure rate, losing about half of the admitted individuals during the course. This work aims to understand the causes associated to low income, through the analysis of the data of the individuals from the application to the end of the course, in order to identify the characteristics of individuals with greater and less likely to succeed in the NA. This work was started in [1]. In an initial phase, some techniques of descriptive statistics of data analysis were used. The first step of this analysis, the candidates are analyzed independently, and the admitted and the finalists are analyzed together comparing the variables at the beginning and at the end of the courses. Simple statistical inference techniques were used [10], namely confidence intervals, parametric tests, contingency tables. Such analysis was completed in [11] using intermediate level inference techniques, namely analysis of variance (ANOVA). In the present manuscript we intend to complete this approach using a general linear models approach [7, 12] and factorial analysis [2, 3, 4, 5]. The study evidences greater success for individuals entering in NA with better grades and for individuals taking notice of the application competition over the internet.


1 Introduction

The Portuguese Naval Academy has an academic failure rate of around 64%.

More than half of the admitted students do not finish the course. This project arises in order to understand the causes associated to the low performance.

For the NA, a public military university, very restricted for budgetary reasons, there is a need to understand the reasons for failure, that would diminish internal attrition.

Common qualified staff, with the means available to carry out studies of this nature, and a commitment to continuous improvement of the quality of it is a challenge to know the predictors of success.

The institution needs to be able to identify the performance, and be awake to possible irregularities during the process of individuals formation or destructive
phenomena of the good performance of the collective and / or individual.

It is mandatory to perform a statistical analysis of individuals from the beginning to the end of the course, to identify the characteristics (denominated variables) of individuals with higher and lower probability to contribute to academic success at the Naval Academy.

The outline of this work consists in five sections. Section 2 describes the motivation and objectives. The statistical models are described in Section 3. Section 4 displays some details about the empirical application. In fifth Section, we present some results and do some discussion. In last Section we get some conclusions.

2 Preliminaries

This work is a case study, that deliberately addresses the Naval Academy of the Portuguese Navy.

The NA, an academy with singular characteristics, is also characterized by its conditions, which shape not only the universe of candidates, but also the success during course attendance.

When competing in NA, the candidate has to face constraints that are not observed in the remaining students in other higher education institutions, detailing:

1. Tests of admission with great physical exigency;
2. Mandatory to wear uniforms on journeys;
3. Freedom inhibition to leave daily;
4. Freedom inhibition to leave if you do not comply with the military regulations;
5. Military discipline, both among peers and with the surrounding environment;
6. Loss the school year if the student fails in a single curricular unit;
7. It is possible to fail a year but its repetition lacks approval and it can only be achieved with exemplary military behavior;
8. Mandatory to spend annually about two months boarded on warships;
9. Accumulates the frequency of an integrated master’s degree course with physical preparation, military and naval training, training leadership-oriented behavioral and scientific research in areas linked to the sea;
10. Does not have daily available time for rest or well-being, due to the great time requirements for the studies and for the component as well as for naval military training.

The aim of this study is to analyze the data of the candidates for EN in the sense to relate the academic success or failure of individuals who are admitted to this educational institution, with socio-economic, cultural and performance details.

Although many variables considered for this study could be equally useful for studies of this nature applied to other institutions, any conclusions taken from this study can not be applied to students who integrate different realities from those of the NA.

This institution presents a singular physical environment in the daily life of those who attend.

In order to carry out this project, it was taken into account of certain issues to evaluate, such as:

- The stereotype of the heterogeneity of social classes, and the gender associated with disciplines, known as bipolarization of knowledge;
- The phenomenon of increasing female participation in services;
- The functionalist perspective of education.

The dedication of any individual in a common context is seen as a predictor of success, however, because it is a very specific context and other factors, such as low rest level and adaptations to hourly demands, inter alia, overlap with the capabilities of the individual, leading to academic failure.

An oscillation of the variables, from year to year or in the same year, between the group of those admitted and the group of finalists is natural, however, a significant or persistent variation of a variable tends to express some phenomenon or event, especially of a social nature, which is a way of influencing the group or generation, motivation, focus or other levels, which goes against the predisposition of students to academic success in the NA. The main phenomenon targeted by this study is the transition of individuals who entering the NA, from the 1st to the 5th year, and the consequent reduction in the number of individuals who were admitted during this five-year period.

Based on data from the admission years from 2007 to 2011, the NA presents a very high academic failure rate, with a success rate always below 50%, even knowing that individuals who complete their studies have a guaranteed job in the Navy.

