Modelling performance of students with generalized linear mixed models

Hildete P. Pinheiro[1], Mariana Rodrigues-Motta[1], Gabriel Franco[1]

[1] University of Campinas, Brazil

E-mail for correspondence: hildete@ime.unicamp.br

Abstract: We propose generalized linear mixed models (GLMM) to evaluate the performance of undergraduate students from the State University of Campinas (Unicamp). For each student we have the final GPA score as well as the number of courses he/she failed during his/her Bachelor's degree. The courses are separated in three categories: Required (R), Elective (E) and Extracurricular courses (Ex). Therefore, for each response variable, each student may have at most three measures. In this model we need to take into account the within student correlation between required, elective and extracurricular courses.
The main purpose of this study is the sector of High School education from which college students come - Private or Public. As some affirmative action programs are being implemented by the Brazilian government to include more students from Public Schools in the Universities, there is a great interest in studies of performance of undergraduate students according to the sector of High School of which they come from. The data set comes from the State University of Campinas (Unicamp), a public institution, in the State of São Paulo, Brazil and one of the top universities in Brazil. The socioeconomic status and academic data of more than 10,000 students admitted to Unicamp from 2000 through 2005 forms the study database.

Keywords: generalized linear mixed models; multivariate analysis; zero inflated models; overdispersion.

1 Introduction

Many authors have been working to study the performance of undergraduate students. Pedrosa et al. (2007) proposed 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 entrance exam grade rank. Maia et al. (2013) tried to find more robust methods to evaluate the performance of students in different groups using nonparametric methods such as those based on quasi U-statistics (Pinheiro et al., 2011).

The class of generalized linear mixed models (GLMM) have been studied by Williams (1982); Zeger et al. (1988); Breslow and Clayton (1993); Breslow and Lin (1995); Schall (1991); Rodrigues-Motta et al. (2013) and others. Here, we will use GLMM to model the GPA scores and the number of failed courses by the students. Note that in this case we have two different distributions for the response variables: a continuous variable (which may be normal distributed) and a count variable (which may be Poisson or negative binomial distributed, zero inflated or not).

For each student we have all the grades in the courses taken in the university as well as the number of courses he/she failed during his/her Bachelor's degree. The courses are separated in three categories: Required (R), Elective (E) and Extracurricular (Ex) courses. Then, we could get the GPA score for each type of course (R, E or Ex). We also have their entrance exam grades (e.g., SAT scores) as well as their socioeconomic status, which are going to be considered as covariates in the models.

The main purpose of this study is the sector of High School education from which college students come - Private (Pr) or Public (Pu). In Brazil, the great majority of middle class students go to Private High Schools (around 70%). The data set comes from the State University of Campinas (Unicamp), a public institution, located in the State of São Paulo and one of the top research universities in Brazil. The socioeconomic data of more than 10,000 students admitted to Unicamp from 2000 through 2005 forms the study database.

2 Statistical Methods

Let $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2})^\top$ be the vector of observations of subject i for courses of type j. Here, j can be up to three, i.e, $j = 1, \ldots, m_i$ (m_i is the number of types of courses taken by individual j), since a student may not have taken extracurricular courses or may have dropped out and not taken extracurricular or elective courses. We will use two independent GLMMs to fit Y_{ij1} and Y_{ij2} separately, using a normal distribution for Y_{ij1} (the GPA score) and a discrete distribution for Y_{ij2} (the number of failed courses). We will consider Poisson, Negative Binomial and zero-inflated models for Y_{ij2}, since we are dealing with count data that can be zero-inflated or have overdispersion. Looking at Figure 1, one can see that there is a high frequency of zero failed courses, specially in Biological, Arts and Social Sciences, which may be an indication of overdispersion in a Poisson model.

Now, let $\mathbf{Y}_{i1}$ be the vector of GPA scores for R, E and Ex and $\mathbf{Y}_{i2}$ be the vector of the number of courses failed by individual i.

In the GLMM notation, we have

$$h(\mu_{ijk}) = \mathbf{x}_{ijk}^\top \boldsymbol{\beta} \text{ and } \mathrm{Var}(Y_{ijk}) = g(\mu_{ijk})\phi, \;\; k = 1, 2, \tag{1}$$

where $\mu_{ijk} = E(Y_{ijk})$, $h(\mu_{ijk})$ is some link function, $g(\mu_{ijk})$ is the variance function and ϕ is the dispersion parameter.

