Statistical Modeling of Students' Academic Performances: A Longitudinal Study

Lionel Establet Kemda[1] & Michael Murray[1]

[1] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Correspondence: Lionel Establet Kemda, School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa.

Received: March 28, 2021 Accepted: June 17, 2021 Online Published: June 23, 2021

doi:10.5430/ijhe.v10n6p153 URL: https://doi.org/10.5430/ijhe.v10n6p153

Abstract

Within students' attrition studies, it is necessary to assess the longitudinal evolution of students within a given course of study, from enrolment to exit from the university through degree completion and academic dropout. Here, the student's academic progress is monitored through the number of courses failed each semester enrolled. The students' failure rate and academic behavior typically provide significant insight into students' exit outcomes from University programs. These programs usually have a maximum time frame required to complete the course. A likelihood-based approach is discussed that conditions on the exit outcome and random effects in adjusting within-subject correlation of longitudinal measurements. Ignoring the number of courses enrolled by a student may produce inadequate results on the actual failure rates. Conditioned on the exit outcomes of the student, we find out that factors such as financial aid, matriculation points, students' race and course type registered, and gender are distinguishing factors that affect students' academic performances, for completers and dropouts. Also, being in university-type accommodation (that often have added services such as transportation and internet connections) does not seem to significantly affect the failure rate within both groups of students. In addition, an increase in matriculation points significantly reduces the failure rate independent of the Quintile school of the student. Several count models such as mixed Poisson, mixed Zero Inflation Poisson, mixed Negative Binomial, and mixed Hurdle Poisson models are fitted and compared. In particular, the mixed Poisson model provides a better fit based on the Bayesian Information Criterion and residues analysis.

Keywords: Longitudinal, dropout, completer, random effects, counts, failure rate, BIC

1. Introduction

In the Budget speech of 2018, the South African government made some commitments towards fee-free education, which has always been an issue in the country, characterized by nationwide strikes in tertiary institutions. These commitments came in the form of over 1 trillion Rand, to be invested over a period of three years. However, the amount aimed to give access to tertiary education and training to all South Africans based on merit and not position. With these investments, the Department of Education and Training desires a national norm of about 80% success rate. According to the Department of Education, back in 2005, 30% of first-year students enrolled in tertiary institutions for the first time in 2000 dropped out after the first year, while 20% dropped out during the subsequent two years. Only a 22% pass rate was observed within the prescribed three-year duration of the undergraduate degrees (Letseka et al., 2010). While most of the dropouts in South Africa may be attributed to families with low economic status, particularly among blacks, the government however laments about the costs within rewards of the amounts invested in tertiary education in South Africa. Scott et al. (2007) also observed similar results, with only 30% of first-year entrants entering higher-level institutions finishing their studies within the first five years. He further questions the readiness of these high school students who eventually get admitted into institutions of higher education. Hence, universities must admit students who satisfy the minimum entry requirements, irrespective of the quality and standard of the high school certificate obtained. Consequently, secondary schools must better equip students to undertake post-secondary education. On the other hand, while these under-prepared students get into universities, the universities must adequately prepare them to succeed (Braxton and Hirschy, 2005). This can be achieved through pre-university evaluations and high school upgrade programs. However, tertiary institutions usually have limited resources and finances, making it difficult to accommodate all those who want to enroll in a given institution. Consequently, these institutions want to make sure that they admit students who can achieve academic success.

One of the earliest models in students' attrition was formulated by Tinto (1975). According to Tinto, students enrolled in university with attributes and characteristics coupled with goals and commitments (towards academic success). As a result of educational and social experiences, they have the potential to integrate the institution academically as well as socially. However, these experiences influence their commitments and goals, thus enforcing their decisions to leave or continue with their academic program. According to Tinto, Students' success is a longitudinal decision-making process. These models have been improved over the years (Cabrera et al. (1993), Ozga and Sukhnandan (1998), Bean and Metzner (1985)). Bean and Metzner (1985) integrate external factors to account for students' retention. These external factors may include finances, commitment to academic goals, stress, family responsibility, and hours of employment, which may hinder academic commitment and integration. While this model was initially formulated for non-traditional students (principally those studying part-time, those older than 24, and those with family responsibilities), it has been applied within a traditional student context. Other theories have been proposed in the literature; Cabrera et al. (1993) "*integrated model of student retention*" combines Tinto (1975) and Bean and Metzner (1985) models into one, Milem and Berger (1997) "*behavior-perception-behavior cycle model*", and Terenzini and Pascarella (1980) "*attrition theory*".

Several studies in the literature focus on a diverse category of factors affecting students' decision to complete an academic program or drop out. These can be classified as demographic, students' academic ability, school characteristics, among others. While most of these studies do not reflect the South African context, some South African-based studies have investigated the factors influencing students' throughput. A qualitative analysis by Moodley and Singh (2015) identifies affordability, lack of academic advice and support, lack of self-discipline and commitment, and being a first-generation student (students whose parents have no post-secondary degree) as important factors contributing towards students' throughput. In particular, it is suggested that higher education institutions should improve open days to include pupils from the tenth grade to make informed decisions before even deciding on post-secondary education. Secondly, higher institutions should improve programs such as career guidance to facilitate the transition from high school into university, particularly for first-generation students.

Demographic factors such as students, age, gender, race are significant factors associated with students' attrition (Charlton et al. (2006), Georg (2009), Pretorius and Prinsloo (2009), Belloc et al. (2011), Johnson (1996)). While there is no clear distinction between gender and academic achievement, some studies have suggested that females generally achieve better results than males. However, when it comes to Science, Engineering, and Technology, males generally outperform females (Gibb et al., 2008). Hyde et al. (2008) refutes these findings and suggests that the differences between males and females are minimal. Johnson (1996) focused on whether a student voluntarily withdrew from their studies or was academically excluded by the institution. It was found that 28% of males were academically excluded from further study in the Faculty of Education. In the Faculty of Arts, proportionally more males were also academically excluded from further studies, whereas more females voluntarily chose to withdraw from their undergraduate studies. In a paper, Belloc et al. (2011) made a distinction between an even more extensive set of dropout scenarios: voluntary dropout, a change of faculty within the same institution, a transfer to another academic institution, and retention within the same faculty. The authors thus recommend caution when interpreting dropout models without adequately considering all the exit outcomes. In particular, being male has a significant positive effect on university dropout but negatively affects the probability of transferring to another faculty. However, the impact is insignificant for the hazard of changing the university.

Within a South African context, Letseka (2010) accounts for the effects of race and poverty in student attrition. In particular, the study consisted of surveying 34548 (14195 graduated and 20353 ended prematurely) students from seven historically institutions (four black institutions and five white institutions) in South Africa. Letseka (2010) suggests that while there is a vast disparity between students with low Socio-Economic Status (SES) who graduate or drop out. Among the African graduates, 71% were from low SES while only 10% White graduates were from low SES. Similarly, only 43% of Indian graduates were from low SES.