To characterize the success profiles in the NA we consider the group of admitted individuals, and within...
the group of individuals who finish the course successfully.

The data analysis is made in order to identify the predictive variables of success and the information available for the accomplishment of this study.

3 Statistical Models

3.1 General linear models

In the classical linear model, a vector $X$ with $p$ explanatory variables $X = (X_1, X_2, \ldots, X_p)$ can explain the variability of the variable of interest $Y$ (response variable), where $Y = Z\beta + \epsilon$. $Z$ is a specification matrix with size $n \times p$ (usually $Z = X$, considering an unitary vector in first column), $\beta$ a parameter vector and $\epsilon$ a vector of random errors $\epsilon_i$, independent and identical distributed to a reduced Gaussian.

The data are in the form $(y_i, x_i)$, $i = 1, \ldots, n$, as result of observation of $(Y, X)$ $n$ times. The response variable $Y$ has expected value $E[Y | Z] = \mu$.

GLM is an extension of classical model where the response variable, following an exponential family distribution [12], do not need to be Gaussian. Another extension from the classical model is that the function which relates the expected value and the explanatory variables can be any differentiable function. $Y_i$ has expected value $E[Y_i | x_i] = \mu_i = x_i^T \beta_i$, $i = 1, \ldots, n$.

It is also defined a differentiable and monotone link function $g$ which relates the random component with the systematic component of response variable. The expected value $\mu_i$ is related with the linear predictor $\eta_i = z_i^T \beta_i$ using the relation

$$\mu_i = h(\eta_i) = h(z_i^T \beta_i), \quad \eta_i = g(\mu_i) \quad (1)$$

where $h$ is a differentiable function; $g = h^{-1}$ is the link function; $\beta$ is a vector of parameter with size $p$ (the same size of the number of explanatory variables); $Z$ is a specification vector with size $p$.

There are different link functions in GLM. When the random component of response variable has a Poisson distribution, the link function is logarithmic and the model is log-linear. In particular, when the linear predictor $\eta_i = z_i^T \beta_i$ coincides with the canonical parameter $\theta_i$, $\theta_i = \eta_i$, which implies $\theta_i = z_i^T \beta_i$, the link function is denominated as canonical link function. Sometimes, the link function is unknown being estimated simultaneously with the linear component of the semi-parametric model for electricity spot prices. A detailed description of GLM methodology can be found in several references such as [12, 7].

3.2 Factorial Analysis

Factor analysis (FA) is technique often used to reduce data. The purpose is to get a reduced number of variables from an initial big set of variables and get easier interpretations [5, 8]. The FA computes indexes with variables that measures similar things. There are two types of factor analysis: exploratory factorial analysis (EFA) and confirmatory factorial analysis (CFA) [13]. It is called EFA when there is no idea about the structure or the dimension of the set of variables. When we test some specific structure or dimension number of certain data set we name this technique the CFA. There are various extraction algorithms such as principal axis factors, principal components analysis or maximum likelihood (see [3, 9] for example). There are numerous criteria to decide about the number of factors and theirs significance. For example, the Kaiser criterion proposes to keep the factors that correspond to eigenvalues greater or equal to one. In the classical model, the original set contains $p$ variables $(X_1, X_2, \ldots, X_p)$ and $m$ factors $(F_1, F_2, \ldots, F_m)$ are obtained. Each observable variable $X_j, j = 1, \ldots, p$ is a linear combination of these factors:

$$X_j = \alpha_{j1} F_1 + \alpha_{j2} F_2 + \cdots + \alpha_{jm} F_m + e_j, j = 1, \ldots, p, \quad (2)$$

where $e_j$ is the residual. The factor loading $\alpha_{jk}$ provides an idea of the contribution of the variable $X_j$, $j = 1, \ldots, p$, contributes to the factor $F_k$, $k = 1, \ldots, m$. The factor loadings represents the measure of association between the variable and the factor [5, 13].

FA uses variances to get the communalities between variables. Mainly, the extraction issue is to remove the largest possible amount of variance in the first factor. The variance in observed variables $X_j$ which contribute to a common factor is defined by communality $h_j^2$ and is given by

$$h_j^2 = \alpha_{j1}^2 + \alpha_{j2}^2 + \cdots + \alpha_{jm}^2, \quad j = 1, \ldots, p. \quad (3)$$

According with the author of [6], the observable variables with low communalities are often dropped off once the basic idea of FA is to explain the variance by the common factors. The theoretical common factor model assumes that observables depend on the common factors and the unique factors being mandatory to determine the correlation patterns. With such objective the factors/components are successively extracted until a large quantity of variance is explained. After the extraction technique be applied, it is needed to proceed with the rotation of factors/components maximizing the number of high loadings on each observable variable and minimizing the number of fac-
tors. In this way, there is a bigger probability of an easier interpretation of factors ‘meaning’.