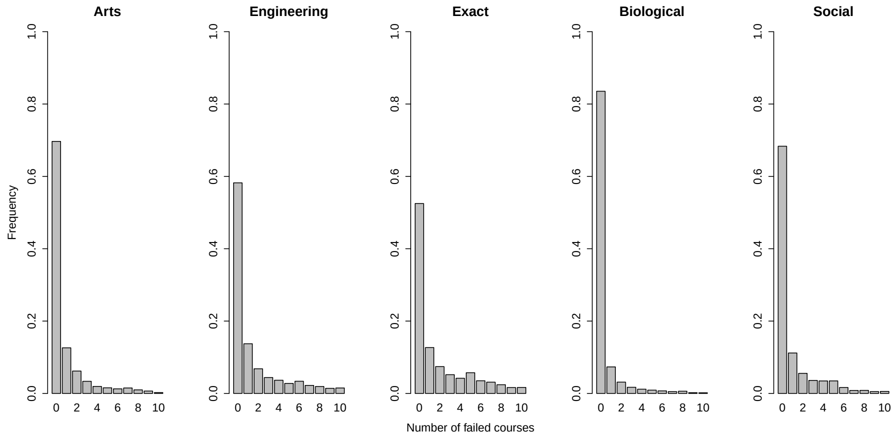

FIGURE 1. Distribution of the number of courses failed according to each area.

For the GPA scores (Y_{ij1}), with a normal distribution, the model is

$$\mu_{ij1} = E(Y_{ij1}) = \mathbf{x}_{ij1}^{\top}\boldsymbol{\beta} \text{ and } \mathrm{Var}(Y_{ij1}) = \sigma^2. \tag{2}$$

For the number of failed courses (Y_{ij2}), with the negative binomial distribution, the model is

$$log(\mu_{ij2}) = log(E(Y_{ij2})) = \mathbf{x}_{ij2}^{\top}\beta + log(N_{ij}) \text{ and } \mathrm{Var}(Y_{ij2}) = \mu_{ij2} + \kappa\mu_{ij2}^2, \tag{3}$$

with $log(N_{ij})$ being the offset, N_{ij} being the total number of courses of type j taken by individual i and κ being the negative binomial dispersion parameter, which will be estimated by Maximum Likelihood. Note that we are modelling here the incidence of failed courses.

Here, we are using the population-average (PA) model described in Zeger et al. (1988). The PA models take into account only the covariance among repeated observations for a subject and their regression coefficients describe the average population response to change in the covariates. The covariance matrix must be positive-definite, but it is unrestricted. In our case, it is reasonable to assume that the GPA scores for elective (E), Extracurricular (Ex) and Required (R) courses are positively correlated as well as the number of failed courses.

The estimation of $\boldsymbol{\beta}$ is done by solving the "generalized estimating equation" (GEE)

$$U(\boldsymbol{\beta}) = \sum_{i=}^{n} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\boldsymbol{\alpha})(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{4}$$

where $\mathbf{V}_i(\boldsymbol{\alpha}) = \mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$, with $\mathbf{A}_i = diag\{g(\mu_{i1}), \ldots, g(\mu_{im_i})\}$, $\mathrm{Cov}(\mathbf{Y}_i) = \mathbf{A}_i\phi$, $\mathbf{R}_i(\boldsymbol{\alpha})$ is the "working" correlation matrix among repeated measures for a subject, which depends on unknown parameters $\boldsymbol{\alpha}$.

Liang and Zeger (1986) show that $\hat{\boldsymbol{\beta}}$ are consistent estimators of $\boldsymbol{\beta}$ with asymptotically normal distribution when $n \to \infty$.

3 Application

The data set is composed by 666,620 observations with all the grades of all the courses taken by the undergraduate students who entered in Unicamp from 2000 to 2005. For each student it was recorded the grade obtained in each course taken during the whole period in the University, the type of course (R, E or Ex), the final result (Passed or Failed) and his/her entrance exam score (e.g., SAT score). There is also information about demographic and socioeconomic status for each student.