Conversely, among the non-completers, 73% were Africans from low SES, while only 12% of the non-completers were White, and only 48% were Indian. The proportions among Africans were similar among Coloured students. Furthermore, the authors identified the persistence of the discrimination and disadvantage cemented during Apartheid as the reasons behind these differences. Scott and Letseka (2010) carried out a study involving students who prematurely dropped out of the University of Witwatersrand in 2002. Several factors were identified as possible reasons for dropout. Amongst the factors, failing courses seemed to be the predominant factor characterizing premature withdrawals from academic programs. Lack of funding and administrative issues seemed dominant among Africans and Coloured students, respectively. Indians and White students seemed unaffected by finances. The factors considered here ranked among a list of possible 31 factors influencing students' academic success. Lemmens (2010) through a questionnaire study from students at the University of Pretoria, observed that the distance from family and the inability of some students to adapt to the university environment was a secondary characteristic contributing towards withdrawals from university. In addition, external pressure and insufficient interaction between lecturers and students were also identified as contributing factors, some students associating it with some form of discrimination or racism. The author also identifies matriculations-scores (low, medium, high), race (White, Coloured, Indian, African), and the total number of credits registered for as statistically significant factors associated with students' withdrawal. In particular, African students tend to choose a lighter credit load, thus lowering their chances of withdrawal. Hence, African students were identified as the racial group most likely to proceed to the second year. This was followed by Afrikaans speaking students and finally English- speaking students. However, White students tend to withdraw voluntarily, especially during the first year of studies. Regarding the total number of credits registered, Lemmens (2010) using a multiway frequency table, suggests that students who register for fewer credits than prescribed by the faculty tend to have a higher risk of withdrawal (possibly voluntary) than those who register the exact amount specified. Indeed, the latter are three times more likely to persist than the former. However, these results were contradicted by the fitted logistic regression.

Students' academic abilities such as admission points, intellectual capacity, and high school grades have had a conflicting influence on future academic performances. While some authors (Harackiewicz et al. (2002), Acato (2007), Geiser and Santelices (2007)) suggest a positive relationship between university admission points and future success, others (Huws et al., 2006) refute such a relationship. Acato (2007) indicates that the number of university admission points plays a significant part in students' future academic performances. Students with a high number of admission points have a very high level of academic performance, enabling them to pursue their studies, hence graduating. On the other hand, students with low admission points tend to have low academic performances and will likely drop out. This view is also shared by Geiser and Santelices (2007), who studied the academic performance of students for a four-year program. Among the variables used, university admission points and Standardized Admissions Tests (which are widely used as a tool to identify students more prone to success in college) were all used as predictors for academic success. However, the authors suggest that these university admission points (high-school grades) are better indicators for future academic success than the standardized tests. Furthermore, they indicate that these standardized tests should be used as additional information that can be used in the selection process. However, Harackiewicz et al. (2002) emphasize that these grades alone are not sufficient, but the development of interest in the discipline is also important. On the other hand, a study by Huws et al. (2006) refutes these relationships between admission points and future success. In a retrospective study assessing students' performance in university, they found that students who obtained Advanced (A) level certificates and obtained higher grades did not necessarily perform better than those who did not.

Johnson (1996) looks at faculty differences among a group of 185 students who had previously withdrawn from an undergraduate study. In particular, these students were classified into one of three Faculties; Education, Science, and Art. Indeed, it was observed that undergraduate students in the Faculty of Science were about ten times more likely to be required to withdraw than students in the Education Faculty. The university required Twenty-nine students to withdraw during or by the end of the first year. On the other hand, Education students were more likely to withdraw only after the second year. Yang (2017) revisits the effects of school characteristics, family background, academic ability, and vocational education to measure high school students' probabilities of dropout or degree completion. This study shows that students with low family income, with more siblings, living in an urban area, and single-parent families, were more likely to drop out. Similarly, students who attended private schools, students with higher abilities in Mathematics, were less likely to drop out. Surprisingly, the model indicated that students were more likely to drop out in their senior years (semester 9). This is highly contradictory, as one would expect students to be at a higher risk of dropout during their early years of university enrolment.

In South Africa, Lemmens (2010) identified residence issues and school transportation from residence to campus as essential contributors towards university withdrawal. Some students also identified their safety, racism, and having issues with lecturers. In South Africa, Scott and Letseka (2010) identified institutional culture integration (characterized by the values, beliefs, and assumptions shared by the university community, expressed by the community's actions) as a factor significant in students' retention. Bengesai and Paideya (2018) examine the effects of institutional factors (such as participation in supplementary instructions (SI)), demographic factors (gender, race), academic characteristics (first-year academic performance and admission point scores or matric points), and institutional factors such as residence and financial aid status, on degree completion of students enrolled for an Engineering degree at the University of KwaZulu-Natal. It is deduced that the required profile for a student to complete an Engineering degree at the University of KwaZulu-Natal is such that; irrespective of gender, the student must be non-African (Indian, Coloured or White), must have a university admission point of above 40, must have financial aid, must attend supplementary instruction regularly (about five times), and must pass at least 75% of all credits in the first year. While it is possible to identify such a student, there are some drawbacks to such a profile. Firstly, the study only tracts students' performances during the first year and does not account for subsequent years of registration. A student may not pass at least 75% of first-year credits and perform better in subsequent years of studies. Thus, the entire longitudinal path describing the students' profile should be accounted for. Secondly, the grouping of race into only two categories (black South Africa=2, other=1) assumes that White, Indian and Coloured have the same academic profile towards graduation, which may be invalid. Finally, the study does not incorporate the competing event of dropout, clearly assuming that students can only graduate after first registration. However, the school of Engineering very often has very high dropout rates. Gershenfeld et al. (2015) also identify the first semester Grade Point Average (GPA) as a distinguishing factor affecting underrepresented students' academic success. Underrepresented students with low first semester GPA do not usually graduate within the prescribed time frame.

In a longitudinal analysis involving count outcomes, a number of repeated measurements is collected over time. Within an educational context, we will count the number of courses failed/passed by a student during each semester from enrollment to university. Here, additional information such as personal, pre-university education, and university characteristics may also be recorded. While some of these factors may vary with time (each semester), others may remain fixed (unchanged) throughout. Amongst those variables that may change over time, we will be considering whether a student received financial aid, the type of student accommodation they stayed in, and the number of courses enrolled in for a given semester. While these longitudinal measurements are nested within students, it is also relevant to note that students are also being nested within a degree (courses). Hence, leading to a hierarchical nature of observation that must be accounted for. In this regard, we may define a model with the following levels;

level 1: longitudinal measurements of the number of courses failed,

level 2: individual student, $i \in 1, \dots, N,$

level 3: degree type registered, $k \in 1, \dots, K.$

This hierarchical clustering of data may suggest that the number of courses failed by a student recorded each semester will be correlated, violating the assumptions of independence in regression analysis. Ignoring this correlation leads to underestimating the standard errors of the regression coefficients, resulting in an incorrect significance test. Secondly, it may also lead to model misspecification because it ignores non-linear relationships. For instance, ignoring students clustering within degrees or courses may result in omission of course-level variation, which influences performance at the student level. Here, we will distinguish between two groups of students; completers and dropouts. Completers refer to students who complete a university academic program, while dropouts are those who voluntarily drop out or are excluded without completing an academic program.

While many studies predict students' performance based on the terminal outcomes of degree completion and academic dropout, few have investigated how students' performance evolves during their studies. Secondly, most of these studies have based their conclusions on descriptive statistics, multiway frequency analysis (Lemmens, 2010), and logistic regression (which quantifies the probability of degree completion or not). Others (Meggiolaro et al., 2017) have focused on competing risk models and machine learning algorithms such as decision trees, Na ïve Bayes (Kabakchieva, 2013). However, as Tinto (1975) noted, a student's academic abilities evolve over time, making it a longitudinal process that should be modeled as such. With a longitudinal approach, one can determine whether students who perform poorly in their first year of study can improve when entering their second year of study. Should one limit the number of courses that a student can enroll for in a given semester, or adjust this limit as a student progresses from one year of study?