4 Empirical application

This study is applied to a universe of 3091 candidates, of which 295 were admitted, and 103 of these successfully completed the course.

Treatment and cleaning of the information was performed, after which analysis of the data concerning the candidates and the individuals who were during the period under study was completed.

The concept of “transition”, often used throughout this work, means the passage from the group of admitted to the success group (finalists) to the over the five-year period that completes the training cycle for officers in the NA.

The decrease of individuals is justified by factors such as disapproval, withdrawal, expulsion, among others.

The simple analysis of collected data conducted to:

- Generally, the age group is between 17 and 26 years;
- Occasionally, individuals with a maximum of 29 years, with no experience except in particular cases, financially dependent;
- They come from public education with daily return to the family home;
- Their origin from all parts of the country and from different social environment;
- Without any kind of preparation for the five years of military education;
- Great attrition during the five year course.

The variables studied in detail were:

- Success( 1 - success, 0 failure)
- Gender (1 - Men, 0 - Female)
- Distance to NA (distance between the student’s residence and the NA)
- Grade at Entry (note of admission of the student in the NA)
- Higher Education Attendance (1 - attended higher education before entering Naval Academy, 0 - otherwise)
- Knowledge of Contest in Internet (1 - was aware of the internet contest, 0 - otherwise)
- Knowledge of Contest through Family (1 - had knowledge of the contest through relatives, 0 - otherwise)
- Know. Contest Visit EN (1 - had knowledge of the contest through a visit to the Naval School, 0 - otherwise)
- Acc. Higher Educat. (1 - went to higher education beyond the Naval School, 0 - otherwise)
- Parent or Mother Active (1 - If the parents get some income: salary / pension, 0 - otherwise)
- Military / Military Parents (1 - If the parents are / were military / militarized, 0 - otherwise)

In [1, 11] was performed a preliminary analysis, the data was organized using descriptive and some simple inference techniques, namely parametric and nonparametric tests, contigency tables, ANOVA Approach (parametric, nonparametric). In present work we have applied logistic modeling, factorial analysis and discriminant analysis.

5 Results and Discussion

The results described in [1, 11] were confirmed by the logistic model which is still being improved.

From the analysis applied to the variables after the end of the course to characterize the transition phase after five years, we can evidence that success rate is better for groups with smaller number of admitted individuals.

Although of the adversities of integrating women into top-class of the armed forces and certain determinations that defend the stereotype of the gender associated with the discipline, there is significant evidence of no prevalence of gender in relation to performance.

No despite the adversities that the NA presents inherent to a Military life, individuals who fail to complete the course are the ones that have worse entry grade. As a rule, individuals who have better grades can be succeeded.

Although of a greater academic experience, students who already had some experience in higher education previously to their ingress in the NA, their level of the performance does not not stand out.

The internet is an excellent way of disseminating NA. Individuals who take notice of the competition through the internet tend to have good performance. Although the family members represent the greater form of disclosure of the NA, this disclosure is not
conducive to success. Visits to the NA, despite being programmed with the specific purpose of the competition and the Navy, have proved to be a form of unfavorable disclosure.

Both parents in the active is unfavorable factor to the success of individuals in the NA. This fact may be justified by the possible financial stability that fuels the intolerance of fatigue, the effort and the spirit of sacrifice inherent in military life. To be son of military or militarized parents is unfavorable to success of individuals in the NA. This may be justified by greater willingness on the part of the parents than on the individual about a career as Navy Officer.

The next step of process applies a FA. Considering the factor analysis approach, we summarize the results in 1. We estimated the communality for the factors, analyzed the significance of R-matrix (test about the multicollinearity or singularity). The Bartlett’s sphericity test provided a strongly significant level $p$, so we confirmed the existence of patterned relationships. Also, the Kaiser-Meyer-Olkin measure (KMO) of sampling adequacy revealed that the data is appropriate to apply an EFA.

In Table 1 is presented the total variance associated to the eigenvalues accounted by each factor by descendent order. The $i^{th}$ line corresponds to cumulative variance percentage explained by the first $i$ factors after extraction and after rotation. Notice that Table 1 also contains information about rotated component matrix and component score component matrix.

