Zero-one augmented beta and zero inflated discrete models with heterogeneous dispersion: an application to students' academic performance

Hildete P. Pinheiro[1], Rafael P. Maia[1]
Eufrásio A. Lima-Neto[2] and Mariana Rodrigues-Motta[1]
[1] *Department of Statistics, University of Campinas, Brazil*
[2] *Departments of Statistics,*
Federal University of Paraíba, Brazil

Abstract

The purpose of this work is to present suitable statistical methods to study the performance of undergraduate students based on the incidence/proportion of failed courses. Some approaches are considered: in one of them the incidence of failed courses is modeled using zero inflated discrete distributions with heteroscedasticity, considering the logarithm of the total number of courses as an offset; in another, the proportion of failed courses is modeled considering a zero-one augmented beta distribution with heterogeneous dispersion parameter. A detailed residual analysis is performed to investigate the best model to fit the data. The zero-one augmented beta model with heteroscedasticity presents a better fit based on residual analysis. Moreover, the model for the proportion brings us more information as it is straight forward to interpret its results. The database consists of records of 3,699 students with Engineering major who entered the State University of Campinas, Brazil, from 2000 to 2005. The entrance exam scores, academic and demographic variables and their socioeconomic status are considered as covariates in the models.

Keywords and phrases: Academic performance Heteroscedasticity Quantile residuals Residual analysis Zero inflated discrete models Zero-one augmented beta models.

1 Introduction

The evaluation of the factors that contribute to improve student performance in the universities is an important research theme, since the results can improve the efficiency of the university systems and reduce the delays or failures that are costly for students and administration. However, the statistical modeling of student performance at university can be challenge due to the complexity of the process. For example, taking into account information about student's high school education is not fully appropriate because of the mismatch between the acquired skills at the high school and those required for a given degree program. Another potential source of information to evaluate the student performance is the pre-enrollment test that many universities consider to select students from high school according to some "best" performance criteria. However, a quick look reveals a low correlation between student's performance in the pre-enrollment test and academic performance in the university.

A wide variety of statistical techniques are considered to model student performance. Some authors evaluate the student performance just in the first academic year using binomial mixture models [7], path analysis [11], multivariate latent growth models [3], quantile regression [4, 8]. Additionaly, [1] recently considered a three-level mixed model to evaluate the student performance in Italian secondary education system.

In Brazil, due to a quota system implemented by the Federal Government and some other affirmative action programs implemented in some state universities, there is special interest on the factors which mostly contribute for the best performance of undergraduate students in the universities. [14] proposed linear regression models to assess the performance of undergraduate students using as response variable the *relative gain* based on the relative rank of student's final (or last) recorded GPA (Grade Point Average) and his/her total entrance exam score (EES) rank. [9] used robust nonparametric methods on quasi U-statistics to evaluate the performance of students in different groups using test statistics developed in [15, 16]. Both papers analyze data from the State University of Campinas (Unicamp), one of the best public universities in Brazil.

Thus, according to the literature one can see that we do not have one technique or model which is more appropriate to model performance of students. In this sense, our goal is to evaluate the performance of the students with Engineering major, during his/her Bachelor's degree, using the incidence or the proportion of failed courses as response variable. In this way, we will consider a wide range of discrete count regression models (binomial, Poisson, negative-binomial, beta-binomial) and their zero-inflated versions, as well as the beta and zero-one augmented beta regression models.

The database consists of 3,699 records of students with Engineering majors who entered in the University from 2000 to 2005. For each student we have the total number of required courses taken and failed, being able to model the incidence or proportion of failed courses during the Bachelor's degree. Models are fitted considering EES, academic variables and socioeconomic status in the regression structure of location, scale and mixture probabilities parameters.

Finally, we believe that the results found in this study can be useful to improve university polices for new students since it was possible to identify student profiles with respect to their academic performance.

The paper is organized as follows: Section 2 presents in details the data set and some descriptive results based on exploratory data analysis. Section 3 brings an overview of the regression models and residual analysis considered for modeling the incidence and the proportion of failed courses. Sections 4 and 5 exhibit the results and a discussion about the models.

2 The data set

The database of this study consists of 3,699 records of students with Engineering majors who entered at Unicamp, Brazil, from 2000 to 2005. For each student i denote Y_i as the incidence and Y_{1i} as the proportion of courses failed during his/her Bachelor's degree. We also have students' EES in each subject (Mathematics, Portuguese, Geography, History, Biology, Chemistry and Physics), some academic variables and socioeconomic status, which are considered as covariates in the models. In Brazil, to get enrolled in a Public University, candidates need to pass a very competitive examination. Unicamp has an entrance examination consisting of two parts. The first exam consisted of Mathematics, Geography, History, Biology, Chemistry and Physics questions and an Essay Writing. If the candidate was approved in the first one, he/she was able to attend the second exam which consists of questions of Mathematics, Portuguese, Geography, History, Biology, Chemistry, Physics and English. In this analysis we are considering only the scores obtained in the second exam.