In this paper, we seek to investigate if there are any significant differences in the average number of courses failed between a student who dropped out and one who completed (completers) a university program, taking into account the number of courses enrolled for and the number of semesters since initial registration. Secondly, we also investigate the effects of students' characteristics such as the availability of financial aid, being in a university residence, matriculation points (Matric points), and the type of school attended prior to university (Quintile), on the number of courses failed for separate groups.

This study is significant for policymakers within the university. It helps them make informed decisions about these students and put in place measures to help prevent academic dropouts. Universities only receive a government-based subsidy if a student passes a given course and complete their degree in a certain length of time - currently minimum time plus one year.

The rest of this paper is summarized as follows; Section 2 describes the construction of the random-effects models and their association with the count outcomes. Section 3 gives a brief description of the data used and the results of the fitted models. Section 4 summarizes the findings and makes recommendations that can be put in place to help students improve their academic performances.

2. Methods

2.1 Poisson Model

Suppose $Y_i = (Y_{i1}, \dots, Y_{ni})$ is a vector of repeated measurements for individual $i \in 1, \dots, N$ such that we can write $Y = (Y_1, \dots, Y_N)$.For each individual i in semester time t, we have a $(p * 1)$ vector of covariates, $X_{it} = \left(X_{it1}, \dots, X_{itp}\right)$, with $X_i = (X_{i1}, \dots, X_{ini})$ the resulting $(n_i * p)$matrix of covariates, and $X = X_1, \dots, X_N$.

Considering the two subgroups of the population that we will be considering in this paper, namely those who have dropped out (dropouts) and those who have completed their degree (completers), let S_i and $(1S_i)$ represent the dropped out and completed groups, respectively. Let Y_{it} denote the number of courses failed by student i in semester t. We can define the random effects that condition Y_{it}. To this end, let $b_i = (b_{1i}, b_{2i})$ be a vector of random effects associated with the dropped out (S_i) and completer groups $(1 - S_i)$ respectively. We define

$$b_{1i|S_i} \sim N\left(\sum_{k=1}^{P_1} \theta_p G_{1k}(\,), \sigma_1^2\right)$$

and

$$b_{2i|1-S_i} \sim N\left(\sum_{k=1}^{P_2} \theta_p G_{2k}(\,), \sigma_2^2\right)$$

where $G_{1k}(\,)$ and $G_{2k}(\,)$ are both functions of the number of semesters registered before exit or simply the semester registered in the dropout and completion groups. Several functions such as dummy variables, the logarithm of time, or linear trend in time can be used based on the behavior of Y_{it}. Similarly, $P_1$and P_2 represent the number of linear terms in each of these groups in quantifying this relationship. For example, if $G_{1k}(\,)$ is a quadratic function of time, then P_1 is 3.

Thus, condition on the random effects b_i, the model describing the distribution of Y_{it} is given by

$$log(\mu_{it}) = \alpha_0 + \alpha_1 S_i + \sum_{k=1}^{P_1} \theta_p G_{1k}(t)(1 - S_i) + \sum_{k=1}^{P_2} \theta_p G_{2k}(t) S_i + (1 - S_i) b_{1i} + S_i b_{2i}$$

$$+ \sum_{j=1}^{l} \beta_j X_{it} S_i + \sum_{m=1}^{n} \alpha_m X_{it}(1 - S_i) \qquad (1)$$

Here, we assume G_{1k} and G_{2k} are functions of semester t such that the log average number of courses failed within the two groups can be modeled with respect to semester t. That is, estimating the change in log average number of courses failed with semester time. The choice of semester time is also based on the distribution of the actual average number of courses failed (and the failure rate) with semester time, as indicated in Figure 2. Also, X_{it} represents the effect of other covariates on the average number of courses failed between completers and dropouts. However, within student attrition, the variation in the number of courses failed may be as a result of the number of courses registered in the particular semester by the student. In this case, we introduce an exposure variable (offset) to account for the number of courses registered for. This exposure variable is constrained to have a coefficient of 1. Thus, Model (1) estimates the log average failure rate. Using the Poisson notation, we may write $Y_{it}|b_i, S_i \sim$ *Poisson* (μ_{it}). Introducing the exposure variable and the semester time variable, *t*, into Model 1, we obtain

$$log(\mu_{it}) = log(d_{it}) + \alpha_0 + \alpha_1 S_i + \alpha_2 t * (1 - S_i) + \alpha_3 t * S_i + (1 - S_i) b_{1i} + S_i b_{2i} + \sum_{j=1}^{l} \beta_j X_{it} S_i + \sum_{m=1}^{n} \alpha_m X_{it} (1 - S_i) \qquad (2)$$

where $log(d_{it})$ is the exposure representing the number of courses registered for by student i in semester time t. Model (2) thus models the log average failure rate, where the failure rate for student i in semester time t is the ratio of the number of courses failed and the number of courses registered for by student i in semester time t. For instance, if a student registers for 5 courses in semester 2 and fails 3 courses at the end of semester 2, then the failure rate is calculated as $(3/5 = 0.6)$ for the student in semester time 2. Additionally, the random effects, $b_i = (b_{1i}, b_{2i})$, capture within student correlations in longitudinal number of courses failed by student i in semester time t for the dropout (S_i) and degree completion $(1 - S_i)$ groups respectively.

2.2 Negative Binomial Model

Unlike the Poisson model that does not account for over-dispersion in the longitudinal counts, the Negative Binomial counterpart accounts for overdispersion (mean smaller than the variance). Thus, let α denote the overdispersion parameter, then we define the model

$$log(\mu_{it}) = \alpha_0 + \alpha_1 S_i + \sum_{k=1}^{P_1} \theta_p G_{1k}(t)(1 - S_i) + \sum_{k=1}^{P_2} \theta_p G_{2k}(t) S_i + (1 - S_i) b_{1i} + S_i b_{2i} + \sum_{j=1}^{l} \beta_j X_{it} S_i + \sum_{m=1}^{n} \alpha_m X_{it}(1 - S_i) + \epsilon_{it} \qquad (3)$$

where $log(E[Y_{it}|b_i, S_i) = log(\mu_{it})$ and $b_i = (b_{1i}, b_{2i})$ is defined as before, and $\epsilon_{it} \sim Gamma(\alpha, \alpha)$.

Condition on the random effects vector b_i and exit pattern S_i, the longitudinal response Y_{it} follows the Negative binomial distribution denoted $Y_{it}|b_i, S_i \sim NB(\mu_{it}, \alpha)$, with density defined by

$$P(Y_{it} = y_{it}|b_i, S_i) = \frac{\Gamma(y_{it}+\alpha)}{\Gamma(\alpha) y_{it}!}\left(\frac{\alpha}{\mu_{it}+\alpha}\right)^{\alpha}\left(\frac{\mu_{it}}{\mu_{it}+\alpha}\right)^{y_{it}} \qquad (4)$$

In particular, $E(Y_{it}|b_i,S_i) = \mu_{it}$ and $Var\ (Y_{it}|b_i,S_i) = \mu_{it} + \mu^2_{it}/\alpha$, where α quantifies the amount of over-dispersion (Booth et al., 2003).