If we consider the Kaiser criterion for simplicity, we retain the first 4 factors (eigenvalues great or equal to one). Other criteria may be applied, for example using the scree plot or using the average of extracted communalities to determine the eigenvalue cut-off. The varimax algorithm which produces orthogonal factors was applied after the extraction process. This technique is adequate when we want to identify variables to create indexes or new variables without intercorrelated components. In the present case, we could get the 'meaning' of each factor. Namely each factor is related with: $F_1$- How took knowledge of admission contest; $F_2$- Economic level of family; $F_3$- Previous attendance of higher education; $F_4$- Combines distance of family residence and knowledge of military environment.

### Table 1: Total variance accounted for each factor and component matrix details.

#### Rotated Component Matrix

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>.105</td>
<td>-.278</td>
<td>.755</td>
<td></td>
</tr>
<tr>
<td>Grade</td>
<td>.368</td>
<td>.529</td>
<td>-.344</td>
<td></td>
</tr>
<tr>
<td>Attendance</td>
<td>-.120</td>
<td>.625</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge Internet</td>
<td>.925</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge Family</td>
<td>-.903</td>
<td>.251</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge Visit</td>
<td>-.617</td>
<td>-.399</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apply Higher</td>
<td>.135</td>
<td>.682</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MotherFather Employed</td>
<td>.258</td>
<td>.763</td>
<td>-.125</td>
<td></td>
</tr>
<tr>
<td>MotherFather Militar</td>
<td>.287</td>
<td>.755</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Component Score Coefficient Matrix

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>.030</td>
<td>.041</td>
<td>.203</td>
<td>.203</td>
</tr>
<tr>
<td>Grade</td>
<td>.226</td>
<td>.448</td>
<td>-.299</td>
<td>-.026</td>
</tr>
<tr>
<td>Attendance</td>
<td>.054</td>
<td>-.065</td>
<td>.496</td>
<td>.073</td>
</tr>
<tr>
<td>Knowledge Internet</td>
<td>.486</td>
<td>.089</td>
<td>.076</td>
<td>.044</td>
</tr>
<tr>
<td>Knowledge Family</td>
<td>.455</td>
<td>.133</td>
<td>.038</td>
<td>.045</td>
</tr>
<tr>
<td>Knowledge Visit</td>
<td>-.019</td>
<td>-.448</td>
<td>-.280</td>
<td>-.081</td>
</tr>
<tr>
<td>Apply Higher</td>
<td>.075</td>
<td>.032</td>
<td>.520</td>
<td>.061</td>
</tr>
<tr>
<td>MotherFather Employed</td>
<td>.081</td>
<td>.571</td>
<td>-.094</td>
<td>.082</td>
</tr>
<tr>
<td>MotherFather Militar</td>
<td>.114</td>
<td>.047</td>
<td>.222</td>
<td>.651</td>
</tr>
</tbody>
</table>

### 6 Conclusion

In present work it is performed a statistical analysis of individuals from the beginning to the end of the course, to identify the characteristics (denominated variables) of individuals with higher and lower probability to contribute to academic success at the Naval
Academy. The data collection corresponds to the time interval 2007 – 2011, it is not complete, more “5 years” data set shall be considered to be analyzed statistically so we can get a more detailed analysis.

This sample can be considered to get some results as an alert to the need and value of ongoing study to a larger horizon.

By first it was performed a descriptive analysis and applied an analysis of variance [1, 11].

In a second stage it is on going the analysis of results from logistic model approach simultaneously with the results from Factorial Analysis.

The statistical approach, estimating a model with relevant information using more elaborate techniques such as generalized linear models was performed. The model improvement and validation is still going on. The statistical process was explained, but not detailed. It is still going on the naming of selected factors, the obtained results and factors scores will be described in detail on an extended version of this article. relationships between socio-demographic variables and success are being evaluated.

Some important issues about academic success in NA were obtained, reinforcing the importance and need of prevention measures implementation. For improved success and quality of performance of the candidates an early diagnosis is necessary and an early and appropriate intervention. Prevention is possible by controlling known and modifiable risk factors.

Prevention requires a higher degree of awareness and possibly the implementation of awareness campaigns promoting regular assessment of the individual performance.

Acknowledgements: This work was supported by Portuguese funds through the Center of Naval Research (CINAV), Naval Academy, Portuguese Navy, Portugal and the Center for Computational and Stochastic Mathematics (CEMAT), The Portuguese Foundation for Science and Technology (FCT), University of Lisbon, Portugal, project UID/Multi/04621/2019.

References:


