Modelling the proportion of failed courses and GPA scores for engineering major students *

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

Abstract

There is special interest on the factors which may contribute for the best academic performance of undergraduate students. Particularly, in Brazil, because of the recent quota system and affirmative action programs implemented by some universities and the Federal Government, this issue has been of great interest. We use here zero-one inflated beta models with heteroscedasticity to model the proportion of failed courses taken by Engineering major students at the State University of Campinas, Brazil. We also model the grade point average score for those students with a heteroscedastic skew t distribution. The database consists of records of 3,549 students with Engineering major who entered in the University from 2000 to 2005. The entrance exam score in each subject, some academic variables and their socioeconomic status are considered as covariates in the models. A residual analysis based on randomized quantile residuals is performed as well. Finally, we believe that the results found in this study can be useful to improve the university polices for new students since it was possible to identify student profiles with respect to their academic performance.

Keywords: *academic performance; beta inflated models; educational data; heteroscedasticity; quantile residuals; skew t distribution.*

*Corresponding author: *E-mail address*: hildete@ime.unicamp.br (H. Pinheiro)

1 Introdution

In Brazil, because of a quota system implemented by the Federal Government and some affirmative action programs implemented in some universities, there is special interest on the factors which contribute more for the best performance of undergraduate students in the universities. Pedrosa et al. (2007) proposed linear regression models to evaluate the performance of students and Maia et al. (2016) studied the performance of groups of undergraduate students using nonparametric methods based on quasi U-statistics.

Using data from the State University of Campinas (Unicamp), Brazil, Pedrosa et al. (2007) proposed linear regression models to assess the performance of undergraduate students using as response variable the *relative gain* which is based on the relative rank of his/her final (or last) recorded GPA (Grade Point Average) and his/her total entrance exam score (EES) rank. Maia et al. (2016) used more robust methods to evaluate the performance of students in different groups using nonparametric methods on quasi U-statistics (Pinheiro et al., 2009; Pinheiro et al., 2011). Here, our goal is to find alternative and more suitable distributions to model the GPA and the proportion of failed courses during the Bachelor's degree of students with Engineering major to evaluate the performance of those students.

The database consists of 3,549 records of students with Engineering major who entered in the University from 2000 until 2005. For each student we have all the grades in the required courses taken in the university as well as the proportion of courses failed during Bachelor's degree. We also have entrance exam scores - EES (e.g., SAT scores) in each subject (Mathematics, Portuguese, Geography, History, Biology, Chemistry and Physics), some academic variables as well as socioeconomic status, which are considered as covariates in the models.

For modelling the proportion of failed courses, we use a zero and one inflated beta regression model (Ospina and Ferrari, 2010) with heteroscedasticity, i.e., we model the proportion of zeros, say ν_1, and ones, say τ_1, with log link function as well as the proportion of failed courses between 0 and 1 with a beta distribution with logit link and the dispersion parameter with logit link.

The model for the GPA score is a heteroscedastic skew t (Azzalini, 1996) with identity link for the mean and log link for the dispersion. Here the GPA scores were standardized within each year and course, i.e., within each year and course the GPA has mean zero and standard deviation one.

2 Statistical Models

GAMLSS is a general framework for regression models where the distribution of the response variable does not have to belong to the exponential family and includes highly skew and kurtotic continuous and discrete distribution. The systematic part of the model is expanded to allow modeling not only the mean (or location) but other parameters of the distribution of Y as, linear and/or non-linear, parametric and/or additive non-parametric functions of explanatory variables and/or random effects. Hence, GAMLSS is especially suited for modeling a response variable which does not follow an exponential family distribution or which exhibit heterogeneity, e.g., where the scale or shape of the distribution of the response variable changes with explanatory variables (Rigby and Stasinopoulos, 2005).

