Chapter 1

Introduction

1.0 Introduction

The lack of discipline at schools is a serious concern nowadays. The problem of learner discipline is given more attention worldwide by researchers who are involved in education since corporal punishment was abolished. In this study the problem and related factors were studied scientifically, using statistical methods such as survival analysis and zero-inflated Poisson models. The interest is on the time taken by a learner to commit an offence again after an initial offence. Zero-inflated models (ZIP) were explored since we expected an inflated number of zero counts of learners as a good number of learners might not offend again. The main focus of this study was on non-parametric, semi-parametric, and parametric survival analysis methods. The time to failure event (learner offence) is described in terms of the survivor functions and the hazard functions as used in studies related to survival analysis. The survivor function is the probability of surviving beyond a particular time *t*, and the hazard function refers to the instantaneous rate of failure. The measure of effect is the hazard ratio. The hazard ratio is defined by Cleves, Gould, and Gutierrez (2004) to be the ratio of the hazard function in the exposed group to the hazard function in the unexposed group. For example, parental involvement as one of the predictor variables can be considered as exposure for those learners under observation. This means, learners who had cases of misconduct or offence were grouped according to those whose parents were involved (exposed group) and those whose parents were not involved (unexposed group). Hence, the hazard ratio for parental involvement was calculated and interpreted. That was done for other predictor variables as well.

Information about those learners under observation who committed offences was collected and the data were processed using statistical software packages, Stata 14 (2015). The outputs from these packages provided information that was interpreted and discussed. That included

the average survival (before a second offence) times, median failure times, average hazard rates, survival probability curves, etc.

1.1 Background of Study

The study used data collected from Fidelitas Comprehensive School located in a formal urban settlement in Soweto. The enrolment at this school has been consistently higher than 1200 for many years. The current enrolment for 2015 is 1390, an increase from 2014 by almost 200 learners. This is a comprehensive school with grade 8 to 12 classes. The school is not classified by the Department of Education as underperforming because the learner performance has been above the target of 80 % in grade 12 results for the past seven consecutive years. It is classified as a non-fee paying school according to the criteria set by the Department of Education. This indicates that it is located in an area of homes with low income levels. This is, however, in contrast to the types of houses around its neighbourhood and other assets that people have. The school environment is conducive for teaching and learning but lack of discipline is always an issue as is the case in schools countrywide and abroad. There are a number of learners who have offence cases reported and some are repeat offenders. Although most of the offences are those learner acts that fall outside the school code of conduct, some are considered more serious and require more attention. Examples include bullying, fighting, suspected substance abuse, disobedience, absence from classes, school work not done, absenteeism, as well as a combination of offences. Past records from this school show that grade 10 is the class that reports the highest numbers of offences.

The main reason for this study of learner misconduct is the fact that at the school where I'm working, several cases are reported to me as a senior manager. Learner misconduct is a challenge which teachers find difficult to handle, especially after corporal punishment was abolished by education authorities in the government. It seems little or no effective measures to deal with the lack of discipline are available to teachers since any form of punishment that might cause physical or psychological pain is not allowed.

Factors that were used to predict learner misconduct were limited to those available and relevant to a school environment such as a learner's grade (class), gender, home, parental involvement, repeating a grade, learner performance, suspected substance abuse, and hostel residence. The survival time for a learner to committing an offence or misconduct becomes important because effective measures of discipline need to be applied appropriately and relevantly. In order to implement severe measures such as suspension and expulsion from school, the behaviour of an affected learner needs to be observed for a specific period of time for the authorities to have back up information before such harsh decisions are made. The subjects observed are learners and the event of interest is when the learner commits an offence again, similar or different from the initial one. A follow-up study is undertaken on those learners who had cases of misconduct reported since their first registration at school.

1.2 Aim and Objectives of Study

The main aim of this study is to model learners' lack of discipline at school using statistical methods in order to understand the problem. This should help in finding better ways and effective measures of managing lack of discipline at schools. The specific objectives are as follows:

1. to describe the length of time that learners survive until they commit or are involved in an offence again.
2. to examine the extent to which related factors contribute to time to a second offence.
3. to estimate and interpret the survivor and hazard functions from survival data using non-parametric survival analysis.
4. to use semi-parametric methods such as the Cox proportional hazards model and calculate the hazard ratios for the covariates in the model.
5. to regress the survival time on group membership using parametric methods such as the Weibull, Gompertz, log-normal, exponential, and other applicable methods.
6. to model the number of offences as a Poisson random variable using;

a. the Poisson model,
b. the Negative Binomial Model, and
c. the Zero-Inflated Poisson Model.

1.3 Research Data Description

This is a study of one metric dependent variable's relationship with categorical and continuous independent variables. The dependent variable is survival time in weeks until an event (repeat offence) occurs. The subjects observed are learners at a school which offers grade 8 to 12. Taking into account learners repeating a grade, the subjects are observed in a five year period. Special attention will be on those learners who had made an offence before. Survival time is the length of follow up time (in weeks) for an individual learner to commit an offence again after an initial offence. The school records provide the names of learners, types of offences committed, and the dates on which the offences were committed.

For analysis the variables are obtained from the school records as follows:

- number of weeks between first offence and second offence by each learner observed
- gender (1 if male, 0 if female)
- grade (to indicate whether a learner is in grade 8, 9, 10, 11, or 12)
- repeating (1 if a learner is repeating a grade, 0 if not repeating)
- home location (1 if learner's home is local, i.e. within a 3 km radius from the school, 0 otherwise)
- hostel (1 if learner is a hostel resident, 0 if not a hostel resident)
- parent involved (1 if the parent was involved in the learner's case, 0 otherwise)
- suspected substance abuse (1 if a learner is suspected to be using drugs, 0 otherwise)
- number of all offences committed after the initial offence by each learner under observation (a random variable taking on one of the values 0, 1, 2, . . .)

- learner performance is included as continuous variable in the three compulsory school subjects, namely, English, Mathematics/Maths Literacy, and Life Orientation. Life Orientation is believed to be an important factor in child social interaction and development.

- the censoring variable is indicating whether the event of interest is reached or not, for example, event = 0 if censored (no offence), event = 1 if failed (offence committed).

1.4 Data Source

This study utilises learner offence data obtained from a public high school in the Gauteng province. The school adopts the policies and standards of the Department of Basic Education. This implies that the data is credible for study purposes. The records of learner offences, marks obtained in the subjects involved, and other information are accessible from the school administration office on request. Appendix A provides the data set used in the analysis.

Chapter 2

Literature Review

2.1 Preliminary Literature Review

There are several studies which have been conducted on learner discipline and corporal punishment but in most of them researchers did not apply statistical techniques such as survival analysis to explain learner behavioural patterns. Hair, Black, Babin, Anderson, and Tatham (2005) indicate that statistical techniques are popular because they enable organisations to create knowledge and thereby improve their decision making. One related study that follows up subjects or cases is the study of parolees by Beck and Shipley (1987) that examined the time until re-arrest. This study was done mainly using survival analysis techniques.

Recommended disciplinary measures as alternatives to corporal punishment at schools seem not to be effective. Learners still continue to commit misconduct or be involved in offences even if one had a case reported before. Lack of discipline at schools has increased. There are related studies concerning this issue. A comparative study by Wolhuter and Russo (2013) indicates that teachers are at a loss as to how to handle learner discipline. Lambert and McCombs (1998) show that disciplinary problems, drug abuse problems, and increasing violence in America's schools have generated an extensive national debate over the past 2 decades. Similarities between that study and the current study is that, of those subjects followed up, there are a number of learner misconduct cases reported on a monthly basis. Researchers are interested in this area of research because discipline directly affects teaching and learning at schools. Researchers also give specific attention to deterioration in discipline since corporal punishment was abolished. Teachers know that discipline is the starting point of learner success to the extent that some still apply corporal punishment even when they know it is illegal. Details of corporal punishment that are not allowed at schools and more issues

concerning learner discipline can be found in the National Education Policy Act (1996), Employment of Educators Act (1998), South African Schools Act (1996), and South African Council for Educators Act (2000). Harrisunker (2014) (pages 1 to 2, paragraph 2) refers to the discipline problems as "a power struggle between students, who have no power in the school system, and adults who do have power". Opinions about educators continuing to use corporal punishment as they believe it to be the most effective tool, are widely discussed in literature. According to Ntuli (2013), principals and educators try to use contemporary disciplinary measures but they are not effective alternatives to corporal punishment. Dirks (2012) stated that educators resort to illegal forms of punishment in a desperate attempt to maintain discipline.

Parental involvement is considered an important factor that is closely related to a child's discipline. Some parents do not respond to calls or letters inviting them to school to discuss matters concerning their children who were involved in a misconduct or offence. Dirks (2012) indicates that educators feel disrespected by parents when they ignore such invitations. There are parents who show little or no interest in their children's school work as well as their behaviour. At Fidelitas School, where this study was based, we had many report cards that were not collected by parents from previous terms and years since they are not given to learners directly but to parents or guardians. Many researchers emphasize parental involvement as a key to discipline at schools. James (2014) pointed out that schools that succeed with discipline, since corporal punishment was abolished are those making the effort to get maximum parental participation. Baumrind (1996) discussed child disciplinary practices by parents in three forms, namely, authoritative, authoritarian, and permissiveness. Based on several parental control variables, her findings indicated that the most effective form is being authoritative because of its nature in affirming the child's present qualities. The study by Singh (2012) found that the causes of learner aggression were also rooted in the family. Dishion and MacMahon (1988) discussed establishment of a link between parental monitoring and problem child behaviour. In agreement with other researchers the findings using longitudinal data

indicated that serious antisocial behaviour can be the result of a progression from relatively trivial behaviours to increasingly dangerous behaviours.

The relationship between survival time and other predictor variables will be investigated. These include gender, grade, whether the learner is repeating or not, area where the learner comes from, and suspected drug abuse. Prevalence of misconduct in boys or girls can be looked at in terms of the ratio of the hazard function in boys to the hazard function in girls. According to Mestry, van der Merwe, and Squelch (2006) there are no differences between male and female learners with regard to involvement in bullying at schools. For the variable, grade, the hazard ratio is the ratio of the hazard function in one grade (e.g., grade 11 learners) to the hazard function in the other grade (e.g., grade 8). The learner's grade is also a factor to be looked at so that special attention could be given to those discovered to be more vulnerable to misconduct. Having noticed that many studies of learner discipline were conducted at secondary schools, Keating (2011) focused on the foundation phase and discovered that despite having younger children in the foundation phase, there are still many cases of learner misconduct challenging the job security of teachers as they are tempted to use illegal forms of punishment. For the repeaters and non-repeaters, the hazard ratio is the ratio of the hazard function in the repeating learners to the hazard function of those not repeating. Age will be considered as a cofactor to repeating learners rather than a direct factor on its own because repeating learners are expected to be older than those who never repeated in a particular grade. Another variable of interest is drug abuse. This is a sensitive factor since it does not only trouble the schools in maintaining discipline but it is a problem affecting the communities at large. Interest will also be on the ratio of the hazard function in the learners exposed to drugs to the hazard function in the unexposed learners.

The factors mentioned above and those brought forward by researchers in studying their impact on learner discipline will be looked at with the aid of statistical software package like Stata 14 (2015).

2.2 Theoretical Review

In this section the theory of the methods used in this study is reviewed.

2.2.1 Survival Analysis

In this work we mainly follow the notation used in Cleves, *et al.* (2004). The event of interest is learner offence, referred to as the failure event in survival analysis. The time to failure event, that is when the learner committed an offence again after the initial offence, is called the survival time. Survival time is the dependent variable for this analysis and is denoted by *T*. It is a random variable $T > 0$ and *t* is a specific value of time *T*. The time to failure event is described in terms of the survivor function $S(t)$ and the hazard function $h(t)$. The survivor function is the probability of surviving beyond time t. It is expressed as $S(t) = Pr(T > t)$.

Nonparametric survival analysis such as Kaplan-Meier estimate, makes no assumptions about the functional form of the survivor function, the hazard function, cumulative hazard, and the effects of the covariates. A Kaplan-Meier estimate of the survivor function $S(t)$ is given by

$$\widehat{S(t)} = \prod_{j \mid t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right) , \qquad [1]$$

where n_j is the number of individuals at risk at time t_j and
d_j is the number of failures at time t_j.

Cleves *et al.* (2004) notes that in survival analysis the median failure time is more relevant than the mean failure time. The time to event data has a positively skewed distribution because subjects may have exceptionally short or long survival times. The problem with mean failure time arises when the last observation is censored because summation is made on those times at which the event of interest occurred.

The hazard function is the instantaneous rate of failure defined as the limiting probability of the event occurring in a given interval, given that the subject has survived up to time t, divided by the width of the interval Δt. It is expressed as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t<T<t+\Delta t \mid T>t)}{\Delta t}. \quad [2]$$

Plots are very important in statistical analyses. The Kaplan-Meier survival plot represents the survival function against time. It is useful when comparing two or more groups of learners involved in the study. For the observations 1, . . . , n, the Kaplan-Meier estimator displays a step function that increases by $\frac{1}{n}$ at each observation in the data. If all the learners were to reach the event of interest (commit an offence) by the end of the study period, fewer learners would be left for observation at later time periods. It indicates that the variance increases as the time increases. The shortcoming of the Kaplan-Meier estimator is that once it is zero, it remains at zero. The alternative which does not suffer from this shortcoming is Nelson-Aalen estimator by Nelson (1972) and Aalen (1978). The Nelson-Aalen estimator is expressed as

$$\widehat{H(t)} = \sum \frac{d_j}{n_j} \text{ for } j \mid t_j \leq t \quad [3]$$

where n_j is the number at risk at time t_j , d_j is the number of failures at time t_j, and the summation is over all different failure times less than or equal to t. The relationship between survivor and cumulative hazard functions is given by

$$S(t) = \exp\{-H(t)\}. \quad [4]$$

In small samples like the learner offence dataset used in this study, the Kaplan-Meier estimator is preferred when estimating the survivor function and the Nelson-Aalen estimator is preferred when estimating the cumulative hazard function.

