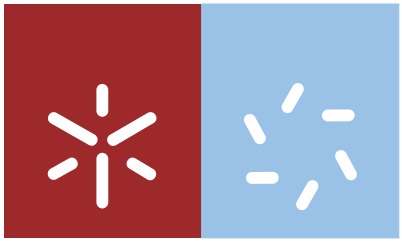
Universidade do Minho
Escola de Ciências

Irina Kislaya

Statistical multivariate approaches for identification of predictors of academic failure for first year students in medical school.

Janeiro de 2012

Statistical multivariate approaches for identification of predictors of academic failure for first year students in medical school.
Irina Kislaya
UMinho|2012



Universidade do Minho
Escola de Ciências

Irina Kislaya

Statistical multivariate approaches for identification of predictors of academic failure for first year students in medical school.

Relatório de Mestrado
Mestrado em Estatística de Sistemas - Ramo Engenharia e Estatística

Trabalho realizado sob a orientação da
Professora Doutora Maria Conceição Serra

Janeiro de 2012

DECLARAÇÃO

Nome Irina Kislaya

Endereço electrónico: irina_teterina@hotmail.com Telefone: 934500028

Número do Bilhete de Identidade: 30862183

Título dissertação □/tese □

Statistical multivariate approaches for identification of predictors of academic failure for first year students in a medical school

Orientador(es):

Professora Doutora Maria Conceição Serra

Ano de conclusão: 2012

Designação do Mestrado ou do Ramo de Conhecimento do Doutoramento: Mestrado em Estatística de Sistemas - Ramo Engenharia e Estatística

Universidade do Minho, ___/___/_____

Assinatura: _____

Acknowledgments

First, I would like to thank my supervisor, Professor Maria Conceição Serra, for encouragement, scientific support, advise, involvement and dedication during this work.

This work was carried out under the Project: "Evaluating the impact of innovation in Higher Education: implementation and development of a longitudinal study in a medical school", supported by the grant of FCT, the Portuguese Foundation for Science and Technology (Project grant reference: $PTDC/ESC/65116/2006$).

I would like to acknowledge Professor Manuel João Costa from the School of Health Sciences of the University of Minho, the coordinator of the Research Project $PTDC/ESC/65116/2006$, who also act as co-supervisor of this work, for receiving me there, for interesting discussions and support and for the opportunity to work in excellent research team.

I am grateful to Professor Miguel Portela from School of Economics and Management, University of Minho for providing me support and advices with statistical software and data management.

I would like to thank also all the support of my close friends and family. Thanks for patience and for all the encouragement.

Abstract

Statistical multivariate approaches for identification of predictors of academic failure for first year students in a medical school

Academic failure is a frequent phenomena in medical education, with huge impact on the students and on the medical schools. Understanding of factors that have the strongest influence on the probability of failure is extremely important to implement procedures for timely assistance and support to students in difficulties.

The main goal of this work was to gain information (factors) that can be useful in predicting, early in the first year of the medical program, academic failure for the students of the School of Health Sciences of Minho University. To determine which factors influence the performance of these students in the 1st year course with the highest failure rates, administrative data of three cohorts of students were analyzed using several multivariate statistical tools, namely: logistic regression, linear discriminant analysis and K-nearest neighbors discriminant analysis.

The results obtained in this study provide substantial empirical evidence that combination of cognitive and non-cognitive characteristics (personality trait "Conscientiousness", change of residence at entry, anticipation of difficulties due to enrollment in a medical program) and of academic achievements in the initial university courses can be useful in the early detection of failure.

Keywords: Logistic regression, Discriminant analysis, Medical students, Academic performance, Predict.

Resumo

Identificação dos preditores de insucesso academico no primeiro ano curricular do curso de medicina: abordagem estatística multivariada.

O insucesso escolar é um fenómeno bastante frequente na educação médica, que tem um enorme impacto sobre os alunos e sobre as escolas de medicina. A compreensão dos factores que mais influenciam a probabilidade de insucesso é extremamente importante para que as escolas possam implementar, em tempo útil, procedimentos de assistência e apoio aos alunos em dificuldades.

O objetivo principal deste trabalho consistia em obter informação (factores) que pudesse ser útil para prever, logo no início do primeiro ano do curso de medicina, o insucesso escolar dos alunos da Escola de Ciências da Saúde da Universidade do Minho. Para identificar os fatores que influenciam o desempenho desses alunos na disciplina do 1 º ano do curso que tem as maiores taxas de reprovação, dados administrativos relativos a três cohortes de estudantes foram analisados usando várias técnicas de estatística multivariada, nomeadamente: regressão logística, análise discriminante linear e análise dos K-vizinhos mais próximos.

Os resultados obtidos neste estudo fornecem evidência empírica de que a combinação de alguns fatores cognitivos e não-cognitivos (a característica pessoal "Conscientiousness, a alteração de residência com a entrada na universidade, a antecipação de dificuldades relacionadas com a frequência de um curso de Medicina) e dos resultados obtidos nas disciplinas iniciais do curso, pode ser útil para a detecção precoce do insucesso escolar.

Palavras-chave: Regressão logística, Análise discriminante, Estudante de medicina, Desempenho académico, Previsão.

Contents

List of Figures

List of Tables

List of abbreviations

AIC	Akaike information criterion
AUC	Area Under Receiver Operating Characteristic curve
BIC	Bayesian information criterion
df	degrees of freedom
ECTS	European Credit Transfer and Accumulation System
FA	First Aid training
FOS1	Functional and Organic Systems 1
GLM	Generalized Linear Model
IMD	Introduction to the Medical Degree
K-NN	K-Nearest Neighbors discriminant analysis
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MC	Molecules and Cells
$N(\mu, \sigma)$	Normal probability distribution with mean value μ and standard deviation σ
OP1	Optional Project 1
ROC curve	Receiver Operating Characteristic curve
SHS-UM	School of Health Sciences of the University of Minho
THC	Training in a Primary Care Unit
UK	United Kingdom
USA	United States of America
$\chi^2(df)$	Chi-squared distribution with df degrees of freedom

1 Introduction

Academic failure is a frequent phenomena in medical education, with huge impact both on students and medical school. For students, failure has high financial and emotional costs and is a cause of personal distress [60]. For the medical school, it is an issue of organizational, financial and academic accountability. Students in difficulty demand more attention, more time and more resources. Therefore, there are potential benefits for both students and institutions on understanding of determinant of academic failure.

The first year in medical school is the most challenging period for students [50, 54]. It is a time of transition, which demands several simultaneous organizational and social adjustments. Suddenly, students enter an unknown and more competitive system. The amount and the complexity of study materials increases sharply, academic standards and teaching methods change remarkably.

The particular difficulties posed by the first year in medical school result in high rates of failure and dropout [2]. Furthermore, students who start fail in first year courses may continue to struggle during the degree and may become poor doctors. Issues related to professional misconduct of practicing physicians can be traced back to behavioral concerns while early years in medical school [49, 68]. The existence of a model that can help the medical school in the identification of students at risk of failure during their first year, could anticipate supporting interventions to decrease the probability of subsequent failures.

The primary motivation for this work derived from an institutional need to gain insight on the factors behind stories of failure in the medical school. Since 2006, the School of Health Sciences of the University of Minho (SHS-UM) develops the longitudinal research project "Evaluating the impact of innovation in Higher Education: implementation and development of a longitudinal study in a medical school" (FCT- PTDC/ESC/65116/2006), with the main goal of understanding the factors that influence performance of medical students and professional competence of SHS-UM graduates.

In the years of existence of the undergraduate medical program at SHS-UM, it became clear that the course with the highest percentage of failures is "Functional and Organic Systems 1" (FOS 1) that is taught in the second part of the first year. Furthermore, failure in FOS1 results in a great cost to students, since success in this course is essential for success in subsequent courses. Hence, in this work, we use the pass/fail score in FOS1 course as an indication of success/failure in the medical school.

The main goal of this study was to develop predictive models for the prospective identification of students at risk of failure in the first year of medical degree, combining variables collected at admission (entry grade point average (GPA), socio-

demographic factors, information of administrative nature, personality traits, expected difficulties during the degree) with the performance on courses taken in the first seventeen weeks in medical school. The underlying goals were to learn how to identify, at admission or very early in the medical school, students at risk of academic difficulties and thus open new opportunities for timely assistance and support.

Traditionally, research problems involving prediction of a dichotomous outcome are addressed either by logistic regression or by linear discriminant analysis.

In medical education, logistic regression is widely used and a popular approach to model binary outcome variables, such as "academic failure" or dropout. For example, Yates and James [67] presented the results of the study of the risk factors for poor performance at different stages of the undergraduate medical course at Nottingham University Medical School. Cleland et al [10] used logistic regression to determine whether poor performance in degree assessments early in medical school is a risk factor for poor performance in later examinations. Arulampalam et al [2] applied logistic regression to analyze the determinants of the probability of dropout, among first year medical students, in the context of changing admissions criteria in the UK.

The other popular methodology for classification problems and for exploring factors that might explain the differences between groups is the discriminant analysis. Literature provides several examples of application of this methodology in the field of educational sciences. For instance, Beeman [4] employed linear discriminant analysis to examine the differences among nursing graduates who failed and who have passed on the national certification license examinations. Vandamme et al [64] presented an example of successful application of linear discriminant analysis to identify first year students at risk of poor academic performance. Morgan [47] used discriminant analysis to determine how cognitive and non-cognitive characteristics of students are related to university attrition.

Both parametric techniques, logistic regression and linear discriminant analysis, depend on strict statistical assumptions. These assumptions include linearity of relationships, lack of multicollinearity among independent variables [26, 32, 44, 63], normality of independent variables, equality of covariance matrices for discriminant analysis [28, 31, 55]. Hence, non-parametric methods of classification appear as an attractive alternative in educational research since they do not require any strict distributional assumptions and can handle easily discrete and mixed data, which is quite common in the field.

To determine which factors influence academic performance of medical students in the FOS1, we used several statistical techniques, namely: multivariate logistic regression, linear discriminant analysis and non-parametric K-nearest neighbors discriminant analisys. We compared the classification results derived from the

application of the three multivariate approaches with data provided by longitudinal research project (FCT- PTDC/ESC/65116/2006).

The remainder parts of this dissertation are structured as follows:

- Section 2 discusses the logistic regression model: the estimation and the interpretation of the parameters, assessment of model fit and classification accuracy;
- Section 3 reviews the theoretical background of discriminant analysis and discusses the non-parametric approach to the classification problem;
- Section 4 discusses the methods of comparison of classification rules;
- Section 5 presents the study settings, the data, displays and interprets the results of practical application of the three classification techniques;
- Section 6 summarizes the conclusions and presents suggestions for future research.

2 Logistic regression

Logistic regression is a statistical tool well suited to describe the relationship between a binary response variable and one or more predictor variables. Predictor variables can be of any type: continuous, discrete, qualitative ordinal or dichotomous.

From a theoretical point of view, logistic regression has been intensively studied during the last decades [26, 32, 41, 44]. Throughout the literature it is possible to find a wide range of examples of its application in behavioral and educational research. It is the most popular statistical method used in medical education research on the issue of undergraduate student performance [10, 36, 67, 68, 69].

There are several alternative ways to introduce the logistic regression model. In this work we consider logistic regression in the framework of generalized linear models (GLM) ..

2.1 Logistic regression model

Before describing the form of a generalized linear model we introduce the exponential family of statistical distributions, which is crucial to understand GLM.
Definition 2.1 (Exponential family of distributions) *A random variable W is said to belong to the exponential family of distributions if the corresponding probability function can be expressed by*

$$f(w;\theta,\phi) = exp\left\{\frac{y\,\theta - b(\theta)}{a(\phi)} + c(w;\phi)\right\}, \quad w \in S_W, \tag{2.1}$$

where S_W is the support of W, a, b and c are some specific functions, and θ and ϕ are some real parameters. θ is known as the canonical parameter of the distribution and ϕ is known as the dispersion parameter.

Many well-known distributions belong to the exponential family; for example, Poisson, Gamma and Normal distribution. We can easily prove that the Bernoulli distribution, a particular case of the Binomial distribution, belongs to the exponential family.
Example 2.2 *Bernoulli distribution as a member of exponential family*
Consider a random variable $Y \sim Bernoulli(\pi)$, with probability mass function given by

$$f(y;\pi) = \pi^y(1-\pi)^{1-y}, \quad y \in \{0,1\}. \tag{2.2}$$

(2.2) *can be written as:*

$$f(y;\pi) = exp\left[ln\left(\pi^y(1-\pi)^{1-y}\right)\right] = exp\left[y\,ln\left(\frac{\pi}{1-\pi}\right) + ln(1-\pi)\right] \tag{2.3}$$

Comparing (2.3) *with* (2.1) *we observe that the canonical parameter is*

$$\theta = ln\left(\frac{\pi}{1-\pi}\right), \tag{2.4}$$

the so-called logit of π*. From* (2.4) *we have*

$$\pi = \frac{exp(\theta)}{1+exp(\theta)} \qquad and \qquad 1-\pi = \frac{1}{1+exp(\theta)}.$$

Rewriting the term $ln(1-\pi)$ *in* (2.3) *as a function of* θ *we get*

$$b(\theta) = ln\left(1+exp(\theta)\right), \qquad a(\phi) = 1, \qquad c(y,\phi) = 0.$$

For members of the exponential family, the expected value and the variance are given by

$$\mathbb{E}(W) = \frac{\partial b(\theta)}{\partial \theta} = b'(\theta), \qquad \mathrm{Var}\,(W) = \frac{\partial^2 b(\theta)}{\partial \theta^2} a(\phi) \tag{2.5}$$

We can easily verify these results for the case of the Bernoulli distribution. In fact, the expected value of the Bernoulli distribution is

$$\mathbb{E}(Y) = b'(\theta) = \left[ln\left(1+exp(\theta)\right)\right]' = \frac{exp(\theta)}{1+exp(\theta)} = \pi$$

and the variance of the Bernoulli distribution is

$$\mathrm{Var}\,(Y) = \frac{\partial^2 b(\theta)}{\partial \theta^2} a(\phi) = \left[\frac{exp(\theta)}{1+exp(\theta)}\right]' \cdot 1 = \frac{exp(\theta)}{(1+exp(\theta))^2} = \pi(1-\pi).$$

Consider a response variable Y, a line vector of p explanatory variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_j, \ldots, X_p)$ and n independent realizations of pair $(Y, \boldsymbol{X})$, $(y_i, \boldsymbol{x}_i)$ for $i = 1, 2, \ldots, n$. Using matrix notation, we have

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{x} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1j} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2j} & \ldots & x_{2p} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ x_{i1} & x_{i2} & \ldots & x_{ij} & \ldots & x_{ip} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nj} & \ldots & x_{np} \end{pmatrix} \tag{2.6}$$

The generalized linear model consists of three elements: the random component, the systematic component and the link function [1]. The random component

of GLM identifies the conditional distribution of the response variable Y_i, given the observed vector of explanatory variables $\boldsymbol{x}_i$. Traditionally, the random component belongs to the exponential family of distributions, defined in (2.1), with

$$\mathbb{E}(Y_i|\boldsymbol{x}_i) = \mu_i = b'(\theta_i), \quad for \quad i = 1, 2, \ldots, n.$$

The systematic component of GLM, also called linear predictor, specifies a linear combination of the explanatory variables, also called covariates or regressors, used in the model

$$\eta_i = \boldsymbol{z}_i\boldsymbol{\beta}, \qquad i = 1, 2, \ldots, n,$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_p]^T$ is a vector of k parameters, with $k = p + 1$, and $\boldsymbol{z}_i$ is a function of the covariates $\boldsymbol{x}_i$. In general, $\boldsymbol{z}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ip})$ is the i-th line of the following matrix

$$\boldsymbol{Z} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1j} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2j} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ 1 & x_{i1} & x_{i2} & \ldots & x_{ij} & \ldots & x_{ip} \\ \vdots & \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nj} & \ldots & x_{np} \end{pmatrix}, \tag{2.7}$$

the so-called design matrix.

The expected value, μ_i, of the random component and the systematic component, η_i, are related by a function h, i.e.

$$\mu_i = h(\eta_i) = h(\boldsymbol{z}_i\boldsymbol{\beta}), \quad i = 1, 2, \ldots, n,$$

where h is a monotonic and differentiable function. Taking $g = h^{-1}$ we get

$$g(\mu_i) = \boldsymbol{z}_i\boldsymbol{\beta}.$$

Function g is known as a link function.

There are many possible choices for the link function. The simplest one is the identity function ($g(\mu) = \mu$), that specifies a linear model for the mean response. When the link function transforms the expected value of the random component into the canonical parameter of the exponential family member, θ, it is designated by canonical link function. For example, the identity function is the canonical link for the Normal distribution.

For $Y \sim Bernoulli(\pi)$, we have $0 \leq \mathbb{E}(Y|\boldsymbol{x}) = \pi \leq 1$, and the link function should be one that maps interval $[0, 1]$ into the whole $\mathbb{R}$. The logit function, $logit(t) = ln\left(\frac{t}{1-t}\right)$ is, therefore, suitable for modeling Bernoulli data.

Performing the logit transformation on π we obtain

$$g(\pi) = logit(\pi) = ln\left(\frac{\pi}{1-\pi}\right) = \theta,$$

thus, the logit function is the canonical link for the Bernoulli distribution.

Consider n independent realizations of a random variable $Y_i \sim Bernoulli(\pi_i)$, $i = 1, 2, \ldots, n$ and line vector $\boldsymbol{z}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ip})$ resultant from vector of covariates $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$. The logistic model relates the value of π_i (the probability that the outcome of interest occurs) with the set of explanatory variables. The formulation of the model is the following

$$\mathbb{P}(Y_i = 1|\boldsymbol{X} = \boldsymbol{x}_i) \equiv \pi(\boldsymbol{x}_i) = \frac{exp(\boldsymbol{z}_i\boldsymbol{\beta})}{1 + exp(\boldsymbol{z}_i\boldsymbol{\beta})}. \tag{2.8}$$

A logistic regression model is a generalized linear model, since:

- the realizations of outcome variables are independent,
- the Bernoulli distribution belongs to the exponential family,
- the expected value $\mu_i = \pi(\boldsymbol{x}_i)$ is related to the linear predictor $\eta_i = \boldsymbol{z}_i\boldsymbol{\beta}$ by

$$\pi(\boldsymbol{x}_i) = \frac{exp(\boldsymbol{z}_i\boldsymbol{\beta})}{1 + exp(\boldsymbol{z}_i\boldsymbol{\beta})},$$

with the link function being the logit function.

2.2 Fitting logistic regression model

To fit a logistic regression (LR) LRLogistic regression model to the data we need to estimate the set of parameters in the linear predictor $\eta_i = \boldsymbol{z}_i\boldsymbol{\beta}$. To estimate unknown parameters the maximum likelihood method is used.

Lets consider the most simple case of LR model, the univariate model. In fitting univariate LR model to the given data two unknown parameters β_0 and β_1 need to be estimated. Denote the vector of parameters by $\boldsymbol{\beta} = [\beta_0, \beta_1]^T$. Observe that in an univariate case the design matrix $\boldsymbol{Z}$ defined in (2.7) is given by

$$\boldsymbol{Z} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix}.$$

To simplify a notation, from now we denote $\pi(\boldsymbol{x}_i)$ by π_i.

Since $Y_i \sim Bernoulli(\pi_i)$, for $i = 1, 2, \ldots, n$, the probability mass function of Y_i is given by

$$f(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \quad i = 1, 2, \ldots, n.$$

and the likelihood function is

$$\mathcal{L}(\boldsymbol{\beta}; y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

The estimation of $\boldsymbol{\beta}$'s requires the maximization of the likelihood function or, equivalentely, the maximization of the natural logarithm of the likelihood function (log-likelihood):

$$\begin{aligned} L(\boldsymbol{\beta}; \boldsymbol{y}) = ln\left(\mathcal{L}(\boldsymbol{\beta}; y_1, y_2, \ldots, y_n)\right) &= ln\left(\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}\right) \\ &= \sum_{i=1}^{n} \left[y_i\, ln(\pi_i) + (1-y_i)\, ln(1-\pi_i)\right] \\ &= \sum_{i=1}^{n} \left[y_i\, ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} ln(1-\pi_i) \end{aligned} \tag{2.9}$$

Since $(1-\pi_i) = [1 + exp(\boldsymbol{z}_i\boldsymbol{\beta})]^{-1}$ and $\eta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{z}_i\boldsymbol{\beta}$, the log-likelihood given in (2.9) can be written as

$$\begin{aligned} L(\boldsymbol{\beta}; \boldsymbol{y}) &= \sum_{i=1}^{n} y_i \boldsymbol{z}_i\boldsymbol{\beta} - \sum_{i=1}^{n} ln[1 + exp(\boldsymbol{z}_i\boldsymbol{\beta})] \\ &= \boldsymbol{\beta}^T\boldsymbol{Z}^T\boldsymbol{y} - \sum_{i=1}^{n} ln[1 + exp(\boldsymbol{z}_i\boldsymbol{\beta})]. \end{aligned} \tag{2.10}$$

To find the value of $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta}; \boldsymbol{y})$ we have now to differentiate (2.10) with respect to β_0 and β_1 and set the two resulting expressions to zero

$$\begin{cases} \dfrac{\partial L}{\partial \beta_0} = 0 \\ \\ \dfrac{\partial L}{\partial \beta_1} = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} (y_i - \pi_i) = 0 \\ \\ \displaystyle\sum_{i=1}^{n} x_{i1} (y_i - \pi_i) = 0. \end{cases} \tag{2.11}$$

This system of equations must be solved by mean of iterative computing methods, since the likelihood equations are non-linear in β_0 and β_1. The solutions of system of equations (2.11), denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, are called the maximum likelihood estimators of β_0 and β_1, respectively. Hence, the likelihood estimator of the conditional probability that the event of interest occurs, denoted by $\hat{\pi}_i$, is given by

$$\hat{\pi}_i = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1})}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1})},$$

and the estimated logit by

$$\hat{g}(\boldsymbol{z}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}.$$

In fitting multiple LR with p explanatory variables the same method of estimation is employed. The log-likelihood function used to obtain the estimators of vector of parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$, is almost identical to (2.10). To obtain the maximum likelihood estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ we must solve the following system of k likelihood equations

$$\begin{cases} \displaystyle\sum_{i=1}^{n} (y_i - \pi_i) = 0 \\ \displaystyle\sum_{i=1}^{n} x_{ij} (y_i - \pi_i) = 0, \qquad \text{for} \quad j = 1, 2, \ldots, p. \end{cases}$$

Like in univariate case, the estimated value of the conditional probability that the event of interest occurs is

$$\hat{\pi}_i = \frac{exp(\hat{g}(\boldsymbol{z}_i))}{1 + exp(\hat{g}(\boldsymbol{z}_i))},$$

where $\hat{g}(\boldsymbol{z}_i)$ is the estimated logit, i.e.

$$\hat{g}(\boldsymbol{z}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}.$$

The estimation of standard errors of estimated coefficients involves the matrix of second partial derivatives of the log-likelihood function; this matrix is known as the observed information matrix and is usually denoted by $\boldsymbol{I}(\boldsymbol{\beta})$. If the model assumptions are correct, it can be shown that $\text{Var}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta}) = (\boldsymbol{Z}^T \hat{\boldsymbol{V}} \boldsymbol{Z})^{-1}$,

where $\boldsymbol{Z}$ is the n by k design matrix defined in (2.7) and $\hat{\boldsymbol{V}}$ is the diagonal matrix, of dimension n, with general element $\pi_i(1-\pi_i)$ on the main diagonal, i.e.

$$\hat{\boldsymbol{V}} = \begin{pmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & \dots & \hat{\pi}_n(1-\hat{\pi}_n) \end{pmatrix}.$$

The estimated standard error of the j-th estimated logistic regression coefficient β_j, associated with explanatory variable x_j for $j = 1, 2, \dots, p$, is given by

$$\hat{SE}(\hat{\beta}_j) = \left[\hat{\text{Var}}\,(\hat{\beta}_j)\right]^{\frac{1}{2}}, \tag{2.12}$$

where $\hat{\text{Var}}\,(\hat{\beta}_j)$ is the j-th diagonal element of the matrix $\hat{\boldsymbol{I}}^{-1}(\boldsymbol{\beta})$.

2.3 Inference for logistic regression

2.3.1 Test for significance of the model

In this subsection we present a brief description of statistical tests for assessment of overall statistical significance of fitted LR models as well as of individual coefficients estimates.

Several methods can be employed to test whether explanatory variables in the model are significantly related to the outcome. In this work we focus on two tests for assessment of the statistical significance of the model and of individual coefficients estimates: the likelihood ratio and the Wald test.

Likelihood ratio test compares the fitted model to a saturated model, the last being a model with as many parameters as subjects. The comparison is based on the likelihood function and indicates how accurately the fitted model represents the data. The test statistic, D, also known as deviance, is defined by following expression

$$D = -2\,ln\left[\frac{\mathcal{L}(fitted)}{\mathcal{L}(saturated)}\right] = -2\sum_{i=1}^{n}\left[y_i\,ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\,ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right]. \tag{2.13}$$

To determine statistical significance of a single explanatory variable, we analyze the difference between the LR model fitted with and without the variable being tested. Suppose that the model MR is obtained from a full model, MU, by setting the restriction that the coefficient estimate for the predictor x_j is equal to zero. In such a situation model MR is called constrained, or restricted, and is said to be nested

in the unrestricted model MU. Thus, we have two models

$$MU : logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{j-1} x_{j-1} + \beta_j x_j + \beta_{j+1} x_{j+1} + \ldots + \beta_p x_p,$$
$$MR : logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{j-1} x_{j-1} + \ldots + \beta_{j+1} x_{j+1} + \ldots + \beta_p x_p.$$

The null hypothesis of the test is that the constrained model (MR) is nested in a unconstrained model (MU), or equivalently,

$$H_0 : \beta_j = 0 | \, \beta_1, \beta_2, \ldots, \beta_p. \tag{2.14}$$

The alternative hypothesis of the test is $H_1 : \beta_j \neq 0 | \, \beta_1, \beta_2, \ldots, \beta_p$. The likelihood ratio test statistics named G is given by

$$G = -2(L_{MR} - L_{MU}), \tag{2.15}$$

where L_{MU} is a value of the log-likelihood associated with the unconstrained model and L_{MR} is a value of the log-likelihood for the constrained model (without predictor x_j). Under the null hypothesis the test statistic $G \sim \chi^2(k_1)$ where k_1 equals to the difference between the degrees of freedom of the unconstrained model and the degrees of freedom of the constrained model. Under the null hypothesis, i.e. that $\beta_j = 0$, likelihood ratio test statistics has a $\chi^2(1)$ distribution.

Likelihood ratio test can be employed to compare any pair of nested models: for instance, it can be used to test whether several coefficients are simultaneously equal to zero.