For the proportion of failed courses a GAMLSS model with zero-one inflated beta distribution was used and its distribution is given by

$$p(y;\alpha,\gamma,\mu_1,\phi) = \begin{cases} \alpha(1-\gamma), & \text{if } y = 0 \\ (1-\alpha)f(y;\mu_1,\phi), & \text{if } y \in (0,1) \\ \alpha\gamma, & \text{if } y = 1 \\ 0, & \text{if } y \notin [0,1] \end{cases} \tag{1}$$

with $f(y;\mu_1,\phi) = \dfrac{\Gamma(\phi)}{\Gamma(\mu_1\phi)\Gamma((1-\mu_1)\phi)} y^{\mu_1\phi-1}(1-y)^{(1-\mu_1)\phi-1},\ y \in (0,1)$.

Note that now $E(Y) = \alpha\gamma + (1-\alpha)\mu_1$ and $Var(Y) = \alpha V_1 + (1-\alpha)V_2 + \alpha(1-\alpha)(\gamma-\mu_1)^2$, with $V_1 = \gamma(1-\gamma)$ and $V_2 = V(\mu_1)/(\phi+1)$. The variance function is $V(\mu_1) = \mu_1(1-\mu_1)$ and ϕ plays the role of a precision parameter in the sense that, for fixed μ_1, the larger the value of ϕ, the smaller the variance of Y. For more details see Ospina and Ferrari (2010).

Here we model μ_1 and $\sigma = 1/(\phi+1)$ with a logit link and $\nu_1 = \alpha(1-\gamma)$ and $\tau_1 = \alpha\gamma$ with a log link, i.e., $log(\mu_1/(1-\mu_1)) = \mathrm{X}_1\boldsymbol{\beta}_1$, $log(\sigma/(1-\sigma)) = \mathrm{X}_2\boldsymbol{\beta}_2$ $log(\nu_1) = \mathrm{X}_3\boldsymbol{\beta}_3$ and $log(\tau_1) = \mathrm{X}_4\boldsymbol{\beta}_4$.

For the GPA scores, we used the GAMLSS model with a skew-t type 1 (ST1) distribution and its distribution is given by

$$F_Y(y \mid \mu_2,\sigma^*,\nu_2,\tau_2) = \frac{2}{\sigma^*} f_{Z_1}(z) F_{Z_1}(\nu_2 z), \tag{2}$$

for $y \in (-\infty,\infty)$, where $\mu_2 \in (-\infty,\infty)$, $\sigma^* > 0$, $\nu_2 \in (-\infty,\infty)$ and $\tau_2 > 0$, and where $z = (y-\mu_2)/\sigma^*$ and f_{Z_1} and F_{Z_1} are the pdf and cdf of $Z \sim TF(0,1,\tau_2)$, a t distribution with τ_2 degrees of freedom with τ_2 treated as continuous parameter and ν_2 is the skewness parameter. For more details of the ST1 distribution see Azzalini (1986). Here we model μ_2 with identity link and σ^* with log link, i.e., $\mu_2 = \mathrm{X}_5\boldsymbol{\beta}_5$ and $log(\sigma^*) = \mathrm{X}_6\boldsymbol{\beta}_6$. Also τ_2 and ν_2 are estimated. Additionally, $\mathrm{X}_1,\ldots,\mathrm{X}_6$ are design matrices (not necessarily the same) and

$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_6$ are the respective vector of parameters. Both models do not consider the presence of random effects in the systematic component.

For the residual analysis of these models, the randomized quantile residuals (Dunn and Smyth, 1996) were computed. In general, the randomized quantile residuals are defined as follows. Let $y_1, \ldots, y_n$ be responses and for each i let $\boldsymbol{x_i}$ be a vector of covariates. Assume the 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 $F(y_i; \mu_i, \sigma_i)$ are uniformly distributed on the unit interval. Then, the quantile residuals are defined by $r_{q,i} = \Phi^{-1}\{F(y_i; \hat{\mu}_i, \hat{\sigma}_i)\}$, where $\Phi()$ is the cdf of the standard normal. The distribution of $r_{q,i}$ converges to the standard normal if $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ are consistently estimated.

If F is not continuous, according to Dunn and Smyth (1996), 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 r.v. on the interval $(a_i, b_i]$. The $r_{q,i}$ are standard normal distributed, apart from sampling variability in $\hat{\mu}_i$ and $\hat{\sigma}_i$.