2.3 Zero-Inflation and Hurdle Models

Count outcomes are sometimes characterized by excessive/extra zeros, which if left unaccounted for, may result in biased and unreliable results Chin and Quddus (2003). The use of two-part models or zero-inflated models have been extensively documented in the literature (Chin and Quddus (2003), Hu et al. (2011)), namely the Hurdle model (Poisson or Negative Binomial) or the Zero-Inflated Poisson (or Negative binomial) models respectively. Here, the Zero-inflated Poisson (ZIP) (Lambert, 1992) and the Hurdle Poisson (Mullahy, 1986) will be considered. While these models play an important role in accounting for excess zeros in count outcomes, some differences exist between the two. Indeed, the zero-Inflated models assume two sources for the extra zeros count; the sampling and structural zeros. The sampling zeros occur by chance due to the Poisson and the Negative Binomial distributions, and the structural zeros arise as a result of some structural behavior of the data (Hu et al., 2011). In a given semester *t*, a large number of students may not fail a single course because there are generally good and high achieving students (thus the need to account for extra zeros in the data when otherwise fitting a Poisson model).

On the other hand, the Hurdle models only assume that all the zero counts result from structural sources only. Thus, the name "two-part" models, in which the non-zero counts are modeled with a truncated Poisson or Negative Binomial model and the extra zero counts using a suitable binary distribution such as the logistic regression.

Consider the distribution of μ_{it}, given in Model (1), the distribution of Y_{it} is Zero-Inflated Poisson (ZIP) when et al., 2015),

$$Y_{it} = \begin{cases} 0, & with\, probability\;\; \phi_{it} \\ Poisson(\mu_{it}), & with\, probability\; (1-\phi_{it}), \end{cases} \tag{5}$$

where φ_{it} is the probability of the zeros arising from the degenerate distribution of zero (point mass). Hence, the distribution function of Y_{it} is given by

$$\begin{aligned} P(Y_{it}=0|W_{it}) &= \phi_{it} + (1-\phi_{it})e^{-\mu_{it}} \\ P(Y_{it}=y_{it}|X_{it}) &= (1-\phi_{it})\frac{\mu_{it}^{y_{it}}e^{-\mu_{it}}}{y_{it}!}, y_{it}=1,2,\ldots \end{aligned} \tag{6}$$

where W_{it} and X_{it} are covariates associated with zero and non zero counts respectively, which may be the same in both sub-models. These covariates may also include random effects such as those on the right-hand side of Model (1). Similarly, the Hurdle Poisson model for Y_{it} is given by

$$Y_{it} = \begin{cases} 0, & with\, probability\;\; \phi_{it} \\ truncated\; Poisson\; (\mu_{it}), & with\, probability\; (1-\phi_{it}), \end{cases} \tag{7}$$

with density function defined by

$$\begin{aligned} P(Y_{it}=0|W_{it}) &= \phi_{it} \\ P(Y_{it}=y_{it}|X_{it}) &= \frac{(1-\phi_{it})\mu_{it}e^{-\mu_{it}}}{y_{it}!}, \quad y_{it}=1,2,\ldots \end{aligned} \tag{8}$$

W_{it} and X_{it} are defined as above. It is important to note that the zero part of Models (6) and (8) are modeled with a logistic regression with a logit link such as

$$logit(\varphi_{it}) = \alpha W_{it} \tag{9}$$

where W_{it} can take the form of the right-hand side of Model (1) with random effects $e_{1i}{\sim}N(0,\sigma_3^2)$ and $e_{2i}{\sim}N(0,\sigma_4^2)$, associated with the degree cpmpletion and dropout groups respectively.

2.4 Model Diagnostics

Four models were fitted, and their Bayesian Information Criterion (BIC) values were compared. The BIC is a generalization of Akaike's information criterion (AIC). A small BIC value corresponds to a good predictive performance of the model (Cavanaugh, 2012), and it measures the fit and the complexity of each model. It is defined by

$$BIC = -2ln(L) + kln(n), \tag{10}$$

where L is the likelihood of the fitted model, k is the number of parameters in the model, and n is the sample size. Unlike other model selection criterion, the BIC tends to favor smaller models, such as when $n \geq 8$, $kln(n) > 2k$. This means that BIC tends to choose more parsimonious models (fits the data more accurately with as few parameters as possible) than those (Cavanaugh, 2012) selected by the AIC.

3. Application

3.1 Data Description

The data used in this paper originates from the Institutional intelligence of the University of KwaZulu-Natal, involving students' records from the College of Agriculture, Engineering and Science (AES). The data involves students' registration information, students' personal information, particularly pre-university characteristics, and faculty information, from first registration into the university until graduation or drop out occurs (whichever comes first). The data is presented in a person-period (semester)/longitudinal format from 2010 to 2017. Since comparisons between dropouts and completers are the main focus of this paper, only these students are considered. Students who did not graduate or drop out post-2017 are considered censored. Thus, the sample consists of 2626 students who completed their studies and 7284 who dropped out, with a total of 41926 observations. Table 1 describes the covariates used in this analysis.

Table 1. Data description

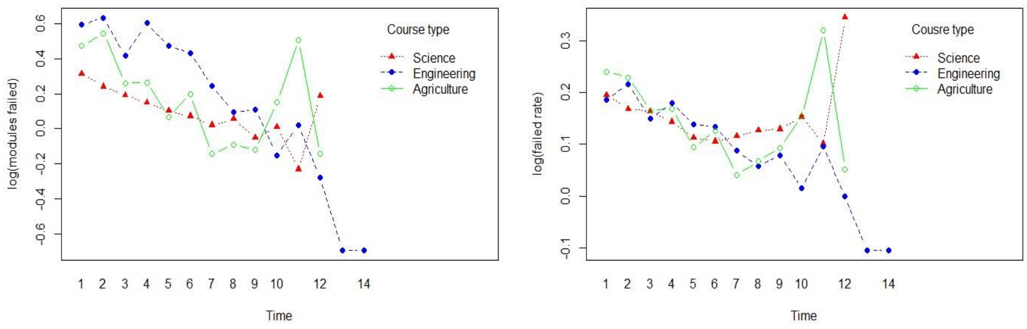
Variable	description	levels/type
S (exit type)	Student exit event	0=completer 1=dropout
Course	Course type registered	0= Science 1= Engineering 2=Agriculture
Financial aid	Whether the student has financial aid	0=No 1= yes
Gender	Students' gender	0=Female 1= Male
Matric	Pre-university matriculation points obtained	0=[15, 25] 1= [26, 37] 2=[38, 48]
Race	Students' race type	0=African 1= Coloured 2=Indian 3=White 4=Other (mostly foreign)
Residence	whether the student is in a university residence	0=No 1=Yes
Quintile	Quality and relative wealth of the surrounding communities where high school was carried out ranging from poorest to the wealthiest communities	1=quintile 1 (poorest) 2=quintile 2 3=quintile 3 4=quintile 4 5=quintile 5

Description: These are the variables that are used in building the models described in Section 2.

Time considered here corresponds to the number of semesters registered until exit from the university. Semester data is deemed appropriate relative to annual records as some students may drop out or complete studies in a particular semester, not the end of the year. The minimum number of semesters registered by a student in the data is 1, and the maximum is 14. Due to the diverse fields of study offered in the college of AES with certain specificities, one may argue that the timing of events may differ across the degree courses registered. As such, one may consider the hierarchical nature of the data, notably, repeated measurements (level 1) of students (level 2) registered for a specific university course (level 3).