There are 58.8% of men and 41.2% of women, the great majority (93.5%) between 17 and 23 years of age divided in four different areas: Medical and Biological Sciences (18.23%), Engineer and Exact Sciences (59.6%), Human Sciences (14.94%) and Arts (7.23%). Also, 70.45% of the students come from Private High Schools (PrHS) against 28.13% coming from Public High Schools (PuHS). In Brazil, the students decide their major before taking the entrance exam to get into the University. Figure 2 shows the Distribution of the standardized by area Entrance Exam Score (EES) and type of High School while Figure 3 shows the distribution of the standardized GPA scores by area, type of courses (Required, Elective or Extra-curricular) and type of High School. The GPA scores and EES were standardized within each entrance year and each major/course (i.e., according to the mean and standard deviation of the students who entered in the same major/course and in the same year). Note that in Engeneering and Exact Sciences there is a greater dispersion in the GPA scores for Required courses than in other areas. Students coming from Private High Schools have greater EES in Arts, Exact and Human Sciences.

Looking at Figure 4, it does not seem to be significant differences in performance of students coming from PuHS compared to those coming from PrHS. In fact, the evolution seems to be greater for PuHS students.

Looking at Tables 1 and 2, one can see that, in all areas, given that all the other variables are fixed, the average GPA scores of female are better than male students; the younger the students, the better is their performance; the greater is their EES the better is their performance. Also, in general (except for Arts), the lower is the family income, the better is their performance. For Arts, the family income was not significant at 5% level . Among type of courses, the bigger correlation is between R and Ex courses. Except for Human sciences, the worst GPA score is for Required courses.

A GLMM model for the incidence of failed courses was also modeled and the distribution that seemed to fit better was the negative binomial (model according to equation (3)). Figure 5 shows the distribution of the incidence of failed courses for each Area, according to type of course and type of

TABLE 1. Main effect coefficients for the population-averaged model for GPA scores of equation (2) with unstructured correlation among observations for a subject

	Engeneering		Exact		Biological	
Coefficient	β	p-value	β	p-value	β	p-value
Intercept	-0.56	< 0.0001	-0.43	< 0.0001	-0.25	0.0008
Type of Course						
Elective	0.52	< 0.0001	0.23	0.0018	0.06	0.4602
Extracurricular	0.23	0.0059	0.24	0.0003	0.16	0.0905
Sex						
Female	0.20	< 0.0001	0.30	< 0.0015	0.27	< 0.0001
High School						
Private	-0.19	< 0.0001	-0.17	0.0001	-0.34	< 0.0001
Family Income						
$< 3\ m.s^*.$	0.06	0.3992	0.31	0.0223	-0.03	0.7507
$3 - 10\ m.s.^*$	0.09	0.0004	0.11	0.1727	-0.003	0.927
Age						
< 18 years	0.82	< 0.0001	0.70	< 0.0001	0.63	< 0.0001
$18 - 20$ years	0.56	< 0.0001	0.36	< 0.0001	0.21	0.0075
EES	0.15	< 0.0001	0.23	< 0.0001	0.15	< 0.0001
Period						
Daytime	-	-	-0.05	0.0692	-	-

$\hat{\alpha}_{E,Ex} = 0.19$, $\hat{\alpha}_{E,R} = 0.08$, $\hat{\alpha}_{Ex,R} = 0.37$ (Engeneering)

$\hat{\alpha}_{E,Ex} = 0.17$, $\hat{\alpha}_{E,R} = 0.15$, $\hat{\alpha}_{Ex,R} = 0.31$ (Exact)

$\hat{\alpha}_{E,Ex} = 0.16$, $\hat{\alpha}_{E,R} = 0.19$, $\hat{\alpha}_{Ex,R} = 0.34$ (Biological)

$^*m.s.$ is minimum salary per month.

Reference cell is Required course, Male, Public HS, $> 10\ m.s.$, > 10 years of age, Night-time.