Logistic regression analysis often includes nominal categorical predictors, such as gender, ethnicity, citizenship. Such variables are introduced in the model by a set of dummy variables, coded either 0 or 1. In general, to represent categorical predictor with m levels it is required that $m - 1$ dummy variables are used. The category of predictor coded by 0 in all dummy variables is called reference category. If we wish to assess the statistical significance of this categorical predictor all $m - 1$ dummies should be removed from the model simultaneously. In such case the likelihood ratio test statistics follows a $\chi^2(m - 1)$ distribution.

The other method to assess the statistical significance of individual coefficients estimates is the Wald test. For a single predictor, the test statistic is given by

$$Z_{W_j} = \frac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)}. \tag{2.16}$$

Under the null hypothesis, provided in 2.14, $Z_{W_j} \sim N(0, 1)$ distribution. We should note that some statistical software packages for fitting logistic regression compute square of the Z_W statistic and compare it to the $\chi^2(1)$ distribution.

2.3.2 Confidence interval estimation

Confidence intervals for the parameters of interest may be more informative than significance tests. An approximate $100 \cdot (1-\alpha)\%$ Wald confidence interval for the j-th coefficient, β_j, is

$$]\hat{\beta}_j - z_{1-\alpha/2}\,\hat{SE}(\hat{\beta}_j), \quad \hat{\beta}_j + z_{1-\alpha/2}\,\hat{SE}(\hat{\beta}_j)[,$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ percentile of the $N(0,1)$ distribution, $\hat{SE}(\hat{\beta}_j)$ denotes the standard error of $\hat{\beta}_j$, given in (2.12).

2.4 Building logistic regression model

This subsection is dedicated to the model building procedure: selection of variables and verification of model assumptions.

2.4.1 Variable selection strategies

An important problem in statistical model building is the choice of an optimal model, that includes a small number of parameters and still has adequate fit and prediction accuracy. We present a brief review of methods for variable selection in LR modeling.

Stepwise is a standard method for variable selection and is based on the procedure of sequential inclusion of predictors into the model, one at a time. This approach has several forms: the main ones are forward selection and backward elimination. The forward selection starts with an empty model and adds predictors, one at a time. In contrast, the backward elimination begins with the full model and successively removes predictors, one at a time, based on specified statistical rule, composed of a combination of selection criteria and stopping criteria. Selection criterion refers to a test statistic (usually likelihood ratio and/or Wald test) used for assessment of statistical significance of the coefficients of explanatory variables. Stopping criterion refers to the p-value, i.e. level of statistical significance, of the coefficients of explanatory variables.

Almost all statistical software packages allow to perform stepwise selection via two basic algorithm or their combination: for instance, stepwise method that starts with forward selection but, at each step, as in the backward elimination, gives the possibility to remove from the model no longer important variables.

The choice of selection and stopping criteria is of crucial importance in stepwise procedure. Hosmer & Lemeshow [26] recommend to use a combination of forward selection and backward elimination. Regarding the selection criterion, the authors give no preference to any particular statistical test, but they recommend p-value

in interval $[0.15, 0.25]$. Wang et al [66] highly recommend to use likelihood ratio test in the stepwise procedure, claiming that it has better statistical properties than the Wald test. For the choice of stopping criterion in stepwise algorithm, literature provides the following guidelines: in exploratory research a conventional p-value of 0.05 may be too small and fail to identify important variables [26, 37, 66]. For the forward selection method, Lee & Koval [37] show, via simulation study, that the overall best p-value varies from 0.15 to 0.20. For the backward elimination, Wang et al [66] recommend p-value in interval $[0.2, 0.4]$ and emphasize that the choice of optimal level of statistical significance depends on the number of potential predictors: for forward selection, it should increase with the number of predictors; for backward elimination the optimal p-value should decrease with the number of predictors.

The other popular variable selection strategy is the so-called best subset selection. According to this strategy the best model, among all possible models, is chosen based on the values of a specified information criterion that penalizes model complexity. The method requires extensive and time consuming computations, because with set of p predictors it is possible to fit 2^p distinct main-effects models. If interactions are considered, the number of possible models becomes very large. Which subset of variables is considered to be the best depends on the information criterion specified. Two commonly used in practice criteria for model selection are Akaike information criterion (AIC) and Bayesian information criterion (BIC). In the general case, AIC is given by

$$AIC = -2\,ln(\mathcal{L}) + 2\,k,$$

and BIC is given by

$$BIC = -2\,ln(\mathcal{L}) + 2\,k\,ln(n),$$

where $\mathcal{L}$ is the model likelihood, $k = p + 1$ is the number of model parameters and n is the number of observations in the sample. The information criteria have the following interpretation:

- $-2\,ln(\mathcal{L})$ is a measure of the lack-of-fit of the chosen model;
- $2\,k$ and $2\,k\,ln(n)$ are penalty terms for model complexity for AIC and BIC, respectively.

Penalty terms increase with the number of estimated parameters.

Given the data, different LR models can be fitted and ranked according to the value of information criterion. The preferred model is the one which achieves the minimum value of information criterion. For more details on the best subset selection algorithm for the class of GLM, see Calcagno & Mazancourt [6].

An alternative approach to standard computer algorithms described above,

variable selection for multiple LR model can be based on univariate analysis. Agresti [1] and Hosmer & Lemeshow [26] recommend, for discrete explanatory variables, to study marginal effects of each potential predictor on the outcome variable, via analysis of contingency tables. To investigate effects of continuous predictors on the outcome, several methods can be employed, namely: graphical analysis, univariate logistic regression modeling, two sample t-test or its non-parametric analog, Mann-Whitney-Wilcoxon test. The level of statistical significance in univariate analysis for inclusion of variables into multiple models should be large enough be able to identify important variables. Hosmer & Lemeshow [26] recommend $p - value < 0.25$.

The criterion for variable selection depends on the problem and on the design of the study. In comparison with stepwise selection, the best subset method has advantage, since allows to compare both non-nested and nested models, but, on the other hand, best subset selection is more time consuming. Both stepwise method and best subset method provide model selection based solely on statistical criteria. Yet, as a part of model building procedure, practical or "clinical" importance of selected variables should be carefully analyzed and taken into account. Variables of special interest may be included in a model even if their estimated effects are not statistically significant at 0.05 level [1].

2.4.2 Verification of logistic regression assumptions

In order to be able to fit a LR model, the following conditions should be satisfied to guarantee valid statistical inference:

- linearity in the logit;
- independence of errors.

LR assumes the existence of a linear relation between any continuous predictor and the logit of the probabilities of the positive outcome of variable Y, meaning that the change in the value of the logit associated with a unit change in the independent variable equals the coefficient in the regression equation.

The simplest way to detect nonlinearity is a graphical analysis. Several scatter plots can be used check the linearity assumption: see, for instance, Kohler & Kreuter [33], Hosmer & Lemeshow [26].

We will focus our attention on two analytical method for verification of the linearity assumption. The first is the so-called Box-Tidwell procedure, that consists in adding into the model the interaction term between the continuous predictor and its logarithmic transformation. If the coefficient for the interaction term is statistically significant, nonlinearity in the relationship between the logit and predictor can be suspected [33, 44, 63].

The other well theoretically established method to examine linearity is fractional polynomial analysis. The technique can be applied to any GLM, in particular

to logistic regression, and it was developed originally by Royston & Altman [57] and extended by Royston & Ambler & Sauerbrei [59] to multivariate models. The method consists of adjusting a number of logit models with power transformations of continuous predictors.

Definition 2.3 (Fractional polynomial of degree m) *For arbitrary set of powers S and single covariate x, a fractional polynomial of degree m is defined as*

$$t(x;\beta,s) = \beta_0 + \sum_{j=1}^{m} \beta_j F_j(x),$$

where for $j = 1, 2, \ldots, m$

$$F_j(x) = \begin{cases} x^{s_j}, & \text{if } s_j \neq s_{j-1}; \\ \\ F_{j-1}(x)ln(x), & \text{if } s_j = s_{j-1} \end{cases}$$

Altman & Royston [57] suggest to consider two particular families of fractional polynomial models: first-order models ($m = 1$) and second-order models ($m = 2$) with powers s selected among the values in the set $S = \{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ where 0 corresponds to a logarithmic transformation of covariate.

To determine a correct functional form of continuous predictor using fractional polynomials approach on the first step the first-order fractional polynomial models are fitted. It is possible to adjust 8 different models. The best is one with the highest value of log-likelihood or equivalently with the lowest deviance. On the second step, second-order fractional polynomial models are fitted. Similarly with the first step the best of 45 possible models is one that maximizes log likelihood.

On the third step, the best 1st-order and 2nd-order models and a linear model are compared by mean of the partial likelihood ratio test. The test statistic is given by difference in model deviances. More complex model is preferred only if it provides significant improvement of fit and in assumption of nonlinearity make sense in practice.

Denote by $D(l)$, $D(p_1)$ and $D(p_1, p_2)$ the deviances, (2.13), of the linear model, of the best 1st-order and of the best 2nd-order fractional polynomials models, respectively. The test statistics for model comparison are given by

$$G(l, p_1) = D(l) - D(p_1), \tag{2.17}$$

$$G(p_1, (p_1, p_2)) = D(p_1) - D(p_1, p_2), \tag{2.18}$$

$$G(l, (p_1, p_2)) = D(l) - D(p_1, p_2). \tag{2.19}$$

According to Altman & Royston [57], under the null hypothesis,

$$G(l, p_1) \sim \chi^2(1), \qquad G(p_1, (p_1, p_2)) \sim \chi^2(2), \qquad G(l, (p_1, p_2)) \sim \chi^2(3).$$

Previously mentioned Box-Tidwell procedure rejects linearity without any indication on the form of an existing relationship between the logit and the continuous predictor. Fractional polynomial approach provides the direct answer for the important question of which transformation better describes the functional relationship and may considerably improve the fit of a model. Graphical analyses are very easy to implement and also useful to understand where the departure from linearity occurs. Hence, in this research, we use both graphical analysis and fractional polynomial methods to check the form of a relationship between logit and independent covariates.

Independence of errors means that, for any two observations in the sample, the error terms should be uncorrelated. To investigate whether errors are independent, both graphical and analytical methods can be used. An autocorrelation plot can be used to check serial correlations between errors. We can also use tests for randomness and the simplest non-parametric one is based on the number of runs.

Definition 2.4 (Run) *Given a sequence of two or more symbols, a run is a sequence of one or more identical symbols which are followed and preceded by a different symbol, or no symbol at all.*

For example, the sequence $S1$ has only 2 runs and the sequence $S2$ has 10 runs.

$$S1 : \overbrace{a\ a\ a\ a\ a}^{run}\ \underbrace{b\ b\ b\ b\ b}_{run} \quad , \quad S2 : a\ b\ a\ b\ a\ b\ a\ b\ a\ b$$

For numerical observations, we need to impose some dichotomizing criterion on the elements of the sequence. Each observation is compared to a specified threshold, commonly the median or mean of the sample, and is coded as "0" or "1", according to whether the observation is larger or smaller than the threshold.

Test for random order is a non-parametric test of the hypothesis that the observations occur in a random order, by counting the number of runs. Too few runs or too many runs are very rare in truly random sequences, therefore they can serve as statistical criteria for the rejection of null hypothesis. For instance, in the first example we observed too many runs, in fact, we observed the maximum number of runs for a given sequence of a's and b's (that is hardly possible to happen in a random sequence). In the second example we have too few runs, because a's and b's are clustered together.

For numerical observations, let n_a and n_b denote the number of observations below and above the threshold, respectively, $n = n_a + n_b$ denote the total number of observations in the sample and r denote the number of runs. Under null hypothesis,

the expectation and the variance of the number of runs are

$$\mu_r = \frac{2n_a n_b}{n} + 1, \qquad \sigma_r^2 = \frac{2n_a n_b(2n_a n_b - n)}{n^2(n-1)},$$

respectively. The test statistic used is

$$Z = \frac{r - \mu_r}{\sigma_r} \tag{2.20}$$

which follows, assintotically, $N(0,1)$.

2.5 Assessment of model fit

LR models are used to explain the effect of covariates on the outcome variable and predict the probability of the occurrence of the event of interest. However, conclusions drawn from a fitted model can be misleading when the model has lack of fit, i.e. when the covariates cannot predict the response accurately. There are many causes of inadequate model fit, for instance, omission of important covariates related to the outcome, incorrect functional form, influential observations and outliers. Several goodness-of-fit measures to assess the adequacy of fitted LR model have been proposed by researchers in recent years; however, none of them can be considered the best, each has advantages and disadvantages. In this work, as recommended by Hosmer & Lemeshow [26], we will use a combination of four goodness-of-fit tests: Pearson Chi-square test, Hosmer-Lemeshow test, Osius-Rojek test and Stukel test.

Before discussing specific goodness-of-fit tests we need to introduce the term "covariate pattern", which is crucial to understand summary measures of goodness-of-fit and other diagnostic measures in logistic regression.

Suppose that, for $i \in \{1, 2, \ldots, n\}$, $(y_i, \boldsymbol{x}_i)$ represent n independent pairs of observations, where y_i is a realization of the Bernoulli variable, with success probability π_i, and $\boldsymbol{x}_i$ is a line vector of values of p explanatory variables associated with y_i. A covariate pattern is an observed vector of values of the p covariate variables used in the model [26]. Denote the number of covariate patterns by J. If each subject in the sample has a unique vector of values for the p covariates, the number of covariate patterns J is equal to the number of subjects, i.e. $J = n$. If some subjects share the same vector of the p covariates, we have $J < n$. Denote the number of subjects in the $j-th$ covariate pattern (i.e. that share the same covariate values $(\boldsymbol{x} = \boldsymbol{x}_j)$) by m_j, $j = 1, 2, \ldots, J$. Obviously, $\sum_{j=1}^{J} m_j = n$. Suppose that the total number of successes $(y_i = 1)$ is n_1, the total number of failures $(y_i = 0)$ is n_0, denote the number of successes in the $j-th$ covariate pattern by y_{j1} and denote the number of failures observed in the $j-th$ covariate pattern by y_{j0}. Then, obviously,

$\sum_{j=1}^{J} y_{j1} = n_1$ and $\sum_{j=1}^{J} y_{j0} = n_0$. To clarify the meaning of "covariate pattern" lets consider the following examples.

Example 2.5 *Number of covariate patterns in given data*

In the first data set, all the 11 *subjects have different vectors of the* 4 *explanatory*

Data 1

Subj	y	x_1	x_2	x_3	x_4
1	0	1	2	0	1
2	0	0	2	0	1
3	1	2	3	1	1
4	0	2	5	0	0
5	0	1	1	1	1
6	1	0	7	1	1
7	0	1	3	0	1
8	1	2	4	0	1
9	1	0	7	1	0
10	0	0	2	0	1
11	1	1	1	0	0

Data 2

Subj	y	x_1	x_2	x_3	x_4
1	0	1	10.2	0	1
2	0	0	8.7	1	0
3	1	0	8.7	1	0
4	1	0	8.7	1	0
5	1	1	10.2	0	1
6	1	0	8.7	1	0
7	1	1	10.2	0	1
8	0	2	9.4	0	1
9	1	2	9.4	0	1
10	0	1	10.2	0	1
11	1	1	10.2	0	1

variables, thus: the number of covariate patterns is $J = n = 11$*, the number of observations within covariate patterns is* $m_1 = m_2 = \ldots = m_{11} = 1$*, the total number of positive outcomes is* $y_1 = 5$ *and total number of null outcomes is* $y_0 = 6$*. In the second data set, some of the* 11 *subjects share the same vector of values for the* 4 *covariates. More precisely: subjects* $1, 5, 7, 10$ *and* 11 *have the same covariate pattern,* $(1, 10.2, 0, 1)$*; subjects* $2, 3, 4$ *and* 6 *also share covariate pattern,* $(0, 8.7, 1, 0)$*; subjects* 8 *and* 9 *share another covariate pattern,* $(2, 9.4, 0, 1)$*. Thus, for the second data set, the number of covariate patterns is* $J = 3$*, the number of observations within covariate pattern is* $m_1 = 5$*,* $m_2 = 4$ *and* $m_3 = 2$ *for the three different patterns, the total number of positive outcomes is* $y_1 = 3 + 3 + 1 = 7$ *and the total number of null outcomes is* $y_0 = 2 + 1 + 1 = 4$*.*

Using the notation introduced above, the Pearson residual for covariate pattern j is computed in the following way

$$r(y_{j1}, \hat{\pi}_j) = \frac{y_{j1} - m_j \hat{\pi}_j}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}}, \quad j = 1, 2, \ldots, J, \tag{2.21}$$

where $\hat{\pi}_j$ is the maximum likelihood estimate of probability π_j associated with covariate pattern j. The Pearson Chi-square test statistic is defined as

$$X^2 = \sum_{j=1}^{J} r(y_{j1}, \hat{\pi}_j)^2. \tag{2.22}$$

Under the null hypothesis, i.e. that the fitted model is correct, $X^2 \sim \chi^2(J-(p+1))$ asymptotically.

Pearson Chi-square test statistic is a classical summary measure of goodness-of-fit, that is available in many statistical software packages for logistic regression modeling. Several authors [1, 26, 25, 32] emphasize, that Pearson Chi-square test works very well under condition that $J < n$, which is often satisfied when the covariates in the model are categorical. However, when $J \approx n$ this test provides incorrect p-values, since the test statistic is no longer $\chi^2(J-(p+1))$.

Hosmer & Lemeshow developed alternative test of goodness-of-fit based on the grouping of observations according to the values of the percentiles of the probabilities estimated from the fitted model. The subjects are placed into c groups, with each group containing approximately n/c subjects. The number of groups is traditionally equal to 10 and these groups are designated by "deciles of risk". If we divide into c groups, we have a $c \times 2$ frequency table with columns corresponding to the two values of the outcome variable Y and rows corresponding to the c groups. The Hosmer-Lemeshow test statistic C is given by

$$C = \sum_{k=1}^{c} \frac{(o_k - v_k\bar{\pi}_k)^2}{v_k\bar{\pi}_k(1-v_k\bar{\pi}_k)}, \tag{2.23}$$

where v_k is the total number of subjects in group k, o_k is the number of subjects with $y=1$ in group k and $\bar{\pi}_k$ is the average estimated probability of π for subjects in group k

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j\hat{\pi}_j}{v_k}, \quad k = 1, 2, \ldots, c,$$

where c_k is the number of covariate patterns in group k. Under the null hypothesis, i.e. that the model fits, the test statistic follows asymptotically a $\chi^2(c-2)$ distribution.

The Osius-Rojek test statistics [48], is obtained by approximation of the Pearson Chi-square test statistics to the standard normal distribution and, therefore it can be applied only when the number of subjects in the sample is sufficiently large. The detailed description of the algorithm to obtain the test statistic is provided by Sarkar et al [**?**] and Hosmer & Lemeshow [26]. The expression of the test statistics is

$$Z = \frac{[X^2 - (J-p-1)]}{\sqrt{A+RSS}},$$

where X^2 is the test statistic from (2.22), A is a correction factor for variance given by

$$A = 2\left(J - \sum_{j=1}^{J} \frac{1}{m_j}\right)$$

and RSS is the residual sum squares of weighted linear regression of $c_j = \frac{1-2\hat{\pi}_j}{v_j}$, on covariates $\boldsymbol{x}_j$, using weights $v_j = m_j\hat{\pi}_j(1-\hat{\pi}_j)$, for $j = 1, 2, \ldots, J$. Under the null hypothesis, i.e. that the model fits, the test statistic Z has, approximately, $N(0,1)$ distribution.
The test proposed by Stukel is based on the comparison of the fitted model

$$logit(\pi) = \boldsymbol{Z\beta} \tag{2.24}$$

with a more general logistic regression model of the form

$$logit(\pi) = \boldsymbol{Z\beta} + \alpha_1 w_1 + \alpha_2 w_2.$$

This general logistic regression model has two additional variables, w_1 and w_2, defined in the following way: for covariate pattern j,

$$w_1 = 0.5 \cdot \hat{g}_j^2 \cdot I(\hat{\pi}_j \geq 0.5), \qquad w_2 = -0.5 \cdot \hat{g}_j^2 \cdot I(\hat{\pi}_j < 0.5),$$

where $I(arg) = 1$ if arg is true and $I(arg) = 0$ otherwise, and $\hat{g}_j$ denotes the estimated logit, i.e.

$$\hat{g}_j = ln\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right).$$

The general and fitted model are compared using a likelihood ratio test. Under the null hypothesis, i.e. $\alpha_1 = \alpha_2 = 0$, the test statistic follows a $\chi^2(2)$ distribution. Hosmer et al [25] claim that Stukel test is not a proper goodness-of-fit test, since it does not compare observed and predicted values, but they agree that it is useful for detecting lack of fit.

2.6 Logistic regression diagnostics

Besides testing goodness-of-fit of a LR model it is also very important to identify influential observations and outliers. Literature suggests several diagnostic measures that, in general, can be classified into two groups:

- basic building blocks,
- measures of the effect of covariate patterns on model fit and parameters estimates.

The basic building blocks measures include Pearson residuals, deviance residuals and Pregibon's leverages. These measures are of interest by themselves for identification of outlying and influential observations and are also used to derive other diagnostic statistics.

Usually residuals are defined as the difference between observed and predicted values of the probability of outcome for each observation in the sample. Yet, as explained by Hosmer & Lemeshow [26], in logistic regression the errors are binomial and error variance depends on the conditional mean of Y, i.e. $\mathrm{Var}\,(Y|\boldsymbol{x}_j) = m_j\pi_j(1-\pi_j)$. Therefore, in logistic regression residuals are standardized as can be seen in Pearson residuals $r(y_{j1}, \hat{\pi_j})$ (2.21). Alternatively to Pearson residuals, we can use the deviance residuals defined, for covariate pattern j, by

$$d(y_{j1}, \hat{\pi}_j) = \pm \left\{ 2 \left[y_{j1} \cdot ln\left(\frac{y_{j1}}{m_j\hat{\pi}_j}\right) + (m_j - y_{j1}) \cdot ln\left(\frac{m_j - y_{j1}}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{1/2}. \quad (2.25)$$

Large values of $r(y_{j1}, \hat{\pi_j})$ and $d(y_{j1}, \hat{\pi}_j)$ suggest that the model does not fit well to the given covariate pattern. There is no absolute standard to define a "large" residual, but, according to Menard [44], for a large sample, Pearson residuals and deviance residuals should have $N(0,1)$ distribution. Hence, according to Menard [44], residuals that in absolute value are greater than 2 can be considered large. On the other hand, according to Hosmer & Lemeshow [26], percentiles of $N(0,1)$ may provide some guidelines for identification of large residuals, but they should be used with caution; they claim "in practice, an assessment of large is, of necessity, a judgment call based on experience and the particular set of data being analyzed".

Observations with large residuals do not necessarily have a strong influence on the estimated parameters of the model, but on the other hand, observations with relatively small residuals can have a large influence. Influential observations are also called high-leverage observations. Pregibon [52] extended to the logistic regression framework the well known diagnostic measure, designated by leverage, in the context of linear regression. Leverage values are derived from the so-called "hat" matrix, i.e.

$$\boldsymbol{H} = \boldsymbol{V}^{1/2}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{V}\boldsymbol{Z})^{-1}\boldsymbol{Z}\boldsymbol{V}^{1/2}, \quad (2.26)$$

where $\boldsymbol{V}$ is the $J \times J$ diagonal matrix, with elements $m_j \cdot \hat{\pi}_j(1-\hat{\pi}_j)$, and $\boldsymbol{Z}$ is the $J \times (p+1)$ design matrix for J covariate patterns formed from the observed values of the p covariates. For covariate pattern j the leverage h_j is given by

$$h_j = m_j\hat{\pi}_j\left[1-\hat{\pi}_j\right]\boldsymbol{z}_j(\boldsymbol{Z}^T\boldsymbol{V}\boldsymbol{Z})^{-1}\boldsymbol{z}_j, \quad with \quad \boldsymbol{z}_j = (1, x_{1j}, x_{2j}, \ldots, x_{pj}).$$

In this formulation $0 \leq h_j \leq 1$, large value of h_j indicates that covariate pattern j has a big effect on the fitted model, even if the corresponding Pearson and deviance residuals are small.

The other type of diagnostic statistics examine the effect of deleting a single observation, or a covariate pattern, on the value of the estimated parameters and on the overall summary measures of fit. Pregibon [52] introduced $\Delta\hat{\beta}_j$, a standardized

measure of the difference in the coefficient vector, β, due to deletion of all observations that share covariate pattern j. Change in vector of estimated coefficients, $\Delta\hat{\beta}_j$, is given by

$$\Delta\hat{\beta}_j = \frac{r(y_{j1}, \hat{\pi}_j)^2 \cdot h_j}{(1 - h_j)^2}.$$

The two following diagnostic statistics allow to identify covariate patterns that are poorly fit:

- The change (decrease) in the value of Pearson X^2 statistic due to deletion of subjects with covariate pattern j is given by

$$\Delta X_j^2 = \frac{r(y_{j1}, \hat{\pi}_j)^2}{(1 - h_j)},$$

- the change in deviance is defined by

$$\Delta D_j = \frac{d(y_{j1}, \hat{\pi}_j)^2}{(1 - h_j)}.$$

A number of different types of diagnostic plots to detect outliers and influential observations have been suggested in literature, namely:

a) plot of leverages *vs* Pearson residuals
b) plot of Pearson residuals *vs* predicted probability
c) plot of deviance residuals *vs* estimated probability
d) plot of leverages *vs* predicted probability
e) plot of change in deviance residuals *vs* predicted probability
f) plot of change in Pearson residuals *vs* predicted probability
g) plot of change in coefficient vector $\boldsymbol{\beta}$ *vs* predicted probability
h) plot of ΔX_j^2 *vs* $\hat{\pi}_j$, where the size of plotting symbols are proportional to $\Delta\hat{\beta}_j$.

Details of these plots and their interpretation can be found in Long [41], Hosmer & Lemeshow [26] and Hosmer et al [27]. In this work we will use, mainly, plots *a*) *e*) and *g*).