3.2 Measuring the Proportion of Clustering in the Data

Preliminary analysis using the Variance Partitioning Coefficient (VPC), which measures the amount of variation in the log average number of courses fails, suggests that only 8.5% of the variation in the log average number of courses failed falls in the course type level. Similarly, 46.7% of the variation in log average number of courses failed within students registered for the same course type. In contrast, 44.8% of the log average number of courses failed by a student lies between the repeated measurements. Since the course type registered for by a student only accounts for about 8% of the variation in the log average number of courses failed, this third level of clustering can be removed and used here as a covariate in the model. Indeed, the VPC enables us to measure the level of clustering of the response at each level of the multilevel model in count data analysis. Within this context, we measure the level of variation of the log average number of courses failed at the students and course-type levels. Thus, enabling us to quantify the relative importance of different sources of clustering in the response. Figure 1 shows the distribution of the average log number of courses failed and the average log failure rate with time among the different course type registered.

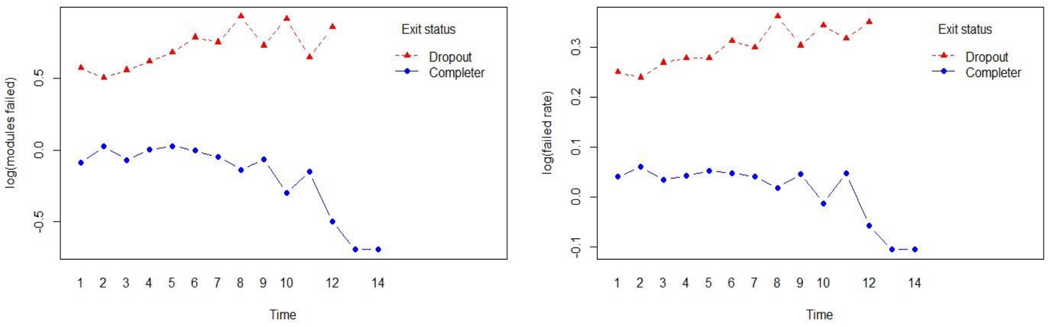


(a) Average log number of courses failed (b) Average log failed rate

Figure 1. Distribution of average log number of courses failed, and the average log failed rate against time with respect to Course type registered.

3.3 Comparing the Average Log Number of Courses Failed between Completers and Dropouts

Figure 2 (a) depicts the distribution of the average log number of courses failed with time between completers and dropouts. Considering the number of courses registered for each semester, Figure 2 (b) shows the distribution of the log average failure rate in the number of courses with time between the two groups.



(a) Average log number of courses failed (b) Average log failed rate

Figure 2. Distribution of average log number of courses failed, and the average log failed rate against time with respect to exit status.

The distinct differences in the average log failure rate in the number of courses between completers and dropouts are evident from Figure 2. In addition, the difference varies linearly in time, especially for the dropouts. This justifies the inclusion of the linear time trend in the model. These distinct differences between the two groups of students may also suggest that when students first enrolled into university, it could be possible to determine students at risk of drop out as soon as their first semester results are available. As such, measures can be put in place immediately to give them a lower chance of academic dropout. It is also relevant to observe that the variation in the log failure rate is fairly constant for both groups of students, with a slight variation after semester 11 for completers. The slight decrease after semester 11 may be attributed to the fact that while some students are still registered after semester 11, only a significantly lower number of courses are left to complete their academic programs. This is very common with students who generally have poor academic performance during the early years of academic registration or students who initially had good performances but along their paths failed some courses which they have to take longer to complete before getting an academic degree.

3.4 Parameter Estimation of Fitted Models

While apparent differences are observed between the two groups, it is relevant to estimate the factors that account for these differences. Several count models were fitted, and only the most significant models, based on convergence and diagnostics (based on residues analysis), are presented here. Furthermore, due to the presence of extra zeros in the number of courses failed, the Zero-inflated and Hurdle Poisson models are presented. The Negative Binomial counterpart of these models is not presented here because the negative binomial regression model results based on the quadratic parametrization indicate no presence of over-dispersion. Indeed, using the quadratic parametrization, as the over-dispersion parameter tends to infinity, the Negative Binomial model simplifies to the simple Poisson model. Here, the over-dispersion parameter is estimated as $2.08E + 07$.

The findings in Table 2 show the parameter estimates and standard errors of the fitted models. The estimated coefficients are those of the significant variables based on p-values, which are not presented in the table. Starting with the simple mixed-effect Poisson model, the log average failure rate in the number of courses is estimated. This is followed by the Negative Binomial model (to account for possible over-dispersion). Based on the estimated over-dispersion parameter, no over-dispersion is present. Thus, the ZIP and Hurdle Poisson models (to account for excess zeros) are fitted and compared. Significant variables were selected based on the p-values and the Likelihood Ratio Test (LRT). A model with all parameters was fitted and compared with the model without insignificant variables (removing insignificant variables one at a time based on the p-value above 0.05). The estimated coefficients with stars (*) represent significant variables, based on p-values < 0.05. The model assessment was carried out by analyzing expected and observed counts through simulated residues based on the work carried out by Hartig (2017). The results of the plots are presented in Figure 3 in the appendix. Here, we can see that the models offer a good fit since a straight line can be drawn through the fitted values.

Table 2. Comparing the estimated conditional models model parts

Conditional	Poisson		NB		ZIP		Hurdle Poisson	
Model	Estimate	std. error	Estimate	std. error	Estimate	std. error	Estimate	std. error
offset(d)								
log(*d*)	1		1		1		1	
Intercept	-0.8298*	0.0426	-0.7951*	0.0603	-0.6596*	0.0590	-0.6844*	0.0390
S(=0)	-0.7872*	0.3831	-0.7660*	0.3931	-0.2968	0.3208	-0.4505	0.3044
S:time 0:time	-0.0487*	0.0032	-0.0486*	0.0032	-0.0248*	0.0050	-0.0252*	0.0050
1:time	0.0456*	0.0031	0.0457*	0.0031	-0.0056	0.0030	-0.0056	0.0030
S:financial aid								
0:1	-0.1766*	0.0137	-0.1768*	0.0137	-0.1094*	0.0129	-0.1093*	0.0129
1:1	-0.1458*	0.0251	-0.1459*	0.0251	-0.0578*	0.0290	-0.0579*	0.0290
S:residence								
0:1			-0.0356	0.0439	-0.0254	0.0454		
1:1			-0.0602	0.0794	-0.1881	0.0971		
S:matric points								
0:1	-0.0520	0.0407	-0.0519	0.0407	0.0946*	0.0379	0.0946*	0.0379
0:2	-0.3192*	0.0456	-0.3191*	0.0455	-0.0037	0.0410	-0.0038	0.0410
1:1	0.0483	0.3808	0.0513	0.3806	-0.2328	0.3010	-0.2373	0.3011
1:2	-0.5126	0.3823	-0.5093	0.3821	-0.4530	0.3024	-0.4592	0.3026
S:race								
0:1	-0.0089	0.0685	-0.0082	0.0685	-0.0630	0.0540	-0.0635	0.0540
0:2	-0.0470*	0.0180	-0.0461*	0.0180	-0.0760*	0.0148	-0.0767*	0.0148
0:3	-0.4001*	0.0461	-0.3992*	0.0461	-0.2288*	0.0417	-0.2295*	0.0417
0:4	-0.4540*	0.1756	-0.4536*	0.1756	-0.2031	0.1757	-0.2038	0.1757
1:1	-0.2010	0.1764	-0.2001	0.1763	0.1966	0.1496	0.1915	0.1497
1:2	-0.0897*	0.0405	-0.0885*	0.0406	-0.0072	0.0352	-0.0128	0.0351
1:3	-0.6321*	0.0714	-0.6311*	0.0714	-0.2299*	0.0699	-0.2350*	0.0699
1:4	-0.4972	0.3522	-0.4962	0.3521	-0.2139	0.3519	-0.2185	0.3522
S:course								
0:1	0.2605*	0.0192	0.2601*	0.0192	0.0457*	0.0147	0.0461*	0.0147
0:2	0.0712*	0.0323	0.0713*	0.0323	0.0659*	0.0265	0.0658*	0.0265
1:1	0.3986*	0.0437	0.3979*	0.0437	0.1492*	0.0355	0.1529*	0.0355
1:2	-0.1595	0.0893	-0.1599	0.0893	-0.0939	0.0826	-0.0911	0.0826
S:gender								
0:1	0.0220	0.0150	0.0220	0.0150				
1:1	-0.0835*	0.0354	-0.0835*	0.0354				
Over dispersion			2.15E+07					
Random effects								
σ_1^2								
σ_2^2	0.4839		0.4835		0.1023		0.1032	
	0.1413		0.1413		6.70E-09		5.35E-07	