Cleves, *et al.* (2004) specify several tests appropriate for testing the equality of survivor functions across two or more groups. The log-rank test of Mantel and Haenszel (1959) is preferred when the hazard functions are thought of as being proportional to one another. The test of Wilcoxon - Breslow (1970) is preferred when the hazard functions are not proportional

but vary in other ways. The test of Peto-Peto (1972) is preferred when there are very big differences in the censoring patterns between groups. These tests compare the expected number of failures with the observed number of failures for each group and then combine the comparisons over all observed failure times. The tests test the null hypothesis (H_o) that there is no difference in the survivor functions between two or more groups of learners. The null hypothesis can be expressed in terms of the hazards as $H_o : h_1(t) = h_2(t)$ and the alternative hypothesis is $H_a : h_1(t) \neq h_2(t)$. The log-rank chi-squared test statistic with $m-1$ degrees of freedom is a quadratic form $\boldsymbol{u}'\boldsymbol{V}^{-1}\boldsymbol{u}$ where $\boldsymbol{V}$ is $m \times m$ variance matrix for m groups. $\boldsymbol{u}'$ is the row vector expressed as

$$\boldsymbol{u}' = \sum_{j=1}^{k} W(t_j)(d_{1j} - E_{1j}, \ldots, d_{mj} - E_{mj})$$

and the $m \times m$ variance matrix $\boldsymbol{V}$, has its elements calculated by

$$V_{il} = \sum_{j=1}^{k} \frac{W^2(t_j) n_{ij} d_j (n_j - d_j)}{n_j(n_j - 1)} \left(\delta_{il} - \frac{n_{ij}}{n_j}\right)$$

V_{il} is the individual element of the $m \times m$ variance matrix $\boldsymbol{V}$ on the i *-th* row and l *-th* column, where $i = 1, \ldots, m$, $l = 1, \ldots, m$, and $\delta_{il} = 1$ if $i = l$ and 0 otherwise. $W(t_j)$ is the weight function equal to zero when the number at risk n_{ij} is zero and is equal to one when the number at risk is nonzero. There are n_j subjects at risk, of which d_j fail and $(n_j - d_j)$ survive. d_{mj} is the number of failures in group m at time t_j, and E_{mj} is the expected number of failures in group m at time t_j.. The Wilcoxon test is constructed in the same way as the log-rank test by setting $W(t_j) = n_j$ in equations [3] and [4]. The Peto-Peto test uses an estimate of the overall survivor function as a weight function, that is, setting $W(t_j) = \tilde{S}(t_j)$.

The Cox proportional hazards regression model (Cox 1972) states that the hazard rate for the *j-th* subject ($j = 1,2, \ldots, n$) in the data is

$$h(t|x_{ij}) = h_0(t)\exp(x_{ij}\beta_i), \qquad (i = 1,2, \ldots, p) \qquad [5]$$

where and $h_0(t)$ is the baseline hazard without a particular parameterization. Cleves, *et al*. (2004) explain this model as having the form $h(time \mid group) = h_0(time)e^{group \times \beta}$.

The exponential function of the covariate, $exp(\beta_i)$, measures the effect of the i predictor variable(x_{ij}). It is a simplified form of the Hazard Ratio *(HR)*. The hazard ratio is the ratio of the hazard function in the exposed group to the hazard function in the unexposed group. When comparing subjects *j* and *m*, we have the hazard ratio expressed as

$$\frac{h(t|x_{ij})}{h(t|x_{im})} = \frac{exp(x_{ij}\,\beta_i)}{exp(x_{im}\beta_i)}.$$

For parametric hazards model

$$h(t) = h_0(t)\, exp\,(\beta_0 + \boldsymbol{x_j\beta}), \qquad [6]$$

we can use the matrix algebra notation, where we have a covariate vector $\boldsymbol{x_j}$ as a vector of predictors and a vector of coefficients $\boldsymbol{\beta}$ in a relationship, $\boldsymbol{x_j\beta} = x_{1j}\,\beta_1 + x_{2j}\beta_2 + \ldots + x_{kj}\beta_k$, where $\boldsymbol{x_j} = (x_{1j}, x_{2j}, \ldots, x_{kj})$ *and* $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_k)$. The regression coefficients $\beta_1, \beta_2, \ldots, \beta_k$ are to be estimated from the data. In terms of the hazard ratios for the covariates, the null hypotheses about the covariates effects that are tested are: $H_0 : \exp(\boldsymbol{\beta}) = 1$ (i.e. the hazard ratios are equal to 1) versus $H_1 : \exp(\boldsymbol{\beta}) \neq 1$. This is equivalent to $H_0 : \boldsymbol{\beta} = 0$, saying that the coefficients are equal to zero versus $H_1 : \boldsymbol{\beta} \neq 0$. $\boldsymbol{\beta}$ is a coefficient vector. The hazard ratio equal to 1 means that there is no change in the response variable (survival time) due to exposure of the covariate. The hazard ratio less than 1 means that exposure is protective or reduces hazard, and the hazard ratio greater than 1 means that exposure increases hazard.

Nonparametric and Semi-parametric models compare learners at the times when failures (offending again) occur. Parametric models do not base their results on such comparisons but depict what occurs over the whole interval, given what is known about the learner during the current time. The assumed distribution of the survival times specifies the hazard functional form. We regress survival time on various groups of learners using the exponential, Weibull, and Gompertz distributions. For parametric models, the baseline hazard $h_0(t)$ in equation [6],

$h(t|x_{ij}) = h_0(t)\exp(x_{ij}\beta_i)$, is specified. As explained by Cleves, *et al*. (2004), fitting the exponential model means assuming the baseline hazard $h_0(t) = \exp(\alpha)$ for some α. The baseline hazard is assumed constant over time, e.g. $h_0(t) = \boldsymbol{\omega}$. It means there is one extra parameter to estimate. So when fitting the exponential model, we are estimating the parameters (ω, β_i). For the Weibull model the baseline hazard is assumed to be

$$h_0(t) = \omega^p p t^{p-1}.$$

That means there are two extra parameters to estimate, i.e. when fitting the Weibull model we are estimating the parameters (ω, p, β_i). For the Gompertz model we assume the baseline hazard

$$h_0(t) = exp\{\alpha + \gamma t\}.$$

In this model we are estimating the parameters $(\alpha, \gamma, \beta_i)$, that is we are estimating two extra parameters. In this study the parametric models are fitted by maximising the likelihood function $L(\theta)$ and using procedures as discussed by Kalbfleisch and Prentice (2002). The likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} \lambda\,(t_i|x_i)^{d_i} S(t_i|x_i)$$

$\lambda(t_i)$ and $S(t_i)$ denote the hazard function and survival function, respectively. The indicator d_i takes the value 1 when a learner is committing a second offence and takes the value 0 for censored observations, and t_i is the survival time to second offence.

In relation to the learner offence study, the actual model for parametric and semi-parametric analyses is

$$h(t) = h_0(t)\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_2 X_6 + \beta_{12} X_2 X_7)$$

where *h_0 (t)* is the baseline hazard, X_1 is Gender (1 if male, 0 if female), X_2 is Grade (whether a learner is in grade 8, 9, 10, 11, or 12), X_3 is Repeat (1 if a learner is repeating a grade, 0 if not repeating), X_4 is Home location (1 if learner's home is local, 0 otherwise), X_5 is Hostel residence

(1 if learner is a hostel resident, 0 if not a hostel resident), X_6 is Parent involvement (1 if the parent was involved in the learner's case, 0 otherwise), X_7 is Substance (1 if a learner is suspected to be using drugs, 0 otherwise), X_8 is English (marks obtained in English), X_9 is Maths (marks obtained in Mathematics), X_{10} is LO (marks obtained in Life Orientation), X_2X_6 is the interaction of Grade and Parent involvement, X_2X_7 is the interaction of Grade and Suspected Substance abuse. The hazard ratios of the learners do not depend on the choice of $h_0(t)$. The regression coefficients $\beta_1, \beta_2, \ldots, \beta_{12}$ are to be estimated from the data.

The null hypothesis about the group effects that are tested using the calculated p-value in testing $H_0: \beta_x = 0$ implies that the group effect is zero, and the alternative hypothesis is $H_\alpha: \beta_x \neq 0$ implying that the group effect is not zero. The effects of interactions between other variables will also be examined. The effect of one factor may increase or decrease with the level of another factor. For example, having survival time explained by only two factors, gender and grade, the Cox model with interaction is

$$LRH = \beta_1 gender + \beta_2 grade + \beta_3 gender * grade$$

LRH is the log relative hazard. To quote Cleves, *et al.* (2004) (page 162) "In a Cox model, the linear predictor $\boldsymbol{x\beta_1}$ is the logarithm of the relative hazard $exp(\boldsymbol{x\beta_1})$ since the baseline hazard function is multiplicatively shifted based on this value". If $\beta_3 > 0$, the effect of gender increases with grade. If $\beta_3 < 0$, the effect of gender decreases with grade. The effect is constant if $\beta_3 = 0$. Alternatively, if the 95% confidence interval used to estimate β_3 includes zero, we cannot reject the null hypothesis at the 5% level that there is no interaction effect.

In survival analysis observations that have not reached the event of interest by the end of study are referred to as censored. Censoring is defined by Kleinbaum and Klein (2012), as, when having some information about individual survival time but not knowing the survival time exactly. These include lost cases, for example, learners who might have withdrawn from school. The basic layout of data will have the censoring indicator equal 0 for censored observations and

1 for failure events. If the event of interest (offence) occurs after the study has ended, that is referred to as right censoring. Left censoring is when the event of interest occurs but the starting time of follow-up is unknown. For example a learner who committed an offence but has not been followed up since he had no previous case reported.

2.2.2 Poisson Regression Model

Poisson regression model can be applied when the response variable Y_i is a count such as a number of offences occurring in a week, month, year, or any relevant time or space. Cameron and Trivedi (2013) explain Poisson regression as the standard model for count data. The discrete response variable Y_i is assumed to follow a Poisson distribution, i.e.,

$$Y_i \sim Poisson\,(\mu_i)\,, \quad i = 1,\ 2,\ \ldots\,, N$$

where, μ_i is the average number of offences per given period e.g. week, and *N* is the number of observations.

The density of Y_i is

$$P[Y = y] = \frac{e^{-\mu t}(\mu t)^y}{y!}\,, \quad y = 0, 1, 2,\ \ldots\,,$$

where, *t* is the exposure representing the length of time during which the events were recorded and the rate parameter $\mu > 0$. By setting the length of the exposure period equal to unity, then we have the probability mass function of a Poisson random variable expressed as

$$P[Y = y] = \frac{\mu^y e^{-\mu}}{y!}\,,\ y = 0\,, 1, 2,\ \ldots\,, \qquad [7]$$

where μ is the average number of offences per specified period. The variance of a Poisson distributed response Y_i is $Var(Y_i) = \mu_i$, so the variance function is $V(\mu) = \mu$

We want to relate the mean (μ_i) to a vector of covariates (regressors). The link function commonly used in Poisson regression is the log function expressed as

$$\log(\mu_i) = \boldsymbol{x_i'\beta}\,, \qquad [8]$$

where $\boldsymbol{x_i}$ is a vector of predictors and $\boldsymbol{\beta}$ is a vector of coefficients. Taking the exponential of $\boldsymbol{x_i'\beta}$ ensures that the parameter μ_i is nonnegative. This model implies that the conditional mean is given by

$$E[y_i \mid x_i] = exp(\boldsymbol{x_i'\beta}). \quad [9]$$

Due to equal dispersion or equality of the conditional variance and conditional mean, we have

$$V[y_i|x_i] = exp(\boldsymbol{x_i'\beta}). \quad [10]$$

The standard estimator for this model is the maximum likelihood estimator (MLE). Given the independent observations, the log-likelihood function is

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} [\boldsymbol{y_i}\, \boldsymbol{x_i'\beta} - exp(\boldsymbol{x_i'\beta}) - ln\boldsymbol{y_i}!]$$

Differentiating equation [9] with respect to $\boldsymbol{\beta}$ yields the Poisson MLE ($\hat{\beta}$) as the solution to the first-order conditions

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} [\boldsymbol{y_i} - exp(\boldsymbol{x_i'\hat{\beta}})\boldsymbol{x_i}] = 0.$$

2.2.3 The Negative Binomial Regression Model

It is best to consider a Negative Binomial Regression Model if we find the response to be overdispersed. Overdispersion is when the correlation in the data is more than it is allowed by the distributional assumptions. We can have a look at the number of offenders as a count out of a possible total (students in a cohort). In this study we look at the number of learners who had offended out of a total number of learners in specific categories of predictors at school. Cameron and Trivedi (2013) indicate that the Negative Binomial model is a standard model that accommodates overdispersion. We assume that data is overdispersed if it has more variation than that allowed for by the Poisson model. However, if we have large totals, like a higher number of learners in a specific category of predictors at school that relates to small probabilities of having learners offending again, then the Poisson model should work well. In the Negative Binomial regression model, instead of assuming that the response Y_i follows a Poisson distribution, we now assume that it follows a Negative Binomial distribution. The Negative Binomial model is expressed as

$$P[Y = y] = \frac{\Gamma\left(\frac{1}{k}+y\right)}{\Gamma\left(\frac{1}{k}\right) y!}\left(\frac{k\mu}{1+k\mu}\right)^{y}\left(\frac{1}{1+k\mu}\right)^{\frac{1}{k}}, \quad [11]$$

and the link function is

$$\log(\mu_i) = o_i + x_i^T\beta \quad [12]$$

where x_i and β are defined as in 2.2.2. The term o_i represents the offset (log of the exposure) and k is the dispersion parameter for the Negative Binomial (NB) distribution, so that NB $\rightarrow$ Poisson as $k \rightarrow 0$. The expected number of offences is $E(Y_i) = \mu_i$ and the variance is $Var(Y_i) = \mu_i + k\mu_i^2$. π_i is the probability that the i-th observation on y takes the realised value y_i and n_i is the i-th nonnegative integer. We test the null hypothesis that data follows a Poisson distribution versus that it follows a Negative Binomial distribution, i.e. (H_0: $k = 0$ versus H_1: $k > 0$). If k is greater than zero, then fit the negative binomial model. If k is approximately zero, then fit a Poisson model. When fitting the Negative Binomial model, the estimated covariance matrix for $\boldsymbol{\beta}$ is$(\boldsymbol{X^TWX})^{-1}$, obtained by Fisher scoring procedure, where

$$W = Diag\left[Var(Y_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2\right]^{-1} \quad [13]$$

is the matrix of weights, $\eta_i = \log \mu_i = x_i^T\beta$, and

$$z_i = \eta_i + \frac{(y_i-\mu_i)}{\mu_i}. \quad [14]$$

By regressing z on X using the μ_i's as weights, we obtain the estimate of β and repeat until the value of β converges.