2.7 Classification accuracy of logistic regression model

The ability of the estimated LR model to describe the response variable, Y, can be evaluated using a classification table. According to the observed values of Y, we can distinguish two sub-groups in the sample: group $Y = 1$ (if the event of interest occurs) and group $Y = 0$. It is usual to call the elements from the first group "positive" and the elements of the second group "negative". From the logistic regression equation (2.8), we obtain the predicted probabilities that can take any values in interval $[0, 1]$. Hence, to be able to compare predicted and observed values by means of 2×2 classification table, we need to select an appropriate threshold,

the so-called cutoff point, to split observations into two groups according to the estimated probabilities. If predicted probability exceeds the established cutoff point, the corresponding observation is classified as "positive", otherwise is classified as "negative". In Table 1, we can see a typical 2×2 classification table. Based in

Table 1: Classification table for LR model

		Observed		
		Y=1	Y=0	Row total
Classified	Y=1	n_{11}	n_{12}	$n_{11} + n_{12}$
	Y=0	n_{21}	n_{22}	$n_{21} + n_{22}$
Column total		$n_{11} + n_{21}$	$n_{12} + n_{22}$	n

the classification table, we are able to compute several measures of classification accuracy:

- *Sensitivity* (or true positive rate) that represents the proportion of "positives" which are correctly classified by the logistic regression model, i.e

$$Sensitivity = P(\hat{Y} = 1|Y = 1) = \frac{n_{11}}{n_{11} + n_{21}};$$

- *Specificity* (or true negative rate) that represents the proportion of "negatives" which are correctly classified by the logistic regression model, i.e.

$$Specificity = (\hat{Y} = 0|Y = 0)\frac{n_{22}}{n_{12} + n_{22}};$$

- *Count* R^2, that is a measure of total classification accuracy,

$$Count\, R^2 = \frac{n_{11} + n_{22}}{n}.$$

Typically, 0.5 is used as a cutoff point for classification. However, several authors point out that it does not always leads to satisfactory results [1, 44]. Classification is highly sensitive to the relative groups sizes and, then the data is unbalanced, further observations are allocated into the larger group. In such cases using of 0.5 provides a high rate of misclassification. As an alternative option, HosmerLemeshow [26] suggest the use of a threshold that maximizes both Sensitivity and Specificity. Such threshold can be easily determined by plotting, on the same graph, Sensitivity and Specificity curves against possible cutoff points, as shown in Figure 1. The optimal cutoff point is the intersection point of the two curves. Sensitivity and Specificity are complementary measures of model performance and both depend highly on the choice of cutoff point. As a general rule, decreasing the value of cutoff point leads, simultaneously, to an increase of Sensitivity and a decrease of Specificity. Some acceptable compromise has to be reached, but it might be difficult

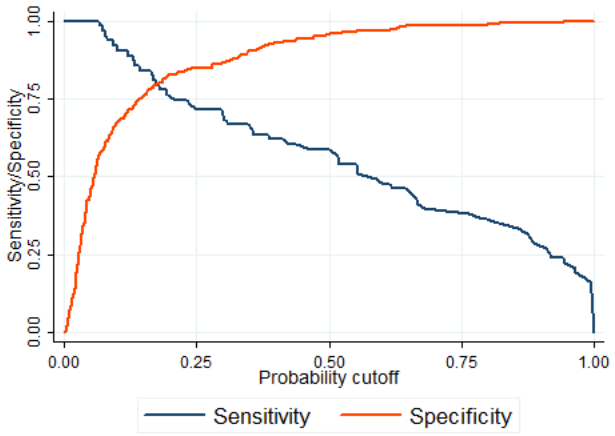


Figure 1: Plot of Sensitivity and Specificity versus all possible cutoff points

to choose which of the two measure is more important in particular circumstances. Additionally, the arbitrary choice of cutoff point brings difficulties when we want to compare different models in terms of the described measures of prediction accuracy. Thus, we need more than the three measures provided by the classification table when we want to compare performance of the classification rules provided by different models.

A plot of $Sensitivity$ versus $(1 - Specificity)$ for the full range of cutoff points is called a Receiver Operating Characteristic (ROC) curve, see Figure 2. It can be

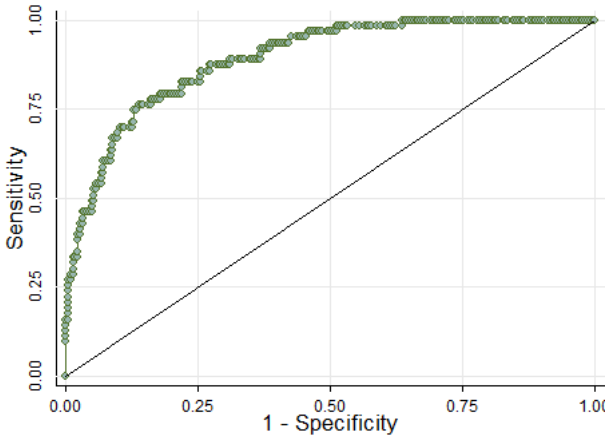


Figure 2: Receiver operating characteristic curve

considered a complete representation of model performance, as the choice of cutoff points varies. The area under the ROC curve characterizes the predictive ability of the model. The area below the reference line, that connects points (0, 0) and (1, 1), is 0.5 and corresponds to prediction by chance. If the area under the ROC curve is larger than 0.5 it means that the model has some predictive power. The higher this area is the better predictions the model provides.

Kleinbaum & Klein [32] propose the following guidelines for assessment of model's classification accuracy via Area Under ROC curve (AUC):

- $0.5 \leq AUC < 0.6$ - failed discrimination;
- $0.6 < AUC < 0.7$ - poor discrimination;
- $0.7 \leq AUC < 0.8$ - fair discrimination;
- $0.8 \leq AUC < 0.9$ - good discrimination;
- $AUC \geq 0.9$ - excellent discrimination.

To compare relative performance of classification methods and their resulting ROC curves, several statistical tests are suggest in the literature. In this work, we are interested in comparing pairs of ROC curves, constructed with the same data set, using the area under the curves. The test statistic for the comparison of areas under two ROC curves, C_1 and $C_\prime{}_2$, is

$$Z = \frac{\hat{AUC}_1 - \hat{AUC}_2}{\sqrt{S_1^2 + S_2^2 - 2rS_1S_2}}, \tag{2.27}$$

[5], where $\hat{AUC}_1$ and $\hat{AUC}_2$ are estimates of area under the corresponding ROC curves, S_1 and S_2 are estimates of the corresponding standard errors and r is an estimate of the correlation between the areas. The value of $\hat{AUC}$ is calculated using a non-parametric method (trapezoidal rule) and the values of S_1, S_2 and r are computed using the algorithm suggested by DeLong et al [14]. For further considerations see DeLong et al [14], Cleves [12] and Braga [5]. Under the null hypothesis, i.e. that the two areas are equal the distribution of Z is approximately $N(0, 1)$.

2.8 Interpretation of logistic regression model

In previous sections we presented and discussed methods for model fitting, for selection of predictor variables, for testing the significance and for checking assumptions of LR. Now we will focus on the interpretation of logistic regression model given in (2.24). Like in the linear regression framework, the logistic regression coefficient can be interpreted as the change in the dependent variable associated with a one unit change in the value of the independent variable. In the framework of multiple logistic regression, with a set of p explanatory variables $x_1, x_2, \ldots, x_j, \ldots, x_p$, the interpretation is the following: for a unit change in explanatory variable x_j the logit of probability of outcome $(Y = 1)$ is expected to change by β_j, if all the other explanatory variables are kept constant. The effect on the logit of probability of outcome of a change in value x_j is constant, since it does not depend on the initial value of x_j (depends only on the amplitude of the change) and does not depend on the values of the other explanatory variables.

Unfortunately, the interpretation of the logistic regression coefficients is not intuitive. In order to discuss and interpret the logistic regression model it is essential to introduce two quantities: odds and odds ratio.

Definition 2.6 (Odds) *Let π be the probability of a success event. The odds of success is defined as*

$$odds = \frac{\pi}{1-\pi}.$$

For example, if the probability of success is 0.25, the odds of success is equal to $0.25/(1-0.25) = 0.33$. The odds compares two probabilities, forming the so-called ratio of probabilities. If $odds > 1$, than "success" is more likely to happen than "failure". On the other hand, $odds < 1$ means that "success" is less likely to happen than "failure". For example, odds of 3 means that the probability of success is 3 times larger that the probability of failure; odds of 0.5 means that the probability of "success" is a $1/2$ of the probability of "failure".

Definition 2.7 (Odds ratio) *Let $\pi(1)$ and $\pi(2)$ be the probabilities that a success event occurs in group 1 and group 2, respectively. The ratio of the odds given by*

$$OR = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(2)}{1-\pi(2)}},$$

is called odds ratio and is denoted by OR.

The odds ratio measures how much more likely it is for elements of group 1 to experience success than for elements of group 2.

To illustrate the relationship between odds ratio and logistic regression model coefficients consider the univariate model that has a single dichotomous explanatory variable. In such case, values of the probabilities provided by the logistic regression model, can be summarized by a 2×2 table as shown below.

The OR is equals to the odds of success for cases with $x = 1$ divided by the

	Predictor	
Outcome	$x = 1$	$x = 0$
$y = 1$	$\pi(1) = \dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \dfrac{e^{\beta_0}}{1+e^{\beta_0}}$
$y = 0$	$1-\pi(1) = \dfrac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0) = \dfrac{1}{1+e^{\beta_0}}$

odds of success for cases with $x = 0$, i.e

$$OR = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} = \frac{\left(\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) / \left(\dfrac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\dfrac{e^{\beta_0}}{1+e^{\beta_0}}\right) / \left(\dfrac{1}{1+e^{\beta_0}}\right)} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}.$$

Thus, for a categorical predictor with dummy coding, there is an exponential relation between the logistic regression coefficient and the odds ratio. For a continuous or qualitative (with more than two categories) explanatory variable, the exponential of the logistic regression coefficient represents a multiplicative factor by which the predicted odds change, given that corresponding explanatory variable changed by one unit and all other explanatory variables are kept constant.

The OR can take any value in $\mathbb{R}$. Odds ratio equal to 1 means that there is no relationship between explanatory variable and the outcome. Odds ratio greater than 1 reflects the increase in odds of success with a one unit increase in the predictor. Odds ratio smaller than one reflects the decrease in odds of success with a one unit change of predictor.

From LR model we can also derive the σ_j effect, i.e. a standard deviation change in predictor x_j, instead of the one unit change. For a standard deviation change in x_j, the odds is expected to change by a factor of $e^{\beta_j \cdot \sigma_j}$, if all the other variables are kept constant.

The percentage change of odds corresponding to one unit change of explanatory variable x_j is given by

$$[e^{\beta_j} - 1] \cdot 100\%.$$

For example, $OR = 1.5$ indicates that the odds of success increases by 50% or, alternativelly, for a one unite change in predictor the odds is expected to change by a factor of 1.5. $OR = 0.4$ means that the odds of success decreases by 60%, when there is a one unit change in the predictor.

Agresti [1], Kleinbaum & Klein [32] and Long [41] provided detailed information on alternative approaches to interpretation of results of LR modeling. In this work, logistic regression results are interpreted in terms of odds ratios, since they are appropriate for all types of explanatory variables.

3 Discriminant analysis

Discriminant analysis is a multivariate statistical technique which is used with descriptive and predictive for purposes[29]. The so-called descriptive discriminant analysis tries to examine differences between particular groups of subjects, or, in other words, tries to "discriminate" between groups on based on a set of characteristics, trying to find out which characteristics are the most powerful discriminators. On the other hand, the primary interest of the so-called predictive discriminant analysis is to define rules that allow to allocate subjects into one of several mutually exclusive groups, based on the set of characteristics exhibited by the subjects.

Both parametric and non-parametric methods can be used in discriminant analysis. Parametric methods, namely linear and quadratic discriminant analysis, are appropriate only when predictors have, approximately, multivariate normal distribution. If this distributional assumption is not met, non-parametric methods, such as kernel method and nearest-neighbor method, can be used to derive classification criteria. In this work, we consider linear discriminant analysis (LDA) and K nearest neighbors discriminant analysis(K-NN).

3.1 Linear discriminant analysis (LDA)

First, let us introduce the key terms used in the framework of LDA. The categorical variable that defines the groups is called grouping or dependent variable. The p characteristics used to distinguish among *a priori* defined groups are usually called independent variables, predictors or discriminating variables, and are usually denoted by $x_1, x_2, \ldots, x_p$. LDA requires several discriminant functions which are linear combinations of $x_1, x_2, \ldots, x_p$. The number of discriminant functions may vary from 1 to $s = min(p, g-1)$ where g is the number of groups. The k-th discriminant function has a general form [55]

$$\zeta_k = u_{k0} + u_{k1}x_1 + u_{k2}x_2 + \ldots + u_{kp}x_p, \quad k = 1, 2, \ldots s, \tag{3.1}$$

where u_{k0} is a constant, $u_{k1}, u_{k2}, \ldots, u_{kp}$ are the so-called discriminant weights, also known as discriminant coefficients.

The discriminant weights are estimated in the following way:

- for ζ_1 the coefficients are derived so that the means of the discriminant function ζ_1 in the different groups are as different as possible;
- The discriminant weights of ζ_2 are derived with the same purpose, under additional restriction that $Cov(\zeta_1, \zeta_2) = 0$;
- the other functions are determined in similar way.

The value of discriminant function for a particular observed vector of the p predictors is called discriminant score.

3.1.1 Assumptions of linear discriminant analysis

The following assumptions for the use of LDA are outlined in literature [31, 55]:

1. Dependent variable must be categorical with two or more mutually exclusive groups ($g \geq 2$) and each subject must belong to one and only one group;
2. There are at least two cases per group;
3. The number of discriminating variables must be such that $1 \leq p < (n-2)$, where n is the total number of subjects in the sample;
4. Discriminating variables should be measured at least at the interval scale;
5. Lack of multicollinearity among discriminating variables: non of the discriminating variable may be a perfect linear combination of other discriminating variables;
6. Each group should be drawn from a population with a p-multivariate normal distribution;
7. Homogeneity of covariance matrices, i.e. covariance matrices of the different groups should be equal.

Assumptions 6 and 7 are particularly important for tests of significance, for computation of probabilities of group membership and for derivation of classification rules. Of all the requirements of the LDA, assumptions 6 and 7 are the most difficult to meet.

To test the hypothesis that covariance matrices are identical, the M-Box test can be employed. For this test we need to compute

$$M = (n-g) \cdot ln|S| - \sum_{i=1}^{g} (n_i - 1) \cdot ln|S_i|, \tag{3.2}$$

and

$$L = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(g-1)} \left(\sum_{i=1}^{g} \frac{1}{(n_i - 1)} - \frac{1}{n-g} \right),$$

where $|S|$ stands for the determinant of the total covariance matrix and $|S_i|$ stands for the determinant of the covariance matrix within group i, $i = 1, 2, \ldots, g$. Under the null hypothesis, i.e. that covariance matrices are identical, $M \cdot L$ has approximately $\chi^2((g-1)p(p+1)/2)$ distribution (see Reis [55] and Huberty [28] for more details). M-Box test has two well-known limitations, mentioned in literature [28, 31, 45, 55]: first, the test is sensitive to multivariate non-normality (that is the null hypothesis could be rejected either due to heterogeneity of covariance structures or non-normality of the data) and second, if groups sizes are large, the test becomes extremely sensitive and even small differences in covariance structures may lead to rejection of the null hypothesis. In such situations natural logarithms of the determinants of the covariance matrices across groups should be additionally examined.

3.1.2 Linear discriminant functions: derivation and assessment of statistical significance

As mentioned before, the basic idea of LDA is to define some linear composites of a set of discriminating variables in order to maximize the difference between the g groups.

In the following, denote by $\boldsymbol{X}$ the $n \times p$ data matrix

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix},$$

denote by $\boldsymbol{T}$ the $p \times p$ matrix of the total sums of squares and cross-products

$$\boldsymbol{T} = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (\boldsymbol{x}_i^{(j)} - \bar{\boldsymbol{x}})(\boldsymbol{x}_i^{(j)} - \bar{\boldsymbol{x}})^T, \tag{3.3}$$

where $\boldsymbol{x}_i^{(j)}$ stands for the observed column vector of the p predictors of the i-th subject in group j and $\bar{\boldsymbol{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)^T$ is the vector of mean values of predictors (grand mean), denote by $\boldsymbol{W}$ the $p \times p$ matrix of the within-groups sums of squares and cross-products

$$\boldsymbol{W} = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (\boldsymbol{x}_i^{(j)} - \bar{\boldsymbol{x}}_j)(\boldsymbol{x}_i^{(j)} - \bar{\boldsymbol{x}}_j)^T, \tag{3.4}$$

where $\bar{\boldsymbol{x}}_j$ is a vector of mean values of p predictors for the group j, and denote by $\boldsymbol{B}$ the $p \times p$ matrix of the between-groups sums of squares and cross-products

$$\boldsymbol{B} = \boldsymbol{T} - \boldsymbol{W}. \tag{3.5}$$

In order to determine the coefficients of the k-th discriminant function we should find the value of vector $\boldsymbol{a}k^T = (a_{k1}, a_{k2}, \dots, a_{kp})$ that maximizes the so-called discriminant criterion, i.e. maximizes

$$\frac{\boldsymbol{a}_k^T \boldsymbol{B} \boldsymbol{a}_k}{\boldsymbol{a}_k^T \boldsymbol{W} \boldsymbol{a}_k}.$$

To find all the vectors, $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_s$, we need to determine the eigenvalues of matrix product $\boldsymbol{W}^{-1}\boldsymbol{B}$. The number of non-zero eigenvalues equals the rank matrix and is equal to $s = min(p, g-1)$. Our vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_s$ are the eigenvectors associated with these eigenvalues. Hence, estimation of the discriminant functions coefficients

implies solving the following equation

$$\left|\boldsymbol{W}^{-1}\boldsymbol{B} - \lambda I\right| = 0.$$

Let λ_1 be the largest eigenvalue. The p-dimensional vector $\boldsymbol{a}_1$ is a solution of equation

$$(\boldsymbol{W}^{-1}\boldsymbol{B} - \lambda_1 I)\boldsymbol{a}_1 = 0.$$

The weights of the first linear discriminant function are obtained through elements of vector $\boldsymbol{a}_1$, within a constant of proportionality. In a similar way,the weights of the second discriminant function are obtained through the eigenvector $\boldsymbol{a}_2$ associated with the second largest eigenvalue λ_2, and so on.

Eigenvalues of matrix $\boldsymbol{W}^{-1}\boldsymbol{B}$ have the following two important properties. The first eigenvalue λ_1 provides the greatest separation between groups, second eigenvalue λ_2 provides the second biggest separation, and so on, and the λ_s provides the smallest group separation. Plus "eigenvalues derived in a such way are mutually uncorrelated "[31].

Although p or $(g-1)$ discriminant functions can be obtained, whether they are significant and how many of them should be considered in the interpretation of resultant group differences are questions that need to be addressed. To evaluate the significance of combination of s discriminant functions we consider a test to the value of the eigenvalues of matrix $\boldsymbol{W}^{-1}\boldsymbol{B}$. The null hypothesis of the test states that $\lambda_1 = \lambda_2 = \ldots = \lambda_s = 0$. The test statistic, known as Lambda Wilks, is given by

$$\Lambda_W = \prod_{k=1}^{s} \frac{1}{1+\lambda_k}.$$

It can be shown that, under the null hypothesis the random variable

$$X^2 = -\left[n - 1 - \frac{p+g}{2}\right] \; ln(\Lambda_W)$$

has approximately $\chi^2(p(g-1)$ distribution. In a similar way, the statistical significance of each one of the s discriminant functions can be evaluated. In such case, for the k-th discriminant function, the null hypothesis is $\lambda_k = 0$ the test statistic

$$X_k^2 = \left[n - 1 - \frac{p+g}{2}\right] ln(1+\lambda_k)$$

has approximately $\chi^2((p-k+1)(g-k))$ distribution.

The eigenvalues of the matrix $\boldsymbol{W}^{-1}\boldsymbol{B}$ provide information regarding the relative contribution of each discriminant function to the group separation. To compare contributions of discriminant functions to separation between groups, eigenval-

ues should be converted into relative percentages, i.e.

$$\frac{\lambda_k}{\sum_{j=1}^{s} \lambda_j} \cdot 100\%, \quad k = 1, 2, \ldots, s, \tag{3.6}$$

that measure the proportion of variance accounted due to the k-th function. Additionally, the canonical correlation associated with each eigenvalue might be used to assess importance of linear discriminant functions. Canonical correlation for the k-th eigenvalue is defined by

$$\eta_k = \sqrt{\frac{\lambda_k}{1 + \lambda_k}}, \quad k = 1, 2, \ldots, s.$$

η_k measures of association between the k-th discriminant function and a set of $g - 1$ dummy variables that define the groups: value zero means no relationship at all and value close to one denote the high degree of association. Squared canonical correlation indicates the proportion of variance shared between groups and predictors on the linear discriminant function [63].

Analysis of canonical correlations and of relative percentages allow us to determine the number of discriminant functions to be considered for interpretation purposes in LDA. Huberty [28] suggests to retain functions on the basis of joint relative percentages (3.6) defined for one, two or more functions until a substantial proportion of the group differences is accounted for

$$100\% \cdot \frac{\lambda_1}{\sum_{j=1}^{s} \lambda_j}, \quad 100\% \cdot \frac{\lambda_1 + \lambda_2}{\sum_{j=1}^{s} \lambda_J}, \quad 100\% \cdot \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_r}{\sum_{j=1}^{s} \lambda_j}.$$

3.1.3 Interpretation of linear discriminant functions

Literature provides several approaches to the interpretation of linear discriminant functions. The vector of coefficients $\boldsymbol{a}_k$, derived in the way described before, can be used to calculate discriminant scores for subjects for the purpose of classification. However it is common practice to standardize the components of the vector $\boldsymbol{a}_k$ in order to obtain a discriminant function ζ as in (3.1), with parameters $u_{k0}, u_{k1}, \ldots, u_{kp}$ given by

$$for \quad i \in \{1, 2, \ldots, p\}, \quad u_{ki} = a_{ki}\sqrt{n - g}, \quad u_{k0} = -\sum_{i=1}^{p} u_{ki}\, \bar{x}_i, \tag{3.7}$$

where $\bar{x}_i$ is the sample mean value of variable i. u_{k0} is the intercept of the discriminant function and coefficients $u_{k1}, u_{k2}, \ldots, u_{kp}$ are referred in the literature as

the raw coefficients. As a result of such transformation the discriminant scores over all cases will have zero mean and a within-group standard deviation of one, which makes possible the comparison with the z-scores. Discriminant score describes the position of a case in the discriminant space along the axis defined by the discriminant function. The transformation provided by (3.7) reallocates the axes defined by the discriminant functions so that the origin (the point corresponding to value zero) coincides with the grand centroid (the point of the space where all discriminating variables have their means). The scores can be interpreted as measures of the distance between the grand centroid and each particular case. For instance, $\zeta_{11} = 1.5$ means that subject 1 is one and a half standard deviations in the positive direction from the center of the axis and $\zeta_{12} = -3$ means that subject 2 is three standard deviations in the negative direction from the center of the axis, quite distant from the origin.

Unstandardized discriminant weights indicate the absolute contribution of each predictor to value of the discriminant score. To assess relative importance of discriminating variables, standardized discriminant function coefficients should be analyzed. Standardized coefficients are derived from (3.7) in the following way

$$u_{ki}^s = u_{ki} \cdot \sqrt{\frac{w_{ii}}{n-g}}, \quad i = 1, 2, \ldots, p, \tag{3.8}$$

where w_{ii} is the i-th main diagonal element of the covariance matrix $\boldsymbol{W}$ defined in (3.4). However, Huberty [28] and Klecka [31] claim that the use of these standardized discriminant function weights to assess the relative importance of a covariate has a serious limitation: if there is multicollinearity, standardized coefficients may have misleading conclusions, since they take into account the joint contribution of all variables. Given that two discriminating variables are highly correlated, two scenarios are possible: first, both predictors may have lower coefficients that do not reflect their true effects (since the predictors share the contribution to the discriminant score), or coefficients may be large but with opposite signs (so that one predictor cancels the effect of the other). Alternatively, Huberty [28] and Klecka [31] suggest to judge relative importance of variables in LDA through the correlation between each discriminating variable and the different discriminant functions. For discriminant function k and discriminating variable i the correlation in question, known as structure coefficient or loading, is given by

$$l_{ki} = \sum_{j=1}^{p} u_{ki}^s r_{ij}, \tag{3.9}$$

where r_{ij} is a pooled within-group correlation coefficient between variables i and j and u_{ki}^s are given by (3.8).

The structure coefficients show how a particular variable and discriminant function are related. A high loading (in absolute value) reveals that variable shares the most variation with a given function, but a loading close to zero tells us that function and variable have less in common. On the basis of structure coefficient we can assign descriptive labels to the linear discriminant functions. On the other hand, the application of structure coefficients in context of LDA has been criticized by some authors since structure coefficients fail to provide multivariate information. Rencher [56] claims that structure coefficients show univariate contribution of each variable for the group separation but ignore the other predictors, since the value of loadings do not change when variables are included or excluded from the model. In order to assess the joint contribution of discriminating variables Rencher [56] recommends the use of standardized coefficients in the interpretation of discriminant functions.

3.1.4 Classification via linear discriminant analysis

In general form, for predictive discriminant analysis the classification rule can be defined in the following way. Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_p]$ denote an observation from dataset measured on p discriminating variables. Let g denote the number of groups, $G_1, G_2, \ldots, G_g$ the different groups, n_i the number of observations in group G_i and π_i the prior probability of membership in group G_i. For $i = 1, 2, \ldots, g$, let $f(\boldsymbol{x}|G_i)$ be the probability density function of $\boldsymbol{x}$, given that the observation was collected in an element of group G_i. Let $P(\boldsymbol{x}|G_i)$ represent the probability of observing vector $\boldsymbol{x}$ conditional on $\boldsymbol{x}$ being collected in an element of group G_i and $P(G_i|\boldsymbol{x})$ denote the posterior probability of group G_i. The relationship between the two conditional probabilities, $\mathbb{P}(G_i|\boldsymbol{x})$ and $\mathbb{P}(\boldsymbol{x}|G_i)$, may be derived through the multiplication rule: for any two events A and B, such that $\mathbb{P}(A) \cdot \mathbb{P}(B) > 0$,

$$\mathbb{P}(A \cap B) = P(B) \cdot \mathbb{P}(A|B) = \mathbb{P}(A) \cdot \mathbb{P}(B|A).$$

Hence, we get

$$P(G_i \cap \boldsymbol{x}) = \mathbb{P}(\boldsymbol{x}) \cdot \mathbb{P}(G_i|\boldsymbol{x}) = \pi_i \cdot \mathbb{P}(\boldsymbol{x}|G_i),$$

Based on the Bayes' theorem the posterior probability of $\boldsymbol{x}$ membership in the group G_i is calculated in following way

$$\mathbb{P}(G_i|\boldsymbol{x}) = \frac{\pi_i \cdot \mathbb{P}(\boldsymbol{x}|G_i)}{\mathbb{P}(\boldsymbol{x})}. \tag{3.10}$$

Total probability theorem yields

$$\mathbb{P}(G_i|\boldsymbol{x}) = \frac{\pi_i \cdot \mathbb{P}(\boldsymbol{x}|G_i)}{\sum_{j=1}^{g} \pi_j \cdot \mathbb{P}(\boldsymbol{x}|G_j)}. \tag{3.11}$$

The substitution of $\mathbb{P}(\boldsymbol{x}|G_i)$ by $f(\boldsymbol{x}|G_i)$ in (3.11) results in

$$\mathbb{P}(G_i|\boldsymbol{x}) = \frac{\pi_i \cdot f(\boldsymbol{x}|G_i)}{\sum_{j=1}^{g} \pi_j \cdot f(\boldsymbol{x}|G_j)}. \tag{3.12}$$

Hence, the classification rule based on Bayesian theorem for posterior probabilities is formulated as following: subject b, with observed vector $\boldsymbol{x}_b$ for the p discriminating variables, is assigned to group G_i if, for all $i \neq j$,

$$\mathbb{P}(G_i|\boldsymbol{x}_b) > P(G_j|\boldsymbol{x}_b), \tag{3.13}$$

with $\mathbb{P}(G_i|\boldsymbol{x}_b)$ calculated as in (3.12).