TABLE 2. Main effect coefficients for the population-averaged model for GPA scores of equation (2) with unstructured correlation among observations for a subject

	Human		Arts	
Coefficient	β	p-value	β	p-value
Intercept	-0.32	< 0.0001	-0.27	0.0096
Type of Course				
Elective	-0.04	0.2335	0.23	0.0541
Extracurricular	-0.05	0.0279	0.17	0.0284
Sex				
Female	0.31	0.0001	0.13	0.0246
High School				
Private	-0.09	0.0167	-0.03	0.8173
Family Income				
< 3 $m.s^*$.	0.48	0.0004	-	-
$3 - 10$ $m.s.^*$	0.18	0.0385	-	-
Age				
< 18 years	0.38	< 0.0001	0.72	< 0.0001
$18 - 20$ years	0.17	0.0411	0.23	0.0905
EES	0.11	< 0.0001	0.17	< 0.0001

$\hat{\alpha}_{E,Ex} = 0.22$, $\hat{\alpha}_{E,R} = 0.08$, $\hat{\alpha}_{Ex,R} = 0.36$ (Human)

$\hat{\alpha}_{E,Ex} = 0.18$, $\hat{\alpha}_{E,R} = 0.06$, $\hat{\alpha}_{Ex,R} = 0.54$ (Arts)

$^*m.s.$ is minimum salary per month.

Reference cell is Required course, Male, Public HS, > 10 $m.s.$, > 10 years of age.

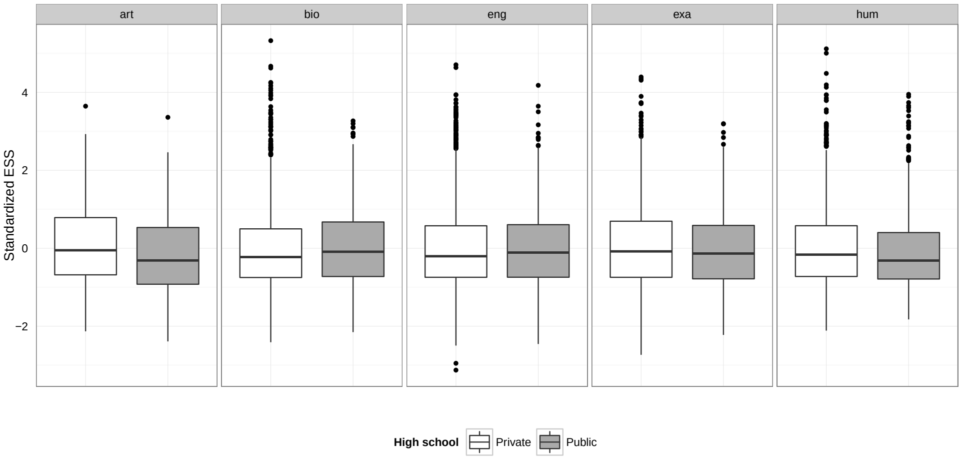

FIGURE 2. Distribution of standardized EES by each area according to type of High School. EES are standardized within each entrance year and each major.

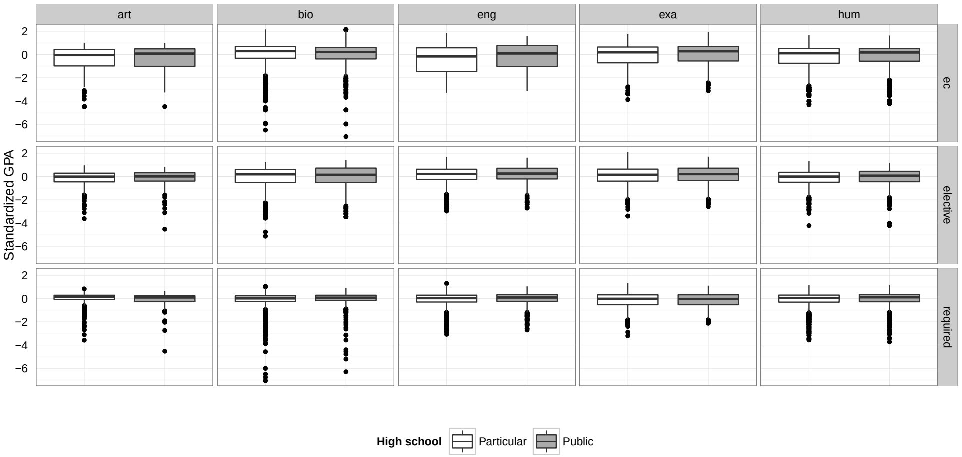

FIGURE 3. Distribution of standardized GPA scores by each area according to type of High School and courses. GPA scores are standardized within each entrance year and each major.