2.1 Motivation

We performed an exploratory data analysis to visualize the shape of the response variable (Y_1) and to evaluate the sample correlation between some explanatory variables and the response variable Y_1. These results guided us toward the most appropriate covariates for the regression models presented and discussed in Sections 3 and 4.

Table 1 shows the mean and variance of the number of failed courses (Y) by year of entrance, showing that there may be over-dispersion and the Poisson might not be an appropriate model.

Table 1: Mean and variance of the number of failed courses by year of entrance

	2000	2001	2002	2003	2004	2005
Mean	4.7	4.0	5.1	5.6	6.6	6.9
Variance	42.8	37.2	45.9	58.4	68.2	57.5

From Figure 1(a), one can see that the distribution of proportions in the $(0, 1)$ interval is asymmetric and there are approximately 30% of zeros (28.35%) and about 1% of ones (1.21%). Figure 1(b) shows the box plots for the proportion of failed courses according to graduation status and number of semesters in the university. Notice that the pattern of the variability of Y_1 varies according to number of semesters and graduation status. In particular, notice the greater dispersion among those who did not graduate compared to those who graduated. This behavior motivates the modeling of the variability as function of covariates. Still regarding Figure 1(b), it is possible to verify that the students who graduated present median and mean (represented by the dots) of the proportion of failed courses close to zero. However, those who did not graduate present median and mean of the proportion of failed courses between 0.30 to 0.4.

We investigated the relationship among quantitative variables in the data set using Spearman correlation, with results given in Figure 2, where Y_1 is the proportion of failed courses.

From Figure 2, one can see that the correlations between ESS and Y_1 are all very low, with all the correlations being less than 0.2, in absolute value. Looking at the correlations related to EES, the highest correlations are between Geography and History (0.38), Physics and Math (0.37), Physics and Chemistry (0.32), respectively in that order.

3 Statistical Models

In our problem the response variable Y_i represents the number of courses that a student i failed in a total of m_i courses ($Y_i \leq m_i$). Although there exists a

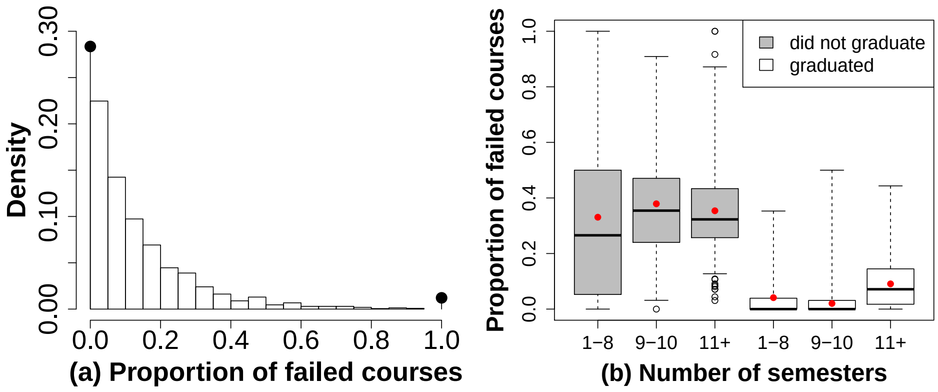


Figure 1: (a) Distribution of proportion of failed courses with probability mass at zero and one and (b) Box plots for the proportion of failed courses according to number of semesters in the university and graduation status.

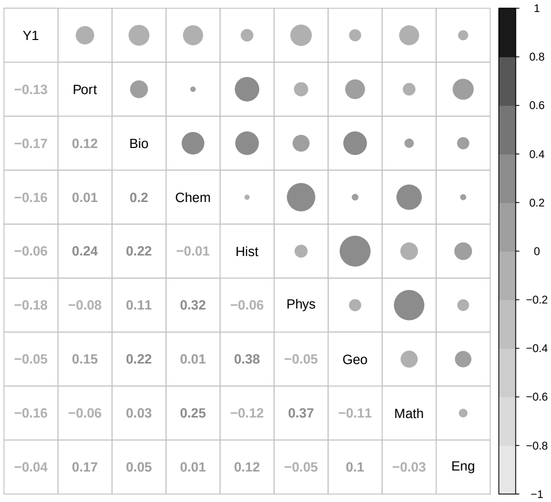
wide range of discrete models in the literature, we model Y_i with binomial, Poisson, negative binomial and beta-binomial distributions. Moreover, as the data set has a lot of students who did not fail any course (about 28.3%), it is possible that the proportion of zeros may be inflated. Therefore, it may be reasonable to fit count models with over-dispersion caused by extra zeros and beyond that. For the count models with Poisson and negative binomial, we consider m_i in the logarithm scale as an offset in the modeling of the mean.