Another popular approach to determine the group membership of subject b is to consider how far the associated vector of observations $\boldsymbol{x}_b$ is from the centroid of each group and place b within the closest group. In the context of LDA the squared Mahalanobis distance, D^2, is used to measure the closeness:

$$D^2_{b,j} = (\boldsymbol{x}_b - \bar{\boldsymbol{x}}_j)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_b - \bar{\boldsymbol{x}}_j), \quad j = 1, 2, \ldots g,$$

where $\bar{\boldsymbol{x}}_j$ is the mean vector for group j and $\boldsymbol{S}$ is the pooled within-groups covariance matrix. After calculating the $D^2_{b,j}$ for all groups, subject b is allocated into the group with the smallest value. Note that the classification rule based on the Mahalanobis distance requires equality of group covariance matrices and equal prior probabilities. Assuming unequal prior probabilities, the following adjustments should be made

$$D^2_{ub,j} = (\boldsymbol{x}_b - \bar{\boldsymbol{x}}_j)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_b - \bar{\boldsymbol{x}}_j) - 2 \cdot ln(\pi_j), \quad j = 1, 2, \ldots g,$$

where π_j represents prior probability for the group j.

To assess the ability of discriminant analysis model to predict group membership the probability of correct classification known in the literature as hit rate, is usually used. Alternatively, we can evaluate its complementary probability known as misclassification rate. The simplest way to estimate misclassification rate is to apply classification procedure to the same data from which the discriminant functions were estimated. In such case, the estimate of misclassification rate is designated by

apparent error rate. When the true group membership of subjects in the sample is known, the results of classification can be summarized in the form of a $g \times g$ classification table, similar to the 2×2 table describe in Section 2.7 for logistic regression. In the framework of discriminant analysis such table is the so-called confusion matrix.

3.1.5 Variable selection in linear discriminant analysis

In similarity with logistic regression, the optimal set of variables for LDA can be selected employing stepwise procedure: forward selection, backward elimination or combination of these two methods. As mentioned previously, variables are eligible for inclusion or elimination from the model on the basis of some statistical criterion. A number of criteria have been suggested in literature in context of discriminant analysis. Perhaps the most widely used criterion is Wilks Lambda, that is defined by following expression

$$\Lambda_W = \frac{|W|}{|W+B|}, \tag{3.14}$$

where W is the within-group matrix of sums of squares and B the between-group matrix of sums of squares defined in (3.4) and (3.5), respectively. At each step of the algorithm, the variable that is added is the one with the smallest value of Λ_W. Minimization of the ratio given in (3.14) implies that the within-groups sum of squares is minimized and between-groups sum of squares is maximized.

Squared Mahalanobis distance and Rao's V statistic can also be used as criteria for stepwise variable selection in LDA; for details see Reis [55].

3.1.6 Linear discriminant analysis: two-group case

In this subsection we give a brief description of LDA procedure for the two group case. Notice that, in this case, we need only one discriminant function.

For two groups of subjects, G_1 and G_2, characterized by p discriminating variables, denote by n_1 and n_2 the number of subjects in G_1 and G_2, respectively, by $\bar{\boldsymbol{x}}_i$ the sample mean vector of group i, by $\hat{\boldsymbol{S}}$ the sample covariance matrix. The estimate of vector $\boldsymbol{a}$ is given by

$$\hat{\boldsymbol{a}} = \hat{\boldsymbol{S}}^{-1} \cdot (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2).$$

The linear discriminant function is estimated by $\hat{Y} = \hat{\boldsymbol{a}}^T \boldsymbol{X}^T$.

The discriminant score of a subject b with observed vector of discriminating variables $\boldsymbol{x}_b$ is given by

$$\hat{Y}_b = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \hat{\boldsymbol{S}}^{-1} \boldsymbol{x}_b,$$

and will be used to classify subject b in one of the two groups. Consider the average scores

$$\bar{Y}_1 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \hat{\boldsymbol{S}}^{-1} \bar{\boldsymbol{x}}_1 \quad \text{and} \quad \bar{Y}_2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \hat{\boldsymbol{S}^{-1}} \bar{\boldsymbol{x}}_2,$$

of G_1 and G_2, respectively. Thus, the average discriminant score for the whole sample is

$$\bar{Y} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \hat{\boldsymbol{S}}^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2).$$

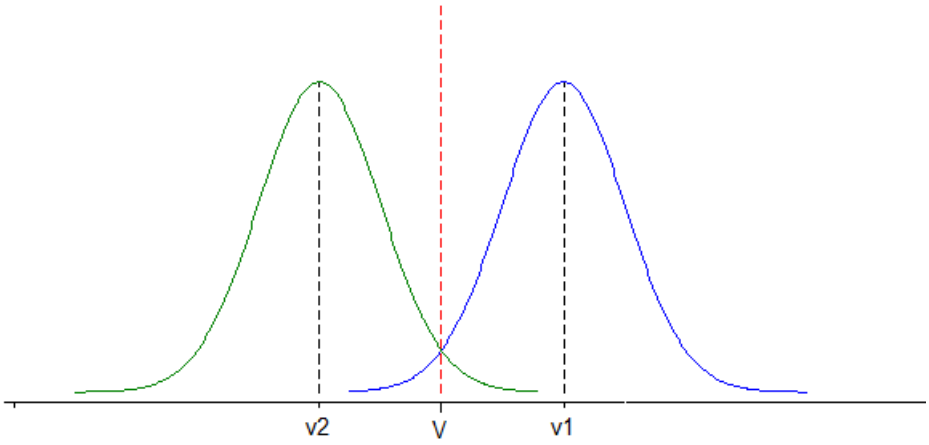
The resultant classification rule is formulated as following: subject b is assigned to group G_1 if

$$\hat{Y}_b > \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \hat{\boldsymbol{S}}^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)$$

and, otherwise, is assigned to the group G_2.

Figure 3 illustrates the classification rule based on cutoff score that assumes equal sample sizes for the two groups. Letters V, $v1$ and $v2$ in the Figure 3 stand for $\bar{Y}$, $\bar{Y}_1$ and $\bar{Y}_2$, respectively. For unequal sample sizes the cutoff score is calculated

Figure 3: Classification rule for two groups with equal dimensions



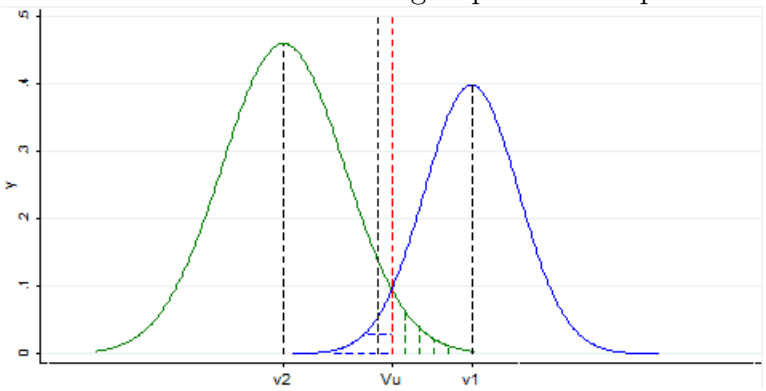
in the following way

$$Vu = \frac{n_2 \bar{Y}_1 + n_1 \bar{Y}_2}{n_1 + n_2},$$

see [55]. Figure 4 shows the classification rule based on cutoff score that assumes unequal sample sizes for the two groups. A cutoff score, Vu, in such case is shifted from the central position, V, to ensure equality of probabilities of misclassification for two groups.

In Section 3.1.4 we presented the general form of classification rule based on the posterior probability. For observation vector $\boldsymbol{x}_b$, posterior probability of group

Figure 4: Classification rule for two groups with unequal dimensions



membership for the two group case is estimated by

$$\hat{\mathbb{P}}(G_i|\boldsymbol{x}_b) = \frac{\hat{\pi}_i \cdot \hat{f}(\boldsymbol{x}_b|G_i)}{\sum_{j=1}^{2} \hat{\pi}_j \cdot \hat{f}(\boldsymbol{x}_b|G_j)}, \quad \text{for} \quad i = 1, 2.$$

where

$$\hat{f}(\boldsymbol{x}_b|G_i) = \frac{1}{\sqrt{(2\pi)^2}\sqrt{|\hat{\boldsymbol{S}}_i|}} exp(-0.5(\boldsymbol{x}_b - \bar{\boldsymbol{x}}_i)^T \hat{\boldsymbol{S}}_i^{-1}(\boldsymbol{x}_b - \bar{\boldsymbol{x}}_i)).$$

Observe that $\hat{f}(\boldsymbol{x}_b|G_i)$ is the density function, assuming that multivariate normal probability model with equal covariance matrices holds. So,

$$\hat{\mathbb{P}}(G_i|\boldsymbol{x}_b) = \frac{\hat{\pi}_i \cdot |\hat{\boldsymbol{S}}_i|^{-1/2} exp(-0.5\, D_{bi}^2)}{\sum_{j=1}^{2} \hat{\pi}_j \cdot |\hat{\boldsymbol{S}}_j|^{-1/2} exp(-0.5 D_{bi}^2)}, \quad i = 1, 2,$$

where $D_{bi}^2 = (\boldsymbol{x}_b - \bar{\boldsymbol{x}}_i)^T \hat{\boldsymbol{S}}_i^{-1}(\boldsymbol{x}_b - \bar{\boldsymbol{x}}_i)$ is the squared Mahalanobis distance between observed vector $\boldsymbol{x}_b$ and $\bar{\boldsymbol{x}}_i$. Therefore, subject b is allocated in group G_1 if

$$\hat{\mathbb{P}}(G_1|\boldsymbol{x}_b) > \hat{\mathbb{P}}(G_2|\boldsymbol{x}_b)$$

and allocated to the group G_2, otherwise.

3.2 K nearest neighbor discriminant analysis

In the previous subsections we dealt with a classification technique involving a set of predictors whose theoretical joint distribution was assumed to be multivariate normal. In this section we a present non-parametric approach to the problem of classification, so-called K-nearest neighbor discriminant analysis (K-NN).

K-NN discriminant analysis is a statistical tool used to predict group mem-

bership of an observation, based on a non-parametric estimator of the distribution of its K-nearest neighbors. This technique was introduced by Fix and Hodges **??** in early 50s. Unlike traditional LDA and LR, K-NN is a quite flexible approach to the classification problem. The technique does not require assumptions of multivariate normality or homogeneity of covariance matrices (like LDA does), or assumptions of linearity and link function specification (like logistic regression does). K-NN discriminant analysis is based on the single assumption that members of the same group have similar characteristics. For example, in the 1-NN rule, the observation is classified into the group corresponding to the membership of the closest observation, according to some metric. With K-NN rule, if observation $\boldsymbol{x}$ is to be assigned, we search through the data for the set of the K-nearest neighbors, according to distance function, and allocate $\boldsymbol{x}$ in the most frequent class among these neighbors.

In Section 3.1.4, we calculated the posterior probability of $\boldsymbol{x}$ membership in the group G_i

$$\mathbb{P}(G_i|\boldsymbol{x}) = \frac{\pi_i \cdot f(\boldsymbol{x}|G_i)}{\sum_{j=1}^{g} \pi_j \cdot f(\boldsymbol{x}|G_j)}. \tag{3.15}$$

and defined the general form of classification rule (3.13) based on the posterior probability of group membership. To allocate an unclassified observation, $\boldsymbol{x}$, to one of mutually exclusive groups according to the maximum posterior probability rule, K-NN uses simply a non-parametric estimator of $f_(\boldsymbol{x}|G_i)$, based on the set of K-nearest neighbors.

A non-parametric estimator of the density function $f_(\boldsymbol{x}|G_i)$ is the relative frequency of observations from the group G_i in the neighborhood of $\boldsymbol{x}$. Let k_i denote the number of the K-nearest neighbors of $\boldsymbol{x}$ that belong to group G_i. Consequently, the formula for posterior probability, given by equation (3.15), transforms in

$$\hat{\mathbb{P}}(G_i|\boldsymbol{x}) = \frac{\frac{\hat{\pi}_i \cdot k_i}{n_i}}{\sum_{j=1}^{g} \frac{\hat{\pi}_j \cdot k_j}{n_j}},$$

where n_i is the size of group G_i. These estimated probabilities are used to classify subjects: subject b is classified into the group for which the posterior probability of membership $\hat{\mathbb{P}}(G_i|\boldsymbol{x}_b)$ is the highest, where $\boldsymbol{x}_b$ is the observed vector of covariates of b.

The performance of the K-NN technique depends crucially on:

- the similarity measure, or distance measure, used for identification of nearest neighbors;
- number of neighbors used in the classification rule.

Definition 3.1 (Distance) *A distance (metric) on a set U is a function*

$$D : U \times U \longrightarrow \mathbb{R}$$

that satisfies the following axioms

1. $D(x, y) \geq 0$, $\forall x, y \in U$;
2. $D(x, y) = 0$ *if and only if* $x = y$;
3. $D(x, y) = D(y, x)$, $\forall x, y \in U$;
4. $D(x, z) \leq D(x, y) + D(y, z)$, $\forall x, y, z \in U$.

Axioms 1 and 2 mean that D is positive definite, axiom 3 means that D is symmetric and axiom 4 is known as triangular inequality.

Lets start by considering the most popular and frequently used distance measure: the Euclidean distance. For two observations in p-dimensional space, $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$ and $\boldsymbol{x}_j = [x_{j1}, x_{j2}, \ldots, x_{jp}]$, Euclidean distance is given by

$$D_{\boldsymbol{x}_i, \boldsymbol{x}_j} = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}.$$

In fact, Euclidean metric is a special case of a more general distance measure, the so-called Minkowski metric. For the two observations in p-dimensional space, $\boldsymbol{x}_i, \pm x_j$ the Minkowski distance is given by

$$D_{\boldsymbol{x}_i, \boldsymbol{x}_j} = \left(\sum_{k=1}^{p} (|x_{ik} - x_{jk}|)^n \right)^{\frac{1}{n}},$$

where n is a fixed natural number. For $n = 1$ the previous formula is simplified to

$$D_{\boldsymbol{x}_i, \boldsymbol{x}_j} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|,$$

and the metric is known as Manhattan or city-block distance.

It can be proved that Euclidean measure of distance is not invariant for scale. In other words, this metric assumes that all p variables are measured with the same metric scale. In practice, it is quite difficult to guarantee this assumption. That's why the common practice is to standardize the variables or, alternatively, compute the Mahalanobis distance, defined in context of LDA. Mahalanobis distance takes into account the correlation among variables, it is invariant for scale and, for the independent variables, it is simply reduces to the Euclidean distance.

The metrics described above are designed for continuous variables. However, some research situations involve other types of variables, for example, binary variables, that measure presence or absence of the characteristic of interest. In this

case, the squared Euclidean distance provides a count of mismatches between observations, however it attributes an equal weight for matching cases. The distance measures designed for binary data are known as matching or similarity coefficients.

Definition 3.2 (Matching coefficient) *For a set X, a similarity coefficient is a function, C, that maps $X \times X$ into $\mathbb{R}$ and, for all x, y in X, satisfies the following conditions:*

- $0 \leq C(x, y) \leq 1$, $\forall x, y \in X$;
- $C(x, x) = 1$, $\forall x \in X$;
- $C(x, y) = 1$, *if and only if* $x = y$;
- $C(x, y) = C(y, x)$, $\forall x, y \in X$.

A large number of similarity coefficients have been proposed in the literature. In order to demonstrate how some of this measures are calculated we will introduce a simple example. Suppose that we have two subjects (items) characterized by p binary variables. In this case, the presence or absence of p attributes can be summarized in a frequency table as shown in a Table 2.

Table 2: Frequency of matches and mismatches for two subjects

	Subject 2	
Subject 1	presence of characteristic(1)	lack of characteristic(0)
(1)	a	b
(0)	c	d

In Table 2, cell a counts the number of attributes, in p, that both subjects have $(1 - 1$ matches). Cell b counts the number of attributes that first subject has but the second subject does not $(1 - 0$ matches), and so on.

We define some metrics in terms of cells of the Table 2:

- Russell matching coefficient, given by

$$D_R = \frac{a}{a + b + c + d}.$$

  In other words, the Russell's metric simply represents proportion of attributes present in both subjects.

- Jaccard matching coefficient, given by

$$D_J = \frac{a}{a + b + c}.$$

  This coefficient ignores the number of $(0 - 0)$ matches; they are considered to be irrelevant.

- Dice matching coefficient, given by

$$D_D = \frac{2a}{2a + b + c}.$$

The Dice coefficient is closely related to the Jaccard coefficient, with the unique difference that $1-1$ matches have additional weight.

Example 3.3 *Calculating the values of matching coefficients*
Consider 6 characteristics of interest for 2 subjects. The number of matches and

Subject	x_1	x_2	x_3	x_4	x_5	x_6
1	0	1	1	0	0	1
2	0	0	1	1	1	1

mismatches is summarized in Table 3.

Table 3: Summary of matches and mismatches for two subjects

	Subject 2	
Subject 1	(1)	(0)
(1)	2	1
(0)	2	1

Russell, Jaccard and Dice matching coefficient for these two subjects are

$$D_R = \frac{2}{2+1+2+2} = 0.33, \quad D_J = \frac{2}{2+1+2} = 0.4, \quad D_D = \frac{2 \cdot 2}{2 \cdot 2+1+2} = 0.57,$$

respectively.

In existent literature we found several guidelines for the choice of the optimal value of K in the nearest neighbor discriminant analysis. For instance, for two groups classification, K should be an odd integer in order to avoid ties. McLahuan [43] cites Enas and Choi work and, according to their recommendations, for the two group classification problem with comparable group sizes K should be selected from interval $[n^{2/8},\ n^{3/8}]$.

According to Huberty [28], the value of K should be large enough in order to obtain consistent density estimates. On the other hand, for data with unbalanced group sizes this suggestion probably not the most appropriate. When there are remarkable differences in group sizes, K must be much smaller than the smallest group. Huberty [28] also suggests that, for each particular situation, the researcher should try several values of K and make his choice based on the classification results.

4 Comparison of classification rules

There are two crucial questions in classification problems:

- if the knowledge of independent variables is helpful in predicting the probability of outcome;
- which of the generated rules is better.

Practical and statistical usefulness of classification rules can be evaluated by several means. In this section we present a brief description of measures for assessment of the practical utility of classification rules and some considerations on the performance of LR, LDA and K-NN.

4.1 Criteria for comparison of classification rules

The statistical criteria described below are suggested by Huberty [28] in the context of discriminant analysis and, therefore, can be applied to compare performance of different classification methods (see [45]). We present the formulation of criteria for the particular case $g = 2$. Consider two groups, G_1 and G_2, and let q_i denote the estimated prior probability of membership in group G_i, $i \in \{1, 2\}$. The results of classification can be summarized in the following table

	Predicted group		Total Row
Actual group	G_1	G_2	
G_1	n_{11}	n_{12}	$n_{1.}$
G_2	n_{21}	n_{22}	$n_{2.}$
Total Column	$n_{.1}$	$n_{.2}$	n

The total observed frequency of correct classifications, denoted by o, is the sum of the elements in the main diagonal, i.e.

$$o = \sum_{i=1}^{2} n_{ii},$$

and observed proportion (rate) of correct classifications is given by

$$H_o = \frac{o}{n}. \tag{4.1}$$

There a number of different ways to determine the chance rate, that is, the rate of correct classification by chance and without knowledge of predictors. When group sizes are equal, the proportion of correct classification due to chance is simply $\frac{1}{2}$. When group sizes are unequal, two different strategies for calculating chance rates can be used. The first is the so-called maximum chance criterion: the proportion of correctly classified subjects is equal to the highest value of prior probability of group membership. This criterion is recommended when the goal of the research

is to maximize overall rate of correct classification. The second approach is the proportional chance criterion, that should be used when the researcher is interested in correctly classify subjects into all of the groups. The chance frequency of correct classifications for Group G_i is given by

$$e_i = \hat{\pi}_i \cdot n_i, \tag{4.2}$$

where $\hat{\pi}$ is the estimated prior probability of group G_i, the total-group frequency of correct classifications due to chance is given by

$$e = \sum_{i=1}^{2} \hat{\pi} \cdot n_i, \tag{4.3}$$

and the overall (expected) rate of correct classifications due to chance is

$$H_e = \frac{1}{n} \sum_{i=1}^{2} \hat{\pi} \cdot n_i. \tag{4.4}$$

The observed total-group correct classification may be compared with the expected (4.3) to decide if we have achieved a classification better than the chance classification. Hence, we perform a test where the null hypothesis stats that the number of subjects correctly classified by model is equal to the number correctly classified by chance. The overall number of correct classifications, o, is the test statistic, that can take any value from zero to n. Since n is generally large enough in classification problems, under the null hypothesis, the distribution of o can be approximated by the $N(0,1)$ distribution. Thus, a statistic defined by

$$Z = \frac{(o-e)}{\sqrt{e(n-e)/n}}, \tag{4.5}$$

may be used to test the null hypothesis. The lower bound of a confidence interval for the true overall frequency of correct classifications is given by

$$o - z_{1-\alpha}\sqrt{e(n-e)/n},$$

where $z_{1-\alpha}$ is the $100 \cdot (1-\alpha)$ percentile of the $N(0,1)$ distribution. Some times the researcher is interested in separating group predictions. In such case, for particular group G_i, the test statistic is

$$Z_{G_i} = \frac{(n_{ii} - e_i)}{\sqrt{e_i(n_i - e_i)/n_i}}, \quad i = 1, 2, \tag{4.6}$$

and corresponding lower bound of a confidence interval is given by

$$n_{ii} - z_{1-\alpha}\sqrt{e_i(n_i - e_i)/n_i}, \quad i = 1, 2.$$

The other measure, suggested by Huberty [28], is the index of improvement over chance, I. This index defined as

$$I = \frac{H_o - H_e}{1 - H_e}, \tag{4.7}$$

takes into account the expected total rate of correct classification, as well as the observed rate of correct classification. The index I provides the percent reduction in error by chance classification if the predictive model is used.

To compare the total-group classification accuracy of two different rules, applied to the same subjects, Huberty [28], Meshbane & Morris [45] recommend to apply McNemar test. The results of group membership prediction should be summarize as shown in Table 4, where n_{11} is a number of subjects correctly classified

Table 4: Comparison of classification rules

		Rule 2		
		Hit	Miss	Total Row
Rule 1	Hit	n_{11}	n_{12}	$n_{11} + n_{12}$
	Miss	n_{21}	n_{22}	$n_{21} + n_{22}$
Total Column		$n_{11} + n_{21}$	$n_{12} + n_{22}$	n

by both rules, n_{12} number of subjects correctly classified by rule 1 and incorrectly classified by rule 2, and so on. The null hypothesis states that the proportion of subjects correctly classified by rule 1 equals the proportion of subjects correctly classified by rule 2. If $n_{12} + n_{21} \geq 25$, the test statistic is given by

$$T = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}. \tag{4.8}$$

Under the null hypothesis, T follows $\chi^2(1)$ distribution. If $n_{12} + n_{21} < 25$, the test statistic is $T_e = n_{12}$ and, under the null hypothesis, it has $Binomial(0.5,\, n_{12} + n_{21})$ distribution.

4.2 Relative performance of classification rules

Classification problems are very common in the field of social and educational sciences. The most frequently used statistical approaches for prediction of group membership are LR and LDA. These methods are employed to predict university drop out, to differentiate between underperforming and successful students

and graduates. Modern development of statistical software brought more flexible non-parametric approaches to the classification problem, such as K-NN discriminant analysis. Non-parametric forms of discriminant analysis are very attractive for educational research since they do not require any distributional assumptions.

In this section, we present a brief review of literature on relative performance classification techniques used in this work. Note that, existent literature focuses mainly on comparison of LR with LDA and relatively little research has been conducted to compare the classification accuracy of K-NN with LDA and with LR.

Performance of classification algorithms can be affected by wide range of factors, namely: data structure, underlining distributional assumptions, dimension of groups and sample size. The classification rule based on linear discriminant functions is derived on the basis of two important assumptions: multivariate normal distribution of explanatory variables within each group and homogeneity of covariance matrices. However, simplicity of computation and interpretation of linear discriminant functions lead to application of LDA in the field of social and educational sciences, where the required assumptions are clearly violated. Hence, the question is: how "good" is the performance of LDA under non-optimal conditions, i.e. when the assumptions do not hold?

A number of studies investigated the behavior of the linear discriminant functions when underlying distributions are non-normal. Evidence of LDA efficiency with mixed continuous and categorical variables is provided in [29, 65]. Lachenbruch [35] summarized results of several studies, concluding that, in general, LDA performs fairly well on discrete data of various types. LDA is recommended by Asparoukhov et al [3] in presence of binary explanatory variables because of "expected stability" when the total number of explanatory variables is large. Regarding continuous distributions, Lachenbruch [35] claims that LDA is moderately robust in the presence of mixture of normal distributions and of heavy-tailed symmetric distributions. Lachenbruch [35] also claims that highly skewed distributions may cause considerable decrement in the performance of this parametric classification method. Seber [62], in review on robustness of LDA, cites studies of Moore, and of Dillon and Goldstein that claime poor performance of LDA in the presence of high correlations between predictors. Krzanowski [34] investigated performance of LDA for mixture of normal and dichotomous variables and also concluded that the ability of LDA to correctly classify individuals is affected if correlations between discrete and continuous variables differ significantly for the two groups. For the assumption of homogeneity of covariance matrices, a number of studies states that it has a deleterious impact on the performance of LDA. Although LR does not make this assumption, it was shown, by empirical research, that LR classification accuracy may also be affected if covariance matrices are unequal [20, 38].

Several studies have compared predictive accuracy of LDA and LR models. Literature provides evidence that LDA is "asymptotically more efficient" than LR than the underlying assumptions are met [38]. Press & Wilson [53] concluded that LR is preferable to LDA when the vectors of explanatory variables do not have multivariate normal distributions within groups.

Recent research finds non, or very little, differences in classification accuracy of the two parametric approaches under normality and equality of covariance matrices. More precisely, Finch Schneider [20] demonstrated, via simulations, that if the assumptions of LDA are met, LDA and LR models have very comparable misclassification rates. Hastie et al [24] claim that, in practice, LDA and LR give very similar results, even when LDA is used "inappropriately", i.e. when underlying assumptions are not met, as is the case with qualitative predictors.