2.2.4 The Zero-Inflated Models

Consider a study of offences in any grade at school where the response is the number offences committed by the learners in a shorter period such as the past three weeks. The data are likely to show an excess of zeros as it is common with count data. Even the most problematic grade might report zero offences during those three weeks of observation. Zero-inflated models are capable of dealing with an excess of zero counts, as explained by Lambert (1992) and Mullahy (1997). They are two component mixture models that combine a point mass at zero with a count distribution such as Poisson or Negative Binomial. They have two sources of zeros, one

coming from the point mass and the other coming from the count component. The point mass at zero is denoted by $I_{\{0\}}(y)$ and the count distribution by $f_{count}(y; x, \beta)$. The probability of belonging to the point mass component is $\pi = f_{zero}(0; z, \gamma)$. The probability of observing a zero count is inflated with $\pi = f_{zero}(0; z, \gamma)$ giving an expression

$$f_{zeroinfl}(y; x, z, \beta, \gamma) = f_{zero}(0; z, \gamma) \cdot I_{\{0\}}(y) + (1 - f_{zero}(0; z, \gamma)) \cdot f_{count}(y; x, \beta), \quad [15]$$

where the response variable y is the number of offences committed by the learner, $I(\cdot)$ is the indicator function, x representing the count component, z representing the zero-inflated component, $\boldsymbol{\beta}$ is the vector of regressors in the count component, and the set of parameters of γ come from the zero-inflated component. The binomial generalised linear model (GLM) with $\pi = g^{-1}(z^T\gamma)$ is used to model the unobserved probability of belonging to the point mass component. The corresponding regression equation for the mean μ_i is given by

$$\mu_i = \pi_i \cdot 0 + (1 - \pi_i) \cdot \exp(\boldsymbol{x_i^T \beta}). \quad [16]$$

The link function $g(\pi)$ in binomial GLM is the logit link. Inference is performed by applying the Negative Binomial model and the parameters β and γ are estimated by the maximum likelihood (ML). Long and Jeremy (2006) explain the numerical methods that are used to find the maximum of the likelihood function and how the slope of the likelihood function and the rate of change in the slope determine the estimates for the parameters.

An assumption of a count response variable to be distributed with point mass of 1 at zero, with mixing probability, is also pointed out by Hall (2000). For the Zero Inflated Poisson (ZIP) model the components of the mixture model mentioned, are those having a zero count with a probability of 1, called (*Always-0 group*) and those having counts predicted by the standard Poisson, called (*Not always-0 group*). Hall (2000) indicates that the random effects are included in the ZIP model to accommodate for the repeated measures in the data set. The overall model combines the following probabilities from the two groups: $P[Y_i = 0|x_i, z_i] = \psi_i \times 1 = \psi_i$,

where ψ_i denotes the probability that observation i is in (*Always-0 group*), and z_i is the vector of covariates. If the zero counts are in (*Not always-0 group*), we have $P[Y_i = 0|x_i, z_i] = (1 - \psi_i) \times \frac{e^{-\mu_i}\mu_i^0}{0!} = (1 - \psi_i)e^{-\mu_i}$. We also have the non-zero counts in (*Not always-0 group*) with the probability given by $[Y_i = y_i|x_i, z_i] = (1 - \psi_i) \times \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$. The overall model is given by

$$P[Y_i = y_i|x_i, z_i] = \begin{cases} \psi_i + (1 - \psi_i)e^{-\mu_i} & if\ y_i = 0 \\ (1 - \psi_i) \times \dfrac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & if\ y_i > 0 \end{cases}$$

The mean and variance are given by

$$E(y_i|x_i, z_i) = [0\ \times\ \psi_i] + [\ \mu_i \times (1 - \psi_i)] = \mu_i(1 - \psi_i),$$
$$V(y_i|x_i, z_i) = \mu_i(1 - \psi_i)(1 + \mu_i\psi_i).$$

Chapter 3

Methodology

3.1 Introduction

The data and analysis methods are outlined in this chapter. The scope of this study is statistical analysis of learner offence in relation to factors around the school environment. The purpose is to analyse the time until the learner commits an offence again and to model the number of offences. The data set was obtained from a high school with a comparatively large enrolment.

As usual the first step in data analysis is to view the graphical representation of data and describe summary statistics to be familiar with the data. Examining the data graphically also helps to check for the incorrect data such as typing errors, recording errors, etc. A detailed analysis focuses on models that are built to describe the distribution of failure times using the survivor and hazard functions. Another focus was on Poisson models, specifically for count data, the number of learner offences in this case.

3.2 Data

The learner offence data was obtained from a high school with an enrolment of about 1390 learners. The records of learner offences and other information are accessible from the school admin office and various class teachers. The learner offence data set used in this study consists of 13 variables and 83 observations made for the period July 2013 to December 2014 with follow up studies until 30 June 2015. The data set showing all the variables and actual values is given in Appendix A. The variable time represents survival time in weeks until the learner offends again or the length of follow up for censored observations. The marks recorded in the three subjects are those obtained by the learner at the end of the year in which an offence was

committed. The variable code, name, and description are given in Table 3.1. The variable code is used to suit the number of characters allowed by the software package, Stata 14 (2015).

Table 3.1 Description of variables in dataset

Variable Code	**Variable Name**	**Description**
Time	Time	Survival time.
Cense	Censoring variable	Whether a failure event was reached or not.
Gender	Gender	Whether the learner is male or female.
Grade	Grade	The grade in which the learner was when committing an offence.
Repeat	Repeating	Whether the learner was repeating a grade or not.
Homeloc	Home location	Whether learner's home is local to school area or not.
Hostelres	Hostel resident	Whether a learner is hostel resident or not.
Parentinvo	Parent involvement	Whether the parent was involved in the learner's case or not.
Substance	Suspected substance abuse	Whether the learner is suspected of using drugs or not.
English	English	Marks scored in English.
Maths	Mathematics	Marks scored in Mathematics or Mathematical literacy.
L O	Life Orientation	Marks scored in Life Orientation.

3.3 Descriptive Statistics

As indicated earlier the researcher must first have a graphical view of data. This was done by drawing and examining basic graphs such as histograms, cumulative frequency curves, and boxplots. Using survival analysis plots we were able to interpret the hazard, hazard rate, and

cumulative hazard. In drawing and interpreting these plots it is also necessary to view basic summary statistics. For instance, the range gives an idea of how data was divided into intervals used to plot the histogram. From the summary statistics, we are able to interpret the defined mean time to failure and the median failure time. These measurements of central tendency allow one to have a sense of the typical time to failure for a particular distribution. As indicated by Cleves *et al.* (2004), some care is required when interpreting the mean failure time as survival-time distributions can display long tails to the right.

3.4 Model Building

3.4.1 Survival Analysis Methods

The main approach in this study of learner offences is survival analysis. Various models suitable for survival analysis were explored. Non parametric methods such as Kaplan-Meier estimator, semi-parametric methods such as Cox proportional hazards regression model, and parametric hazards models such as exponential, Weibull, and Gompertz were used. Parental involvement is considered as an important influential factor to learner discipline. The study investigated parental involvement as an exposure in terms of the Cox proportional hazards model. That enabled the calculation of the ratio of the hazard function in the group whose parents were involved to the hazard function in the group whose parents were not involved. Conclusions on whether exposure was protective or increased hazard were made.

Other parametric methods make an assumption of normality and lack the ability to work with censored data. The assumption of normality of time to event is unreasonable, for example, an event cannot have instantaneous risk of occurring which is constant over time. Parametric survival analysis is concerned with substituting the normality assumption appropriately and is able to work with the censored data. Information about survival time which is our dependent variable, was obtained from school records of learner offences or misconduct cases. A follow-up study was only done on those subjects (learners) who had been previously reported for certain offences. Some learners might be susceptible or prone to be involved in offences and other forms of misbehaviour, due to influences arising from a variety of factors. For instance, a

group of learners coming from one area might have more cases reported than another group from a different area.

The time until the learner make an offence again is observed. Not all types of offences were considered for the purpose of this study because some might be very common and regarded as minor. For example, late coming in the morning may be considered to be a minor case and therefore be excluded from the analysis unless the subject was already under observation from a previous case. The predictor variables include learner's gender, grade/class, whether repeating a grade or not, parental involvement, home area or location. These are the variables to explain the dependent variable (survival time). The censoring indicator was also included in the basic layout of data. Those subjects (learners) who have not reached the event of interest by the end of the study are referred to as censored. For censored observations, the recorded survival time in weeks is the length of follow up until the study ends. Unlike for subjects who reach the event of interest, number of days won't be a relevant time unit for censored observations as it would be too big because the study was undertaken for a longer period. The basic layout of data was entered into a statistical software package Stata 14 (2015). The outcome was used to explain and discuss the relationships between survival time and the predictor variables. Having the dataset with survival times, number at risk (with one offence), number failed (committed second offence), and number censored (not committed second offence), the Kaplan-Meier estimator of the survivor function was used to calculate the probabilities of surviving beyond the particular time periods. The corresponding unconditional probabilities of surviving beyond specified times give estimates of the survivor functions, $S(t)$. Learners were also be grouped, for example, according to group 1 (punished before) and group 2 (not punished). This is analogous to common grouping of subjects according to treatment group and control group.

The full model that was used to investigate the effects of predictor variables and interactions is

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_2 X_6 + \beta_{12} X_2 X_7 + \epsilon$$

where y is the survival time, X_1 is Gender (1 if male, 0 if female), X_2 is Grade (whether a learner is in grade 8, 9, 10, 11, or 12), X_3 is Repeat (1 if a learner is repeating a grade, 0 if not repeating), X_4 is Home location (1 if learner's home is local, 0 otherwise), X_5 is Hostel residence (1 if learner is a hostel resident, 0 if not a hostel resident), X_6 is Parent involvement (1 if the parent was involved in the learner's case, 0 otherwise), X_7 is Substance (1 if a learner is suspected to be using drugs, 0 otherwise), X_8 is English (marks obtained in English), X_9 is Maths (marks obtained in Mathematics), X_{10} is LO (marks obtained in Life Orientation), X_2X_6 is the interaction of Grade and Parent involvement, X_2X_7 is the interaction of Grade and Suspected Substance abuse, and ϵ is the error term.

The log-rank test, Wilcoxon test, and Peto-Peto test were applied to learner offence data to compare various groups of learners based on categorical variables.

3.4.2 Modelling the Number of Repeat Offences

An attempt was made to model the number of repeat offences using the Poisson Model, Negative Binomial Model and the Zero Inflated Models. The actual model for investigating the effects of variables on the number of repeat offences committed by learners is

$$\log(E(y)) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7 + \beta_8X_8 + \beta_9X_9 + \beta_{10}X_{10}$$

where y is the number of repeat offences, X_1 is Gender (1 if male, 0 if female), X_2 is Grade (whether a learner is in grade 8, 9, 10, 11, or 12), X_3 is Repeat (1 if a learner is repeating a grade, 0 if not repeating), X_4 is Home location (1 if learner's home is local, 0 otherwise), X_5 is Hostel residence (1 if learner is a hostel resident, 0 if not a hostel resident), X_6 is Parent involvement (1 if the parent was involved in the learner's case, 0 otherwise), X_7 is Substance (1 if a learner is suspected to be using drugs, 0 otherwise), X_8 is English (marks obtained in

English), X_9 is Maths (marks obtained in Mathematics), X_{10} is LO (marks obtained in Life Orientation).

Since the number of offences are counts of $0, 1, 2, \ldots$, the distribution is assumed to follow a Poisson distribution. In this investigation we might expect an excess of zero counts as we expect the bulk of the learners to be good or to not offend again. If subjects are observed for a shorter period like one month, there would be many zero counts of offences even in a grade which is more prevalent or susceptible to offences like grade 10. In such cases the Zero Inflated Poisson Model could be the alternative model to consider. The null hypothesis (H_0: κ =0) that the data are Poisson was tested against the alternative (H_a: κ > 0) that they are Negative Binomial. The Negative Binomial Model is considered to account for overdispersed observations while the Zero Inflated Poisson Model only accounts for the extra zeros relative to the Poisson Model.

Chapter 4
Analysis

4.0 Introduction

In this chapter the data is summarised and described. The number of offences committed by learners in the study period are also analysed. Survival analysis methods and methods of analysing count data are applied to the time to offence and number of offences respectively. The analysis is done using the statistical software package Stata 14 (2015). Outputs from the model building and analysing processes in this chapter are given in the appendices.

4.1 Descriptive Statistics

To learn a bit more about the data we see in Appendix A that all the variables have 83 observations except Maths with 82 observations, i.e. there is only one missing value. For the variable Cense, 0 indicates a censored observation and 1 indicates a failure event. Of the 11 covariates, 3 are continuous variables and all the other 8 are categorical. Table 4.1 provides descriptive statistics for the continuous variables.