High School. Looking at Tables 3 and 4, one can see that, the incidence of Required courses failed is greater than elective or extracurricular courses. For all the areas, there is no difference with respect to incidence of failed courses between students coming from PrHS or PuHS nor for family income. The greater the EES, the less courses they fail. The younger the students,

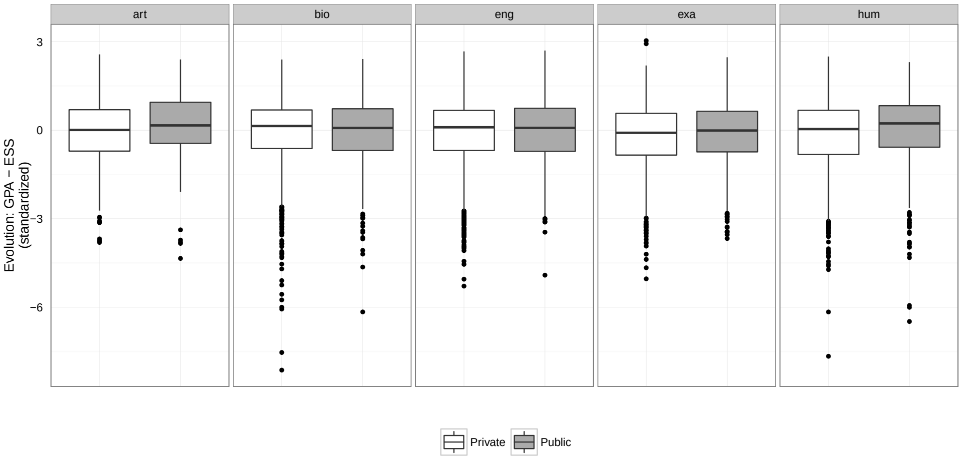

FIGURE 4. Distribution of the standardized GPA minus standardized EES by each area according to type of High School . GPA scores and EES are standardized within each entrance year and each major.

the less courses they fail. Female students fail less courses than Male.

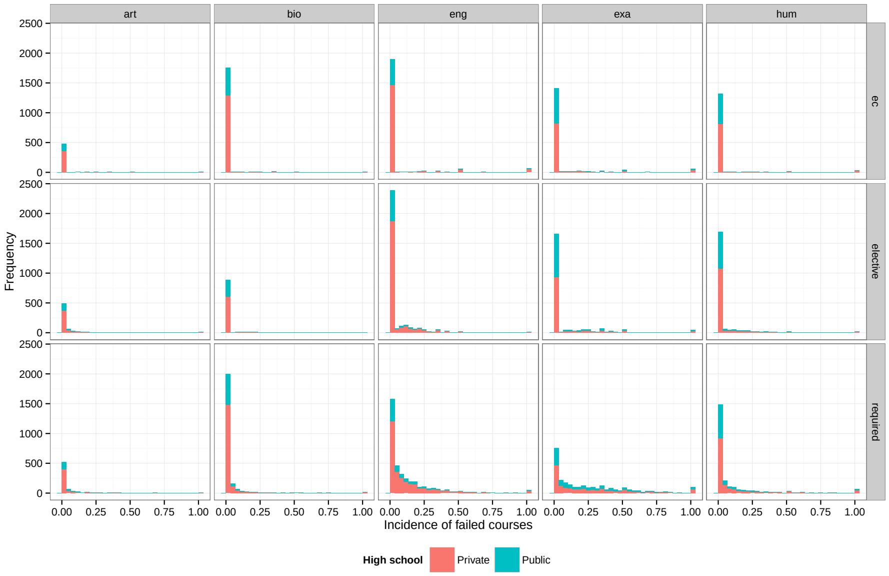

FIGURE 5. Incidence of failed courses according to each area and type of course.

TABLE 3. Main effect coefficients for the population-averaged model for incidence of failed courses of equation (3) with unstructured correlation among observations for a subject

	Engeneering		Exact		Biological	
Coefficient	β	p-value	β	p-value	β	p-value
Intercept	-1.08	< 0.0001	-1.41	< 0.0001	-3.06	< 0.0001
Type of Course						
Elective	-2.44	< 0.0001	-2.13	< 0.0001	-1.30	< 0.0001
Extracurricular	-3.76	0.0059	-2.97	< 0.0001	-2.76	< 0.0001
Sex						
Female	-0.27	0.0125	-0.36	< 0.0001	-0.55	0.0039
High School						
Private	-0.02	0.8552	0.08	0.1563	0.32	0.1787
Family Income						
< 3 $m.s^*$.	0.19	0.5126	0.19	0.1311	0.34	0.5047
$3 - 10$ $m.s.^*$	-0.01	0.8904	0.07	0.2383	0.25	0.2104
Age						
< 18 years	-1.19	< 0.0001	-0.63	< 0.0001	-1.20	0.0002
$18 - 20$ years	-0.58	0.0033	-0.19	0.0045	-1.02	0.0001
EES	-0.33	< 0.0001	-0.30	< 0.0001	-0.36	0.0002
Dispersion	2.53	-	1.52	-	8.01	-