Description: Parameter estimates of each of the fitted models.

The variable Quintile was insignificant in all fitted models, irrespective of the exit status. This is contrary to the belief that students who carried out pre-university education within the wealthiest communities (higher quintile schools) generally have lower failure rates than students from more impoverished communities (lower quintiles, 1 and 2). However, this may be justified because students admitted into programs in the college of AES generally are better achieving high school students. Similarly, being in university residence seems to have an insignificant effect on the failure rate based on the mixed Poisson model and mixed Negative Binomial models. However, it is included in the mixed Negative Binomial model since excluding it resulted in convergence issues. This is also unusual as students in university-type residences are often provided with additional facilities such as transportation and internet facilities, which may have a significant effect in reducing students' failure rates. According to the mixed Poisson model, for a degree completer, having financial aid lowers the log average failure rate by about 0.1766, relative to someone without financial aid. Conversely, for a student who drops out, the log average failure rate only decreases by 0.1458 for a student on financial aid. Based on the mixed Poisson model, a student who completed his program admitted with matric points in the interval [38, 48] has a significantly lower average log failure rate than a student with matric points in the interval [15, 25]. Similar results are visible for a random student who eventually dropped out. Indeed, the log average failure rate is reduced by 0.5126 for a student with a matric point in the interval [38,48] than a student with a matric point in the interval [15, 25]. However, a student with a matric point in the interval [26, 37] has a higher failure rate than a student with matric points in the interval [15, 25], given that they dropped out. Regarding students' race, it is observed that for a student who dropped out, the log average failure rate is a lot lower by 0.6321 and 0.4972 for whites and Other students respectively, compared to African students. Similarly, for degree completers, the average log failure rate decreases by 0.4001 and 0.4540 for whites and Others respectively, relative to African students. Regarding the course type registered, it is evident that a degree completer registered for an Engineering course has a higher average log failure rate than a student registered for a Bachelor of Science degree. Based on the results in Table 2, a student who completed a degree and is registered for a Bachelor of Science degree generally has a lower average log failure rate than the degree completers registered for other degrees. This may also suggest that Bachelor of Science students finish their programs a lot quicker than the other students in the college of AES. This is because the lower the failure rate, the faster a student completes their program and subsequently graduates. On the other hand, for a student who drops out, the log average failure rate increases by 0.3986 for an Engineering student relative to a Science student. However, there is a decrease in the log failure rate for an Agricultural student relative to a Science student. Finally, gender has opposite effects on the log average failure rate for completers and dropouts. A male student who has completer his studies had a higher log average failure rate than a female student who completer her studies. On the other hand, the log average failure rate was lower for a male dropout than a female dropout.

Table 3. Comparing the estimated Zero part of the Zero-Inflated Poisson and Hurdle Poisson models

	ZIP		**Hurdle Poisson**	
Variable	**Estimate**	**Std. Error**	**Estimate**	**Std. Error**
Intercept	-1.8570*	0.2400	-1.6898*	0.1654
S(=0)	1.5666*	0.6761	1.3398*	0.6277
S:time 0:time	0.1275*	0.0069	0.1274*	0.0068
1:time	-0.1998*	0.0137	-0.1989*	0.0136
S:financial aid				
0:1	0.6239*	0.0567	0.6230*	0.0567
1:1	0.2724*	0.0462	0.2723*	0.0462
S:residence				
0:1	0.1729	0.1795		
1:1	-0.0611	0.1856		
S:matric point				
0: 1	0.4636*	0.1600	0.4646*	0.1601
0: 2	1.4055*	0.1842	1.4057*	0.1842
1: 1	-0.1934	0.6058	-0.1934	0.6058
1: 2	0.6113	0.6080	0.6110	0.6080
S:race				
0:1	-0.3334	0.3148	-0.3301	0.3147
0:2	-0.1608*	0.0780	-0.1569*	0.0779
0:3	1.0322*	0.1701	1.0362*	0.1700
0:4	1.2331*	0.5671	1.2370*	0.5671
1:1	0.6320*	0.2719	0.6308*	0.2719
1:2	0.1883*	0.0643	0.187*1	0.0642
1:3	1.0038*	0.1081	1.0027*	0.1081
1:4	0.8719	0.5321	0.8708	0.5320
S:course				
0:1	-2.1057*	0.0939	-2.1073*	0.0939
0:2	-0.6066*	0.1351	-0.6062*	0.1351
1:1	-1.0723*	0.0659	-1.0716*	0.0659
1:2	0.1737	0.1356	0.1738	0.1356
Random effects				
σ_3^2	1.058		1.058	
σ_4^2	2.567		2.567	

Description: Parameter estimates of the probability of zero counts resulting from the ZIP and Hurdle Poisson models from Table 2.

The ZIP model from Table 3 suggests that the probability of failing no course increases with time for a student who eventually completer their studies as time increases. The opposite relationship occurs between a student who eventually drops out and the probability of having failed no course. This is expected as the lower the number of courses failed, the faster the student completes the registered program.

3.5 Model Comparison

Further model assessment and selection were based on the computed values for the goodness of fits. Models with a smaller BIC value are preferred. The findings from Table 4 indicate that the mixed Poisson model has the smaller BIC and is therefore considered as the preferred model

Table 4. Comparison of model fit using Bayesian Information Criteria (BIC)

	Poisson	NB	ZIP	Hurdle Poisson
BIC	115356	115386.7	116579.2	116541.5
Loglik	-57539.7	-57539.0	-58012.9	-58015.3
Deviance	115079.3	115078.1	116025.7	116030.6

Thus, the predictive model representing the log average failure rate for a student who eventually completes a selected degree can be written as

$$\begin{aligned}\log(\mu_{it}) = \log(d_{it}) &- 0.8298 - 0.7872 * S - 0.0487 * time - 0.1766 * financial\_aid \\ &- 0.0520 * matric\_point_{[26,37]} - 0.3192 * matric\_point_{[38,48]} \\ &- 0.0089 * Coloured - 0.0470 * Indian - 0.4001 * White - 0.4540 * Other \\ &+ 0.2605 * Engineering + 0.0712 * Agriculture + 0.0220 * male + b_{1i} \qquad (11)\end{aligned}$$

where $\log(d_{it})$ represents the logarithm of the number of courses enrolled by student i in semester t and $b_{1i} \sim N(0, 0.4839)$. This means that the failure rate for a student who completes a university degree is a lot lower if the student had a very high pre-university/matriculation points, especially within the interval [38,48]. Thus, more emphasis should be put in place to encourage and motivate high school pupils to perform better, as this will go a long way to better prepare them for university studies. Secondly, even though the government has facilitated higher education through grants for some students, many students still fail because of insufficient financial aid. Males and Engineering students should also be motivated towards the benefits and opportunities of degree completion, which are a direct consequence of lower failure rates.