Table 4.1 Summary Statistics for Continuous variables

Variable	Obs	Mean	Std. Dev.	Min	Max
Time	83	41.46988	28.76242	1	98
English	83	38.81928	14.23509	5	76
Maths	82	19.7561	10.14166	2	43
LO	83	40.39759	17.45792	3	81

The mean and standard deviation of categorical variables as are meaningless since these variables indicate the presence or absence of an attribute, not the amount. The frequency tables and plots will be used to analyse the categorical and ordinal variables. On average we see

a worst performance in Mathematics for learners who committed an offence. No one obtained a mark above 43%. Their overall performance in English and Life Orientation (LO) indicates a moderate achievement. As for LO the moderate achievement is not satisfactory because it is comparatively considered an easy subject for learners to pass because most of the assessment tasks are based on physical activities. So these learners might have not been serious when they were given the tasks for assessment as we see the minimum mark of 3% which is highly unlikely for LO. The response variable has the mean survival time of 41.46988 with minimum and maximum values of 1 and 98, respectively. The mean survival time is the total failure time divide by the number of observations. Care should be taken when interpreting the mean survival time which can be underestimated since some observations are censored.

Appendix B.1 provides more details about the response variable. The median survival time (40) is more sensible because survival data is often very positively skewed. It indicates that half of the offenders survive up to 40 weeks before committing the second offence. The four observations with the longest survival times are between 92 and 98 weeks. More of the overall information about data is provided in Table 4.2 produced by Stata command "stdes" and the corresponding output is provided in Appendix B.2.

Table 4.2 Summary of time (in weeks) to second offence

Category	Total	Mean	Min	Median
No. of subjects	83			
No. of records	83	1	1	1
(First) Entry time		0	0	0
(Final) Exit time		41.46988	1	40
Subjects with gap	0			
Time at risk	3442	41.46988	1	40
Failures	30	0.3614458	0	0

In Table 4.2, we see that the number of subjects and number of records are both equal to 83. This is because there is only one observation per subject. In cases where there are more than one observation per subject, the number of records will be more than the number of subjects. Zero or no delayed entry is reported because everyone entered the study at time 0. The

average exit time was 41.46988 which is equal to the average survival time. Altogether 30 failures (committed second offence) were reported.

The average hazard rate is

$$\bar{h} = \frac{number\ of\ failures}{total\ time\ at\ risk}$$

$$= \frac{30}{433} = 0.069284.$$

The total time at risk in the denominator is not necessarily the number of weeks in five years. Although the learners are expected to stay at school for five years, they did not enter the study at the same time. The learner was not observed until he/she committed the first offence. So the total time at risk used in the denominator is the total of the survival times for those 30 learners who committed the second offence, that is who reached the failure event. The reciprocal $(\frac{433}{30} = 14.4)$ of the hazard rate, indicates that we would expect to wait for 14.4 weeks for the failure event (second offence), if the hazard rate stayed constant. If the expected time between failures has constant hazard rate, then the number of learner offences that occur in a given interval is a Poisson random variable (that is failures are repeatable by subjects in this study since some learners have committed offences many times during the period of observation).

As indicated earlier in Chapter 2, parent involvement and suspected substance abuse are considered factors of some higher level of concern. A quick summary of the frequencies of parent involvement and suspected substance abuse is given in Table 4.3 and 4.4, respectively.

Table 4.3 Frequency of parent involvement

Parentinvo	**Freq.**	**Percent**	**Cum.**
No	63	75.90	75.90
Yes	20	24.10	100.00
Total	83	100.00	

In Table 4.3 we see that roughly 76% of the learners are those whose parents were not involved in the handling of offence cases, whereas only 24% are those whose parents were involved. In Table 4.4 we see that about 55% of the learners are suspected of substance abuse and those who are not suspected account for 45% of learner offence cases.

Table 4.4 Frequency of suspected substance abuse

Suspected substance abuse	**Freq.**	**Percent**	**Cum.**
No	37	44.58	44.58
Yes	46	55.42	100.00
Total	83	100.00	

Figure 4.1 gives proportions of learner grades who committed an offence.

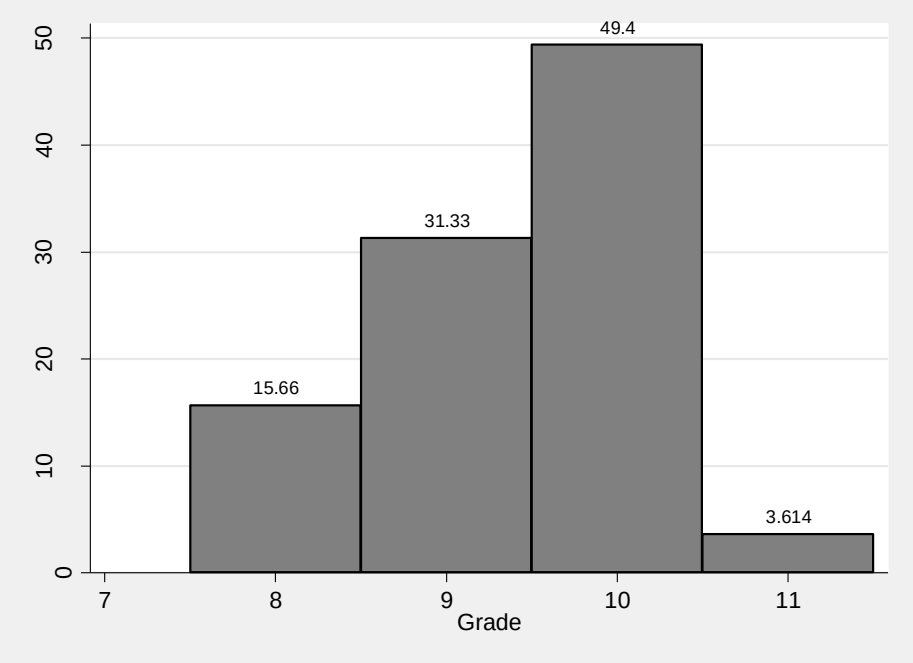


Figure 4.1 Bar graph for learner grade

We see that roughly 16% of the learners who committed an offence are in grade 8. Grade 9 and 10 account for higher proportions of learner offence cases of 31% and 49.4%, respectively. Grade 11 accounts for only 3.6% of learner offences. It seems that grade 11 learners are more disciplined as compared to learners in other grades involved in this study.

Now looking again at the response variable, the histogram in Figure 4.2 displays the frequencies of survival times.

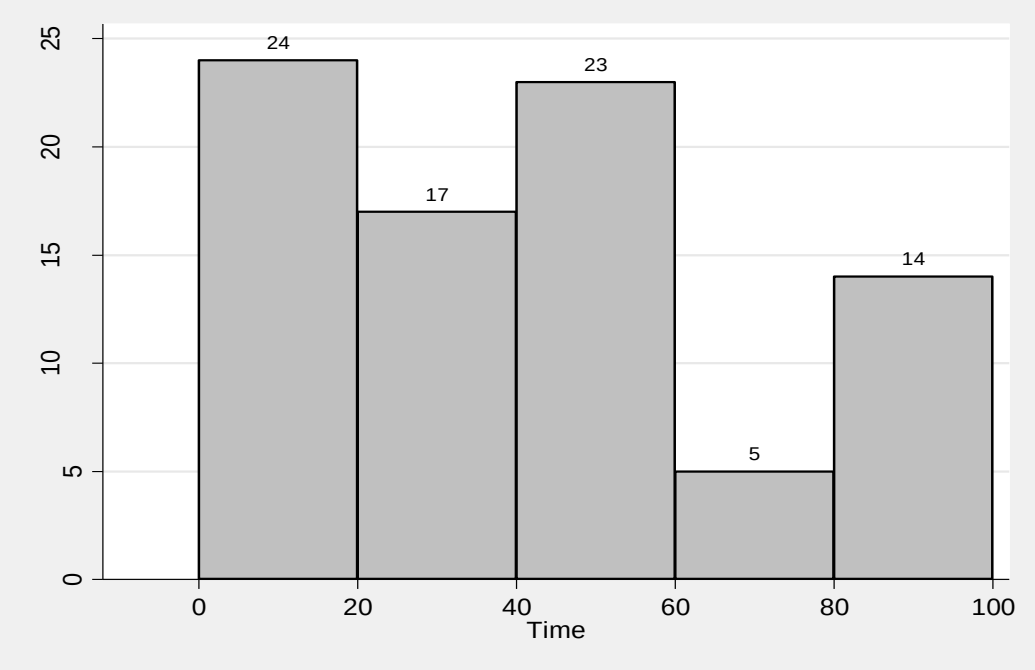
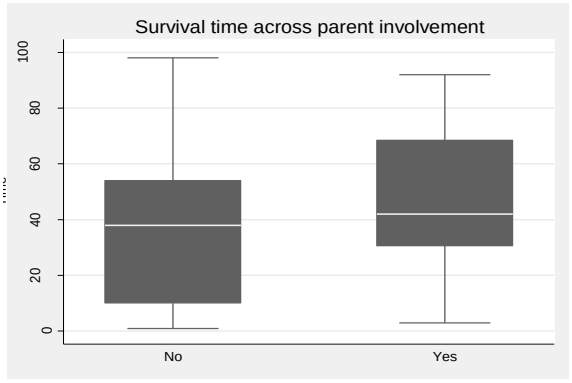
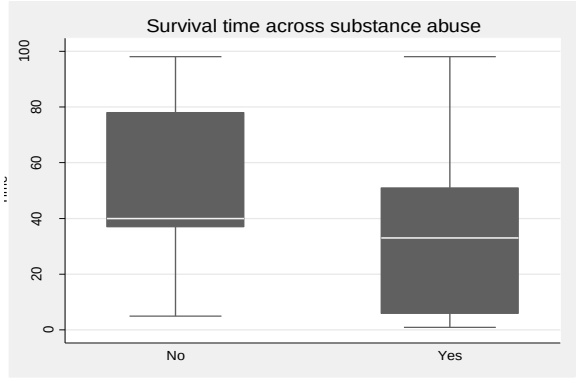


Figure 4.2 Histogram showing frequency of survival times (in weeks)

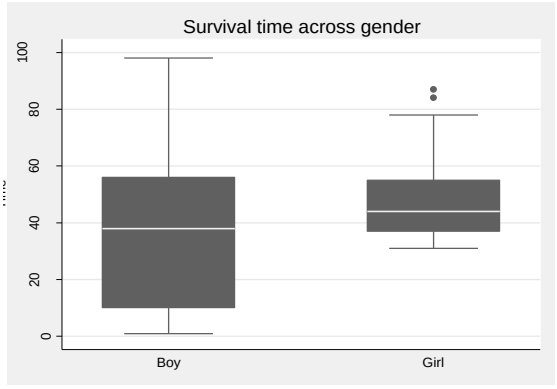
Figure 4.2 is displayed to indicate that the overall distribution of survival times to commit a second offence cannot be explained by the histogram since the survival times include censored observations and the subjects (learners) entered the study at different times. We have already observed in Table 4.2 that 30 learners committed a second offence. The boxplots in Figure 4.3(a) to (c) are used to observe the survival time differences across groups by parental involvement, suspected substance abuse, gender, and can further be observed for other categorical variables.



(a)



(b)



(c)

Figure 4.3: Survival times by (a) Parental involvement (b) Substance abuse (c) Gender

In Figure 4.3(a) we see that there are some differences in survival time across the two groups concerning parental involvement. The major difference is that the group of learners whose parents were involved in learner offences generally have longer survival time than the group without parent involvement. This means that parental involvement reduces the likelihood of a second offence. We also see in Figure 4.3(b) that the group of learners suspected of substance abuse have shorter survival time than the group not suspected of substance abuse. This means that substance abuse is more likely to lead to committing an offence. In Figure 4.3(c) we see the gender differences of survival time that the boys have substantially more variability in terms of the time to second offence than the girls. Furthermore, the boxplot is showing that there are

boys with survival time less than 30 weeks, whereas the lowest survival time for girls is above 30 weeks. It means the girls are less likely to commit a second offence.

The boxplot in Figure 4.4 provides information about the grade differences of survival time to second offence across two groups concerning parent involvement and the boxplot in Figure 4.5 provides information about the grade differences of survival time to second offence across two groups concerning suspected substance abuse.

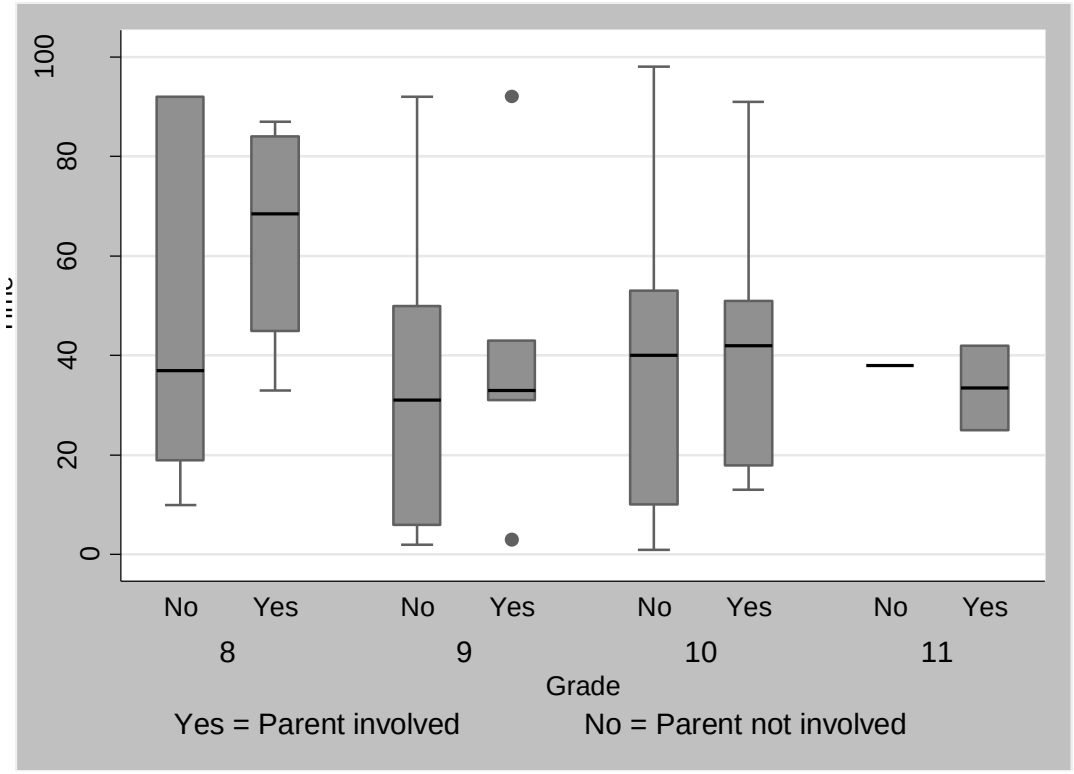


Figure 4.4 Survival time across Parental Involvement within Learner Grade

From the boxplots in Figure 4.4 we see that there are major differences in survival time to second offence between grade 8 versus grade 9 and 10. The wider box and whisker plot for grade 8 shows substantially more variability in time to second offence as compared to other grades. We can see that there are very few offenders in grade 11 for uninvolved parents.