$\hat{\alpha}_{E,Ex} = 0.19$, $\hat{\alpha}_{E,R} = 0.07$, $\hat{\alpha}_{Ex,R} = 0.15$ (Engeneering)
$\hat{\alpha}_{E,Ex} = 0.48$, $\hat{\alpha}_{E,R} = 0.29$, $\hat{\alpha}_{Ex,R} = 0.36$ (Exact)
$\hat{\alpha}_{E,Ex} = 0.29$, $\hat{\alpha}_{E,R} = 0.26$, $\hat{\alpha}_{Ex,R} = 0.45$ (Biological)
$^*m.s.$ is minimum salary per month.
Reference cell is Required course, Male, Public HS, > 10 $m.s.$, > 10 years of age.

TABLE 4. Main effect coefficients for the population-averaged model for incidence of failed courses of equation (3) with unstructured correlation among observations for a subject

	Human		Arts	
Coefficient	β	p-value	β	p-value
Intercept	-1.90	< 0.0001	-2.20	< 0.0001
Type of Course				
Elective	-0.67	< 0.0001	-0.53	0.0003
Extracurricular	-2.37	< 0.0001	-1.92	< 0.0001
Sex				
Female	-0.95	< 0.0001	-0.65	0.0025
High School				
Private	-0.01	0.9049	-0.005	0.9844
Family Income				
$< 3\ m.s^*.$	0.10	0.6191	-0.78	0.1458
$3 - 10\ m.s.^*$	-0.11	0.2546	-0.12	0.5908
Age				
< 18 years	-0.43	0.0025	-0.73	0.0145
$18 - 20$ years	-0.25	0.0337	-0.52	0.0434
EES	-0.28	< 0.0001	-0.42	0.0001
Dispersion	3.68	-	3.20	-

$\hat{\alpha}_{E,Ex} = 0.39$, $\hat{\alpha}_{E,R} = 0.40$, $\hat{\alpha}_{Ex,R} = 0.58$ (Human)

$\hat{\alpha}_{E,Ex} = 0.15$, $\hat{\alpha}_{E,R} = 0.36$, $\hat{\alpha}_{Ex,R} = 0.77$ (Arts)

$^*m.s.$ is minimum salary per month.

Reference cell is Required course, Male, Public HS, $> 10\ m.s.$, > 10 years of age.

Acknowledgments: H.P. Pinheiro's research was partially supported by CNPq-Brazil (305988/2012-3) and FAPESP (2011/15047-7). We thank the following people for their direct support regarding databases used for this study: Antonio Faggiani, Coordinator of the Academic Records Division (DAC) at Unicamp, Renato Hirata, database specialist at COMVEST, Flavio Ferraresi, database specialist at COMVEST and Silvio de Souza, database specialist at DAC.

References

Breslow, N. E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association.* **88** (421), 9 – 25.

Breslow, N. E and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika.* **82** (1), 81 – 91.

Liang, K-Y and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika.* **73**, 13 – 22

Maia, R.P., Pinheiro, H.P. and Pinheiro, A. (2013). Academic performance of students from entrance to graduation via quasi U-statistics: a study at a Brazilian research university. Universidade Estadual de Campinas, Brazil. Technical Report. RP 09/2013.

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.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika.* **78**, 719 – 727.

Rodrigues-Motta, M., Pinheiro, H.P., Martins, E.G., Araújo, M.S. and dos Reis, S. (2013) Multivariate models for correlated count data. *Journal of Applied Statistics*, **1**, 1 – 11.

Williams, D. A. (1982) Extra-binomial variation in logistic linear models. *Applied statistics*, **31**(2), 144 – 148.

Zeger, S.L., Liang, K-Y and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics.* **44**(4), 1049 – 1060.