Meshbane & Morris [45] compared the leave-one-out classification performance of LDA and LR for 29 real data sets. The authors considered rates of correct classification for each group and for the total sample with data of many different types. The comparisons were made using McNemar test. For 28 data sets, Meshbane & Morris [45] claim not to have found statistically significant differences in total hit rates between LDA and LR.

The results of Monte Carlo simulation study on relative efficiency of LDA and of LR for a two group classification problem, conducted by Fan & Wang [16], indicate that LR and LDA have similar performance when covariance matrices are equal and groups have approximately equal sizes.

Fan & Wang [16], Meshbane & Morris [45] and Finch & Schneider [20] suggested that different group sizes have impact on the performance of both methods. For the two group case, Finch & Schneider [20] demonstrated that, if the assumption of equal covariance matrices holds, the misclassification rate was very high for the smaller group, very low for the larger group in both models. According to Fan & Wang [16], if two groups have very different proportions, like $10 : 90$ or $25 : 75$, LR minimizes the error rate for the smaller group, but LDA, appears to minimize the error rate for the larger group, independently of covariance matrices. On the other hand, Meshbane & Morris [45] present a study that concludes "superior performance of the LR model in classifying the larger group was offset by superior performance of the LDA model in classifying the smaller group", for unbalanced group sizes. Hence, it is not consensual which of the two parametric methods performs better in a situation of unbalanced group sizes.

Not much information is available on the comparison of traditional LR e LDA with non-parametric methods of classification. K-NN discriminant analysis is very flexible and, according to Asparoukhov et al [3], is very efficient in classification problems with large or moderate number of explanatory variables (greater than 5)

and with small or moderate sample size. Hastie et al [24] provide some examples of classification problems for which K-NN technique outperforms traditional parametric methods such as LR and LDA. Theoretically, K-NN is expected to perform better than LDA in the precence of heterogeneous covariance matrices and without multivariate normality. However, in [19], one example of superiority of LDA over K-NN, for the two group classification problem, with non-normal data, unequal covariance matrices and different groups sizes, is presented.

Based on our findings in the literature reported above, LR is expected to slightly outperform LDA if group sizes are different, there is mixture of quantitative and qualitative variables. On the other hand, since there is not much information available in literature on the issue of relative performance of K-NN and LR, it is difficult to set expectations. Hence, we we decided to use the three methods and compare the results obtained with the available data.

5 Application of logistic regression, linear discriminant analysis e K-NN discriminant analysis to the data

5.1 Academic performance of medical students: literature review

In this section we present a brief review of existent literature on the issue of academic performance of medical students. To conduct a literature search, we used MEDLINE and ERIC databases combining several search terms, such as "medical student", "predict", "performance", "success", "failure", "first year", "high education", "at-risk student", "demographic characteristics", "admission", "student selection" and "personality". MEDLINE and ERIC are databases of published research articles in medical sciences and educational sciences, respectively. The results highlighted a series of explanatory factors associated with academic performance in medical school.

A relation between a pre-university performance and performance in medical school was studied extensively [36, 42, 67, 69]. Although a variety of measures of previous academic performance were considered in the studies, according to systematic review of predictors of success in medical school [17], prior academic abilities account for a relatively low percentage of variance (up to 23%) in undergraduate medical student performance.

Among non-academic factors, the most explored are: personality, age, gender and information of administrative nature from admission records. Psychological theories have identified five dimensions, the so-called "*Big* 5" or Five Factor Model, that describe essence of personality: *Neuroticism*, *Extroversion, Openness, Agreeableness* and *Conscientiousness* [8]. Table 5 shows the most important characteristics in each of the 5 dimensions. More detailed information on personality Five Factor Model is provided by Chamorro-Premuzic [8], Lievens et al (2002) [39] Lievens et al (2009) [40]. Some personal traits of "Big 5", such as low levels of conscientiousness [15, 18, 39, 40], high levels of extroversion [15, 39] and of neuroticism [9, 15], have been shown to predispose students to poor academic outcomes.

There are socio-demographic "factors" internationally identified as being associated with failure of undergraduate medical students, such as male gender [10, 42, 67, 69] and non-caucasian ethnicity [42, 67, 69]. Regarding age, some studies provide empirical evidence that being older at entrance is a risk factor for poor performance in medical school [23], others conclude that mature students are more likely to be successful [30] and some find no association between age and academic achievements [10, 67, 69]. Other characteristics, such as family socio-economic sta-

Table 5: Characteristics of five personality dimensions

Factors	**Characteristics**
Extroversion	Warmth, Assertiveness, Humor, Activity, Excitement seeking
Conscientiousness	Competence, Order, Achievement striving, Self-discipline, Deliberation
Openness	Fantasy, Curiosity, Imagination
Agreeableness	Altruism, Trust, Modesty
Neuroticism	Anxiety, Hostility, Depression, Impulsiveness, Vulnerability

tus, parents' level of education and student employment responsibilities, may also be helpful in predicting academic outcomes (see [2, 13, 42, 46]).

Among admission factors, the presence of negative comments in the academic reference [67, 69], low interview scores [36] and the late offer of a place [67, 69] are also recognized as predictors of poor academic achievements in the UK. Other factors suggested in the literature are personal preference for the degree and commitment to the university [46].

Trying to uncover the reasons underlying academic failure, qualitative studies [11, 60] pointed to mental health problems, stress, personal problems and financial concerns. A comprehensive descriptive study from USA (36 medical schools involved) claims that the most prevalent learning difficulties of students are associated with "organizing and integrating large amounts of information" and time management [60].

This brief review of literature allows us to delineate a range of factors internationally recognized to be associated with academic performance of medical students, providing us with guidelines for this type of research in the Portuguese University of Minho.

5.2 The study

5.2.1 Data collection

Since 2006, the SHS-UM develops the longitudinal research project "Evaluating the impact of innovation in Higher Education: implementation and development of a longitudinal study in a medical school" ((FCT- PTDC/ESC/65116/2006), with the main goal of investigating the factors that influence the performance of students and the professional competence of SHS-UM's graduates.

All SHS-UM's first year students are invited to participate in the longitudinal study in an annual briefing session delivered by the Medical Education Unit staff

during freshman welcome week. The project collects multiple data, always with the participants' informed consent.

The constitution of this longitudinal is an ongoing project. It contains information, of diverse nature, about 857 students (representing ten cohorts) and overall, 39.769 observations on 326 variables. The data comprises:

- *socio-demographic variables*, for instance, gender, age at entrance, district of residence prior to enrollment in medical studies, residence during the studies, level of parents education, parents occupation, student civil status, level of education and previous qualifications;
- *admission variables*, for example, pre-university grade point average, regime of admission, preference for university, degree preference, commitment to university for the subsequent years, factors that determine the choice of degree and university;
- *students'perceptions*, for instance, anticipation of difficulties due to enrollment and perceptions about degree program and teaching methods;
- *personality variables* neuroticism, extroversion, openness, agreeableness and conscientiousness;
- *academic performance variables*, that resulted from several types of assessments targeting different aspects of the student performance in medical school, namely written test scores (knowledge assessments), practical tests scores (skills assessments), attitudinal scores (behaviors assessments) and measures of continuous evaluation of professionalism and clinical competence.

Socio-demographic information and students' perceptions about the training program are collected with a home-made survey. Students' perceptions are defined as self-reported expected difficulties that the admission to the medical degree might cause (for instance, financial difficulties, difficulties in interpersonal relationships, in time and stress management, in development of effective learning methods and strategies).

Personality is measured with NEO-FFI inventory. NEO-FFI is a short version of the Portuguese NEO-PI-R questionnaire designed to assess five dimensions of personality: conscientiousness, neuroticism, extroversion, agreeableness and openness to experience. The NEO-FFI inventory contained 60 items, 12 for each of the five dimensions. The item response is measured in a 5-points Likert scale, ranging from 0 (strongly disagree) to 4 (strongly agree). Portuguese version of NEO-FFI is a reliable instrument that reflects the universality of basic dimensions of personality, with the Cronbach's Alphas c 0.69 for openness, 0.8 for neuroticism, 0.74 for extroversion, 0.71 for agreeableness and 0.81 for conscientiousness.

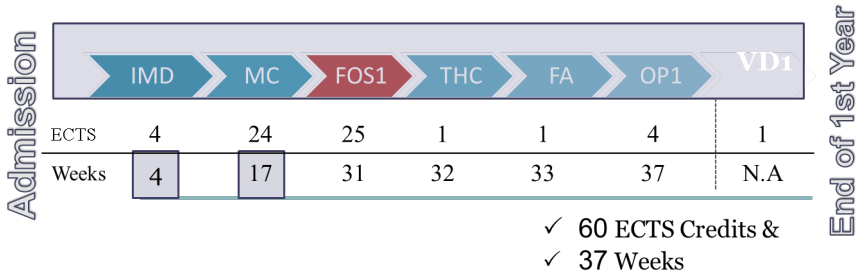
The data this work were extracted from SHS-UM longitudinal database, and consist of individual measures of academic performance in first year courses, pre-

university Grade point average (GPA), as well as non-cognitive and socio-demographic information. Overall, we consider 24 variables, factors that could be of "clinical" importance to explain phenomenon of academic failure among SHS-UM medical students.

5.2.2 First year of medical degree in SHS-UM

The curriculum of SHS-UM is designed in horizontally integrated multidisciplinary courses. The first year corresponds to 60 ECTS and consists of seven courses: one is a year-long course, Vertical Domains 1 (VD1), and the remaining six, Introduction to the Medical Degree (IMD), Molecules and Cells (MC), Functional and Organic Systems 1 (FOS1), Training in a Primary Care Unit (THC), First Aid training (FA) and an elective, Optional Project 1 (OP1), are sequentially distributed along the academic year. General description of the first year courses structure, corresponding ECTS and finishing times (in weeks) are represented in Figure 5.

Figure 5: First year in SHS-UM



In this study, the first three courses are of particular interest. Students start the training with the introductory course, IMD, that is organized around the following themes: learning by modules of objectives methodology, basic laboratory procedures, foundations of biostatistics and the essential molecular mechanisms in biology. After 4 weeks, they take the MC course, that integrates biochemistry and foundations of genetics. After 17 weeks, students start the FOS1 course, that focuses on general organization of the skeletal-muscular system and digestive system.

For the purpose of this study, first year failure was defined in terms of academic performance in the course with the highest failure rates. Students marks range between 0 and 20 (20 is the maximum score) and students fail a course when they scored below 9.5 points. In the years of existence of the medical program at SHS-UM, FOS1, constantly have the highest percentage of failure. Taking 9.5 as cutoff point, we defined a dichotomous variable that takes value 1 if the student fails FOS1 and takes value 0, otherwise.

Our analysis strategy consists of descriptive investigation of available data, univariate analysis and multivariate analysis, that will be described in subsequent sections. The underlying goal of the study is to learn how to predict academic failure very early in medical school. Thus, multivariate analysis methods, such as LR, LDA and K-NN, were used at two distinct instances in time: at admission and after 17 weeks in medical school. Models adjusted at admission are labeled by "Model week 0", or abbreviated as "Model 0". Models adjusted for data available after 17 weeks are labeled by "Model week 17", or simply "Model 17".

5.2.3 Sample characteristics

The study sample consists of 288 first year students from 3 subsequent cohorts who consented to take part in the study. The sample represents 77% of all matriculants during the period under consideration. Of the total of participants, 30% were male. The respondents' ages ranged from 17 to 22 years, with mean age of 18.46 years and a standard deviation of 0.68 years. Of the 288 students considered in the current study, 62 failed the FOS1 course at the first attempt, and 226 passed sucessfully. Table 6 summarizes categorical variables considered in the research.

Table 6: Summary statistics for categorical variables

Variable	Total N (%)	Failure %
Cohort		
1	78(27)	26.9
2	106(37)	16.0
3	104(36)	23.0
Gender		
Male	86(30)	24.4
Female*	202(70)	20.3
Regime of admission		
General*	255(89)	17.6
Special	33(11)	51.5
Preference for University		
1^{st} option*	209(72)	17.2
2^{nd} option	34(12)	44.1
$3^{rd} - 6^{th}$ option	45(16)	24.4
Matriculation[1]		

[1]Whether a participant was enrolled with any high educational institution.

Table 6: Summary statistics for categorical variables

Variable	Total N (%)	Failure %
1^{st} matriculation*	214(74)	20.1
Other	74(26)	25.7
IMD course		
Pass*	257(89)	15.2
Fail	31(11)	74.2
Level of education of mother		
No degree*	144(50)	19.4
High education degree	144(50)	23.6
Level of education of father		
No degree*	175(61)	20.6
High education degree	113(39)	23.0
Father's career$^{\diamond}$		
Higher managerial, administrative*	36(12)	33.3
Intellectual professions	92(32)	23.9
Intermediate managerial, administrative	26(9)	19.2
Clerical and junior managerial	20(7)	25.0
Service and sales workers	41(15)	12.2
Skilled manual workers	37(13)	21.6
Unskilled manual workers	22(7)	26.7
Mother's career$^{\diamond\diamond}$		
Higher managerial, administrative*	21(7)	23.8
Intellectual professions	124(43)	25.8
Intermediate managerial, administrative	13	23.1
Clerical and junior managerial	42(15)	14.3
Service and sales workers	22(8)	13.6
Skilled manual workers	26(9)	19.2
Unskilled manual workers	17(6)	29.2
Change of residence at entry: leaving home		
Yes	144(50)	27.7
No*	144(50)	15.3
AD:[2] effective learning		
Yes	93(32)	15.1

[2]AD: anticipation of difficulties.

Table 6: Summary statistics for categorical variables

Variable	Total N (%)	Failure %
No*	195(68)	24.6
AD: time management		
Yes	231(80)	16.7
No*	57(20)	36.8
AD: family relationship		
Yes	41(14)	26.8
No*	247(86)	20.6
AD: financial		
Yes	47(16)	21.3
No*	241(84)	21.6
AD: physiological (anxiety, loneliness)		
Yes	57(20)	17.5
No*	231(80)	22.5

Note: * Default category for logistic regression analysis (coded as 0);
$^{\diamond}$ Missing observations for 14 participants;
$^{\diamond\diamond}$ Missing observations for 24 participants;

Table 6 indicates that rate of failure among male students is higher than among female, 24.4% vs. 20.3%. Different failure rates are observed across different admission groups. Students admitted through the general national process are more successful in their first year in medical school: in this group the failure rate is 17.6%; in contrast, those admitted through the special system experienced much more difficulties in their first year of training program (51.5% of these students failed FOS1 course).

Additionally, failure rates vary with students' preferences for the degree. During the application to the university, candidates have the possibility to run, simultaneously, for several degrees, indicating the order of preference on a scale from 1 to 6. For this study we make a clear distinction between three groups of students: the reference group is formed by those who chose the SHS-UM as their first option, the other group consists of those who indicated the school as second option and the third group puts together all other students. For the reference group, the rate of failure is 17.2%. Table 6 shows that the proportion of underperforming students is different for the 2nd and the 3rd group: the proportion of underperformers is about 44% and 24%, respectively.

Table 6 illustrates that the failure rate is remarkably higher among those who

had history of failure, i.e. those who failed the IMD course: over 74% of students that failed IMD course failed the FOS1 course too.

The summary statistics (mean and standard deviation) for collected quantitative covariates are displayed in Table 7.

Table 7: Summary statistics for qualitative variables

	Failure		Success	
	Mean	SD	Mean	SD
Conscientiousness	29.74	6.49	33.99	6.10
Neuroticism	25.38	7.62	23.79	8.28
Extroversion	31.27	5.42	31.48	5.55
Openness	30.11	5.58	30.24	5.48
Agreeableness	33.45	5.31	32.77	5.80
MC score	8.09	3.79	12.47	1.76
GPA	176.86	12.78	184.15	6.19
Age	18.43	0.79	18.61	0.69

We notice that underperforming students have average lower GPA grades. Turning to personality traits, underperforming students scored on average 25.38 points of Neuroticism while successful students scored on average 23.79 points. In contrast, scores for Conscientiousness were higher for successful students, with average 33.99, while the average for the underperformers group was 29.74. We notice also that failing students were slightly older than successful ones.

5.3 Univariate analysis

To investigate marginal effects of the 24 potential predictors extracted from the longitudinal database and to determine which variables should be involved in the multivariate modeling, we carried out univariate analysis. The analysis consisted of Pearson Chi-square and Mann-Whitney-Wilcoxon tests. To avoid the risk of losing any relevant information at this stage, we did not use the traditional significance levels, but decided to use a higher significance level of < 0.4. The choice of Mann-Whitney-Wilcoxon test, instead of the parametric t-test, was motivated by the fact that distributional assumptions were not assured for variables in consideration. The results of Pearson Chi-square and Mann-Whitney-Wilcoxon tests are summarized in Tables 8 and 9.

According to the univariate analysis, the factors with statistically significant ($p-value < 0.4$ impact on students performance in the first year course with highest failure rates (FOS1) comprised: Conscientiousness, GPA, Neuroticism, grades on MC course, change of residence at entry, preference for degree, matriculation, regime of admission, level of education of mother, pass/fail classification on IMD course,

Table 8: Chi-square test results

Variable	$X^2(df)$	$p - value$
Gender	0.607 (1)	0.436
Change of residence at entry	6.659(1)	0.010
Preference for University	12.618(2)	0.002
Matriculation	1.012(1)	0.314
Regime of admission	19.839	0.000
Father's career	5.541(6)	0.477
Mother's career	4.304(6)	0.636
Level of education of mother	0.740(1)	0.390
Level of education of father	0.242(1)	0.623
IMD course	57.037(1)	0.000
AD: time management	9.866(1)	0.002
AD: family relationship	0.795(1)	0.372
AD: effective learning	3.408 (1)	0.065
AD: financial	0.002(1)	0.963
AD: physiological	1.054(1)	0.305
Cohort	3.38 (2)	0.184

Table 9: Mann-Witney-Wilcoxon test results

Variable	Z	$p - value$
Conscientiousness	4.59	< 0.001
GPA	5.15	< 0.001
Agreeableness	0.666	0.505
Extroversion	-0.179	0.858
Openness	-0.035	0.972
Neuroticism	-1.217	0.224
Age	-1.622	0.105
MC score	9.391	< 0.001

anticipation of difficulties in time management, family relationship, effective learning and psychological, and cohort. Level of education of father, parental career, expected financial difficulties and health problems, Agreeableness, Extroversion and Openness did not exhibit univariate statistical significance ($p-value > 0.4$), thus, mentioned factors were excluded from further multivariate modeling. We opt to test the effect of variable Gender in a multivariate analysis because of consistent evidence of its "clinical" importance provided by international research in medical education. Thus, the total number of variables candidates was 17.

5.4 Results of logistic regression

This section is dedicated to the presentation of results obtained in the LR modeling.

Several univariate and multiple models were fitted. Two distinct methods (stepwise selection and best subset selection) were employed to select the best model for both instances in time (admission and after 17 weeks).

Application of stepwise selection method, with likelihood ratio test as selection criterion and ($p-value < 0.05$) as stopping criterion, to the subset of data formed exclusively by pre-admission factors yielded a model containing 3 explanatory variables: Conscientiousness, GPA and AD: time management. The next step of model building procedure was to test statistical significance of Age and Gender, was examined, since these factors are internationally recognized as important for prediction of medical student performance. Rerunning the regression with Gender included, likelihood ratio criteria found no evidence that the model can be improved using this independent variable ($G(1) = 0.29; p-value = 0.59$). Including Age as predictor we obtained $G(1) = 3.8$ with corresponding $p-value = 0.054$. Accepting statistical significance of predictor at the level of 10%, we included variable Age in the model. The resultant main effect model is described in Table 10

Table 10: Logistic regression: Model week 0

	$\hat{\beta}$	SE	Z_W	$p-value$	95% CI	
Conscientiousness	-0.115	0.03	-4.455	0.000	-0.166,	-0.065
GPA	-0.089	0.02	-4.731	0.000	-0.126,	-0.052
AD: time management	-1.151	0.37	-3.072	0.002	-1.885,	-0.417
Age	0.415	0.215	1.93	0.053	-0.006,	0.836
Constant	11.757	5.13	2.292	0.022	1.702,	21.812

Employing *glmulti* function [6], we performed exhaustive screening of all subsets of data formed by up to 6 explanatory variables. Based on AIC, we obtained the 10 best subsets of covariates. The subset that provided model with the lowest

value of AIC (AIC= 245.5) contains: Conscientiousness, GPA, AD: time management, Age, Preference for university and AD:effective learning. However, the model described by Table m0 was also in top 10 (with $AIC = 247,2$). Since Taking difference in values of information criterion for two models is relatively small and these two models are nested, we performed the likelihood ratio test for comparison. For the model described in Table 10, the value of the log-likelihood is -118.607. The log-likelihood for the "best subset" model is -116.759. Thus, the value of the likelihood ratio test statistic is $G(3) = -2[(-118.606) - (-116.759))] = 3.694$ and $p - value = 0.29$, which is not significant at the 0.05 level. Thus, we preferred the parsimonious model resulting from the combination of variable selection based on univariate analysis and stepwise procedure (given by Table 10).

For the second instance in time (after 17 weeks), stepwise algorithm, based on the likelihood ratio test with stopping criterion $p-value = 0.05$, and the best subset selection, based on AIC with constraints for model complexity (maximum number of parameters less or equal to 6), yield the same model equation. The inclusion of the additional predictors, Age and Gender, provided no improvement (likelihood ratio test statistic was $G(2) = 0.67$ with $p-value = 0.715$). Hence, we preferred the parsimonious model resulting from the best subset selection and stepwise procedure.

Based on the literature of educational research and in our univariate analysis, we believe that there is a possibility of interaction between two categorical predictors in our model: Change of residence and AD: family relationship. To test such possibility, we included in the model the interaction term obtained by multiplying AD: family relationship by Change of residence. Testing the hypothesis that the regression coefficient for interaction is equal to zero, we obtained $Z_W = -1.07$ and $p - value = 0.283$. When interaction term was added, the associated change in the model deviance was $G = (163.96 - 162.94) = 1.02$, leading to a non-significant $\chi^2(1)$ value ($p - value = 0.313$). Hence, we concluded that there is no evidence of interaction and we return to main effect model obtained previously. Estimates of parameters, standard errors, Wald test statistic Z_W, corresponding p-values and 95% confidence intervals for $\hat{\beta}$'s of the final model are shown in Table 11.

The next step of the modeling process was to check linearity of the covariates Conscientiousness and GPA (for Model 0) and MC score and Conscientiousness (for Model 17).

To investigate linearity in logit we used two graphical methods (a univariate smoothed scatter plot on the logit scale and plot of regression coefficients for dummy variables) and one analytical test (method of fractional polynomials).

The lowess smoothed logit plots and plots resultant from dummy variables analysis for the model described in Table 10 are shown in Figures 13 and 15 in Appendix A. The lowess smoothed logit plots and results of dummy variables analysis

Table 11: Logistic regression: Model week 17

	$\hat{\beta}$	SE	Z_W	$p-value$	95%	CI
Conscientiousness	-0.109	0.03	-3.372	0.001	-0.172,	-0.046
MC score	-0.638	0.11	-5.998	0.000	-0.847,	-0.430
IMD course	2.007	0.70	2.858	0.004	0.631,	3.383
AD: family relationship	-1.791	0.73	-2.467	0.014	-3.214,	-0.368
Change of residence	0.908	0.44	2.087	0.037	0.055,	1.761
AD: time management	-1.180	0.48	-2.484	0.013	-2.112,	-0.249
Constant	9.589	1.78	5.385	0.000	6.099,	13.079

for model given by Table 11 are shown in Figures 14 and 16 in Appendix A. In the lowess smoothed logit plots for all covariates in both models, we observed an S-shaped curve.

Dummy variables were created according the distributional quartiles of Conscientiousness, GPA and MC score. The first quartile was considered to be a reference group with null regression coefficient. For variables Conscientiousness and GPA, we observed a monotone decrease of LR coefficients that support treating mentioned variables as linear in logit. For variable MC score it was not possible to apply the technique due to numerical problems in the estimation of regression coefficients of dummy variables.

Finally, we present a detailed description of fractional polynomial method for the two models (Model 0 and Model 17). Recalling that the set of powers considered in this research is $\{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ where 0 corresponds a logarithmic transformation of covariate, for each of covariates, we built 8 fractional polynomial models of the 1st-order, 45 fractional polynomial models of the 2nd-order and the linear model.

- Model 0

Regarding variable GPA we reached the following conclusion: none of the fractional polynomials transformations of covariate was significantly better than the linear model. In particular, the best 1st-order nonlinear model was the one with cubic transformation of covariate (the corresponding test statistic (2.17) was 0.799 and $p-value = 0.371$). For test of the best 1st-order nonlinear model versus the best 2nd-order nonlinear model, we obtained for test statistic (2.18) value 0.985 and $p-value = 0.419$. For the test of linear model versus the best 2nd-order model, the test statistic (2.19) was 1.784 with $p-value = 0.409$.

Application of fractional polynomials method for Conscientiousness yielded the same conclusion. The best 1st-order fractional polynomials model contained $Conscientiousness^2$, the best 2nd-order model included $Conscientiousness$ and $Conscientiousness^{-3}$, however in three tests we have not rejected the null hypothesis

($p-values$ for three likelihood ratio tests were $0.840, 0.403$ and 0.593, respectively).

- Model 17

For variable Conscientiousness the best 1st-order fractional polynomials model contained $Conscientiousness^3$, that achieved a deviance 160.743. The best 2nd-order model included $Conscientiousness^3$ and $Conscientiousness^{-3}$, correspondent deviance was 158.84. The likelihood ratio test statistic (2.18) was 1.9 with $p-value = 0.387$. Thus, we conclude that between the two models, the 1st-order one should be chosen. Consulting the p-value of the partial likelihood ratio test, we found that the best 1st-order non-linear model was significantly different from the linear at the level of 0.1. However the improvement that the model with cubic transformation provided over the linear model was insufficient to proceed with non-linear transformation.

For variable MC score the best of fitted 2nd-order fractional polynomials models, that contained $(MC\ score)^3$ and $(MC\ score)^{-3}$, reached deviance equal to 159.445. For the best 1st-order model (that included $(MC\ score)^3$)) deviance was equal to 160.715. Consulting the corresponding p-values of likelihood ratio tests, we conclude that the covariate MC score should be treated as linear in logit.

Summarizing, we conclude that the fractional polynomials approach and the plots support the decision to treat variables as linear in the logit.

To assess model adequacy, we performed a combination of four goodness-of-fit test, as recommended in literature [26]. In our case, the number of covariate patterns, J, is approximately equal to the sample size $n = 288$, since continuous predictors were used. More precisely, for the set of explanatory variables in Model 0 $J = 280$, and for the set of explanatory variables in Model 17 $J = 285$. The goodness-of-fit test statistics and corresponding $p-values$ are reported in Table 12.