On the other hand, one may define the previous response variable as $Y_{1i} = Y_i/m_i$ (the proportion of failed courses). Therefore, to model $Y_{1i} \in [0, 1]$ one may consider a beta regression model augmented in 0 and 1 ([13]). Parameter estimation was carried out by maximum likelihood in the context of mixture models, modeling the mean, variance related parameters and mixture probabilities as functions of explanatory variables.

3.1 Modeling the incidence of failed courses

Let Y_i be the number of courses student i failed in a total of m_i courses, $i = 1, \ldots, n$. In the first approach, we model the mean of Y_i, say μ_i, as the probability of student i fail a course when a binomial model and a beta-binomial model is considered as the distribution of Y_i [10].

In the second approach we model Y_i as a Poisson and a negative binomial model, and μ_i/m_i is the mean incidence of failed courses by student i. Here, we model μ_i considering the offset as the logarithm of m_i.



Y1								
−0.13	Port							
−0.17	0.12	Bio						
−0.16	0.01	0.2	Chem					
−0.06	0.24	0.22	−0.01	Hist				
−0.18	−0.08	0.11	0.32	−0.06	Phys			
−0.05	0.15	0.22	0.01	0.38	−0.05	Geo		
−0.16	−0.06	0.03	0.25	−0.12	0.37	−0.11	Math	
−0.04	0.17	0.05	0.01	0.12	−0.05	0.1	−0.03	Eng

Figure 2: Spearman's correlations between the proportion of failed courses (Y_1) and all EES.

If Y_i follows a zero inflated beta-binomial (ZIBB) model, the distribution function may be given by

$$\begin{aligned} p(y_i|\mu_i,\sigma_i,\tau_i) &= \tau_i I(y_i = 0) \\ &+ (1-\tau_i)\binom{m_i}{y_i}\frac{B\left(y_i+\frac{\mu_i}{\sigma_i}, m_i - y_i + \frac{1-\mu_i}{\sigma_i}\right)}{B\left(\frac{\mu_i}{\sigma_i}, \frac{1-\mu_i}{\sigma_i}\right)} I(y_i \in \{0,1,\ldots,m_i\}), \end{aligned} \tag{1}$$

where $\tau_i = P(Y_i = 0)$, $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, and μ_i is the probability of a random selected student i fail. $E(Y_i) = (1-\tau_i)m_i\mu_i$ and $V(Y_i) = (1-\tau_i)(V_i - (m_i\mu_i)^2) - (1-\tau_i)^2(m_i\mu_i)^2$, where $V_i = \frac{m_i\mu_i(1-\mu_i)(1+m_i\sigma_i)}{1+\sigma_i}$.

When Y_i follows a zero inflated negative binomial (ZINB) model, the distribution function may be given by

$$\begin{aligned} p(y_i|\mu_i,\tau_i,\sigma_i) &= [\tau_i + (1-\tau_i)(1+\sigma_i\mu_i)^{-1/\sigma_i}]I(y_i = 0) \\ &+ (1-\tau_i)\frac{\Gamma(y_i + 1/\sigma_i)}{\Gamma(1+y_i)\Gamma(1/\sigma_i)}\left(\frac{\sigma_i\mu_i}{1+\sigma_i\mu_i}\right)^{y_i}\left(\frac{1}{1+\sigma_i\mu_i}\right)^{1/\sigma_i} I(y_i \in \{1,2,\ldots\}), \end{aligned} \tag{2}$$

with mean and variance given by $E(Y_i) = (1-\tau_i)\mu_i$ and $V(Y_i) = (1-\tau_i)\mu_i^2(\sigma_i+1)$, respectively. Now, μ_i is the mean number of failed courses for a random selected student i.

3.2 Modeling the proportion of failed courses

We modeled the proportion of failed courses $Y_{1i} = Y_i/m_i$ using a zero-one augmented beta distribution, whose distribution is given by

$$\begin{aligned} p(y_{1i}; p_{0i}, p_{1i}, \mu_{1i}, \phi_i) &= p_{0i} I(y_{1i} = 0) + p_{2i} f(y_{1i}; \mu_{1i}, \phi_i) I(y_{1i} \in (0,1)) \\ &+ p_{1i} I(y_{1i} = 1) \qquad (3) \end{aligned}$$

with $f(y_{1i}; \mu_{1i}, \phi_i) = [B(\mu_{1i}\phi_i, (1-\mu_{1i})\phi_i)]^{-1} y_{1i}^{\mu_{1i}\phi_i - 1}(1-y_{1i})^{(1-\mu_{1i})\phi_i - 1}$, for $y_{1i} \in (0,1)$ and $p_{2i} = 1 - p_{0i} - p_{1i}$. Moreover, $E(Y_{1i}) = p_{1i} + p_{2i}\mu_{1i}$ and $Var(Y_{1i}) = \frac{p_{0i}p_{1i}}{p_{0i}+p_{1i}} + p_{2i}\frac{\mu_{1i}(1-\mu_{1i})}{(\phi_i+1)} + \frac{p_{2i}}{p_{0i}+p_{1i}}[p_{1i} - \mu_{1i}(p_{0i}+p_{1i})]^2$.