3 Application and Results

Initially, we will consider exploratory data analysis techniques to visualize the shape of the response variables: proportion of failed course (Y_1) and GPA score (Y_2), as well as the evaluation of the correlation between some explanatory variables and the response variables. Then, the *gamlss* package, available in the software R, will be used to fit the models for the data (Stasinopoulos et al., 2006, Stasinopoulos and Rigby, 2007).

From Figure 1, one can see that the distribution of the GPA scores for the required courses of the Engineering major students are skewed to the left and for the proportion of failed scores, there is 27.75% of zeros and 1.24% of ones. Therefore, we tried to model the GPA with normal, skewed normal and skewed t distributions. For the proportion of failed courses we used a zero and one inflated beta model.

In order to understand better the quantitative variables of the data set, we computed Spearman correlations between the quantitative variables as can be seen on Table 1, where Y_1 is the proportion of failed courses, Y_2 is the GPA score, W_1 is EES-Physics (EES in Physics), W_2 is EES-Math (EES in Math), W_3 is EES-Biology (EES in Biology), W_4 is EES-Chemistry (EES in Chemistry), W_5 is EES-Portuguese (EES in Portuguese), W_6 is EES-Geography (EES in Geography), W_7 is EES-History (EES in History).

From Table 1, one can see that the highest correlation is between the proportion of failed courses (Y_1) and the GPA score (Y_2), which

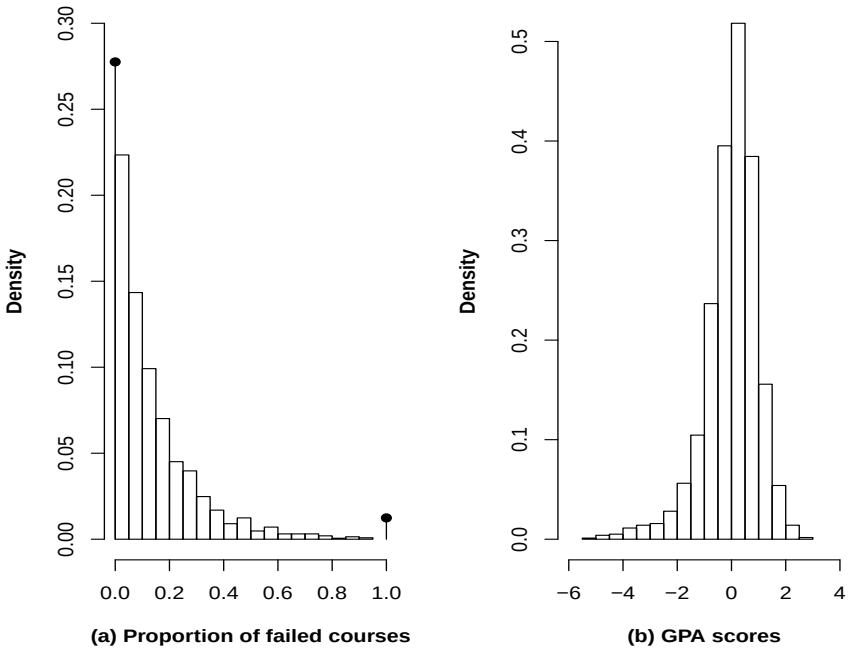

Figure 1: (a) Distribution of proportion of failed courses with probability mass at zero and one and (b) Distribution of GPA scores for engineering major students.

is expected. The correlations between the explanatory variables (EES in all subjects) are all very low, with the higher correlations being between Physics (W_1) and Math (W_2), Physics (W_1) and Chemistry (W_4), Geography (W_6) and History (W_7). All the correlations between the EES's and Y_1 and between the EES's and Y_2 are less than 0.2.

Table 2 shows the best model for the mean proportion of failed courses (μ_1) and for the proportion of zeros (ν_1). Table 3 shows the results of the model for the dispersion parameter (σ). Note that here the larger the $\hat{\sigma}$ value, the larger is the variance of Y_1.

The course/major codes on Table 2 are: 8 = Agricultural Engineering; 9 = Chemical Engineering (daytime); 10 = Mechanical Engineering; 11 = Electrical Engineering (daytime); 12 = Civil Engineering; 13 = Food Engineering (daytime); 34 = Computational Engineering, 39 = Chemical Engineering (night); 41 = Electrical Engineering (night); 43 = Food Engineering (night); 49 = Automation and Control Engi-

Table 1: Spearman correlations.