Table 12: Assessment of models goodness-of-fit

	Model 0		Model 17	
Test	Statistic (df)	p-value	Statistic (df)	p-value
Pearson X^2	279.13(275)	0.4527	201.16 (278)	0.999
Hosmer-Lemeshow C	9.85(8)	0.276	8.38(8)	0.397
Osius-Rojek Z	0.044	0.964	-0.191	0.848
Strukel test	0.05(2)	0.9769	3.48(2)	0.175

Literature warns that Pearson goodness-of-fit statistic may give misleading results if the number of distinct patterns is large. However, given that the results of four tests agree, we have no statistically significant evidence against a satisfactory model fit. Hence, we expect to have few covariate patterns with poor fit.

First we present summary statistics for the so-called basic building blocks measures of LR diagnostic and the diagnostic plots discussed in Section 2.6 (plot of leverage versus Pearson residual, plot of ΔD versus estimated probabilities and plot of $\Delta\beta$ versus estimated probabilities), for models in Table 10 and in Table 11. Despite the fact that are no fixed cut points that may be used to identify

Table 13: Model 0: summary statistics for basic building block diagnostic measures

	Min	Max	Mean	SD
Deviance residuals	-1.88	2.45	-0.14	0.9
Pearson residuals	-2.20	4.38	-0.001	0.99
Leverage	0.0006	0.1632	0.024	0.029

an exceptionally large value of residuals in LR, the values of percentiles of $N(0,1)$ distribution may provide some guidance to assess whether residuals are large. For Model 0, Pearson residuals appear with magnitude less than -1.96 or greater than 1.96, deviance residuals have relatively lower magnitude, but some exceed 1.96, which definitely deserves closer inspection.

Table 14: Model 17: summary statistics for basic building block diagnostic measures

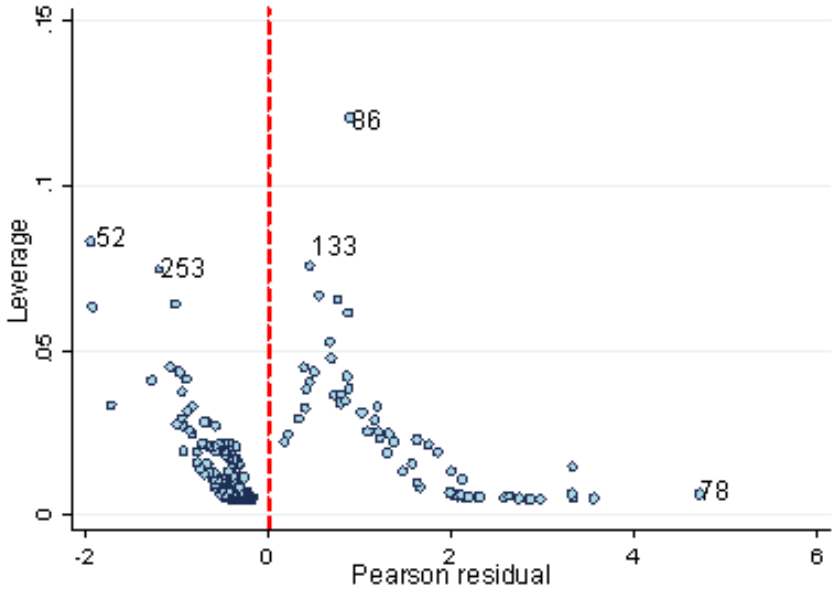
	Min	Max	Mean	SD
Deviance residuals	-2.42	2.48	-0.01	0.74
Pearson residuals	-4.2	4.55	-0.02	0.84
Leverage	0.0043	0.1335	0.0183	0.019

In Model 17 we note that both Pearson and deviance residuals for some covariate patterns are outside interval $[-1.96,\ 1,96]$ and should be examined in detail.

From the graph displayed in Figure 6 we identified subjects within covariate patterns with large residuals (Model 0: 62, 253, 78; Model 17: 76, 6, 62) and with relatively high values of leverage (Model 0:160 2, 260; Model 17: 287).

Figure 7 shows the plots of ΔD's versus estimated probabilities. According to Hosmer & Lemeshow, the value of quantile 0.95 of the $\chi^2(1)$ distribution, that is 3.84, may provide some guidance to assess whether a value of ΔD for a particular covariate pattern is large. Remembering that ΔD is diagnostic statistics that measures of the effect of covariate pattern deletion on the model fit, we note that, for model in Table 10, elimination of subjects 78, 149, 76 49 and 219 can result in improvement of model fit, since the corresponding values of ΔD exceed 3.84. Regarding Model 17 subjects 76, 6, 150 and 62 are problematic, since corresponding ΔD points lie in the top corners of the graph. Except mentioned subjects, the plot illustrate that model fits quite well.

Figure 6: Plot of Leverage vs Pearson residual
Model 0



Model 17

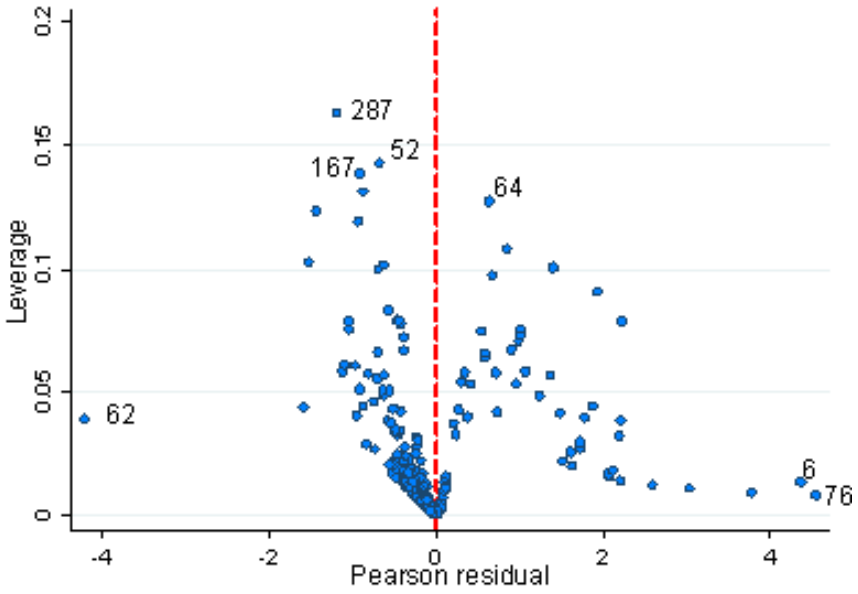
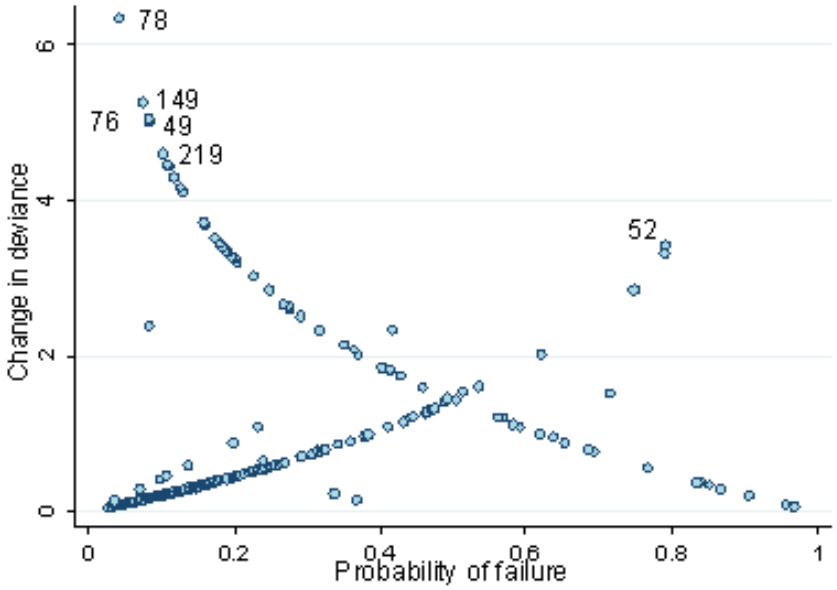


Figure 7: Plot of ΔD vs estimated logistic probability

Model 0



Model 17

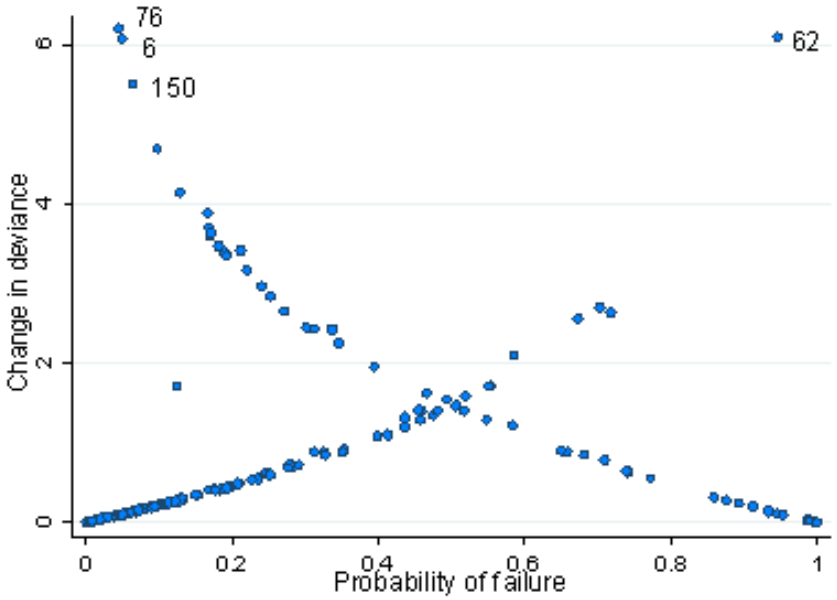
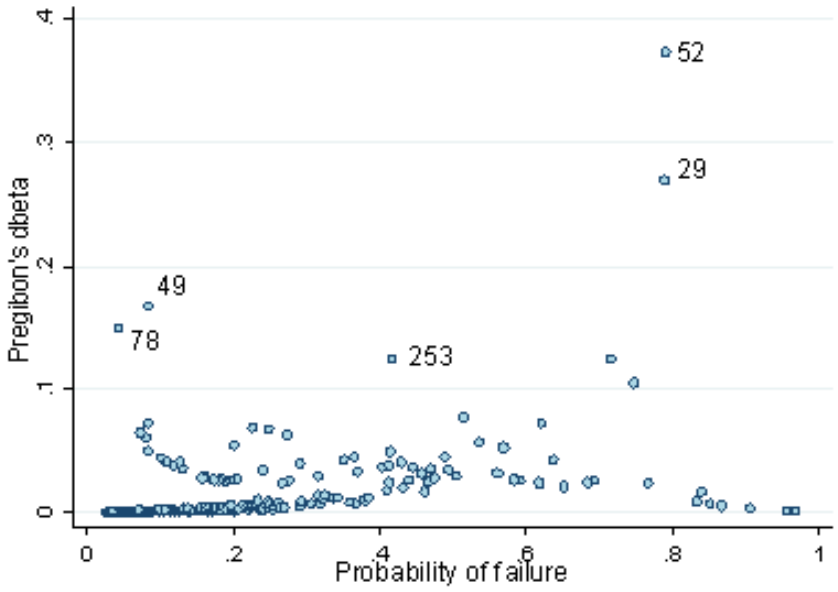


Figure 8: Plot of $\Delta\beta$ vs estimated logistic probability

Model 0



Model 17

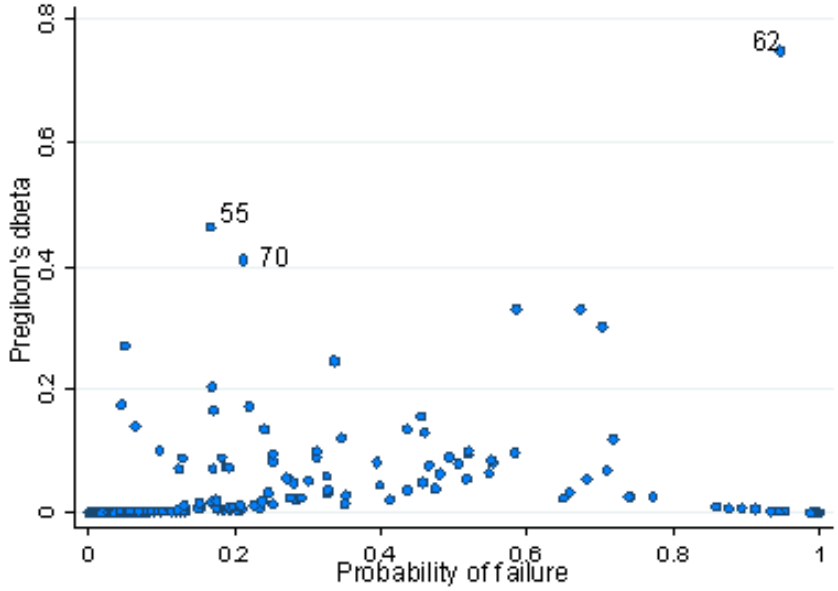


Figure 8 shows the plots of $\Delta\beta$ versus estimated probability. We observed that, for Model 0, the points corresponding to subjects 52 and 29 are located a bit away from the others. The largest value, $\Delta\beta = 0.4$, is associated with observation 52. The remaining values belong to interval $[0, 0.3]$. For Model 17 we noticed that all $\Delta\beta$´s were below 0.75. Three points, corresponding to observations 62, 55 and 70, are positioned far from the others, and so these subjects have the greatest influence on values of estimated parameters of Model 17. According to Hosmer & Lemeshow [26], to have a notable effect on parameters estimates the covariate patterns must have $\Delta\beta$ larger than 1.

Table 15 summarizes diagnostic measures for 11 "outlying" covariate patterns identified in Model 0. Note, that each problematic covariate pattern detected in

Table 15: LR diagnostic measures for problematic subjects under Model 0

Sub	j	$\hat{\pi}_j$	$\hat{r_j}$	$\hat{d}_j$	$\hat{h}_j$	ΔX_j^2	ΔD_j	$\Delta\hat{\beta}_j$
78	278	0.0494	4.399	2.451	0.0078	19.35	6.06	0.153
149	155	0.0585	4.022	2.382	0.0053	16.18	5.71	0.087
76	203	0.0658	3.776	2.332	0.0058	14.26	5.47	0.083
219	225	0.0664	3.759	2.328	0.0067	14.13	5.46	0.097
206	150	0.0822	3.350	2.235	0.0056	11.22	5.03	0.064
49	273	0.0921	3.166	2.184	0.0168	10.02	4.85	0.172
13	201	0.0981	3.043	2.155	0.0067	9.25	4.68	0.063
241	140	0.1065	2.906	2.116	0.0062	8.44	4.51	0.052
276	148	0.1292	2.605	2.023	0.0064	6.78	4.12	0.044
267	86	0.1314	2.579	2.015	0.0066	6.65	4.09	0.045
282	212	0.1429	2.488	1.972	0.0311	6.18	4.02	0.19

Sub indicates the position of student in the sample
j indicates the position in the covariate patterns list

Model 0 contains observations referred to a single subject, that belongs to the failing group ($Y = 1$). Table 15 indicates that all identified subjects, except one, had relatively low value of leverage, $\hat{h}_j$ (below the average value for the sample) and also relatively low value of $\Delta\hat{\beta}_j$ statistic. Hence, we can conclude that the observations under consideration had no strong influence on parameters estimate and their influence is mainly due to lack of fit.

Deleting from the sample the 11 subjects with the largest residuals, resulted in improvement of the fit of the model. However, deletion reduces the size of the group of failing students by 17%. Additionally, identified "outlying" covariate patterns did not have strong influence on estimated parameters. Taking into account the arguments stated above, we decided to retain all subjects in the analysis.

For the Model 17, based on regression diagnostic measures, we identified 5 problematic covariate patterns (each one containing observation of a single subject),

5 of them with $Y = 1$. Table 16 displays summary of diagnostic statistics for problematic observations.

Table 16: LR diagnostic measures for problematic subjects under Model 17

Sub	j	$\hat{\pi}_j$	$\hat{r_j}$	$\hat{d}_j$	$\hat{h}_j$	$\Delta\chi^2_j$	ΔD_j	$\Delta\hat{\beta}_j$
6	225	0.049	4.399	2.449	0.0138	19.34	6.083	0.271
21	144	0.098	3.055	2.157	0.0107	9.33	4.702	0.101
76	202	0.046	4.569	2.481	0.0083	20.87	6.207	0.176
149	143	0.129	2.615	2.024	0.0125	6.84	4.148	0.086
150	69	0.065	3.805	2.337	0.0096	14.47	5.513	0.139
62	278	0.947	-4.291	-2.420	0.0391	18.41	6.094	0.748

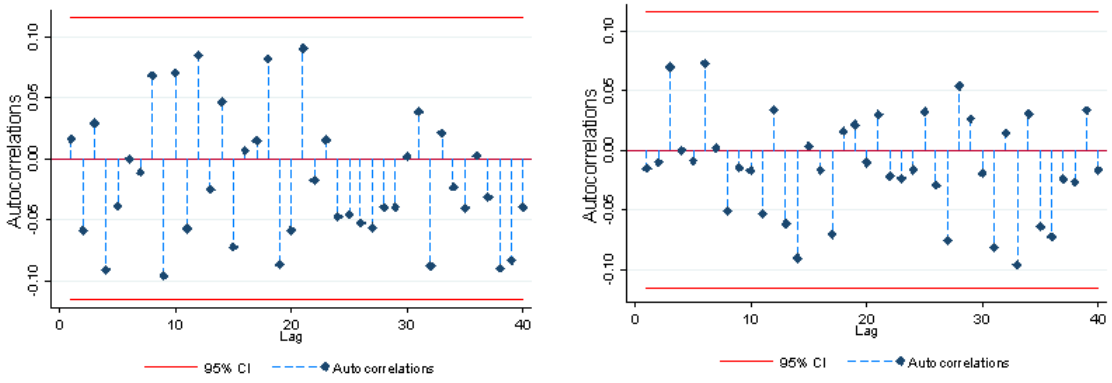
We note that 5 of these subjects had relatively low leverage and low value of $\Delta\beta_j$, indicating that the influence is ma due to lack of fit. Subject 62 has high negative values of residuals, relatively high leverage value and the highest value of $\Delta\beta_j$.

Given that few subjects had large values of diagnostic statistics and that all these observations except one were associated with the outcome "failure" we proceeded in both models with the analysis without deleting any subjects from the sample.

In order to check LR errors for serial correlations, graphical and analytical methods were employed. Figure 9 displays the sample autocorrelation for the stan-

Figure 9: Autocorrelation plots

Model 0 Model 17



dardized residuals of the two models, until lag 40. The red horizontal lines in Figure 9 (plotted at $-1.96/\sqrt{288} = -0.115$ and $1.96/\sqrt{288} = 0.115$) provide critical values for the test of whether or not the autocorrelation coefficients are significantly different from zero. Given that all values of the sample autocorrelation are within the horizontal red lines it is reasonable to infer that error terms are uncorrelated.

For Model 0, when performing the tests for random order described in Section 2.4.2, we obtained $Z = 1.18$ and $p - value = 0.24$. Since $p - value$ is not significant at the 5% level, we do not reject the null hypothesis of independence of errors. For Model 17, the same tests provided $Z = 0.96$ and $p - value = 0.34$. Hence, based on the analysis of the autocorrelation plots and of the runs tests, we conclude that in both models errors should be treated as random sequences of observations.

Turning to prediction accuracy, we computed, for both models, 3 measures based on classification tables as described in Section 2.7. In this study, Sensitivity is the ability of model to correctly predict failure and Specificity is the ability to correctly predict success.

For Model 0 the resulting classification is shown in Table 17. The measures of

Table 17: Classification table for LR: Model 0

		Observed	
		Y=1	Y=0
Classified	Y=1	37	43
	Y=0	25	183
Column total		62	226

prediction accuracy obtained were:

$$Sensitivity = \frac{37}{37 + 25} = 0.597;$$

$$Specificity = \frac{183}{29 + 183} = 0.809;$$

$$Count\ R^2 = \frac{37 + 183}{288} = 0.7638;$$

meaning that 59.7% of failing students were correctly classified by the model and the percentage of correct classification for the group of sucessful students, 80.9%, was remarcably high.

For Model 17 the resulting classification is shown in Table 18. The overall rate

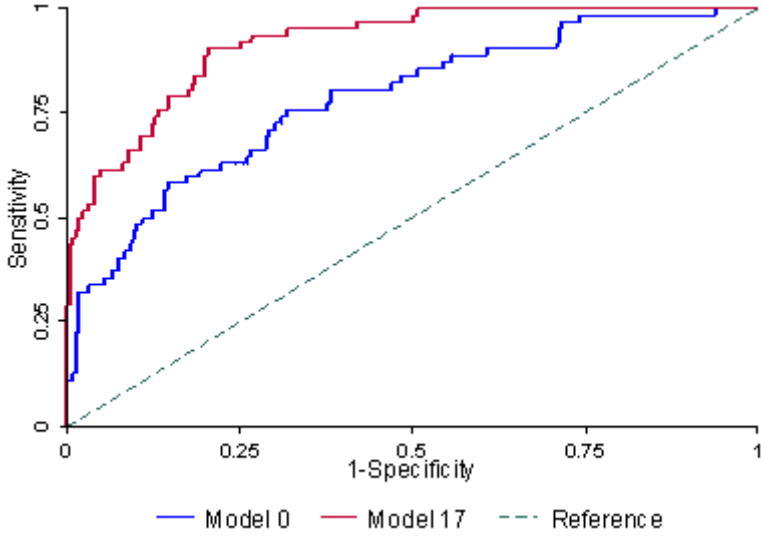
Table 18: Classification table for LR: Model 17

		Observed	
		Y=1	Y=0
Classified	Y=1	46	21
	Y=0	16	197

of correct classification ($Count\ R^2$) is $100[(46 + 197)/288]\% = 84.4\%$, Sensitivity is $100[46/(46 + 16)]\% = 74.19\%$ and Specificity is $100[197/(197 + 21)]\% = 87.61\%$.

The Figure 10 presents the ROC curves for the two models. Model 0 had AUC

Figure 10: Comparison of ROC curves for two LR models



of 0.7820, with 95% CI equal to $]0.7156, 0.8483[$, and Model 17 had greater AUC, of 0.92, with 95% CI equal to $]0.8807, 0.9541[$.

According to the guidelines of assessment of model's prediction accuracy via area under ROC curve, proposed by Kleinbaum & Klein [32], Model 0 provides "fair discrimination", while Model 17 provides "excellent discrimination". We carried out the test for the hypothesis that the difference in area under the ROC curve for two non-nested models is equal to zero, and obtained for test statistic (2.27),

$$Z = -4.26, \quad p - value < 0.001.$$

Since the p-value is extremely significant, we conclude that two models have different predictive ability.

Do decide if predictions provided by fitted models are better than the ones obtained by chance, we turn to the "by chance classification". To determine the chance rate, we applied the proportional chance criterion, because in this work we are mainly interested in maximizing rate of correct classifications for the smaller group of failing students. For our sample, the estimated prior probabilities of group membership are 0.215 for the group of failing students and 0.785 for the group of successful students. According to the criteria described previously, we get:

- the chance frequency of correct predictions for group of failing students is $e_{failure} = 0.215 \cdot 62 = 13.33$;
- the chance frequency of correct predictions for group of successful students is $e_{success} = 0.785 \cdot 226 = 177.41$;
- total-group chance frequency of correct predictions is $e = 13.33 + 177.41 = 190.74$;
- overall expected rate of correct predictions is

$$H_e = \frac{1}{288}\left(0.215 \cdot 62 + 0.785 \cdot 226)\right) = \frac{190.74}{288} = 0.662$$

Hence, 66.2% of students from our sample could be correctly classified by chance.

From Table 17 we compute, for Model 0, the total observed frequency of hits

$$o = 37 + 183 = 220$$

and the value of overall statistic Z (4.5)

$$Z = (220 - 190.74)/\sqrt{190.74 \cdot (288 - 190.74)/288} = 3.65.$$

The lower bound for a 99% confidence interval for the true frequency of total-group correct classifications is $220 - 2.326 \cdot 8.02 = 201.35$. For individual groups we get the following results:

- for failing students,

$$Z_f = \frac{37 - 13.33}{\sqrt{13.33 \cdot (62 - 13.33)/62}} = 7.31,$$

  with the lower bound for a 99% confidence interval for the true frequency of correct classification equals to $37 - 2.326 \cdot 3.23 = 29.5$.

- for successful students,

$$Z_s = \frac{183 - 177.41}{\sqrt{177.41 \cdot (226 - 177.41)/226}} = 0.905,$$

  with the lower bound for the 99% confidence interval for the true frequency of correct classifications equals to $183 - 2.326 \cdot 6.17 = 168.6$

Results indicate that the rate of correct classifications for the group of failing students and the total rate of correct classifications are slightly better than expected by chance rate.

For Model 17, from Table 18, we have $o = 46 + 197 = 243$ and the test statistic (4.5)

$$Z = (243 - 190.74)/\sqrt{(190.74 \cdot (288 - 190.74)/288)} = 6.51, \quad p - value < 0.01,$$

which clearly indicates a better than chance result. The lower bound for the 99% confidence interval for the true frequency of total-group correct classifications is $243 - 2.326 \cdot 8.02 = 224.35$. For the separate groups the results are the following In

Group	n_{jj}	n_j	e_j	Z_{G_i}	p-value
Failure	46	62	13.34	10.1	< 0.001
Success	197	226	177.41	3.17	< 0.001

this case, rate of correct classification for the group of failing students, as well as the rate of correct classification for the group of successful students, is better than what may be expected by chance. In other words, information given by the five predictors enable us to classify students into failing and successful groups statistically better than by chance.

Finally, we present interpretation of the LR results of Model 17, which has higher prediction accuracy.

Table 19: Odds ratios for LR: Model week 17

	β	OR
Conscientiousness	-0.109	0.90
MC score	-0.638	0.52
IMD course	2.007	7.44
AD: family relationship	-1.791	0.17
Change of residence	0.908	2.48
AD: time management	-1.180	0.31

According to Table 19 for 3 additional points (correspond to approximate value of standard deviation) in the MC score estimated odds of failure are expected to change by factor of 0.25, if all other variables in the model are kept constant. On the other hand, the odds of failure are 7.44 times larger for students that failed IMD course. Living away from home (change of residence at entry) increases the odds of failure by a factor of 2.48. For students who anticipate difficulties in family relationship due to the enrollment in the medical degree, the odds of failure are 0.17 times smaller. In the same way, for students who anticipate difficulties in time management the odds of failure are 0.31 times smaller. A standard deviation increase in Conscientiousness (6.42 points) is expected to change the odds of failure by factor 0.55.