Here, ϕ_i plays the role of a precision parameter in the sense that, for fixed μ_{1i}, the larger the value of ϕ_i, the smaller the variance of Y_{1i}. For more details see [13] and [19]. In our study, $p_{0i} = P(Y_{1i} = 0)$ is the probability of student i be approved in all courses, $p_{1i} = P(Y_{i1} = 1)$ is the probability of student i fails all courses, and $p_{2i} = P(Y_{1i} \in (0,1))$ is the probability of student i fails at least one but not all the m_i courses.

Note that this parametrization induces a restriction in the parameter space given by $0 < p_{0i} + p_{1i} < 1$, with the maximization of the likelihood subject to the restriction $p_{0i} + p_{1i} + p_{2i} = 1$.

We model $\log(\mu_{1i}/(1-\mu_{1i})) = \mathrm{x}_{1i}\boldsymbol{\beta}_1$, $\log(\sigma_i/(1-\sigma_i)) = \mathrm{x}_{2i}\boldsymbol{\beta}_2$, with $\sigma_i = (\phi_i + 1)^{-1/2}$, $\log(\nu_{1i}) = \mathrm{x}_{3i}\boldsymbol{\beta}_3$ and $\log(\tau_{1i}) = \mathrm{x}_{4i}\boldsymbol{\beta}_4$, where $\nu_{1i} = p_{0i}/p_{2i}$ and $\tau_{1i} = p_{1i}/p_{2i}$. Here x_{ji} represents the i^{th} row of the design matrix X_j and $\boldsymbol{\beta}_j$ is the respective vector of fixed effects for $j = 1, \ldots, 4$.

3.3 Residual analysis

We computed randomized quantile residuals [6] for all fitted models. In general, the randomized quantile residuals are defined as follows. Let $y_1, \ldots, y_n$ be a random sample and for each i let $\boldsymbol{x_i}$ be a vector of covariates. Assume y_i's to be independent following a distribution $\mathcal{P}(\mu, \sigma)$. Let $F(y; \mu, \sigma)$ be the cumulative distribution function (cdf) of $\mathcal{P}(\mu, \sigma)$. If F is continuous, then $F(y_i; \mu_i, \sigma_i)$ is uniformly distributed on the unit interval. Then, the quantile residual is defined as $r_{q,i} = \Phi^{-1}\{F(y_i; \hat{\mu}_i, \hat{\sigma}_i)\}$, where $\Phi(.)$ is the cdf of a standard normal distribution if $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ are consistently estimated.

If F is not continuous, according to [6], a more general definition of quantile residuals is required. Now let $a_i = \lim_{y \uparrow y_i} F(y; \hat{\mu}_i, \hat{\sigma}_i)$ and $b_i = F(y_i; \hat{\mu}_i, \hat{\sigma}_i)$. The randomized quantile residual for y_i is defined by $r_{q,i} = \Phi^{-1}(u_i)$, where u_i is a uniform random variable on the interval $(a_i, b_i]$. Therefore, when y_i follows a discrete distribution, at each generation of u_i, the randomized residuals may have different values. In any case, $r_{q,i}$

follows a standard normal distribution apart from sampling variability of $\hat{\mu}_i$ and $\hat{\sigma}_i$. .

In our study, we deal with completely discrete distributions when modeling the count data (incidence of failed courses) and a semi-continuous distribution (the zero-one augmented beta distribution) when modeling the proportion of failed courses. To perform a residual analysis, we generated 100 vectors of randomized residuals based on the model of interest and for each vector of residuals we applied a Shapiro Wilks' test to check for normality of the residuals. Afterwards, we calculated the proportion of times that the null hypothesis (presence of residual normality) was rejected at a significance level of 0.05. This proportion was used as an indicator of lack of normality of the residuals and lack of model fitting.

4 Results

Model inference comprehended parameter estimation, model selection and residual analysis. Parameter estimation was performed by maximum likelihood via the existing routine `gamlss` [18] of `R` software ([17]). Model selection of the discrete models were based on AIC, BIC and residual analysis as well. Details of the residual analysis and model selection are described below.

According to AIC and BIC (Table 2) the best fitting is achieved by the ZIBB model, followed by the Beta-Binomial and ZINB models, respectively. However, following the residual criteria described in Section 3.3 from 100 simulated random residuals, results in Table 2 indicate the ZINB model as the best one because it produces the smallest percentage of median points out of the residual envelopes (10.9%) followed by ZIBB (18.9%) and Beta-Binomial (47.7%) models, respectively. Nonetheless, those percentages are high and leads to a high rate of rejection of residual normality hypothesis. Except for the ZINB model, the percentage of times we rejected the Shapiro-Wilks' test of normality is equal to 100%. In the ZINB model the percentage was 75%.