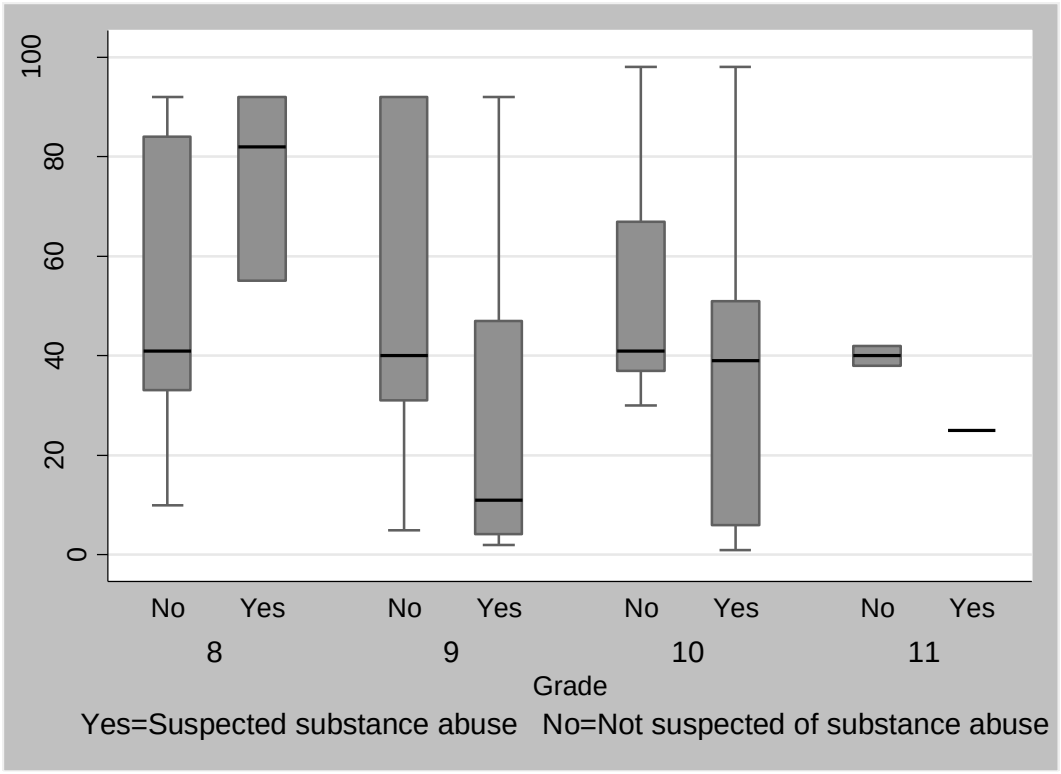


Figure 4.5 Survival time across Suspected Substance Abuse within Learner Grade

Looking at the group without suspected substance abuse in Figure 4.5, there are no major differences in the median survival times between the learner grades. In the group with suspected substance abuse we see substantial differences between grade 8 and other grades. The boxplot is showing shorter survival times for grade 9 and 10 due to suspected substance abuse. This means that suspected substance abuse seems to increase the likelihood of committing a second offence.

The boxplots provided important basic information about the learner differences in survival time to second offence according to gender, parental involvement, and suspected substance abuse. We have seen that boys are more likely to commit a second offence than girls. Parental involvement seems to reduce the likelihood of the learner to commit a second offence. Lastly, the boxplots suggested that the learners who are suspected of substance abuse are more likely to commit a second offence. To learn more about the learner differences in survival time, statistical tests and analyses are performed and discussed in the following sections.

4.2 Nonparametric Analysis

4.2.1 Applying Kaplan-Meier Estimate of the Survivor Function

The nonparametric estimate of the survivor function $S(t)$ gives the probability of survival past time t. For the dataset in this study, we have the observed failure times $t_1, \ldots, t_{30}$ since we have 30 failure times as seen in Table 4.2. To estimate $S(t)$ at any time *t,* we apply equation 1 given in Chapter 2, i.e. Kaplan-Meier estimate of $S(t)$. Consider the subset from our dataset summarised in Table 4.5 showing the survival times from 1 to 5 weeks.

Table 4.5 Subset of learner offence data for 5 weeks

Time (weeks)	No. at risk	No. failed	No. censored
1	83	2	0
2	81	2	0
3	79	3	0
4	76	4	0
5	72	1	0

The earliest time in our data is 1 week where all the 83 subjects (learners) were at risk of failure (committing offence again). At that time (week) *t* = 1, two subjects failed, i.e. two learners committed offences again. At the next time (week) *t* = 2, 81 learners were at risk since 2 already failed. At time (week) 2, two other learners committed offences again. At time (week) *t* = 3, 79 learners were at risk and 3 failed. At time (week) *t* = 4, 76 learners were at risk and 4 failed, leaving 72 at risk at time (week) *t* = 5. One learner failed at time (week) *t* = 5. The probability of survival beyond a particular time given in Table 4.5 is calculated using the Kaplan-Meier estimate

$$\widehat{S(t)} = \prod_{j \mid t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right)$$

To calculate the probability of survival beyond time t = 1 week, we notice that n_1 = 83 and d_1=2. So the estimate $\widehat{S(t)}$ is $\frac{81}{83}$. The probability of surviving beyond time t = 2 weeks, given survival right up to t = 2, is $\frac{79}{81}$. Therefore the unconditional probability of surviving beyond time t = 2 weeks is $\left(\frac{81}{83}\right)\left(\frac{79}{81}\right) = \left(\frac{79}{83}\right)$. The probability of surviving beyond time t = 3 weeks, given survival right up to t = 3, is $\left(\frac{76}{79}\right)$. The unconditional probability of surviving beyond time t = 3 weeks is $\left(\frac{79}{83}\right)\left(\frac{76}{79}\right) = \left(\frac{76}{83}\right)$. The probability of surviving beyond time t = 4 weeks, given survival right up to t = 4, is $\left(\frac{72}{76}\right)$ and the unconditional probability of surviving beyond t = 4 weeks is $\left(\frac{76}{83}\right)\left(\frac{72}{76}\right) = \left(\frac{72}{83}\right)$. Lastly, the probability of surviving beyond time t = 5 weeks, given survival right up to t = 5, is $\left(\frac{71}{72}\right)$ whereas the unconditional probability of surviving beyond this time is $\left(\frac{72}{83}\right)\left(\frac{71}{72}\right) = \left(\frac{71}{83}\right)$. These probabilities are summarised in Table 4.6 with conditional probabilities in the column denoted by p and unconditional probabilities in the column under $\widehat{S(t)}$. The estimate of the survivor function will be calculated using Stata 14 (2015) for the rest of the times. The manual calculations of the Kaplan-Meier estimate based on the subset in Table 4.6 are used to confirm that the results obtained are similar to those produced by Stata in Appendix C.

Table 4.6 Subset of learner offence data with conditional probabilities (p) and unconditional probabilities $\widehat{S(t)}$

At time	No. at risk	No. failed	No. censored	p	$\widehat{S(t)}$
1	83	2	0	$\frac{81}{83}$	$\frac{81}{83}$
2	81	2	0	$\frac{79}{81}$	$\frac{79}{83}$
3	79	3	0	$\frac{76}{79}$	$\frac{76}{83}$
4	76	4	0	$\frac{72}{76}$	$\frac{72}{83}$
5	72	1	0	$\frac{71}{72}$	$\frac{71}{83}$

The overall estimated survival function for the learner offence data is displayed graphically in Figure 4.6.

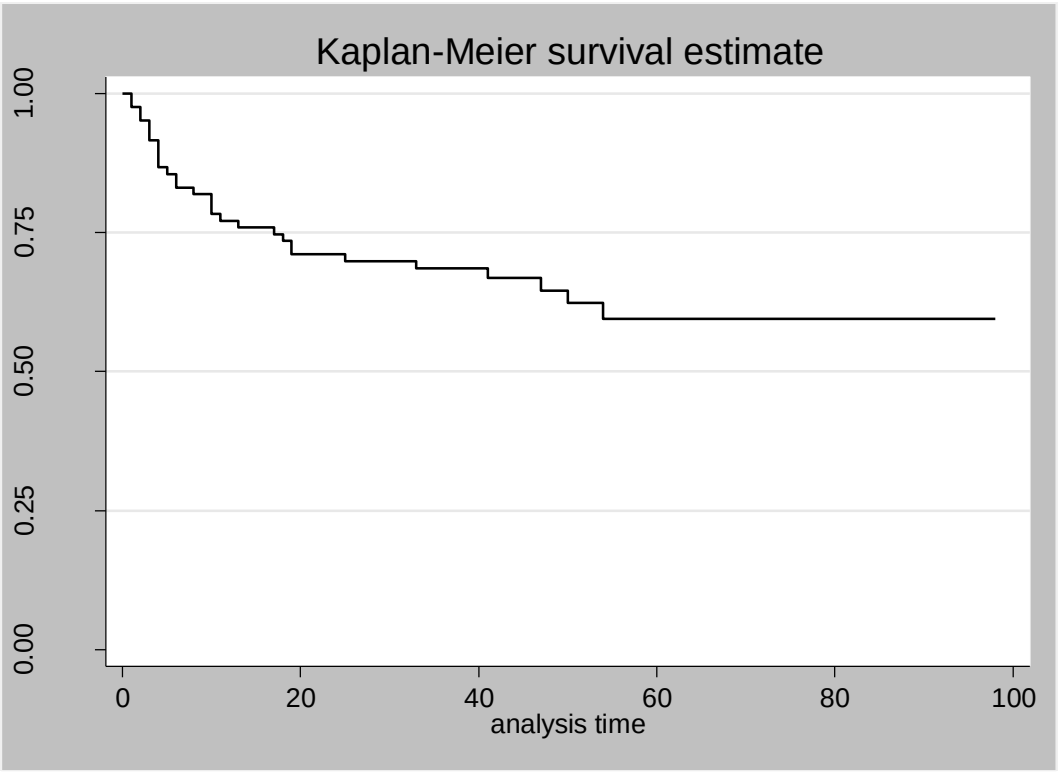


Figure 4.6: Kaplan –Meier survival estimate of learner offence

The graph indicates that learners in general had a better survival experience especially after 18 weeks. That means surviving for 18 weeks before a second offence, the learner is less likely to commit a second offence. The graph also suggests that in general there is no learner at risk of committing a second offence beyond 50 weeks. It was indicated in Chapter 2 that the Kaplan-Meier estimator has the shortcoming that once it is zero, it remains at zero. The Nelson-Aalen estimator which does not suffer from this shortcoming will be used to estimate the cumulative hazard $H(t)$ since the hazard function and survivor function are related by the equation

$$H(t) = -\ln\{S(t)\}.$$

The Nelson-Aalen estimate of the hazard is represented in Figure 4.7.

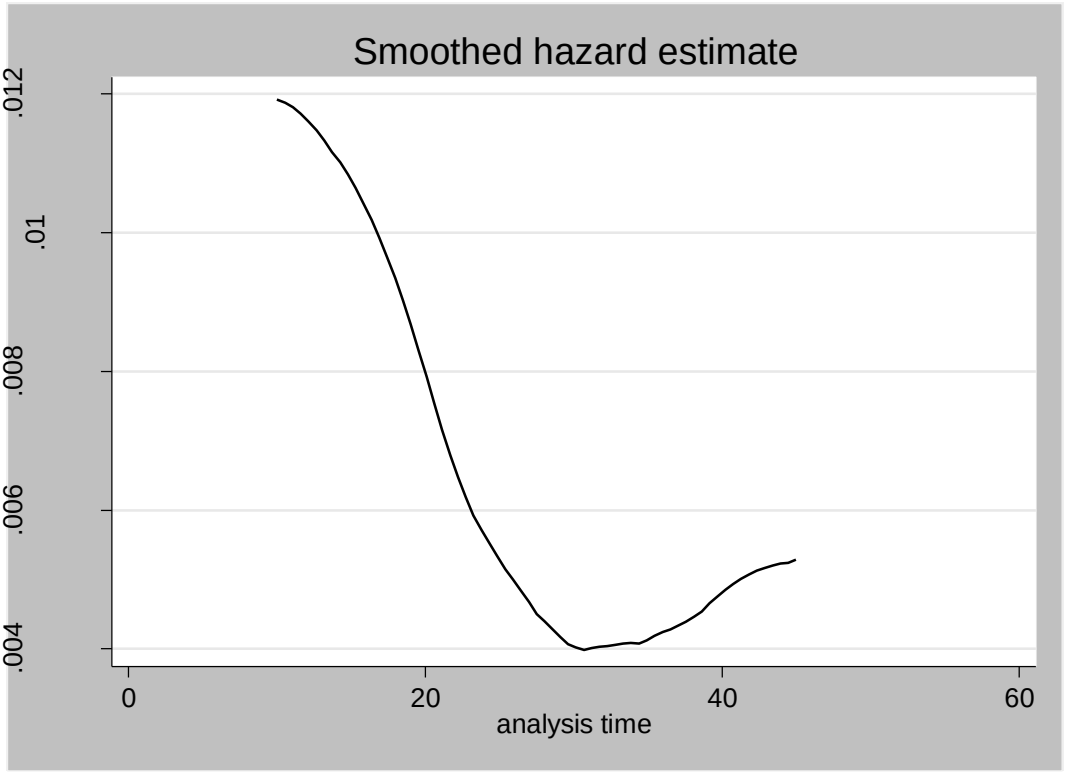


Figure 4.7: Nelson-Aalen estimate of the hazard

It appears in Figure 4.7 that the hazard decreases between 10 and 30 weeks. It means that the chances of learners to commit a second offence decrease between the times 10 and 30 weeks. After 30 weeks the hazard increases meaning that during this time we need to worry because it is likely for learners to commit the second offence. These findings are related to the results of Nelson-Aalen estimate of the cumulative hazard in Figure 4.8.