Similarly, the model representing the log average failure rate for a student who would eventually drop out can be written as

$$\begin{aligned}\log(\mu_{it}) =\log(d_{it}) &- 0.8298 + 0.0456 * time - 0.1458 * financial\_aid + 0.0483 * matric\_point_{[26,37]} \\ &- 0.5126 * matric\_point_{[38,48]} - 0.2010 * Coloured - 0.0897 * Indian - 0.6321 * White \\ &- 0.4972 * Other + 0.3986 * Engineering - 0.1595 * Agriculture \\ &- 0.0835 * male + 0.1413 + \mathrm{b}_{2\mathrm{i}} \qquad (12)\end{aligned}$$

where $b_{2i} \sim N(0, 0.1413)$. Here, we can see that having financial aid actually lowered the failure rate for a student who dropped out, even though it did not prevent the student from dropping out. The same observation can be made about having higher matriculation points. This may seem illogical as one would expect these students with higher matriculation points to complete their degree. However, this may be more common to students who decide to change institutions or voluntarily drop out even though they were academically not at risk of exclusion. Unlike a male degree completer who has a higher failure rate, females should be warned about the disadvantages of having high failure rates as they are a direct consequence of academic dropouts

4. Conclusion

This study has shown the importance of modeling hierarchical count data with mixed models—especially the longitudinal process characterizing students' academic behavior. Using VPC, we were able to account for the variation in the log average number of courses failed by a student registered for different degree types in the college of AES. It is found that 8.5% of the variation in the log average number of courses failed between students registered for different courses. Similarly, 46.7% of the variation in log average number of courses failed between students registered for the same course type. In comparison, 44.8% of the log average number of courses failed by a student lies between the repeated measurements. This enabled us to ignore the level 3 cluster and only focus on a two-level model. Where the course-type registered was incorporated as a covariate in the subsequent models. It is also evident that distinct differences exist between students who eventually drop out and those who complete a university degree, and these differences vary with time. Condition on the random effects, accounting for within-subject variations in the number of courses failed and the exit status of a given student, we build a series of count models to measure the log average failure rate. Using the BIC and residues plots, the mixed Poisson model was selected as the best fit model. Based on the results, we can conclude that while being in university residence and the type of high school attended (Quintile) seems to affect a student's failure rate directly, other factors such as financial aid and matriculation points are essential. Indeed, for a student who completes a university degree, the failure rate is a lot lower if the student had very high matriculation points, especially within the interval [38,48]. Thus, more emphasis should be put in place to encourage and motivate high school pupils to perform better, as this will go a long way to better prepare them for university studies. Secondly, even though the government has facilitated higher education through grants for some students, many students still fail because of insufficient financial aid. Males and Engineering students should also be motivated towards the benefits and opportunities of degree completion, which are a direct consequence of lower failure rates. Similarly, for a student who dropped out, having financial aid actually lowered the failure rate, even though it did not prevent the student from dropping out. The same observation can be made about having higher matriculation points. This may seem illogical as one would expect these students with higher matriculation points to complete their degree. However, this may be more common to students who decide to change institutions or voluntarily drop out even though they were academically not at risk of exclusion. Unlike a male degree completer who has a higher failure rate, females should be warned about the disadvantages of having high failure rates as they directly consequence academic dropouts.

These findings can assist policymakers to target students at risk of University dropouts or students who may take longer to complete their university degrees. Firstly, target students at the high school level, motivate them and explains the benefits of high academic achievements towards future university achievements. Secondly, providing more financial assistance to students. Emphasis could also be put in place to counsel and motivate female students, especially those enrolled in an engineering degree, against high failure rates, which directly affect academic dropouts. One important limitation observed in this study is that we did not pay specific attention to students' academic behavior in the first year to see how it impacts their studies in the future. The apparent assumption was that some may still complete their degrees, even those who performed poorly during the first year but will take longer. Future studies involve building a model that jointly models the exit outcomes and the particular exit time. These models have not been adequately documented and discussed in the literature. Such models may help identify the specific time during which the students are at a higher risk of academic dropout, especially in discrete time.

References

Acato, Y. (2006/2007). Quality assurance vital. Technical report, New vision, university guide.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of non-traditional undergraduate student attrition. *Review of Educational Research, 55*(4), 485-540. https://doi.org/10.3102/00346543055004485

Belloc, F., Maruotti, A., & Petrella, L. (2011). How individual characteristics affect university students drop-out: A semiparametric mixed-effects model for an Italian case study. *Journal of Applied Statistics, 38*(10), 2225-2239. https://doi.org/10.1080/02664763.2010.545373

Bengesai, A. V., & Paideya, V. (2018). An analysis of academic and institutional factors affecting graduation among engineering students at a south african university. *African Journal of Research in Mathematics, Science and Technology Education, 22*(2), 137-148. https://doi.org/10.1080/18117295.2018.1456770

Booth, J. G., Casella, G., Friedl, H., & Hobert, J. P. (2003). Negative binomial loglinear mixed models. *Statistical Modeling, 3,* 179-191. https://doi.org/10.1191/1471082X03st058oa

Braxton, J., & Hirschy, A. S. (2005). *College Student Retention: Formula for Student Success*, chapter Theoretical Developments in the Study of College Student Departure, 61-87. Greenwood Press.

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence:structural equations test of an integrated model of student retention. *The Journal of Higher Education, 64*(2), 123-139. https://doi.org/10.1080/00221546.1993.11778419

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *WIREs Comp Stat, 4*, 199-203. https://doi.org/10.1002/wics.199

Charlton, J. P., Barrow, C., & Atkinson, P. H. (2006). Attempting to predict withdrawal from higher education using demographic, psychological and educational measures. *Research in Post-Compulsory Education*, 31-47. https://doi.org/10.1080/13596740500507904

Chin, H. C., & Quddus, M. A. (2003). Modeling count data with excess zeroes: An empirical application to traffic accidents. *Sociological Methods & Research, 32*, 90-115. https://doi.org/10.1177/0049124103253459

Geiser, S., & Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year:high-school record vs. standardized tests as indicators of four-year college outcomes. *Center for Studies in Higher Education*.