As an alternative we considered modeling the proportion of failed courses given by $Y_{1i} = Y_i/m_i$ and fitted a zero-one augmented beta distribution to Y_{1i} as now the response belongs to the $[0, 1]$ interval.

As we cannot compare AICs and BICs originated from models whose nature of the response differs, we compared goodness of fit between discrete models (ZIBB and ZINB) and the zero-one augmented beta model by means of a residual analysis. The zero-one augmented beta model presents a median of points out of the residuals envelopes equal to 0.48% and rejects the Shapiro-Wilks' null hypothesis (normality of residuals) in 23% of the cases. The percentages 0.48% and 23% are much lower than percentages produced by the discrete models and they are displayed on Table 2. Therefore, we think it is more appropriate to fit a model for proportion instead of incidence of failed courses. From now on we present model estimates and interpretations based on a zero-one augmented beta model for Y_{1i}.

Table 2: Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) for discrete models.

Distribution	**Regression**	**logL**	**g.l.**	**AIC**	**BIC**	**p.out**[1]
Binomial	μ	-10,457	22	20,958	21,095	99.9%
Poisson	μ	-11,247	22	22,538	22,675	99.9%
Negative-binomial	μ,σ	-8,752	36	17,576	17,800	53.7%
Beta-binomial	μ,σ	-8,366	36	16,805	17,029	47.7%
ZIB	μ,τ	-9,303	42	18,691	18,952	99.1%
ZIP	μ,τ	-9,973	42	20,031	20,292	99.4%
ZINB	μ,σ,τ	-8,656	52	17,417	17,741	10.9%
ZIBB	μ,σ,τ	-8,285	54	16,678	17,014	18.9%

μ : location parameter; σ : dispersion parameter; τ : probability of zeros.
[1] median percentage of points out of the envelopes for the half normal plot from 100 simulated random residuals

Tables 3 and 4 show the estimates of the parameters for the zero-one augmented beta model. In Table 3 we have the estimates of the effects on the proportion of failed courses (μ_1) when $Y_1 \in (0, 1)$, and the estimates for the logarithm of the ratio between the probability of being approved in all courses (p_0) and the probability of failing at least one course but not all of them (p_2), i.e., $\nu_1 = p_0/p_2$. Table 4 shows the results of the model for the dispersion parameter (σ) with a logit link when $Y_1 \in (0, 1)$.

To group levels of courses/majors we performed a cluster analysis based on the Euclidean distance with the agglomeration method being the Ward's minimum variance [12], using the proportion of failed courses as response variable. Three clusters were found, namely *Cluster 1* (Agricultural Engineering, Mechanical Engineering, Civil Engineering), *Cluster 2* (Chemical Engineering (night), Automation and Control Engineering, Electrical Engineering (night), Food Engineering (night)) and *Cluster 3* (Chemical Engineering (daytime), Electrical Engineering (daytime), Food Engineering

Table 3: Final model for the mean proportion of failed courses (μ_1) when $Y_{1i} \in (0,1)$ ($\nu_1 = p_0/p_2$ and $\tau_1 = p_1/p_2$).

	Model for μ_1 with logit link			Model for $\nu_1 = p_0/p_2$ with log link		
	Estimate	Std. Error	P-value	Estimate	Std. Error	P-value
Intercept	-0.65	0.11	< 0.001	-3.23	0.37	< 0.001
Year 2001	-0.02	0.05	0.749	-0.05	0.14	0.711
Year 2002	0.16	0.05	0.002	-0.16	0.14	0.262
Year 2003	0.11	0.05	0.024	-0.09	0.14	0.514
Year 2004	0.17	0.05	0.001	-0.41	0.14	0.005
Year 2005	0.08	0.05	0.119	-0.51	0.15	< 0.001
sex Male	0.23	0.03	< 0.001	-0.31	0.09	< 0.001
age < 17	-0.17	0.03	< 0.001	0.68	0.09	< 0.001
age > 21	0.27	0.06	< 0.001	-0.73	0.24	0.002
Public HS	-0.16	0.04	< 0.001	0.58	0.11	< 0.001
Physics	-0.08	0.01	< 0.001	0.23	0.05	< 0.001
Math	-	-	-	0.19	0.05	< 0.001
Biology	-0.06	0.02	0.005	0.22	0.04	< 0.001
Chemistry	-0.04	0.01	0.003	0.14	0.04	0.002
Portugues	-0.03	0.01	0.038	0.19	0.04	< 0.001
Cluster 2	-0.38	0.04	< 0.001	0.94	0.13	< 0.001
Cluster 3	-0.13	0.03	0.0003	0.76	0.10	< 0.001
graduated	-2.54	0.10	< 0.001	2.68	0.34	< 0.001
1 to 8 semesters	-0.17	0.11	0.111	1.40	0.34	< 0.001
$\geq$ 11 semesters	-0.09	0.11	0.395	-1.53	0.10	< 0.001
grad.*(1 to 8 sem.)	1.37	0.33	< 0.001	-	-	-
grad.*($\geq$ 11 sem.)	1.12	0.11	< 0.001	-	-	-

(daytime), Computational Engineering).