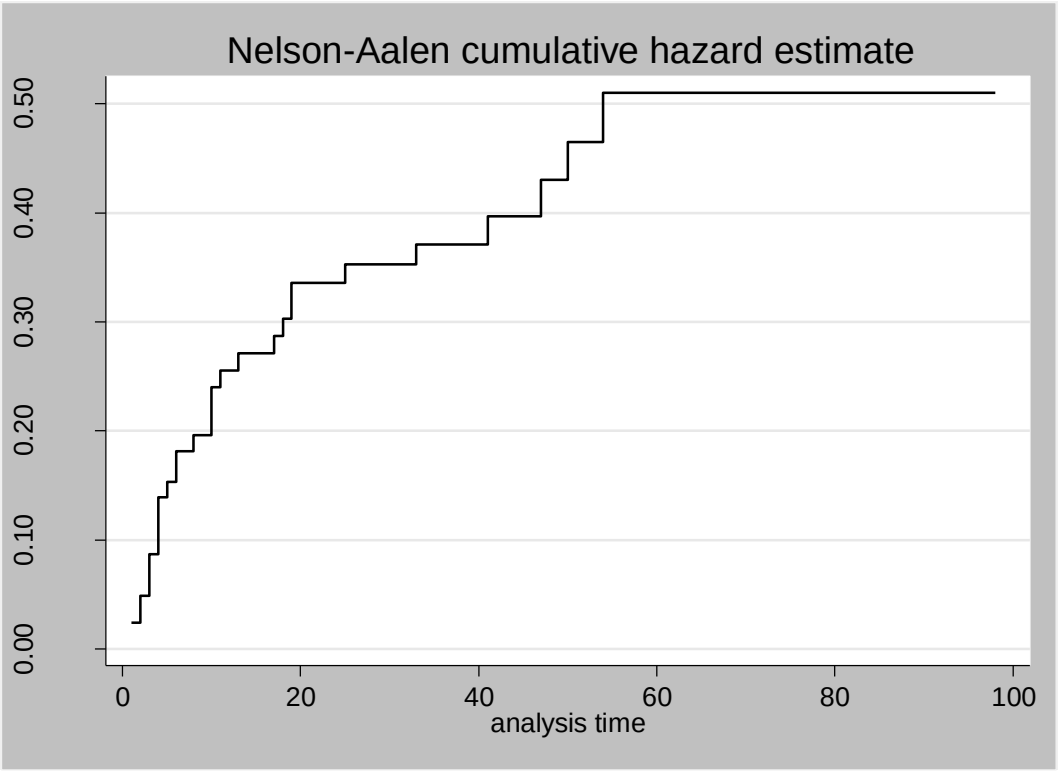


Figure 4.8 Nelson-Aalen estimate of the cumulative hazard

In Figure 4.8 we see that the Nelson-Aalen estimator increases with committing of a second offence by the learner. It appears that the cumulative hazard for learners to commit a second offence increases at a decreasing rate up until about 30 weeks. It means we have similar results as in Figure 4.7 that before 30 weeks the hazard itself is decreasing because the hazard is the derivative (the rate of change) of the cumulative hazard. After 30 weeks the cumulative hazard increases at an increasing rate meaning that the hazard itself is rapid. The cumulative hazard becomes constant at about 55 weeks which means there is no more hazard (that is the derivative of a constant is zero). When the cumulative hazard is constant it means that we don't have to worry because it is unlikely for the learners to commit a second offence.

4.2.2 Comparisons between groups according to Treatment versus Control

To understand the likelihood of survival, further comparisons of the estimated survival function can be made based on grouping data according to an important aspect such as parental involvement. The results of parental involvement in handling of learner offence are displayed in Figure 4.9 and Table 4.7.

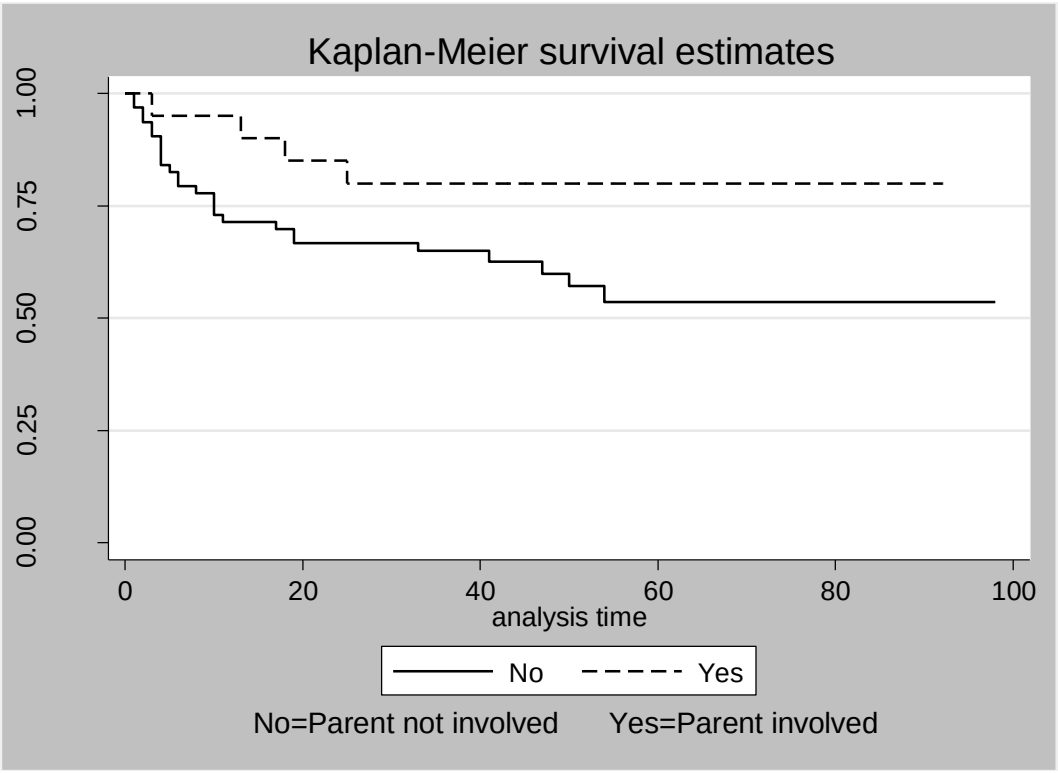


Figure 4.9: Kaplan-Meier survival estimates: Parent involved vs Parent not involved

The group of learners whose parents were involved in handling of the learner offence cases is regarded as treatment group and the one whose parents were not involved is the control group. From the graph in Figure 4.9 we see the dotted line representing the treatment group and the control group represented by solid line. The graph shows that the learners whose parents were involved in handling of offence cases had a better survival experience to second offence than the group of learners without parental involvement. This is also seen in Table 4.7 where a further comparison of the estimated survivor functions of the two groups is done.

Table 4.7 Survival probability

	Parent involved	
Time	No	Yes
1	0.9683	1
13	0.7143	0.9
25	0.6667	0.8
37	0.65	0.8
49	0.5987	0.8
61	0.5358	0.8
73	0.5358	0.8
85	0.5358	0.8
97	0.5358	.
109	.	.

We see from Table 4.7 that at each survival time the value in the column for the group where parents do not participate is less than the one for the group where parents participate. We further observe that from the group of learners having parents involved in handling of offence cases, there are no learners who committed a second offence beyond 85 weeks whereas the group of learners whose parents were not involved in handling of offence cases, had offences committed until 97 weeks. However, we must also take note that learners did not enter the study at the same time. The learners were observed only after committing the first offence, so there is a number of learners who entered the study a few weeks towards the end of observation period. That might have led to having more censored observations than failure events. In this study, censored observations are those learners who did not commit a second offence. We have 53 censored observations and 30 failure observations. Censored observations could further be analysed if there are learners who reached the failure event (committed second offence) after the study had ended, where more records would be required to pursue such analyses.

It is of interest to study learner offences based on suspected substance abuse. The results are shown in Table 4.8 and Figure 4.10.

Table 4.8 Survival Probability

	Suspected substance abuse	
Time	No	Yes
1	1	0.9565
13	0.9459	0.6087
25	0.9189	0.5217
37	0.9189	0.5
49	0.9189	0.4482
61	0.9189	0.3782
73	0.9189	0.3782
85	0.9189	0.3782
97	0.9189	0.3782
109	.	.

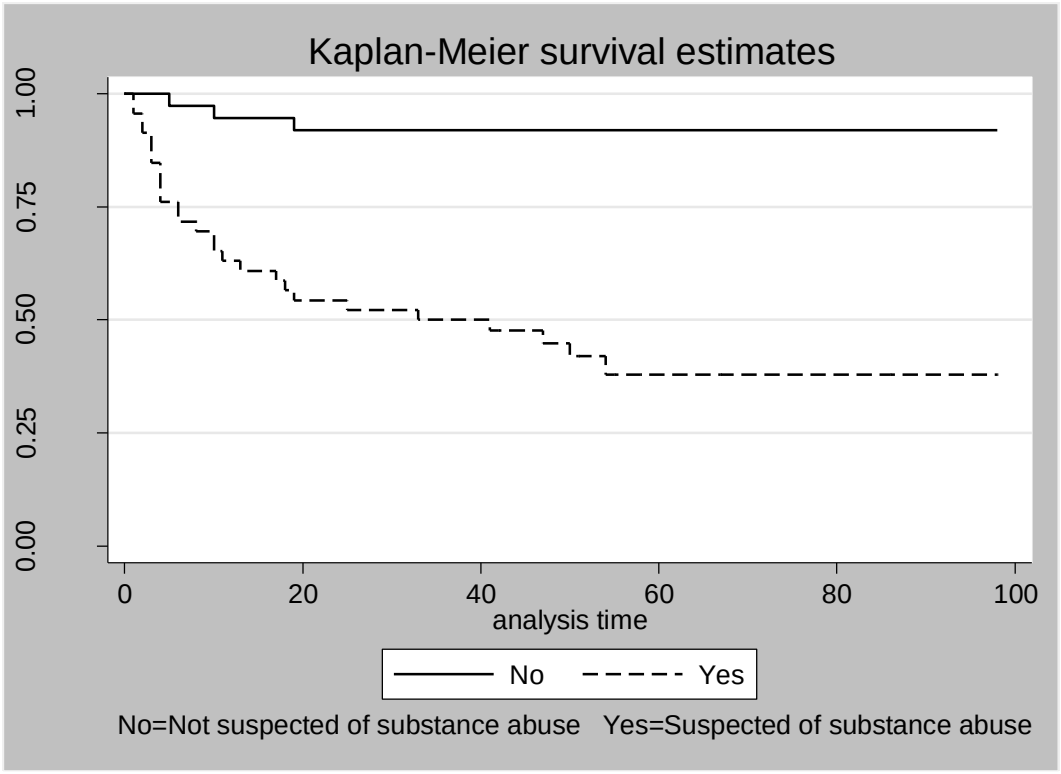


Figure 4.10: Survival estimates: Suspected substance abuse vs not suspected

Although substance abuse is analogous to a treatment as discussed in the previous paragraph, its outcome for interpretation should not be regarded as a treatment. This is because treatment usually refers to what was done or given to subjects under study in order to address the problem, and thus, expecting positive results, whereas substance abuse is related to worsening the problem and thus, expecting negative or undesired results. Similar analyses and interpretations arising from treatment and substance abuse are found in studies related to medical purposes. In Table 4.8 and Figure 4.10 we compare the group of learners who are suspected of substance abuse with the group not suspected of substance abuse. The dotted line in Figure 4.10 represents the group with suspected substance abuse and the solid line represents the group not being suspected. The two survival curves are too apart from each other, having the survival curve for learners not suspected of substance abuse above the survival curve for learners suspected of substance abuse. Since Table 4.8 provides the analysis time at intervals of 12, the critical times are clear in Figure 4.10. From Figure 4.10 we observe the critical times at week 20 and week 54. Learners not suspected of substance abuse do not commit a second offence once they pass week 20. The learners who are suspected of substance abuse are at risk of committing a second offence until after week 54. Thus, as expected the learners not suspected of substance abuse have far much better survival experience than those suspected of substance abuse.

4.2.3 Grouping of subjects based on other Categorical Variables.

The survival time of learner offences according to gender is displayed graphically in Figure 4.11.