	Y_1	Y_2	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Y_1.	1.000	-0.858	-0.178	-0.159	-0.171	-0.154	-0.135	-0.057	-0.058
Y_2	-	1.000	0.186	0.163	0.188	0.169	0.158	0.092	0.080
W_1	-	-	1.000	0.371	0.113	0.322	-0.075	-0.046	-0.054
W_2	-	-	-	1.000	0.027	0.257	-0.060	-0.112	-0.121
W_3	-	-	-	-	1.000	0.197	0.133	0.231	0.226
W_4	-	-	-	-	-	1.000	0.009	0.021	-0.004
W_5	-	-	-	-	-	-	1.000	0.157	0.239
W_6	-	-	-	-	-	-	-	1.000	0.385
W_7	-	-	–	-	-	-	-	-	1.000

neering.

The parameters for the categories: Year 2000, Female, Age $[17, 21]$, Private High School, Course 10, did not graduate and stayed 9 to 10 semesters are set to zero for the model of the proportion (μ_1) of failed courses (Table 2).

In the model for the proportion of zeros, ν_1, (Table 2) the reference cell is Year 2000, Female, Age $[17, 21]$, monthly income < 3 minimum salaries (m.s.), Course 10, did not graduate and stayed 9 to 10 semesters.

Reference cell for the dispersion (σ) model of Table 3 is Age ≤ 21; Courses: (10, 11, 13, 34, 39, 41, 43, 9); did not graduate and stayed 9 to 10 semesters.

The model on Table 2 shows that there is not much difference on the proportion of failed courses among the years, but the proportion of zeros seems to be smaller in 2005, followed by 2004 compared with the other years. Male students have greater proportion of failed courses than Female, but the proportion of zeros is smaller for Male than Female students. The younger the students the fewer courses they fail and, of course, the proportion of zeros is higher for younger students. The significant EES in the model are Physics (W_1), Biology (W_3), Chemistry (W_4) and Portuguese (W_5). The higher the score in the Entrance Exam the less courses they fail. For the model of proportion of zeros, the EES were not significant. The students with lower income have a larger proportion of zeros. The lowest proportion of failed courses is of the Automation and Control Engineering students followed by Food Engineering (night) students. The course/major with the higher proportion of failed courses is Mechanical Engineering. When looking at the model for ν_1, the highest proportion of zeros is for the Food Engineering students and the smallest is for Civil Engineering

Table 2: Final models for the proportion (μ_1) of failed courses and for the proportion of zeros (ν_1).

	Model for μ_1 with logit link			Model for ν_1 with log link		
	Estimate	Std. Error	$Pr(>\|t\|)$	Estimate	Std. Error	$Pr(>\|t\|)$
Intercept	-0.42	0.11	0.0002	-2.35	0.46	< 0.001
Year 2001	-0.04	0.05	0.500	-0.02	0.15	0.886
Year 2002	0.11	0.05	0.026	-0.20	0.14	0.146
Year 2003	0.08	0.05	0.108	-0.14	0.14	0.299
Year 2004	0.11	0.05	0.021	-0.43	0.14	0.002
Year 2005	0.01	0.05	0.786	-0.54	0.15	< 0.001
sex Male	0.18	0.04	< 0.001	-0.26	0.11	0.014
age < 17	-0.17	0.03	< 0.001	0.64	0.09	< 0.001
age > 21	0.33	0.06	< 0.001	-0.72	0.24	0.003
Public HS	-0.16	0.04	< 0.001	-	-	-
W_1	-0.08	0.01	< 0.001	-	-	-
W_3	-0.04	0.01	0.006	-	-	-
W_4	-0.04	0.01	0.003	-	-	-
W_5	-0.04	0.01	0.011	-	-	-
3 to 10 m.s.	-	-	-	-0.53	0.26	0.040
> 10 m.s.	-	-	-	-0.73	0.26	0.004
course 11	-0.38	0.06	< 0.001	0.65	0.19	< 0.001
course 12	-0.19	0.05	< 0.001	-0.05	0.18	0.788
course 13	-0.38	0.06	< 0.001	0.57	0.18	0.002
course 34	-0.10	0.05	0.063	0.74	0.17	< 0.001
course 39	-0.38	0.07	< 0.001	1.09	0.22	< 0.001
course 41	-0.43	0.08	< 0.001	1.01	0.23	< 0.001
course 43	-0.52	0.08	< 0.001	1.32	0.22	< 0.001
course 49	-0.64	0.07	< 0.001	0.33	0.21	0.115
course 8	-0.28	0.05	< 0.001	-0.04	0.21	0.839
course 9	-0.22	0.06	< 0.001	0.56	0.18	0.001
graduated	-2.57	0.10	< 0.001	2.72	0.33	< 0.001
1 to 8 semesters	-0.15	0.10	0.142	1.43	0.34	< 0.001
≥ 11 semesters	-0.14	0.11	0.182	1.57	0.10	< 0.001
grad.*(1 to 8 sem.)	1.50	0.34	< 0.001	-	-	-
grad.*(≥ 11 sem.)	1.16	0.11	< 0.001	-	-	-
$\hat{\tau}_1 = exp(-4.05) = 0.017$ (s.e.=0.15 with p-value < 0.0001)						