The year of student's entrance in the university stayed in all models as a control variable, since there may be differences in the entrance exams among the years. The reference cells in Table 3, models for μ_1 and ν_1, are: Year 2000, Female, Age $[17, 21]$, Private High School, *Cluster 1*, did not graduate and stayed from 9 to 10 semesters.

The reference cells for the dispersion (σ) model of the proportion of failed courses in Table 4 are: Year 2000, *Cluster 1*, did not graduate and stayed from 9 to 10 semesters in the University. Additionally, it is important to point out that the effect of one variable in model is adjusted by all the other variables, i.e., when we interpret the results for one variable is considering that all the other variables are fixed.

Table 4: Final model for dispersion parameter σ with logit link.

	Estimate	Std. Error	P-value
Intercept	-0.64	0.12	< 0.001
Year 2001	0.05	0.06	0.426
Year 2002	0.13	0.06	0.039
Year 2003	0.09	0.06	0.130
Year 2004	0.12	0.06	0.036
Year 2005	0.16	0.06	0.011
Biology	-0.05	0.02	0.029
Cluster 2	0.001	0.05	0.974
Cluster 3	0.09	0.04	0.022
graduated	-1.31	0.12	< 0.001
1 to 8 semesters	0.15	0.12	0.211
$\geq$ 11 semesters	-0.22	0.13	0.098
grad.*(1 to 8 sem.)	0.91	0.33	0.006
grad.*($\geq$ 11 sem.)	0.88	0.14	< 0.001

Consider the estimates in Table 3 (left column) for the final model on the mean of proportion of failed courses μ_1, when $Y_1 \in (0, 1)$. Estimates show that there is not much difference on the mean of the proportion of failed courses among the years, except for years 2002, 2003 and 2004. Compared with the other years, the mean proportion of failed courses seems to be greater in 2004, being followed by year 2002 and 2003. The mean proportion of failed courses is greater among males than females and greater among older students. Students who went to public high schools tend to have lower mean proportion of failed courses compared to students who went to private high schools. The effects of the entrance exam scores are significant for scores on the subject Physics, Biology, Chemistry and Portuguese, with Physics being the subject with the greater impact on the proportion of failed courses, meaning that students with higher scores in the entrance exams tend to have a smaller mean proportion of failed courses when $Y_1 \in (0, 1)$. Students which majors/courses belong to *Cluster 2* followed by those which majors are in *Cluster 3* presented smaller proportion of failed courses than those students which majors are in *Cluster 1.* The final estimates for effects on $\nu_1 = p_0/p_2$ are in Table 3 (right column). Estimates show that, except for years 2004 and 2005, there is no significant difference on the probability of being approved in all courses and the probability of failing at least one course but not all of them. The probability of being approved in all courses is smaller for males and older students. This probability is the lowest for students which majors belong to *Cluster 1*, followed by those which majors are in *Cluster 3* and in *Cluster 2*.

Now, consider the estimates in Table 4 for the final model on σ when $Y_1 \in (0,1)$. There is an interaction effect between graduation status and number of semesters in the university affecting σ, which makes sense when we look at Figure 1. Consequently, for those who graduated, the smallest variance of the proportion of failed courses is for those who stayed from 9 to 10 semesters in the university, followed by those who stayed at least 11 semesters and then for those who stayed from 1 to 8 semesters. On the other hand, for those who did not graduate, the smallest variance of the proportion of failed courses when $Y_1 \in (0,1)$ is for those who stayed at the University at least 11 semesters, followed by those who stayed from 1 to 9 semesters and then for those those who stayed from 1 to 8 semesters. The variance of the proportion of failed courses is lower for those who graduated. Finally, the smallest dispersion is for those students who graduated and stayed in the University from 9 to 10 semesters, which courses belong to *Cluster 1* and joined Unicamp in 2000. The biggest dispersion is for those who did not graduate, which majors belong to *Cluster 3*, stayed from 1 to 8 semesters at the university and joined Unicamp in 2005.
Finally, the logarithm of the ratio between the probability of failing all courses (p_1) and the probability of failing at least one but not all of them (p_2) is estimated to be -4.06 (with s.e.=0.15), i.e., $\hat{\tau}_1 = \exp(-4.06) = 0.017$ and it is significantly different from zero (p-value < 0.0001). This result implies that the probability of failing at least one but not all courses is about 58 times larger than the probability of failing all of them.