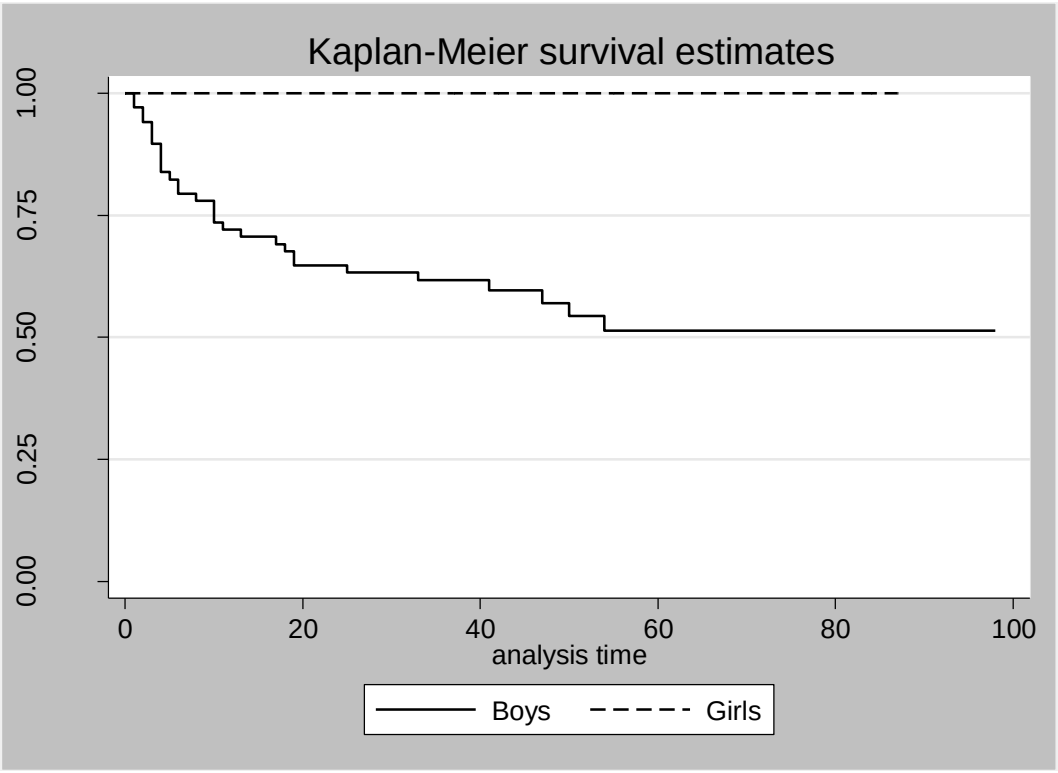


Figure 4.11: Survival estimates for Gender

The graph is showing a better survival experience for girls than for boys. Actually no girl learners experienced failure event (offended again) as seen from the dotted line. The boys are at risk of committing the second offence until after 54 weeks.

The graph in Figure 4.12 compares two groups of learners according to whether they were repeating or not repeating a grade.

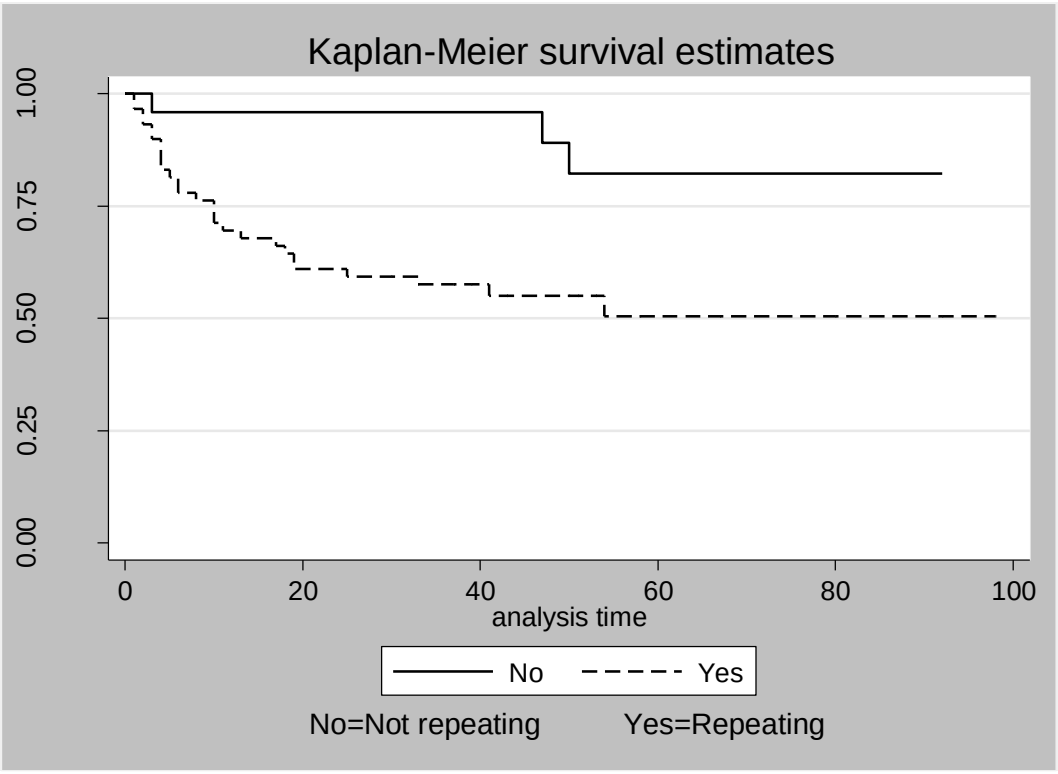


Figure 4.12: Survival curves: repeaters vs non-repeaters

It indicates that there was a lower survival experience for repeaters than for non-repeaters. After 50 weeks non-repeaters are no more at risk while repeaters are still at risk of committing a second offence. Repeaters are no more at risk of committing a second offence beyond 54 weeks of survival time.

In Figure 4.13 we compare the two groups of learners according to whether their home location is local to the school area or not, as described in Chapter 3.

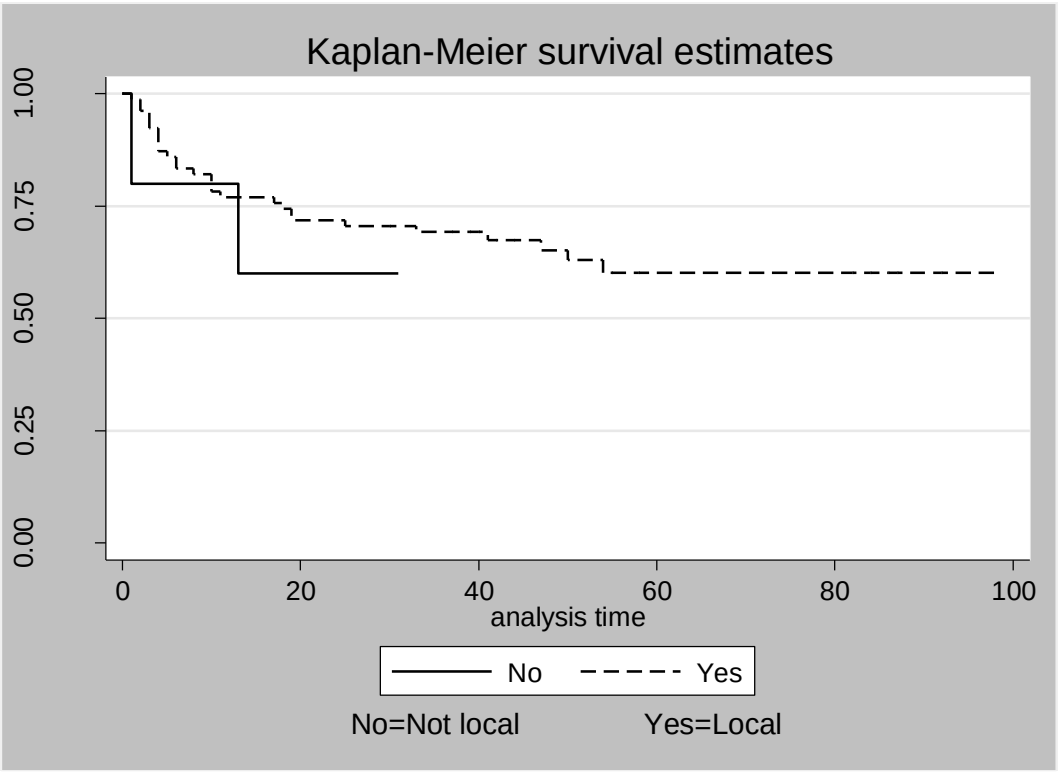


Figure 4.13: Survival curves: Local home location vs non-local home location

We observe a small gap between the two graphs. The graph for local learners is slightly above the one for non-local learners. That means the survival experience to second offence by local learners does not differ much from the survival experience by learners not coming from local home location. However after 12 to 15 weeks, the non-local learners are no more at risk of committing the second offence whereas the local learners are still at risk of committing the second offence until after 54 weeks. Most of the learners are local with very few none local learners, making comparison difficult. The issue about where the learners live in relation to offences, is further investigated by comparing learners residing at a hostel with those not residing at a hostel. The results are displayed graphically in Figure 4.14.

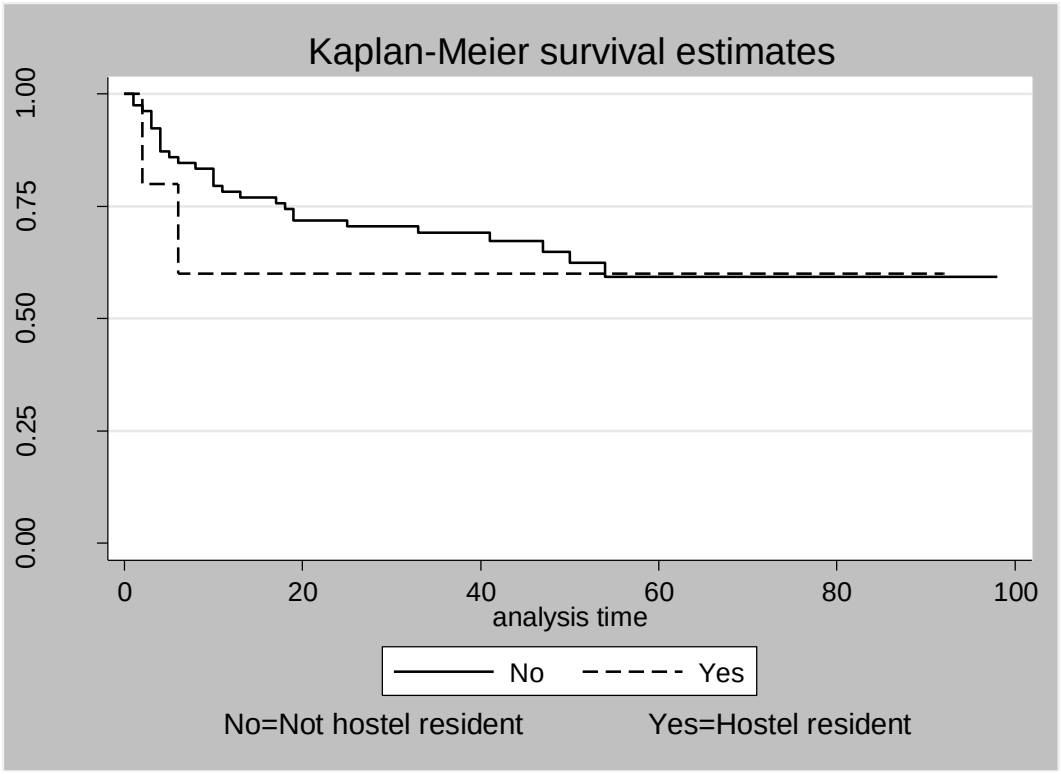


Figure 4.14: Survival curves: Hostel resident vs non-hostel resident

In Figure 4.14 we observe that before week 54, learners residing at the hostel had survival experience to second offence slightly lower than the learners not residing at the hostel. After 54 weeks the learners had more or less the same survival experience to second offence irrespective of whether they live at the hostel or not. It seems that the survivor probabilities for learners to commit a second offence do not differ according to learners' place of residence after surviving for 54 weeks.

In Figure 4.15 the survival curves are compared according to learner grades.

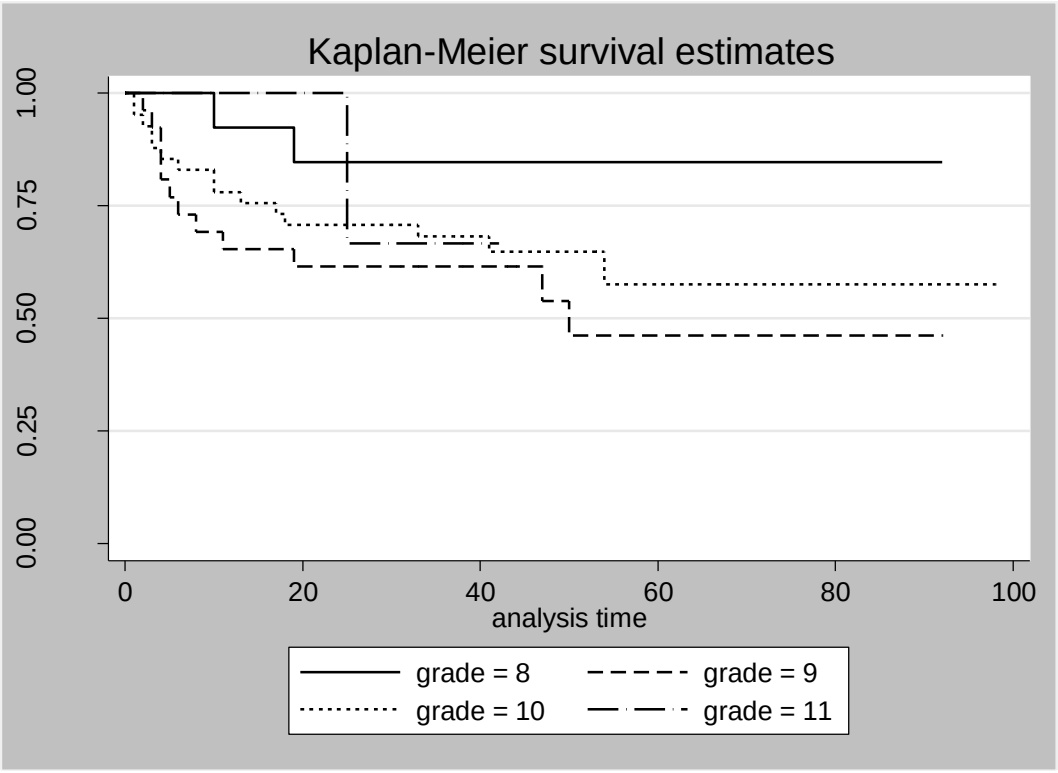


Figure 4.15: Survival curves for learner grades

There were four groups of learner grades participated in this study, namely grade 8, 9, 10, and 11. We observe that grade 8 had a better survival experience as compared to other grades. After 20 weeks grade 8 learners are no more at risk of committing a second offence. Grade 9 had the worst survival experience amongst all the grades. The major difference is between grade 8 versus grade 9 and 10. Grade 11 cannot be reasonably compared with other grades since it had only three observations (learners), one reached a failure event (offended again) and two were censored.