5.5 Results of Linear discriminant analysis

As mentioned before, our data set contain mixture of quantitative e qualitative variables. Therefore, we do not expected to meet the assumption of multivariate normality required for the use of LDA. Evaluating the distribution of continuous variables across groups, we found that the skewness coefficient ranged from -1.83 to 0.09, the kurtosis coefficient ranged from 2.4 to 5.48 and variables GPA and MC score, exhibited considerable departure from normality in both groups.

To test the assumption of homogeneity of covariance matrices we used the M-Box test. We should recall that this test is quite sensitive to multivariate non-normality. Hence, for our sample, there is very strong possibility to reject the

null hypothesis. Nevertheless, the M-Box test was routinely used for testing the homogeneity of covariance matrices across groups. The logarithms of determinants of covariance matrices were also determined. For the model that explored pre-enrollment factors exclusively (Model 0), M-Box test statistic $X^2(6) = 73.03$ yield $p-value < 0.001$. However the log-determinants of covariance matrices were similar: 5.23 and 7.35 for the large group and small group, respectively. For the model obtained with data available at week 17 (Model 17), M-Box test statistic $X^2(15) = 183.72$ and corresponding p-value indicated that there is strong evidence in the sample to reject the null hypothesis of equality of covariance matrices across groups. The values of log-determinants of covariance matrices were clearly not in the same: -2.89 for the large group and 1.23 for the small group.

Several empirical studies [29, 35, 65] have reported robustness of LDA to violation of assumptions 6 and 7 mentioned previously in Section 3.1.1. Hence, although the required assumption are not satisfied, we proceed with application of LDA.

To formulate a classification rule, equal prior probabilities of group membership were used. A choice of equal prior probabilities was based on the ideas of Huberty [28], that with groups of unequal sizes use of unequal priors increases the hit rates for the larger group and decreases the hit rates for the smaller group. Ferrer & Wang [19] give additional support to Huberty [28] under conditions very similar to ours (namely with a real data set of size $n = 244$ characterized by departures from multivariate normality, unbalanced group size and heterogeneity of covariance matrices).

Like in LR, the stepwise method was employed to select the best set of discriminating variables to appear in the discriminant function. A stepwise discriminant analysis was performed based on following criteria:

- Wilks Lambda statistic used as selection criterion, i.e in each step a variable was added or removed from discriminant function according to the value of Λ_W;
- a probability levels 0.05 and 0.10 were used for entering and removing of variables, respectively.

For the subset of data formed exclusively by pre-enrollment factors (Model 0), the first step of stepwise procedure was to include variable GPA in the discriminant function, since it provided the maximum separation of two groups according to the selection criterion $\Lambda_W = 0.878, F(1, 286) = 39.77, p - value < 0.001$. At step 2, variable Conscientiousness entered because it minimized the overall Wilk's lambda: $\Lambda_W = 0.817, F(1, 285) = 31.82, p - value < 0.001$. After step 2 non of the variables was removed from the discriminant function. Then, variable AD: time management was eligible for inclusion in the discriminant function with overall $\Lambda_W = 0.785, F(1, 284) = 25.94, p - value < 0.001$. Again, none of variables was

deleted form the model and, of the variables that were not in the model so far, none was a candidate for inclusion. Thus, the stepwise procedure terminated and the resultant discriminant function included 3 variables, namely: GPA, Conscientiousness and AD:time management. Canonical correlation and the Wilks lambda statistic Λ_W for obtained discriminant function are reported in Table 20. Λ_W is significant

Table 20: Linear discriminant analysis: Model week 0

Canonical correlation	Eigenvalue	Λ_W	X^2	df	p-value
0.464	0.274	0.785	68.89	3	0.000

at 0.001 level. The squared canonical correlation, 0.215, indicates that about 22% of the variation between groups of failing and successful students is accounted for by these three discriminating variables.

The standardized and unstandardized discriminant function coefficients, as well as structure coefficients, are given in Table 21. The absolute value of the standardized coefficient u^s provides the index of the importance of predictors for the group separation. The greater in absolute value the standardized coefficient is, the greater is the relative importance of corresponding variable for the group separation. The sign of standardized coefficient indicates the direction of the relation and whether the contribution of variable is positive or negative. The structure coefficients measure the importance of contribution of variables to the discriminant function. Thus, using the standardized coefficients we infer that the separation

Table 21: Model week 0: Standardized, Unstandardized and Structure coefficients of linear discriminant function

Variable	u^s	u	l
Conscientiousness	0.597	0.096	0.541
GPA	0.732	0.091	0.712
AD:time management	0.432	1.100	0.360
Constant	-	-20.637	

u^s=Standardized coefficient; u=Unstandardized coefficient; l=Structure coefficient

among two groups may be attributed mainly to variable pre-university GPA, that measures cognitive abilities. Conscientiousness score was the second strongest predictor while AD: time management contributed less for allocation of subjects to the failing or to the successful group. The analysis of structure coefficients leads to the same conclusion: the discriminant function is most closely related to variable GPA.

The unstandardized coefficients u are used to calculate individual score on the discriminant function and to describe each one of the groups in terms of its profile. Let consider one participant who scored 28 points for personality dimension of

Conscientiousness, had GPA of 160.5 and did not anticipate difficulties due to enrollment, that is AD: time management= 0, the discriminant score for this particular individual is

$$-20.673 + 0.096 \cdot 28 + 0.091 \cdot 160.5 + 1.1 \cdot 0 = -3.38. \tag{5.1}$$

Using the vectors of group means of the discriminating variables ((33.99, 184.15, 0.84) for the successful group and (29.74, 176.86, 0.66) for the failing group), we determine the location of the group centroids: for the group of successful students (the larger group) we have

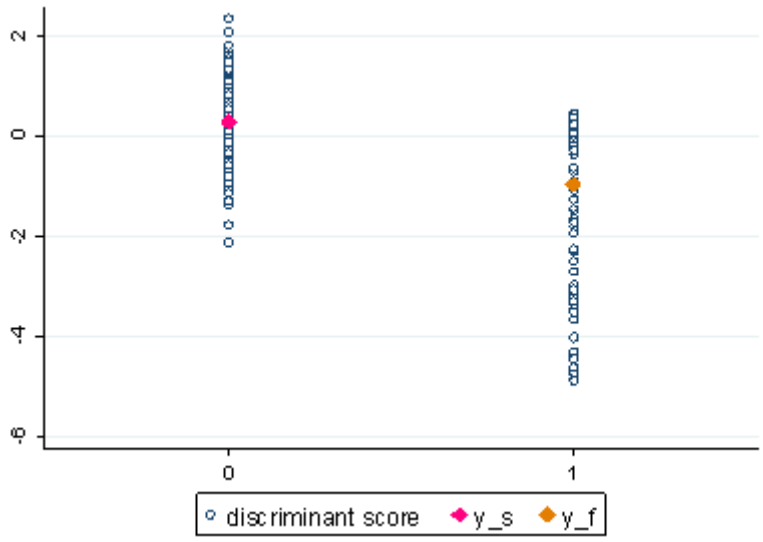
$$y_s = -20.673 + 0.096 \cdot 33.99 + 0.091 \cdot 184.15 + 1.1 \cdot 0.84 = 0.273$$

and for the group of failing students we have

$$y_f = -20.673 + 0.096 \cdot 29.74 + 0.091 \cdot 176.86 + 1.1 \cdot 0.66 = -0.996.$$

In Figure 11 we can see the centroids for the two groups. We observe that the individual score (5.1) is closer to the centroid of the smaller group, and so the subject is allocated by the LDA model to the smaller group. The confusion matrix

Figure 11: Plot of group centroids: Model week 0



of linear discriminant model is given in Table 22. The classification results indicate, that 23 of the failing students have been misclassified as belonging to the group of successful students and 42 of the successful students have been wrongly classified into the group of failing students. Correct classification rates in the failing group and successful group were 69.2% and 81.4%, respectively, and overall hit rate was

Table 22: LDA classification table: Model 0

		Observed	
		Y=1	Y=0
Classified	Y=1	39	42
	Y=0	23	184

of 77.4% The value of the overall statistic Z given by 4.5 is

$$Z = (223 - 190.74)/\sqrt{(190.74 \cdot (288 - 190.74)/288)} = 4.02,$$

with associated $p - value < 0.01$, which clearly indicates a better than by chance. The lower bound for the 99% confidence interval for the true frequency of total-group classifications is $223 - 2.326 \cdot 8.02 = 204.35$. For individual groups we get

Group	n_{jj}	n_j	e_j	Z_{G_i}	p-value
Failure	39	62	13.34	7.93	< 0.001
Success	184	226	177.41	0.905	0.183

We may also state that the overall hit rate is approximately 33% better than what may be expected by chance. The hit rate for larger group is no better than what may be expected by chance, but, in contrast, the hit rate for the smaller group is about 53% better than what may be expected by chance. Hence, if we use the derived linear classification rule prediction of failure is about 53% better than by chance prediction.

Application of stepwise LDA model after 17 weeks, yields the following: MC score, IMD course pass/fail score, Conscientiousness, AD: family relationship and AD: time management were included in the final model. The summary of stepwise discriminant analysis is shown in Table 23. On the first step, variable MC score was included in the discriminant function, since it provided the maximum separation of two groups according to the selection criterion. Table 24 illustrates the decrease in Λ_W statistic in the remaining steps of the stepwise procedure. The resultant linear discriminant function reached statistical significance ($\Lambda_W = 0.562, \quad p - value < 0.001$). The squared canonical correlation (see Table 5.5) was slightly higher than previously. The squared canonical correlation indicates that almost 44% of the variation between the two groups of students is accounted for by these discriminating variables.

Standardized, unstandardized and structure coefficients of linear discriminant function are displayed in Table 26. We observed that, in general, standardized and structure coefficients are similar in magnitude. The considerable difference was de-

Table 23: LDA: summary of stepwise variable selection for Model 17

Step	Variables	Tolerance	p-value of F to Remove	Λ_W
1	MC score	1.000	0.000	
2	MC score	0.896	0.000	0.802
	IMD course	0.896	0.001	0.628
3	MC score	0.896	0.000	0.761
	IMD course	0.895	0.001	0.603
	Conscientiousness	0.997	0.002	0.602
4	MC score	0.897	0.000	0.736
	IMD course	0.894	0.001	0.592
	Conscientiousness	0.987	0.001	0.593
	AD:time management	0.987	0.025	0.581
5	MC score	0.882	0.000	0.731
	IMD course	0.859	0.000	0.588
	Conscientiousness	0.980	0.001	0.587
	AD:time management	0.986	0.028	0.572
	AD: family relationship	0.921	0.039	0.570

Table 24: LDA: summary of Λ_W in stepwise procedure

Step	Λ_W	df1	df2	df3	F	df1	df2	p-value
1	0.628	1	1	286	169.693	1	286	0.000
2	0.602	2	1	286	68.366	3	284	0.000
3	0.581	3	1	286	68.366	3	284	0.000
4	0.570	4	1	286	53.269	4	283	0.000
5	0.562	5	1	286	43.978	5	282	0.000

Table 25: Linear discriminant analysis: Model week 17

Canonical correlation	Eigenvalue	Λ_W	X^2	df	p-value
0.662	0.78	0.562	163.43	5	0.000

tected for covariate AD: family relationship. Analysis of standardized discriminant function coefficients, reported in Table 26, leads to the following conclusion: measures of previous academic performance, namely MC score and pass/fail classification in the IMD course, contribute the most for the resulting group differences. On the basis of structured coefficients, we may label the discriminant function as "measure of cognitive abilities", since these two variables share with discriminant function more variation than other predictors. Figure 12 displays the plot of centroids for

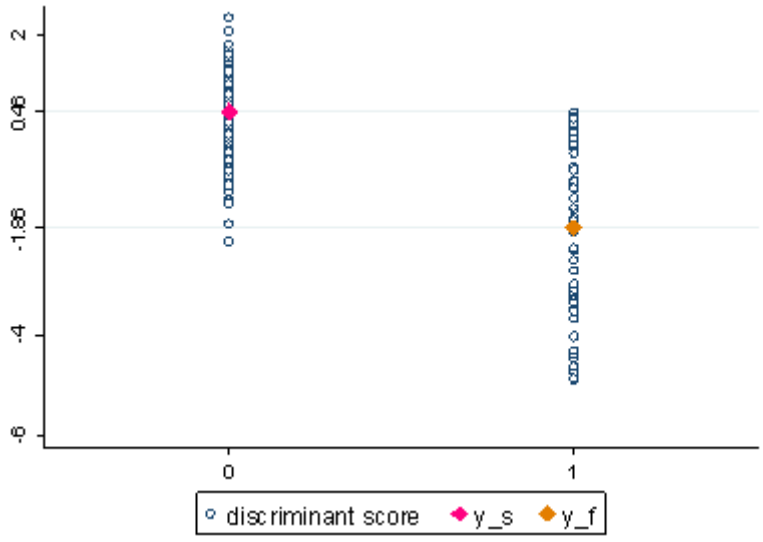
Table 26: Standardized, Unstandardized and Structure coefficients of linear discriminant function: Model 17

Variable	u^s	u	l
Conscientiousness	0.313	0.051	0.321
MC score	0.774	0.329	0.872
AD:time management	0.198	0.504	0.213
AD: family relationship	0.193	0.552	-0.06
IMD course	-0.345	-1.239	-0.563
Constant	-	-5.821	

u^s=Standardized coefficient; u=Unstandardized coefficient; l=Structure height

two groups, located in 0.461 and -1.68.

Figure 12: Plot of group centroids: Model week 17



Finally, the 2×2 confusion matrix is given in Table 27. Overall Model 17 predicts correctly the group membership of 247 (85.8%) of students, in the failing group and sucessful group the hit rates are 67.7% and 90.7%, respectively. The value of the overall Z statistic given by (4.5) is

$$Z = (247 - 190.74)/\sqrt{(190.74 \cdot (288 - 190.74)/288)} = 7.01,$$

Table 27: LDA classification table: Model 17

		Observed	
		Y=1	Y=0
Classified	Y=1	42	21
	Y=0	20	205

with $p - value < 0.001$, which is clearly better than classification by chance. The lower bound for a 99% confidence interval for the true frequency of total-group correct classifications is $247 - 2.326 \cdot 8.02 = 228.35$. For separate groups the results are the following: both separate group hit rates are better than what may be expected

Group	n_{jj}	n_j	e_j	Z_{G_i}	p-value
Failure	42	62	13.34	8.86	< 0.001
Success	205	226	177.41	4.47	< 0.001

by chance.

5.6 Results of K-NN discriminant analysis

This section represents results of non-parametric K-NN discriminant analysis. For the two instances of time (admission and week 17) in current study, we applied non-parametric discriminant analysis to subset of covariates selected previously for LR modeling and for LDA modeling, with the number of discriminating variables $p \in \{3, 4, 5, 6\}$. Euclidean metric and Jaccard's matching coefficient were used as distance functions to delineate the set of nearest neighbors. To allow the use of matching coefficient, we transformed quantitative variables into categorical ones, based on the quartiles. Recalling that Euclidean distance is sensitive to the scale of measurement, predictors GPA, MC score and Conscientiousness were standardized. The nearest neighbor discriminant analysis was performed for three values of the parameter K, namely $K = 3, 4, 7$. We choose the best number of neighbors based on classification error rates estimated from the data. All models were fitted using prior probabilities estimated from the sample.

Table 28 displays rates of correct classification that resulted from application of K-NN using the set of predictors formed by GPA, Conscientiousness and AD: time management. We found that hit rates differ remarkably with the value of K and with the metric used to formulate the rule. In Table 28 the total hit rate decreased when K pass from 3 to 7. The non-parametric method of classification was also sensitive to the choice of distance function. Better results for the smaller group were obtained with Euclidean metric.

Table 29 summarizes rates of correct classification for the K-NN discriminant

Table 28: Accuracy of KNN classification rules with 3 predictors

		% of correct classifications		
Distance	K	total	failure	success
Jaccard	3	81.6	37.1	93.8
Jaccard	4	80.9	30.6	94.7
Jaccard	7	79.5	19.4	96
Euclidean	3	88.2	66.1	94.2
Euclidean	4	86.1	72.6	89.8
Euclidean	7	83.7	48.4	93.4

analysis obtained using the following set of 4 discriminating variables: GPA, Conscientiousness, AD: time management and Age.

Table 29: Accuracy of KNN classification rules with 4 predictors

		% of correct classifications		
Distance	K	total	failure	success
Jaccard	3	81.6	37.1	93.8
Jaccard	4	80.9	30.6	94.7
Jaccard	7	80.2	25.8	95.1
Euclidean	3	87.8	64.5	95.1
Euclidean	4	87.8	75.8	91.2
Euclidean	7	84	51.6	93.3

Analyzing Table 29 we observe that hit rates for rules formulated with Jaccard matching coefficient and $K = 3, 4$ do not differ from hit rates estimated under identical conditions with three predictors. When $K = 7$, we observe a slight increase of in both, the total hit rate and in the hit rate for the smaller group (failing group). Again, non of the hit rates estimated for the smalled group with Jaccard matching coefficient were greater than 40%. Additionally, when the Euclidean distance is used to determine the set of the K-nearest neighbors, the loss of classification accuracy for the smaller group diminishes for $K = 4$ and increases for $K = 7$ (similar with the previous results for 3 predictors). Finally, these results suggest that, for the first instance of time (at entrance), the set of four discriminating variables performs better in predicting group membership when the Euclidean distance and $K = 4$ were used.

The expected frequencies of correct classifications are: 13.33 for group of failing students, 177.41 for group of successful students and 190.74 for the total sample. Hence, the value of the overall Z statistic given by (4.5) is

$$Z = (253 - 190.74)/\sqrt{(190.74 \cdot (288 - 190.74)/288)} = 7.56,$$

with $p - value < 0.001$, that is clearly better than classification by chance. For separate groups the results are the following. The results presented above show that

Group	n_{jj}	n_j	e_j	Z	p-value
Failure	47	62	13.34	10.4	< 0.001
Success	206	226	177.41	4.62	< 0.001

hit rates of non-parametric discriminant analysis model developed using only pre-enrollment factors differ significantly from the rates of correct classification expected by pure chance.

Table 30 shows hit rates for K-NN classification rules derived for the following 5 discriminating variables, available after 17 weeks: MC score, Conscientiousness, AD:time management, AD:family relationship and IMD course. Table 31 displays

Table 30: Accuracy of KNN classification rules with 5 predictors

		% of correct classifications		
Distance	K	total	failure	success
Jaccard	3	85.4	56.5	93.4
Jaccard	4	84.7	53.2	93.4
Jaccard	7	84	46.8	94.2
Euclidean	3	90.9	67.7	97.3
Euclidean	4	90.6	79	93.8
Euclidean	7	88.2	54.8	97.3

hit rates for K-NN classification rules with 6 independent variables, available after 17 weeks: MC score, Conscientiousness, AD:time management, AD:family relationship, IMD course and Change of residence. In general, for Model 17 all

Table 31: Accuracy of KNN classification rules with 6 predictors

		% of correct classifications		
Distance	K	total	failure	success
Jaccard	3	88.5	74.2	92.5
Jaccard	4	87.2	62.9	93.8
Jaccard	7	86.4	53.2	95.6
Euclidean	3	89.8	61.3	97.3
Euclidean	4	89.6	77.4	92.9
Euclidean	7	87.2	51.6	96.1

non-parametric discriminant functions performed alike with respect to the overall rates of correct classifications. The differences were observed in terms of separate-group hit rates. Comparing the models developed with Jaccard matching coefficient we found that composite of 6 predictors provided better classification accuracy in

all considered values of K with respect of prediction of smaller group membership. In contrast, when Euclidean distance is used, relatively larger classification accuracy for the smaller group is achieved by models fitted with 5 predictors. Of all non-parametric models considered the 4-NN classification rule based on MC score, Conscientiousness, AD:time management, AD:family relationship and IMD course variables, provided the smallest separate group misclassification rates.

In order to evaluate the effectiveness of classification rule discussed above, we used Z statistics from (4.5). The results indicate that separate group hit rates, as

Group	n_{jj}	n_j	e_j	Z	p-value
Failure	49	62	13.34	11.02	< 0.001
Success	212	226	177.41	5.6	< 0.001
Total	261	288	190.74	8.75	< 0.01

well as overall hit rate, were significantly better than may be expected by chance.

5.7 Comparison of classification rules

To develop a multivariate model for the prospective identification of students at risk of failure in the first year of medical degree in SHS-UM, we applied three statistical tools. As was mentioned previously, all of them provided total-group rates of correct classification significantly better that could be expected by chance for our data set. In this section we discuss, in detail, results of assessment of statistical and practical significance of classification rules developed by LR, LDA e K-NN. It should be recalled that to compare classification rules we used estimates of apparent hit rates.

Table 32 presents summary of measures of prediction accuracy of models built at the entrance time (at the beginning of academic year).

Table 32: Model week 0: comparison of classification rules

	% of correct classifications			
Method	total	failure	success	I_h
LR	76.4	59.7	71.8	16.79%
LDA	77.4	69.2	81.4	33.22%
K-NN	87.8	75.8	91.2	63.9%

To answer the question whether statistical models could predict a group membership of students, we calculated an index of improvement over chance. Our results suggest that

- application of LR model the entrance time reduces an error over chance classification by 16.79%,

- application of LDA allows to reduce error of group membership predictions by 33.22%,
- non-parametric K-NN discriminant analysis provides the reduction of error of 63.9%.

hence, of the three methods considered, K-NN provided the greatest index. In general, the three statistical methods for predicting of students group membership based, on the set pre-admission factors used in this research, exhibit satisfactory classification accuracy.

Table 33 presents summary of measures of prediction accuracy of models built at second instance of time (after 17 weeks in medical school). With the informa-

Table 33: Model week 17: Comparison of classification rules

	% of correct classifications			
Method	total	failure	success	I_h
LR	84.7	74.2	87.6	54.7%
LDA	85.8	67.7	90.7	57.9%
K-NN	90.6	79	93.8	72.2%

tion available after 17 weeks in medical school, our models provided considerable reduction of overall misclassification, 54.7%, 57.9% and 72.2%.

First of all, comparing the results of Tables 32 and 33, we can conclude that the percentage of correct classifications is considerably higher for the group of successful students (the larger group) in all three classification methods. It is of note that, for both subsets of data, in parametric and in non-parametric discriminant analysis the observed differences in separate group rates of correct classification were considerably of higher magnitude when compared with LR.

Finch & Schneider [20] claim that, for LR and LDA, under condition of extremely unequal group dimensions, the misclassification rates tend to be very high for smaller group, but very low for the larger group. The two subgroups in our study sample are clearly of the different size, (ratio 21.5 : 78.5) and our conclusions agree accordance with existent studies on comparison of effect of sample size ratio on error rates of different classification rules [20, 66].

Fan & Wang [66] show that drastically unequal prior probabilities, under condition of heterogeneity of covariance matrices, affect performance of LDA and LR in different ways. In such case, the techniques have very similar overall rates of correct classification, but differ remarkably on separate group rates of correct classifications: LR favors the smaller group and discriminant analysis favors the larger group. In accordance to these results, in our study the application of the LDA and LR to 17 weeks data set yielded very similar overall rates of correct classifications: 85.7% and

84.6%, respectively. With respect to separate group hit rates we observed considerable differences. For group of failing students (the smaller group - 21.5% of the sample) the rate of correct classifications was provided by LR was higher.

The best improvement over chance in classification, for both subsets of data, was obtained with K-NN.

To compare performance of different classification rules we used McNemar test. A summary of the 288×2 matrix is given in the Table 34. Comparing LR vs LDA we

Table 34: Comparison of rules for Model 0: LR vs LDA

		LDA		
		Hit	Miss	Row total
LR	Hit	213	7	220
	Miss	10	58	45
Column total		223	65	288

have 17 misclassified by one of the rules, $7+10 < 25$, hence the value of the McNemar test statistic is $T_e = 7$. Under the null hypothesis, $T_e \sim Binomial(0.5, 17)$ yield p-value of 0.629. Thus, we may concluded that there are no statistically significant differences in apparent hit rates of LR and LDA models adjusted at at week 0.

Table 5.7 summarize classification results of LDA and LR models developed with the data available after 17 week in medical school.

Table 35: Comparison of rules for Model 17: LR vs LDA

		LDA		
		Hit	Miss	Row total
LR	Hit	238	5	243
	Miss	9	36	45
Column total		247	41	288

A number of misclassification of interest was $9 + 5 < 25$, as such the McNemar test statistic is $T = 5$. Under the null hypothesis $T_e \sim Binomial(0.5, 14)$, with $p - value = 0.424$. Hence, there is no evidence in the sample to claim the difference in the total hit rates of LR and LDA models.

Comparing classification accuracy of K-NN and LR in two instances of time, we observe that number of misclassification of interest was greater than 25. Hence, for week 0 models, the test statistic was $T = 19.1$. Under the null hypothesis, i.e that hit rates are equal, T is approximately $\chi^2(1)$, with corresponding $p - value < 0.001$. For week 17 $T = 8.1$ with $p - value = 0.004$ yielded rejection of null hypothesis. In fact, in instances of time (admission and week 17) the KNN model yielded a significantly higher rate of correct classifications.

The number of misclassification resulted from application of LDA and K-NN was greater than 25 at both pre-defined time instances. Hence, we use McNemar test statistic defined in (4.8), that under the null hypothesis has approximately $\chi^2(1)$ distribution. For week 0 $T = 16.07$ and $p - value < 0.001$; for week 17 $T = 4.90$ and $p - value = 0.027$. Hence, for both instances of time 4-NN total hit rates were significantly higher than the LDA hit rates.

It is also of interest to compare separate group hit rates of discussed techniques. Recalling the main goal of this work, we focus our attention on the accuracy of predicting the membership in the smaller group (failing students). The results of comparisons are displayed in Table 36 and suggest that while 4-NN discriminant analysis hit rate for the group of failing students is statistically higher than the LDA hit rate and the LR hit rate for the models that explore pre-enrollment factors exclusively, no statistically significant differences in hit rates were detected when pre-enrollment data was combined with measures of academic performance in the early courses in medical school. Hence, the three methods have comparable performance in predicting a membership in underachieving group.

Table 36: Comparison of efficiency of classification rules for group of underachievers

Rules	T $(T_e)^3$	Distribution of T	p-value
First timing point models			
LR vs LDA	1	Binomial (0.5, 4)	0.625
LR vs K-NN	3	Binomial (0.5, 16)	0.021
LDA vs K-NN	4	Binomial (0.5, 16)	0.077
Second timing point models			
LR vs LDA	1	Binomial (0.5, 6)	0.218
LR vs K-NN	5	Binomial (0.5, 12)	0.774
LDA vs K-NN	4	Binomial (0.5, 14)	0.179

6 Conclusions

To determine which factors influence academic performance of medical students in the first year course with the highest failure rates (FOS1) at the SHS-UM, we used multivariate LR, LDA and non-parametric K-nearest neighbors discriminant analysis. Predictive accuracy of the multivariate statistical models was assessed under conditions of unequal group sizes (62:226), mixture of non-normal continuous, ordinal and binary predictor. Regarding performance of parametric and non-parametric discriminant analysis and LR we concluded the following:

- although the distributional assumptions were not satisfied for our data set, apparent overall hit rates of LR, with proportional priors, and overall hit rates of LDA, with equal priors, were relatively close;
- regarding the separate hit rates of the two groups, we found no statistically significant differences
- with unbalanced group sizes, the percentage of correct classifications were considerably higher for the larger group, for all the classification methods used.