Figure 3 displays the Q-Q plot of 100 randomized residuals from (a) the proportion of failed courses model using a heteroscedastic zero-one augmented beta distribution, (b) the incidence of failed courses using a zero inflated beta-binomial model and (c) a zero inflated negative binomial model. The jittering technique [5] is used here, i.e., we added random noise to data in order to prevent over-plotting in statistical graphs. In Figure 3(a), the grey dots are the residuals from the proportion of failed courses (Y_{1i}) in $(0,1)$, while the black dots are the residuals for the proportion of failed courses in $\{0,1\}$.

In the zero-one augmented beta model we model a response whose nature is a mixture of discrete and continuous random variables, as described in Section 3.2. The residuals in Figure 3(a) are concentrated on the bisector while residuals on Figure 3(b) and (c) tend to deviate from it specially on the upper tail. Additionally, the zero-one augmented beta model presented a median of points out of the residuals envelopes equals to 0.48% and the percentage of times we rejected the Shapiro-Wilks' test is equal to 23%, which are both quite low and much lower when compared to fitting results of the discrete models (see Table 2 and Figures 3(b) and (c)).

5 Discussion

It is important to point out that there is a great competition among students to join Public Universities in Brazil, with entrance exams highly selective. Also, most of middle-class students in Brazil go to Private Schools (Elementary, Middle and High Schools). Therefore, the socioeconomic status could be measured indirectly by looking at the High School system of the students. Unicamp, located in the state of São Paulo, Brazil, is one of the top research universities in Brazil with a highly competitive entrance exam (more than 20 candidates per place). In 2005 Unicamp implemented an affirmative action program, where students who studied all High School years in Public Schools are allowed to receive a bonus in the final score of their entrance exam.

In this work we focus on the use of alternative distributions to model the incidence/proportion of failed courses in order to look for suitable methods to evaluate the performance of undergraduate students. We model heteroscedasticity of incidence and proportion of failed courses of Engineering major students by means of discrete distributions and a zero-one augmented beta distribution, respectively. The residual analysis for the chosen model of the response variable Y_1 suggests the adequacy of the distribution proposed for the response variable, as well as the selection variable process. Heteroscedasticity of variance related parameters was modeled by means of covariates. Overall, the model for proportion of failed courses may bring researchers more insight toward public policies as it is straight forward to interpret results. In particular, under this model we can have a direct estimate of the proportion of students approved in all courses, the proportion of students who failed 100% of the courses and the proportion of students who failed at least 1 but not all courses.

It is important to point out that those students who come from public high schools have presented, on average, a lower proportion of failed courses compared to those who come from private schools. This is a surprising result, since the public high school system in Brazil is known to be worst than the private system. We conjecture this may be due to the fact that many of the public high school students come from technical schools, which are more selective and present better student performance than regular public schools. Also, one needs to remember that the effect of high school in the models is based on the fact that all the other variables are fixed.

Acknowledgements

H.P. Pinheiro's research was partially supported by CNPq-Brazil (308583/2015-9) and FAPESP (2011/15047-7, 2016/07226-2). We thank the following people for their direct support regarding databases used for this study: Renato Hirata, database specialist at the Admissions Committee (Comvest) at Unicamp and Silvio de Souza, database specialist at the Academic Records Division (DAC) at Unicamp.

References

[1] Agasisti, T., Leva, F. and Pagononi, A. M. Heterogeneity, school-effects and the North/South achievement gap in Italian secondary education: evidence from a three-level mixed model. *Statistical Methods & Applications*, **26**, 157–180 (2017).

[2] Azzalini, A. Further results on a class of distributions which includes the normal ones. *Statistica*, **46**, 199–208, (1986).

[3] Bianconcini, S. and Cagnone, S. Multivariate latent growth models for mixed data with covariate effects. *Communications in Statistics - Theory and Methods*, **41**, 3079–3093, (2012).

[4] Birch, E. R. and Miller, P. W. Student outcome at university in Australia: a quantile regression approach. *Australian Economic Papers*, **45**, 1–17, (2006).

[5] Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. Tutorial on methods for interval-censored data and their implementation in R. *Graphical Methods for Data Analysis*, Wadsworth, Belmont, California, (1983).

[6] Dunn, P.K. and Smyth, G.K. Randomized quantile residuals. *Journal of Computational Graphical Statistics*, **5**, 236–244, (1996).

[7] Grilli, L., Rampichini, C. and Varriale, A. Binomial mixture modelling of university credits. *Communications in Statistics - Theory and Methods*, **44**, 4866–4879, (2015).

[8] Grilli, L., Rampichini, C. and Varriale, A. Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: An approach based on quantile regression for counts. *Statistical Modelling*, **16**, 47–66, (2016).

[9] Maia, R. P., Pinheiro, H. P. and Pinheiro, A. Academic performance of students from entrance to graduation via quasi U-statistics: a study at a Brazilian research university. *Journal of Applied Statistics*, **43(1)**, 72–86, (2016).

[10] McCullagh, P. and Nelder, J.A. *Generalized Linear Models*, Chapman & Hall, Boundary Row, London, (1992).