Table 3: Final model for dispersion parameter σ for the proportion of failed courses with logit link.

	Estimate	Std. Error	$Pr(>\|t\|)$
Intercept	-0.45	0.11	< 0.001
age > 21	0.17	0.07	0.018
course 12	-0.11	0.05	0.040
course 49	-0.19	0.07	0.007
course 8	-0.28	0.06	< 0.001
graduated	-1.31	0.12	< 0.001
1 to 8 semesters	0.18	0.12	0.157
≥ 11 semesters	-0.28	0.13	0.033
grad.*(1 to 8 sem.)	1.00	0.33	0.003
grad.*(≥ 11 sem.)	0.92	0.14	< 0.001

students. There is an interaction effect between the status of graduation and the number of semesters in the university in the model for μ_1, which makes sense when we look at Figure 2. For those who graduated, the lower proportion of failed courses is for those who stayed 9 to 10 semesters in the university, followed by those who stayed at least 11 semesters and then those who stayed 1 to 8 semesters. On the other hand, for those who did not graduate, the lower proportion of failed courses is for those who stayed 1 to 8 semesters, followed by those who stayed at least 11 semesters and then those who stayed 9 to 10 semesters. The proportion of zeros is higher for those who graduated. The proportion of ones (τ_1) is estimated to be 0.017 and it is significantly different from zero (p-value < 0.0001).

From Table 3, one can see that the smallest dispersion is for those who graduated and stayed 9 to 10 semesters, are at most 21 years of age and from Agricultural Engineering (course 8). The biggest dispersion is for those who did not graduate, are over 21 years old, are from the baseline courses (10, 11, 13, 34, 39, 41, 43, 9) and stayed 1 to 8 semesters at the university.

Figure 3 displays the Q-Q plot of the (normalized quantile) residuals from the proportion of failed courses model. The grey dots are the residuals from the proportion of failed courses (μ) $\in (0, 1)$, while the black dots are the residuals for the proportion of failed courses $\in \{0, 1\}$, where the ν_1 model estimates the proportion of zeros and $\hat{\tau}_1$ is the estimated proportion of ones. Note that in this case, the random variable is mixed (discrete and continuous), as described in (1). Therefore, following Dunn and Smith (1996), since we have a mixture of continuous and discrete distributions, the randomized residuals for