Unexpectably, the best improvement over chance in overall classification was reached with non-parametric discriminant analysis, for both subsets of data. This should be interpreted with caution. It is mentioned in literature, that K-NN procedure is very flexible and has the tendency to overfit data [3]. Additionally, higher classification accuracy of the 4-NN discriminant analysis may be explained, in part, by the fact that apparent hit rates were used to compare rules. Huberty [28] warns that apparent hit rates, based on an internal analysis, are not as good as those based on cross-validation. Hence, to compare classification rules, external results should be used. The performance of this method should be examined in more detail in further research.

This study illustrates the potential of multivariate statistical approaches for early identification of cognitive and non-cognitive factors, that predispose undergraduate medical students to fail in the first year. The results of this study indicate that multivariate models, with high levels of classification accuracy, can be obtained combining pre-university GPA, academic performance in early courses, the personality trait Conscientiousness, change of residence in the transition to medical school, age and self-declared anticipation of difficulties with family relations and with time management.

The results of this study showed that, in SHS-UM, the influence of pre-university academic achievements on first year academic performance is moderate. Models based exclusively on combination of pre-enrolment factors (before admission) were able to classify correctly $71 - 86\%$ of students.

The substantial improvement in the predictive ability was achieved when pre-admission variables were combined with performance indicators of first courses in medical school. The results reveal that failure in the first year can be accounted, at week 17, by lower levels of Conscientiousness, leaving home, poor academic achievements in introductory courses and unanticipated difficulties due to enrolment.

The association of higher levels of Conscientiousness and reduced a risk of academic difficulties, is understandable since conscientious individuals are self-disciplined, persistent, organized and goal-oriented. This association has been described previously [15, 39, 40] and supports the value of personality characteristics to succeed in the first year of medical school.

The conclusion that the anticipation of difficulties is associated with smaller probabilities of academic failure are a novelty and of interest. The research in high education suggests that many incoming students have inadequate views regarding the university life, teaching and assessment styles and the required learning strategies. Hence, they tend to underestimate the amount of study that the medical program expects from them. Poor time management and interpersonal problems are recognized in literature to be the most frequently experienced deficiencies among medical students [51, 60]. Hence, students who express concern about probable difficulties during the first year of medical degree seem to be aware of the challenges of transition phase and of special demands of the medical training program. Recognition of probable difficulties has positive impact on the academic achievements since in phase of task analysis and development of strategic plan to pursuit the academic goals it induce to selection of adequate coping techniques to attain desired outcomes.

Hence, students who express concerns about possible difficulties during the first year of the medical degree seem to be aware of the challenges of the phase-transition and of the special demands of a medical training program. Therefore, recognition of possible difficulties has positive impact on the academic achievements because students can develop strategies to overcome the difficulties.

Concerning the influence of student accommodation in their academic experience, some studies claim that on campus residential accommodation is beneficial for students [2, 50]. However, literature provides insufficient evidence to claim an association between type of accommodation and academic failure. For our data set, we concluded that living away from home is a risk factor for underperformance in the first year of medical degree. One possible explanation for the apparent link between change of residence and academic failure, might be that without family assistance, first year students are overloaded with academic tasks, housework and social activities, and are unable to spend an adequate amount of time studying. Association between high probability of first year failure in medical school and leaving parental home also, may also be explained, in part, by the decrease of emotional support

from family and friends.

Factors mentioned in previous studies such as gender, preference for the degree and parental education were not included in our multivariate models. Since these factors were in fact analysed in the study, our conclusion is that, in the context of the SHS-UM, they do not contribute for prediction of first year academic failure.

Our study has, obviously, some limitations. First, it was conducted with a relatively small sample size. Small sample size may affect the accuracy of parameters estimates. Second, it is common, for research based on voluntary response questionnaires, that volunteers differ from the other members of population on several features so that bias may limit the generalization of conclusions. The third limitation is the method chosen to estimate hit rates and compare performance of the three multivariate techniques. In this study, we used apparent rates of correct classification with the intention of performing cross-validation of the models in the future, as additional data becomes available. Finally, while variable selection in LR was performed employing two distinct methods, in LDA we used only stepwise procedure. It is desirable to explore the best subset selection for LDA in future research.

Despite these limitations, our study provides substantial empirical evidence that personality characteristics, such as Conscientiousness, and anticipation of difficulties due to enrolment in medical program, as well as academic achievements are important predictors of first year failure. The study highlights the importance of non-academic factors for prediction of students failure in the first year of medical degree. The existence of a statistical model with adequate levels of Sensitivity and Specificity for prospective identification of students that struggle to perform well in the medical program, offers interesting opportunities for early remediation.

References

[1] Agresti A. *Categorical data analysis.* 2 ed, Wiley 2002.

[2] Arulampalam W, Naylor R, Smith JP. *Dropping out of medical school in the UK: explaining the changes over ten years.* Med Educ 2007; 41: 385-94.

[3] Asparoukhov OK, Wojtec J, Krzanowski J. *A comparison of discriminant procedures for binary variables.* Computational Statistics and Data Analysis 2001; 38: 139-60.

[4] Beeman PB, Waterhouse JK. *NCLEX-RN Performance: Predicting success on the computerized examination.* Journal of professional nursing 2001; 17-4:158.65

[5] Braga AC. *Curvas ROC: aspectos funcionais e aplicações.* [PhD dissertation] Universidade do Minho, Braga, 2000.

[6] Calcagno V, Mazancourt C. ***glmulti****: An R package for easy automated model selection with generalized linear models.* Journal of Statistical Software. 2010; 34:12.

[7] Challis M, Fleet A, Basyone G. *An accident waiting to happen? A case for medical education.* Med Teach 1999; 21: 582-5.

[8] Chamorro-Premuzic, T. *Personality and individual differences.* Oxford: Blackwell. 2007.

[9] Chamorro-Premuzic T, Furnham A. *Personality predicts academic performance: evidence from two longitudinal university samples.* Journal of Research in Personality 2003; 37: 319-38.

[10] Cleland J, Milne A, Sinclair H, Lee AJ. *Cohort study on predicting grades: is performance on early MBChB assessments predictive of later undergraduate grades?* Med Educ 2008;42: 676-83.

[11] Cleland J, Arnold R, Chesser A. *Failing finals is often a surprise for the student but not the teacher: identifying difficulties and supporting students with academic difficulties.* Med Teach 2005; 27: 504-8.

[12] Cleves MA, *From the help desk: Comparing areas under receiver operating characteristic curves from two or more probit or logit models.* Stata Journal. 2002, 2:3, 301-13.

[13] Cooter R, Erdman JB, Gonnella JS, Hojat M, Xu G. *Economic Diversity in Medical Education: The Relationship between Students' Family Income and Academic Performance, Career Choice, and Student Debt.* Eval Health Prof 2004; 27: 252-64.

[14] DeLong ER, DeLong DM, Clarke-Pearson DL. *Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach.* Biometrics 1988, 44: 837-45.

[15] Doherty EM, Nuget E, *Personality factors and medical training: a review of the literature.* Med Educ 2011; 45:132-40.

[16] Fan X, Wang L. *Comparing linear discriminant function with logistic regression for the two-group classification problem.* Journal of Experimental Education 1999; 67: 265286.

[17] Ferguson E, James D, Madeley L. *Factors associated with success in medical school: systematic review of the literature.* BMJ 2002;324: 952-7.

[18] Ferguson E, James D, O'Hehir F, Sanders A. *Pilot study of the roles of personality, references, and personal statements in relation to performance over the five years of a medical degree.* BMJ 2003; 326: 429-31.

[19] Ferrer AJ, Wang L. *Comparing the classification accuracy among nonparametric, parametric discriminant analysis and logistic regression methods.* Paper presented at the Annual meeting of the American Educational Research Association, Quebec, 1999, Available from http://www.eric.ed.gov

[20] Finch WH, Schneider MK *Misclassification Rates for Four Methods of Group Classification Impact of predictor distribution, effect size, sample size and group size ratio.* Educational and Psychological Measurement 2006, 66(2): 240-57.

[21] Fix E, Hodges JL. *Discriminatory analysis: Nonparametric discrimination, consistency properties.* In Technical Report 4, Project 21-49-004. Randolph Field, Texas: Brooks Air Force Base, USAF School of Aviation Medicine, 1951.

[22] Gonnella JS, Erdmann JB, Hojat M. *An empirical study of the predictive validity of number grades in medical school using 3 decades of longitudinal data: implications for a grading system.* Med Educ 2004;38: 425-34.

[23] Haist S, Wilson J, Elam C, Blue A, Fosson S. *The effect of gender and age on medical school performance: the importance of interaction.* Advances in Health Sciences Education: theory and practice 2000; 5: 197-205.

[24] Hastie T, Tibshirani RJ, Friedman J. *The Elements of Statistical Learning. Data mining, Inference, and Prediction.* Springer-Verlag. 2001

[25] Hosmer DW, Hosmer T, Le Cessie S, Lemeshow SA. *Comparison of goodness-of-fit tests for the logistic regression model.* Stat in Med, 1997; 16: 965-80.

[26] Hosmer DW, Lemeshow S. *Applied Logistic Regression.* 2 ed. New York: Wiley 2000.

[27] Hosmer DW, Taber S, Lemeshow S. *The importance of assessing the fit of logistic regression models: a case study.* Am J Public Health 1991; 81(12):1630-5.

[28] Huberty CJ. Olejnik S. *Applied MANOVA and Discriminant Analysis.* Wiley. 2006.

[29] Huberty CJ, Wisenbaker M, Smith J D, Smith J C. *Using categorical variables in discriminant analysis.* Multivariate Behavioral Research 1986; 21: 479-496.

[30] James D, Chilvers C. *Academic and non-academic predictors of success on the Nottingham undergraduate medical course 1970-1995.* Med Educ 2001;35: 1056-64.

[31] Klecka WR. *Discriminant Analysis.* Beverly Hills, CA: Sage, 1980

[32] Kleinbaum Dg, Klein M. *Logistic Regression: A Self-Learning Text.*3rd ed. Springer- Verlag, New York. 2010.

[33] Kohler U, Kreuter F. *Data Analysis Using Stata.* 2 ed. Texas: Stata Press 2009.

[34] Krzanowski WJ. (1977). *The performance of Fisher's linear discriminant function under non-optimal conditions.* Technometrics. 9(2), 191-200.

[35] Lachenbruch PA.*Robustness of discriminant functions.* SAS Conference Proceedings: SAS Users Group International (SUGI), 1982, San Francisco, California USA

[36] Lambe P, Bristow D. *Predicting student performance from attributes at entry: a latent class analysis.* Med Educ 2011; 45:308-16.

[37] Lee KI, Koval JJ. *Determination of the best significance level in forward stepwise logistic regression.* Communications in Statistics - Simulation and Computation 1997, 26: 2, 559 - 75.

[38] Lei PW, Koehly LM. *Linear discriminant analysis versus logistic regression: A comparison of the classification errors in the two-group case. The Journal of Experimental Education.* 2003; 72(1): 25 - 49.

[39] Lievens F, Coetser P, De Fruyt F, De Maeseneer J. *Medical Students' personality characteristics and academic performance: a five-factor model perspective.* Med Educ 2002; 36: 1050-6.

[40] Lievens F, Ones DS, Dilchert S. *Personality Scale Validities Increase throughout Medical School.* Journal of Applied Psychology 2009;94:6: 1514-35.

[41] Long J. *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks: Sage 1997.

[42] Lumb AB, Vail A. *Comparison of academic, application form and social factors in predicting early performance on the medical course.* Med Educ 2004;38: 1002-5.

[43] McLachlan GJ. *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley.1992.

[44] Menard S. *Applied logistic regression analysis.* Sage 1995.

[45] Meshbane A, Morris, J. D. *Predictive discriminant analysis versus logistic regression in two-group classification problems.* Paper presented at the

annual meeting of the American Educational Research Association, New York. 1996.

[46] Mills C, Heyworth J, Rosenwax L, Carr S, Rosenberg M. *Factors associated with the academic success of first year health science students.* Advances in Health Sciences Education: theory and practice 2009;14: 205-17.

[47] Morgan MK. *Male university attrition: a discriminant analysis.* Research in Higher Education 1974; 2:281-9.

[48] Osius G, Rojek D. *Normal goodness-of-fit tests for parametric multinomial models with large degrees of freedom.* Journal of the American Statistical Association; 1992. 87:1 145 - 52.

[49] Papadakis MA, Teherany A, Banach MA, Knetter TD, Rattner SL, Stern DT, et al. *Disciplinary action by medical boards and prior behavior in medical school.* N Engl J Med 2005; 353: 2673-82.

[50] Pascarella E, Terenzini P. *How College Affects Students. Vol. 2. A Third Decade os Research.* San Francisco, CA: Jossey-Bass.

[51] Paul G, Hinman G, Dottl S, Passon J. Academic development: a survey of academic difficulties experienced by medical students and support services provided. Teach Learn Med 2009; 21: 254-60.

[52] Pregibon D, 1981. *Logistic regression diagnostics.* Ann. Statist., 9: 705-724.

[53] Press SJ, Wilson S. *Choosing between logistic regression and discriminant analysis.* Journal of the American Statistical Association. 1978; 73:364:699 - 705.

[54] Reason RD, Terenzini PT, Domingo RJ. *First things first: developing academic competence in the first year of college.* Research in Higher Education 2006; 47.2: 149-75.

[55] Reis E. *Estatística multivariada aplicada.* Lisboa: Edições Sílabo, 1997.

[56] Rencher AC. Methods of multivariate analysis 2nd ed. Wiley, 2002.

[57] Royston P, Altman DG. *Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling.* Journal of the Royal Statistical Society, Series C (Applied Statistics), 1994, 43(3) . 429-67.

[58] Sarkar SK, Midi H. *Importance of assessing the model adequacy of binary logistic regression.* Journal of Applied Sciences. 2010; 10(6):479-86.

[59] Sauerbrei W, Royston P, Binder H. *Selection of important variables and determination of functional form for continuous predictors in multivariate model building.* Statist. med. 2007; 26:5512-5528.

[60] Sayer M, De Saintonge MC, Evans D, Wood D. *Support for students with academic difficulties.* Med Educ 2002;36:643-50.

[61] Stata Corp. 2009 Stata: Release 11. Statistical Software. College Station, TX: StataCorp LP.

[62] Seber GAF. *Multivariate observations.* New York: Wiley, 1984.

[63] Tabachnick B, Fidell LS. *Using multivariate statistics.* 5th ed. USA Pearson Education, 2007.

[64] Vandamme J-P, Meskens N, Superby J-F. *Predicting academic performance by data mining methods.* Education Econometrics 2007; 15(4): 405-19.

[65] Vlachonikolis IG, Marriott FHC, *Discrimination with mixed binary and continuous data.* Appl.Statist. 1982; 31(1): 23 - 31.

[66] Wang Q, Koval JJ, Mills CA, Lee D. *Determination of the selection statistics and best significance level in backward stepwise logistic regression.* Communications in Statistics-Simulation and Computation; 2008, 37: 62-72.

[67] Yates J, James D. *Risk factors for poor performance on the undergraduate medical course: cohort study at Nottingham University.* Med Educ 2007;41: 65-73.

[68] Yates J, James D. *Risk factors at medical school for subsequent professional misconduct: multicenter retrospective case-control study.* BMJ 2010;340 20-40.

[69] Yates J, James D. *Predicting the "strugglers": a case-control study of students at Nottingham University Medical School.* BMJ 2006; 332: 1009-13.

Appendix

Appendix A

Figure 13: Model 0: Smoothed scatter plots on the logit scale

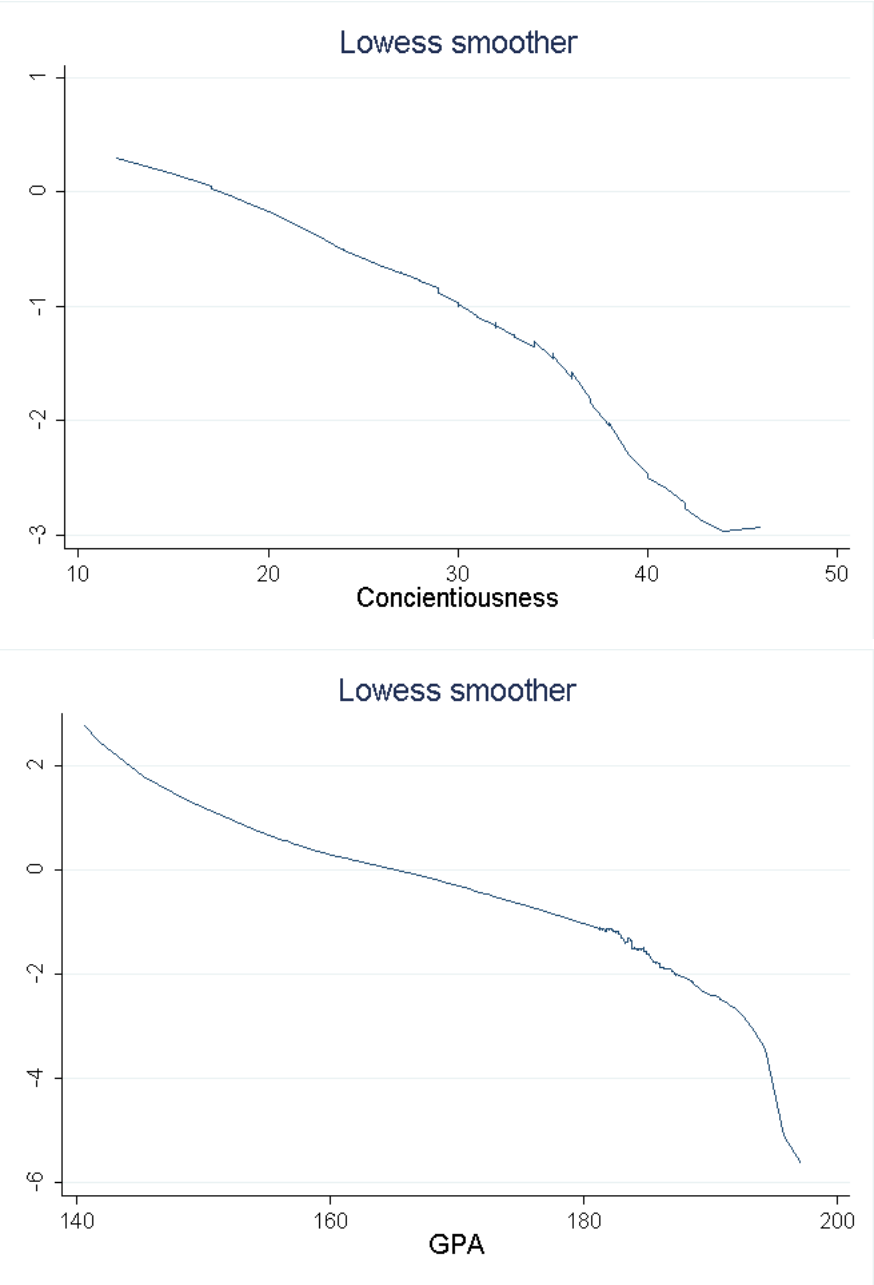


Figure 14: Model 17: Smoothed scatter plots on the logit scale

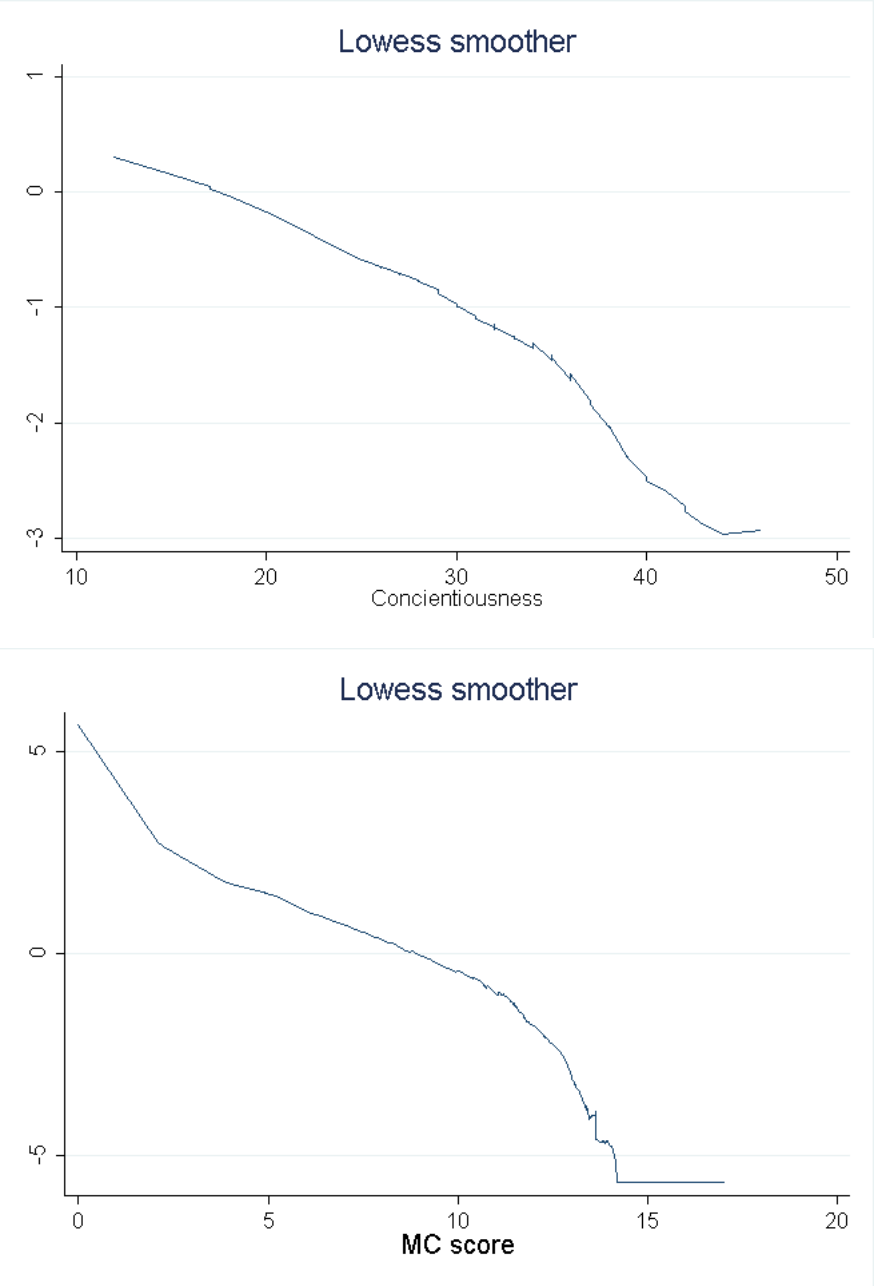
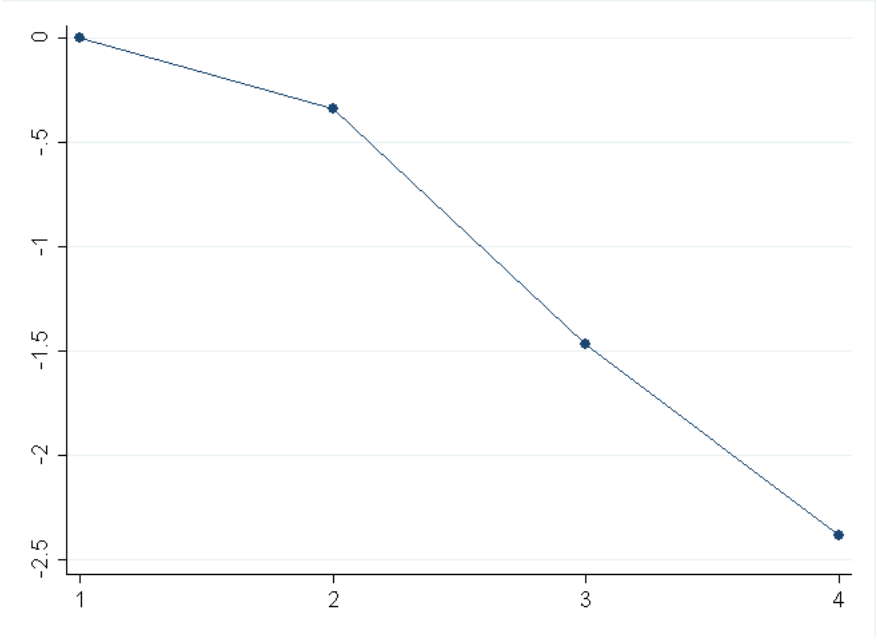


Figure 15: Model 0: dummy variables analysis of linearity
Conscientiousness



GPA

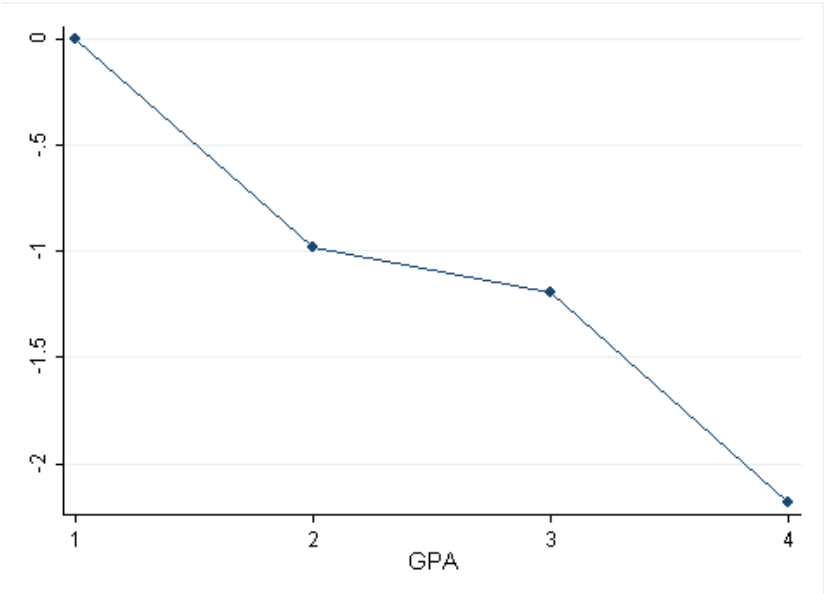


Figure 16: Model 17: dummy variables analysis of linearity