Georg, W. (2009). Individual and institutional factors in the tendency to drop out of higher education: Multilevel analysis using data from the konstanz student survey. *Studies in Higher Education, 34*(6), 647-661. https://doi.org/10.1080/03075070802592730

Gershenfeld, S., Hood, D. W., & Zhan, M. (2015). The role of first-semester gpa in predicting graduation rates of underrepresented students. *Journal of College Student Retention: Research, Theory and Practice, 17*(4), 469-488. https://doi.org/10.1177/1521025115579251

Gibb, S. J., Fergusson, D. M., & Horwood, J. L. (2008). Gender differences in educational achievement to age 25. *Australian Journal of Education, 52*(1), 63-78. https://doi.org/10.1177/000494410805200105

Harackiewicz, J. M., Barron, K. E., & Tauer, J. M. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational psychology, 94*(3), 562-575. https://doi.org/10.1037/0022-0663.94.3.562

Hartig, F. (2017). Dharma: residual diagnostics for hierarchical (multilevel/mixed) regression models. *R package version 0.1, 5*(5).

Hu, M. C., Pavlicova, M., & Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: Examples from an hiv-risk reduction trial regression with an application to defects in manufacturing. *The American journal of drug and alcohol abuse, 37*(5), 367-375. https://doi.org/10.3109/00952990.2011.597280

Huws, N., Reddy, P., & Talcott, J. B. (2006). Predicting university success in psychology: Are subject-specific skills important? *Psychology Learning and Teaching, 5*(2). https://doi.org/10.2304/plat.2005.5.2.133

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities in mathematics and science. *Science, 321*(5888), 494-495. https://doi.org/10.1126/science.1132154

Johnson, G. M. (1996). Faculty differences in university attrition: a comparison of the characteristics of arts, education and science students who withdraw from undergraduate programmes. *Journal of Higher Education Policy and Management, 18*(1), 75-91. https://doi.org/10.1080/1360080960180107

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies, 13*(1), 61-72. https://doi.org/10.2478/cait-2013-0006

Lambert, D. (1992). Zero-inflated poisson regression with an application to defects in manufacturing. *Technometrics, 34*, 1-14. https://doi.org/10.2307/1269547

Lemmens, J. C. (2010). *Students' readiness for university education*. PhD thesis, College of Humanities.

Letseka, M. (2010). *Student Retention and Graduate Destinations: Higher Education and Labour Market Access and Success*, chapter Poverty, race and student achievement in seven higher education institutions, 25-40. HSRC Press. https://www.asclibrary.nl/docx/337893861.pdf

Meggiolaro, S., Giraldo, A., & Clerici, R. (2017). A multilevel competing risks model for analysis of university students's carreres in italy. *Studies in Higher Education, 42*(7). https://doi.org/10.1080/03075079.2015.1087995

Milem, J. F., &Berger, J. B. (1997). A Modified Model of College Student Persistence: Exploring the Relationship Between Astin's Theory of Involvement and Tinto's Theory of Student Departure. *Journal of College Student Development, 38*(4), 387-400. https://scholarworks.umass.edu/cie_faculty_pubs/11

Moodley, P., & Singh, R. J. (2015). Addressing student dropout rates at south african universities. *Alternation Special Edition*, 17, 91-115. http://hdl.handle.net/10321/1648

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics, 33*, 341-365. https://doi.org/10.1016/0304-4076(86)90002-3

Ozga, J., & Sukhnandan, L. (1998). Undergraduate non-completion: developing an explanatory model. *Higher Education Quarterly, 52*(3), 316-333. https://doi.org/10.1111/1468-2273.00100

Pretorius, A., & Prinsloo, P. (2009). Students performance in introductory microeconomics at an african open and distance learning institution. *Africa Education Review, 6*(1), 140-158. https://doi.org/10.1080/18146620902857574

Scott, G., & Letseka, M. (2010). *Student Retention and Graduate Destinations: Higher Education and Labour Market Access and Success*, chapter Student inclusion and exclusion at the University of the Witwatersrand, 41-52. HSRC Press.

Scott, I., Yeld, N., & Hendry, J. (2007). A case for improving teaching and learning in south african higher education. *The Council on Higher Education*, 6.

Terenzini, P. T., & Pascarella, E. T. (1980). Toward the validation of tinto’s model of college student attrition: A review of recent studies. *Research in Higher Education, 12*(3), 271-282. https://doi.org/10.1007/BF00976097

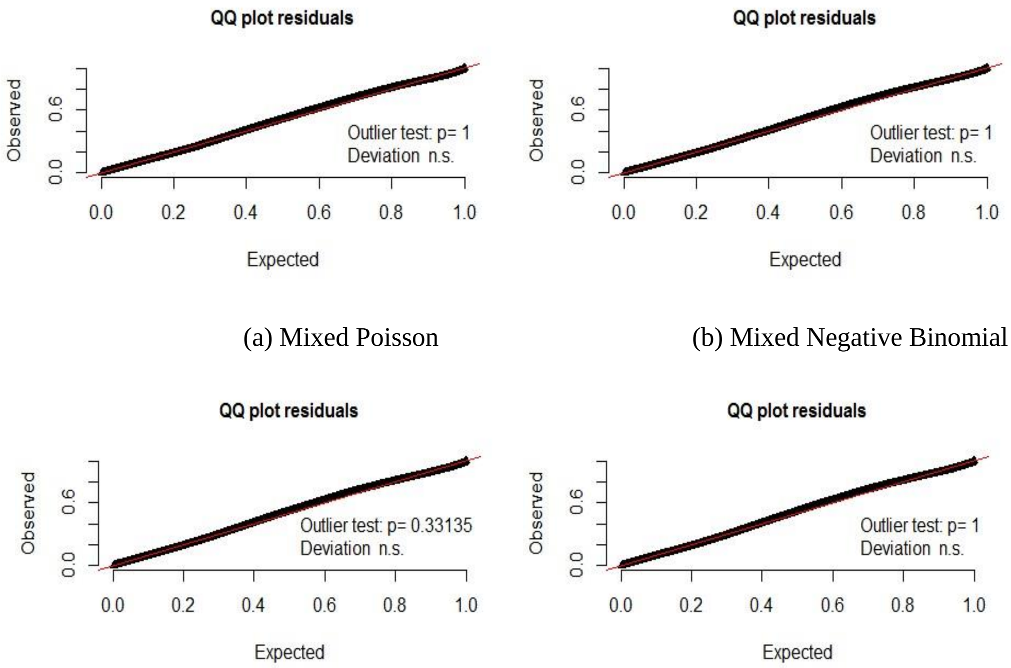
Tinto, V. (1975). Dropout from higher education: a theoretical synthesis of recent research. *Review of Educational Research, 45*, 89-125. https://doi.org/10.3102/00346543045001089

Yang, F. (2017). *A competing risks survival analysis of high school dropout and graduation: a two-stage model specification approach*. PhD thesis, Psychological and Quantitative Foundations.

Zhu, H., Luo, S., & DeSantis, S. M. (2015). Zero-inflated count model for longitudinal measurements with heterogenous random effects. *Statistical Methods in Medical research*, 0, 1-16. https://doi.org/10.1177/0962280215588224

Appendix

The model assessment was carried out by analyzing expected and observed counts through simulated residues based on diagnosis for hierarchical mixed models elaborated by Hartig (2017). The plots of these fit are presented in Figure 1.



(a) Mixed Poisson (b) Mixed Negative Binomial

(c) Mixed ZIP (d) Mixed Hurdle Poisson

Figure 3. *Quantile-Quantile plots of estimated residues*

Evidently, from Figure 3, the straight-line relationship between the observed and expected log counts suggests that the models capture the log counts.

However, while the relationship is evident around the tails, some discrepancies may be observed around the center, especially for the mixed Negative Binomial model.

Copyrights