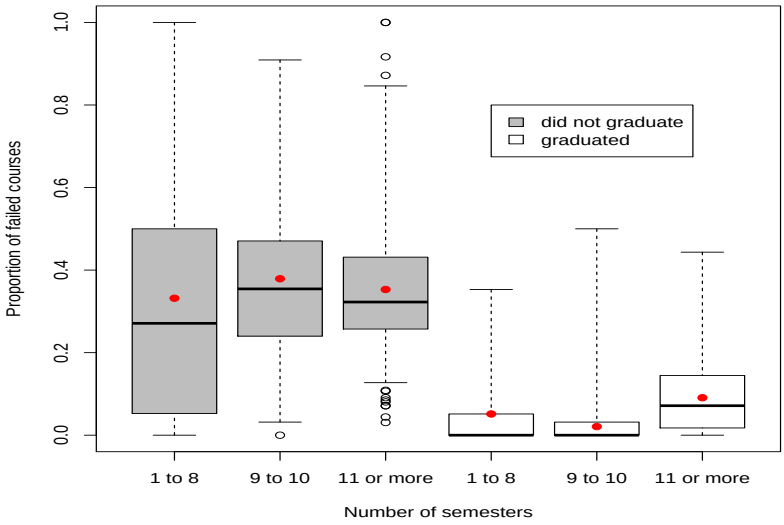

Figure 2: Box plots for the proportion of failed courses according to number of semesters in the university and status of graduation.

the discrete component are not the same each time they are generated for the same model. For this reason we generate 1,000 vectors of randomized residuals based on the same model, and for each vector of residuals we applied a Shapiro Wilks' normality test. In 91.36% of the tests the p-values were greater than 0.05 and the normality hypothesis were not rejected. Figure 3 displays the first one hundred generated randomized residuals. The jittering technique (Chambers et al., 1983) is used here, i.e., the act of adding random noise to data in order to prevent overplotting in statistical graphs.

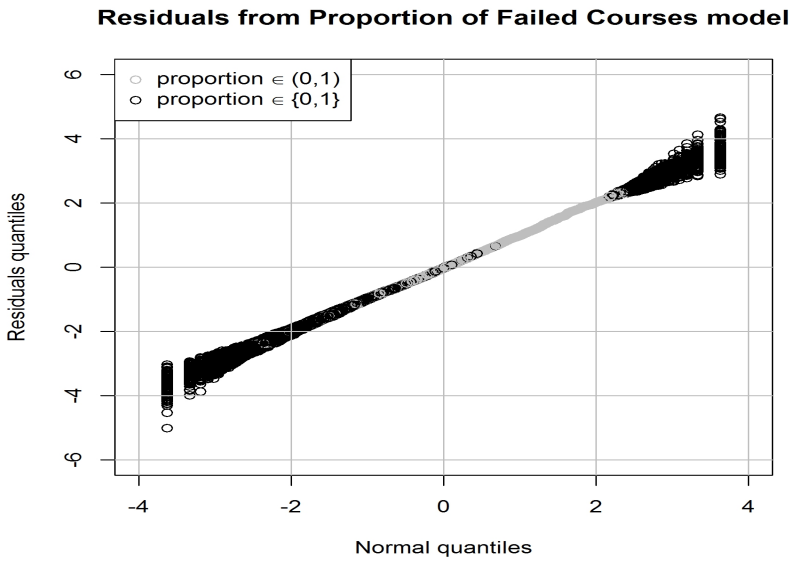

Figure 3: Q-Q Plot of the residuals of the proportion of failed courses model (generated 100 times from the same model).

The best model for the GPA score is a heteroscedastic skew t with identity link for the mean and log link for the dispersion. Table 4 shows the results of the models for GPA mean (μ_2) and the dispersion (σ^*). One can say that the younger the student and the greater his/her EES's, the greater is his/her GPA. Students from Public High Schools (PuHS) and Female have greater GPAs. There is an interaction between the status of graduation and the number os semesters, which confirms the results seen in Figure 4. For those who graduated and stayed 9 to 10 semesters, the GPA score is greater than those who drop out or were still active. Also, the more semesters they stayed in the University, the worst is their GPA. The skewness parameter (ν_2) was found to be negative and significantly different from zero, which makes sense, as one can see from the distribution of the GPA in Figure 1. The model for the dispersion parameter (σ^*) showed that only the status of graduation and the number of semesters was found to be significant, but with an interaction between them. The greater variability was found to be for those who did not graduate and stayed 1 to 8 semesters in the university, which makes sense as these are the students who drop out for various reasons. On the other hand, the smallest variability was found to be for those who graduated in 9 to 10 semesters.