[11] Murray-Harvey, R. Identifying characteristics of successful tertiary students using path analysis. *Australian Educational Researcher*, **20**, 63–81, (1993).

[12] Murtagh, F. and Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, **31**, 274–295, (2014).

[13] Ospina, R. and Ferrari, S.L.P. Inflated beta distributions. *Statistical Papers*, **51**, 111–126, (2010).

[14] Pedrosa, R.H.L., Dachs, J.N.W., Maia, R.P., Andrade, C.Y. and Carvalho, B.S. Academic Performance, Student's Background and Affirmative Action at a Brazilian Research University. *Higher Education Management and Policy*, **19(3)**, 1–20, (2007).

[15] Pinheiro, A., Sen, P.K. and Pinheiro, H.P. Decomposability of High-Dimensional Diversity Measures: Quasi U-Statistics, Martingales and Nonstandard Asymptotics. *Journal of Multivariate Analysis*, **100(8)**, 1645–1656, (2009).

[16] Pinheiro, A., Sen, P.K. and Pinheiro, H.P. A class of asymptotically normal degenerate quasi U-statistics. *Annals of the Institute of Statistical Mathematics*, **63**, 1165–1182, (2011).

[17] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (2016). URL https://www.R-project.org/.

[18] Rigby, R. A. and Stasinopoulos D. M. Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54(3)**, 507–554, (2005).

[19] Rigby, R. A., Stasinopoulos D. M., Heller, G. and Voudouris, V. The Distribution Toolbox of GAMLSS, (2014). (see also http://www.gamlss.org/).

[20] Stasinopoulos, D. M., Rigby R.A. and Akantziliotou, C. Instructions on how to use the GAMLSS package in R. Accompanying documentation in the current GAMLSS help files, (2006). (see also http://www.gamlss.org/).

[21] Stasinopoulos, D. M. and Rigby, R.A. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23(7)**, 1–46, (2007).

[22] Zelleis, A. Kleiber, C. and Jackman, S. Regression Models for Count Data in R. *Journal of Statistical Software*, **27(8)**, 1–25, (2008).

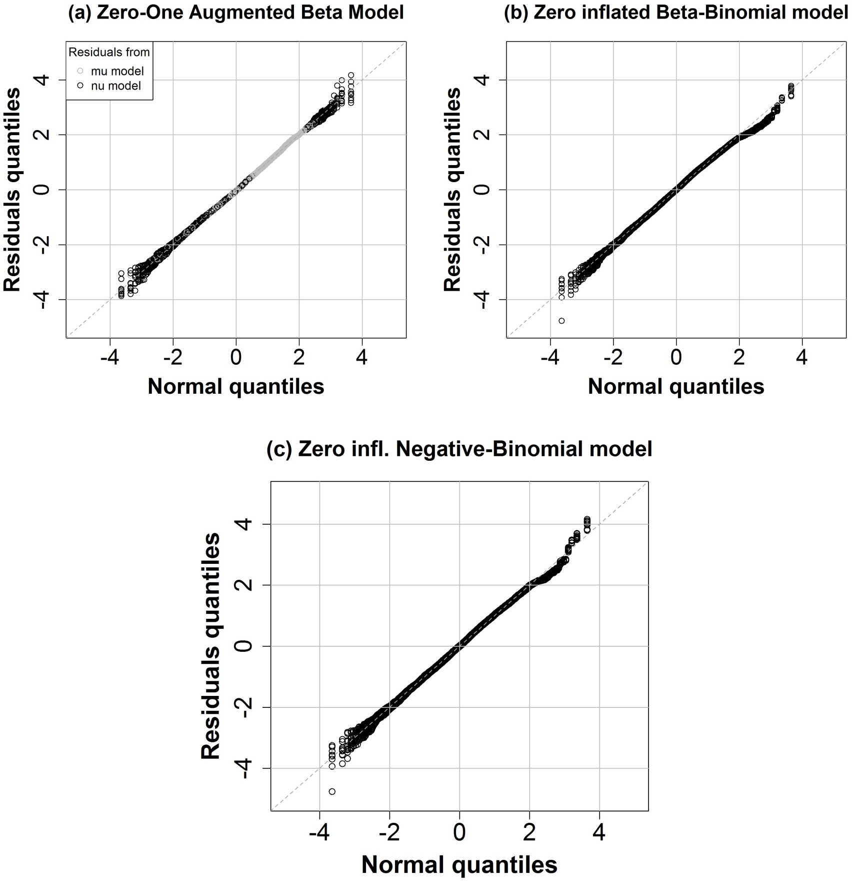


Figure 3: Q-Q plot of the quantile residuals (generated 100 times from the same model) of: (a) zero-one augmented beta model (p.out =0.48% and s.test=23%); (b) zero-inflated beta-binomial model (p.out =18.9% and s.test=100%); and (c) zero-inflated negative binomial model (p.out =10.9% and s.test=75%).