Table 4: Final model for GPA score with skew t distribution.

	Model for μ_2 with identity link			Model for σ^* with log link		
	Estimate	Std. Error	$Pr(>\|t\|)$	Estimate	Std. Error	$Pr(>\|t\|)$
Intercept	-0.81	0.14	< 0.001	-0.15	0.10	0.116
sex Male	-0.15	0.03	< 0.001	-	-	-
age < 17	0.22	0.02	< 0.001	-	-	-
age > 21	-0.21	0.05	< 0.001	-	-	-
Public HS	0.26	0.03	< 0.001	-	-	-
W_1	0.09	0.01	< 0.001	-	-	-
W_2	0.05	0.01	< 0.001	-	-	-
W_3	0.08	0.01	< 0.001	-	-	-
W_4	0.05	0.01	< 0.001	-	-	-
W_5	0.08	0.01	< 0.001	-	-	-
courses (13, 43)	-0.13	0.03	< 0.001	-	-	-
course 9	-0.24	0.03	< 0.001	-	-	-
graduated	1.47	0.11	< 0.001	-0.53	0.10	< 0.001
1 to 8 semesters	0.40	0.14	0.003	0.47	0.10	< 0.001
≥ 11 semesters	-0.04	0.12	0.739	-0.06	0.11	0.595
grad.*(1 to 8 sem.)	-0.42	0.19	0.029	-0.24	0.19	0.214
grad.*(≥ 11 sem.)	-0.36	0.13	0.005	0.25	0.12	0.033
$\hat{\nu}_2 = -0.32$ (s.e.=0.13 with p-value=0.018) and						
$\hat{\tau}_2 = exp(2.51) = 12.305$ (s.e.=0.20 with p-value < 0.0001)						

The reference cell for the μ_2 model of Table 4 is Female sex; Age 17 to 21; Private High School (PrHS); Courses: (8, 10, 11, 12, 34, 39, 41, 49); did not graduate; 9 to 10 semesters. The reference cell for the σ^* model of Table 4 is did not graduate; 9 to 10 semesters.

Figure 5 shows the Q-Q plot of the (normalized quantile) residuals from the GPA model. The residuals are all within the confidence band and the Shapiro-Wilks test of normality gives a p-value of 0.6064, showing that we could not reject the hypothesis of normality of the residuals. Here, since the distribution of GPA is continuous there is no sampling variability, therefore, the residuals do not vary as in the model for proportion of failed courses.

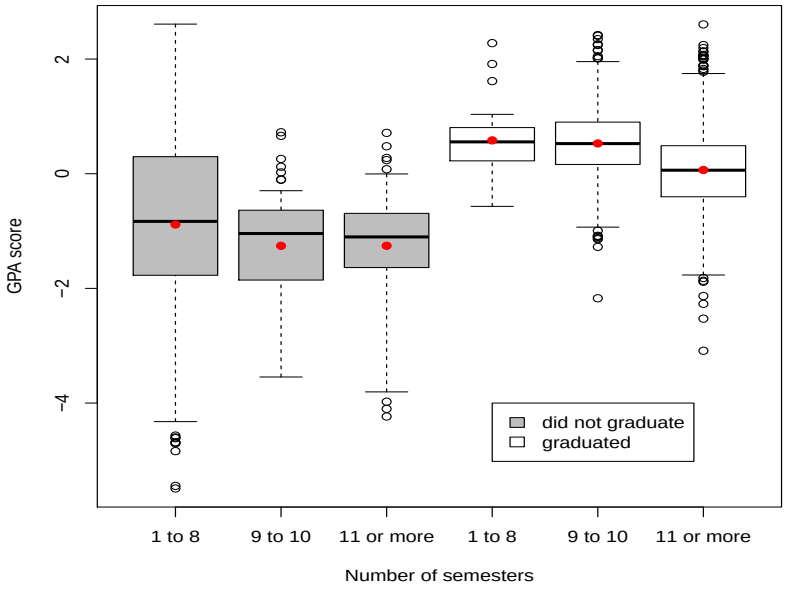

Figure 4: Box plots for GPA scores according to number of semesters in the university and status of graduation.

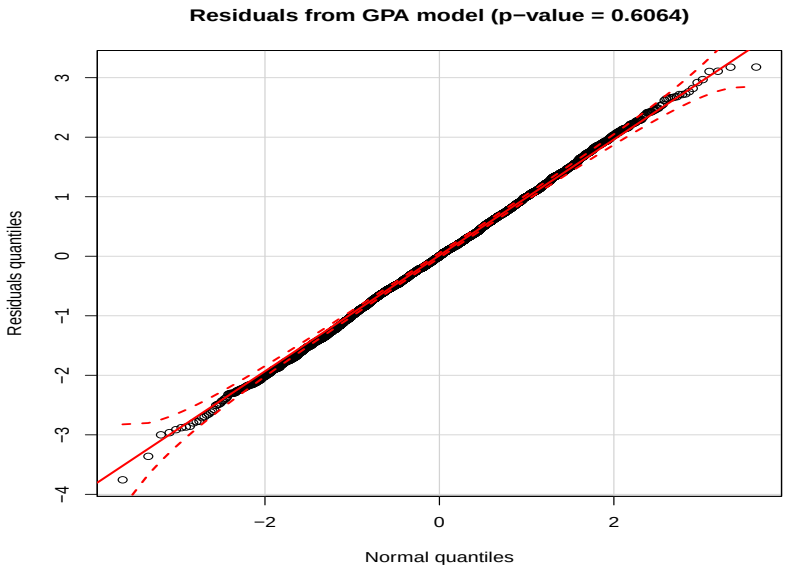

Figure 5: Q-Q plot of the residuals of the GPA model.

4 Conclusion

The implementation of a quota system in Public Federal Universities in Brazil and some affirmative action programs in Public State Universities in São Paulo has brought more interest in the study of performance of undergraduate students. It is important to point out that there is a great competition to enter in Public Universities in Brazil, with entrance exams (e.g., SAT) highly competitive. Also, most of middle-class students go to Private Schools (Elementary, Middle and High Schools). Therefore, the socioeconomic status can be measured indirectly by looking at the High School system of the students.

The State University of Campinas (Unicamp), located in the state of São Paulo, Brazil, is one of the top research universities in Brazil with a highly selective Entrance Exam. 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 order to look for suitable methods to evaluate the performance of undergraduate students, in this work we focus on the use of alternative distributions for the GPA score and the proportion of failed courses. We considered a zero-one beta inflated model with heteroscedasticity to model the proportion of failed courses in Engineering major students as well as the GPA score for those students from Unicamp with a heteroscedastic skew-t distribution. The EES, some academic variables and their socioeconomic status were considered as covariates in the models.

With respect to the proportion of failed courses model, this study highlights that Male students have greater proportion of failed courses than Female. The younger the students the fewer courses they failed and, of course, the proportion of zeros is higher for younger students. The EES of Physics, Biology, Chemistry and Portuguese were significant for the proportion of failed courses model. Moreover, the higher is the performance of the students in these exams the lower is the proportion of failed courses. The students with lower income presented a larger proportion of zeros.

For the GPA models, we would like to highlight that the younger is the student and the greater is their EES's, the greater is their GPA. The EES of Physics, Biology, Chemistry, Portuguese and Mathematics were significant for this model. Moreover, the more semesters they stayed in the university the worst is their GPA. It is important to point out that students from Public High Schools have presented a higher GPA and a lower proportion of failed courses.

The residual analysis obtained in both model suggests the adequacy of the distribution proposed for the response variables Y_1 and Y_2 as well as the selection variable process and the accuracy of the parameter

estimates.

Finally, this study can be useful to improve the university policies for new students since it was possible to identify student profiles with respect to their academic performance.

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: Antonio Faggiani, former Coordinator of the Academic Records Division (DAC) at Unicamp, Renato Hirata, database specialist at the Admissions Committee (Comvest) at Unicamp and Silvio de Souza, database specialist at DAC.

References

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

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Wadsworth, Belmont, California.

Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Comput. Graph. Statist.*, **5**, 236 – 244.

Maia, R. P., Pinheiro, H. P. and Pinheiro, A. (2016). 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.

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

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

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

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

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

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