4.2.4 Tests of Hypothesis

In Table 4.9 we see the Log-rank test, Wilcoxon test and Peto-Peto test for equality of survivor functions between two groups of learners: parent involved and parent not involved. The tests test the null hypothesis that there is no difference in the survivor functions between the two groups of learners.

Table 4.9: Parent Involvement: Tests for equality of survivor functions

	Events observed		Events expected		χ^2-value	P-value
	Parent not involved	Parent involved	Parent not involved	Parent involved		
Log-rank	26	4	22.03	7.97	2.74	0.0978
Wilcoxon	26	4	22.03	7.97	2.53	0.1119
Peto-Peto	26	4	22.03	7.97	2.85	0.0915

The p-value of 0.0978 from the log-rank test implies that the null hypothesis is not rejected at 1% and 5% levels of significance. We also see that the Wilcoxon test clearly fails to reject the null hypothesis at 10%, 5%, and 1% levels of significance. Peto-Peto test also agrees with other tests for not rejecting the null hypothesis at 5% and 1% levels of significance. We conclude that the survivor probabilities of the two groups are equal meaning that according to these tests, the survivor probabilities to commit a second offence by the leaners whose parents were involved in handling of offence cases are the same as for those whose parents were not involved. That is parental involvement has no significant effect on survival time of learners to commit a second offence. Care should be taken when making conclusions from these tests since they are based only on 30 failure observations. The Kaplan-Meier estimate of the survivor function that suggested some difference between the two groups, analyses the survival time for all 83 observations (learners) including those censored.

The log-rank test, Wilcoxon test, and Peto-Peto test were also used to test the equality of survivor functions between the group of learners suspected of substance abuse and those not suspected of substance abuse. The results are given in Table 4.10.

Table 4.10: Suspected substance abuse: Tests for equality of survivor functions

	Events observed		Events expected		χ^2-value	P-value
	Suspected substance abuse		Suspected substance abuse			
	No	Yes	No	Yes		
Log-rank	3	27	15.17	14.83	20.19	0.0000*
Wilcoxon	3	27	15.17	14.83	18.72	0.0000*
Peto-Peto	3	27	15.17	14.83	19.85	0.0000*

We see that all the three tests clearly reject the null hypothesis at the 10%, 5%, and 1% levels of significance in favour of the alternative hypothesis that the survivor probabilities of the two groups are not the same. That means suspected substance abuse does have an effect on the survival time to a second offence experienced by learners. The tests agree with the findings from the Kaplan-Meier survival plots where the survival experience to second offence by learners who are suspected of substance abuse is less than that of learners who are not suspected.

The tests for equality of survivor functions of boy and girl learners were also performed and the results are shown in Table 4.11.

Table 4.11: Gender: Tests for equality of survivor functions

	Events observed		Events expected		χ^2-value	P-value
	Girls	Boys	Girls	Boys		
Log-rank	0	30	6.45	23.55	8.36	0.0038*
Wilcoxon	0	30	6.45	23.55	7.84	0.0051*
Peto-Peto	0	30	6.45	23.55	8.18	0.0042*

We observe that there were no girls who had a failure event (offended again). In all the three tests we have a p-value less than 5% and 1% level of significance. The tests reject the null hypothesis that the survivor functions of boy and girl learners are the same. That means gender has an effect on the survival time of learners to commit a second offence. We have already seen from the Kaplan-Meier plots that girls had a better survival experience to second offence than boys.

In Table 4.12 the Log-rank test, Wilcoxon test, and Peto-Peto test were used to test the equality of survivor functions of learners who live close to school and those who live more than 3km away from school.

Table 4.12: Home location: Tests for equality of survivor functions

	Events observed		Events expected		χ^2-value	P-value
	Local	Not local	Local	Not local		
Log-rank	28	2	28.65	1.35	0.33	0.5643
Wilcoxon	28	2	28.65	1.35	0.38	0.5357
Peto-Peto	28	2	28.65	1.35	0.41	0.5224

We see greater p-values in all the three tests. The null hypothesis that the survivor functions between these two groups of learners are the same, is not rejected at 1%, 5%, and 10% levels. It means that proximity to school does not have significant effect on survival time of learners to commit a second offence. The Kaplan-Meier plots also suggested no major differences in survival probabilities between local and non-local learners. The issue of learner's home place is further looked at by grouping learners according to whether they are hostel residents or not. The results of tests for equality of survivor functions of hostel residents and non-hostel residents are given in Table 4.13.

Table 4.13: Hostel residence: Tests for equality of survivor functions

	Events observed		Events expected		χ^2-value	P-value
	Hostel resident	Not hostel resident	Hostel resident	Not hostel resident		
Log-rank	2	28	1.62	28.38	0.09	0.7584
Wilcoxon	2	28	1.62	28.38	0.34	0.5571
Peto-Peto	2	28	1.62	28.38	0.24	0.6267

The p-values 0.7584, 0.5571, and 0.6267 clearly indicate that the null hypothesis that the survivor functions of the two groups are equal cannot be rejected at 10%, 5%, and 1% levels of significance. That is, hostel residence does not have a significant effect on survival time of learners to commit a second offence. It was seen from the Kaplan-Meier plots that the survivor probabilities of learners residing at the hostel differ slightly from those who are not hostel residents. We can conclude that the survival time of learners to commit a second offence does not differ according to learner's place of residence.

The learners were grouped according to whether they were repeating a grade or not and the tests were performed to produce the results in Table 4.14.

Table 4.14: Repeating: Tests for equality of survivor functions

	Events observed		Events expected		χ^2-value	P-value
	Repeating	Not repeating	Repeating	Not repeating		
Log-rank	27	3	19.65	10.35	8.17	0.0043*
Wilcoxon	27	3	19.65	10.35	9.13	0.0025*
Peto-Peto	27	3	19.65	10.35	8.38	0.0038*

The tests in Table 4.14 test the null hypothesis that the survivor functions of repeaters and non-repeaters are equal. We can see from the p-values 0.0043, 0.0025, and 0.0038 that the three tests reject the null hypothesis at 5% and also 1% level of significance. That means repeating a grade does have an effect on the survival time of learners to commit a second offence. The Kaplan-Meier plots indicated a lower survival experience for repeaters than for non-repeaters.

With regard to grade, the tests, test the equality of survivor functions among the four groups of learners, namely, grade 8, 9, 10, and 11. By observing the p-values from Table 4.15 we can see that the null hypothesis that the survivor functions of the four grades are the same, is not rejected.

Table 4.15: Grade: Tests for equality of survivor functions

	Events observed				Events expected				χ^2-value	P-value
	Grade				Grade					
	8	9	10	11	8	9	10	11		
Log-rank	2	12	15	1	5.66	8.60	14.62	1.12	3.81	0.2828
Wilcoxon	2	12	15	1	5.66	8.60	14.62	1.12	3.43	0.3305
Peto-Peto	2	12	15	1	5.66	8.60	14.62	1.12	3.94	0.2685

The four grades survival distributions do not differ significantly at the 1%, 5% and 10% levels. That means the learner grade does not have significant effect on survival time of learners to commit a second offence. The Kaplan-Meier plots indicted a major difference in survival probabilities between grade 8 versus grade 9 and 10.

The differences in the tests and Kaplan-Meier plots may be explained by confounding from some of the variables. Perhaps fitting models may explain these differences.

4.3 Semi-Parametric Analysis

4.3.1 Applying the Cox Proportional Hazards Model

In this section the Cox proportional hazards regression model is applied to the learner offence dataset. The model does not assume a parametric form of the survivor function. The assumption is that the model is correctly specified and additional variables will add little or no explanatory power. As indicated in Section 2.2.1 the Cox proportional hazards regression model to be fitted is given by

$$h(t) = h_0(t)\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_2 X_6 + \beta_{12} X_2 X_7),$$

where, $h_0(t)$ is the baseline hazard, X_1 is Gender (1 if male, 0 if female), X_2 is Grade (whether a learner is in grade 8, 9, 10, 11, or 12), X_3 is Repeat (1 if a learner is repeating a grade, 0 if not repeating), X_4 is Home location (1 if learner's home is local, 0 otherwise), X_5 is Hostel residence (1 if learner is a hostel resident, 0 if not a hostel resident), X_6 is Parent involvement (1 if the parent was involved in the learner's case, 0 otherwise), X_7 is Substance (1 if a learner is suspected to be using drugs, 0 otherwise), X_8 is English (marks obtained in English), X_9 is Maths (marks obtained in Mathematics), X_{10} is LO (marks obtained in Life Orientation), X_2X_6 is the interaction of Grade and Parent involvement, X_2X_7 is the interaction of Grade and Suspected Substance abuse . The hazard ratios of the learners do not depend on the choice of $h_0(t)$. The regression coefficients $\beta_1, \beta_2, \ldots, \beta_{12}$ are to be estimated from the data.

4.3.2 Examining the Categorical Variables

The Cox proportional hazards model was fitted using each of the categorical variables independently. The corresponding Stata outputs are given in Appendices D1 to D10. The results relating to parental involvement are provided in Table 4.16.

Table 4.16: Parental Involvement Likelihood Ratio (LR) test

Variable	$\exp(\beta)$	P-value	LR (P-value)
Parent Involvement	0.4245	0.111	0.0778

In Table 4.16 we see that the p-value = 0.1111, which is greater than 0.05 level of significance. The null hypothesis that the hazard ratio is equal to 1 is not rejected at the 5% and also 10% level. The hazard ratio of 0.4245 for the variable (Parent involvement), means that the hazard faced by those learners whose parents were involved is 0.4245 times the hazard of those whose parents were not involved, which is substantially less. We can say that parent involvement is protective since the hazard ratio is less than 1. This means that parent involvement reduces the risk of learner to offend again although it's not statistically significantly. We have seen similar results when testing the equality of survivor functions that parental involvement has no significant effect on survival time of learners to commit a second offence. The LR P-value (0.0778) is greater than 0.05 meaning that parental involvement is not a useful predictor variable.

The results relating to suspected substance abuse are provided in Table 4.17.

Table 4.17: Suspected substance abuse LR test

Variable	$\exp(\beta)$	P-value	LR (P-value)
Suspected substance abuse	9.3207	0.000*	0.0000*

The variable (Suspected substance abuse) has a hazard ratio of 9.3207. This means that substance abuse is not protective to learner offending again, instead it increases hazard since the hazard ratio is greater than 1. The hazard faced by learners who are suspected of substance abuse is 9.3207 times the hazard faced by those who are not suspected of substance abuse. That is moving from a group not suspected of substance abuse to a group of suspected substance abusers, increases hazard by 832%. The lowest p-value (0.000) indicates that the

effect of substance abuse is highly significant. The significant LR P-value (0.0000) implies the usefulness of the presence of this predictor variable.

The effect of learner repeating a grade was tested and the results are given in Table 4.18.

Table 4.18: Repeating LR test

Variable	$\exp(\beta)$	P-value	LR (P-value)
Repeating	4.8006	0.010*	0.0017*

The p-value of 0.010 is less than 0.05. The null hypothesis that the hazard ratio is equal to 1 is rejected at the 5% level of significance. We observe that the hazard ratio is 4.8006. It means that the hazard faced by repeaters is 4.8006 times the hazard of non-repeaters. It means that being a repeater increases hazard by 380%, alternatively being a non-repeater decreases hazard by 20%. This hazard ratio is greater than 1 which implies that repeating a grade is not protective to learner offending again. The LR P-value (0.0017) < 0.05 meaning the presence of this variable is useful in predicting learner offence.

As for the variable (Home location), we observe the p-value of 0.571 and the hazard ratio of 0.6582 in Table 4.19.

Table 4.19: Home location LR test

Variable	$\exp(\beta)$	P-value	LR (P-value)
Home location	0.6582	0.571	0.5924

Appendix D.4 also gives the 95% confidence interval (0.1551; 2.7927) for the hazard ratio which includes one. The null hypothesis that the hazard ratio is equal to 1, is not rejected at the 5% level of significance. This means that the hazard faced by learners who live near the school does not differ significantly from the hazard of those living far from school. The non-significant LR P-value (0.5924) implies that home location is not a useful predictor variable. As mentioned

earlier, the other variable related to learner home is (Hostel residence). The results for testing the effect of Hostel residence are given in Table 4.20.

Table 4.20: Hostel residence LR test

Variable	$\exp(\beta)$	P-value	LR (P-value)
Hostel residence	1.2502	0.761	0.7678

The hazard ratio of 1. 2502 means that the hazard faced by learners residing at the hostel is 1.2502 times the hazard of those not residing at the hostel. That is, residing at the hostel increases the hazard by 25%. The null hypothesis that the hazard ratio is 1, is not rejected at the 5% level since the p-value is 0.761. It means that the hazard faced by learners who are hostel residents does not differ significantly from the hazard of those not residing at the hostel. The LR P-value (0.7678) > 0.05 implies that the presence of the variable hostel residence is not useful in predicting learner offence.

The effect of Gender was also examined with the corresponding results given in Table 4.21.

Table 4.21: Gender LR test

Variable	$\exp(\beta)$	P-value	LR (P-value)
Gender	$2.38e^{+15}$	1.000	0.0001*

In Table 4.21 we observe a very big hazard ratio which indicates that the hazard faced by male learners is extremely much higher than the hazard faced by female learners. In other words the male learners are at a higher risk of offending again than the female learners. Since the p-value is 1.000 the null hypothesis that the hazard ratio is 1, is not rejected at the 10%, 5%, and 1% levels of significance. We have noticed from Table 4.11 that there was no girls who committed a second offence. The tests for the equality of survivor functions rejected the null hypothesis that the survivor functions of boy and girl learners are the same. The difficulty in comparison might be due to the fact that in this study most of the learners are boys with very few girls. The

LR P-value (0.0001) < 0.05 implies that the presence of gender is useful in predicting learner offence.

For the variable (Grade), the results of testing the effect of a particular grade to learner offence are given in Table 4.22.

Table 4.22: Grade LR test

Variable	$\exp(\beta)$	P-value	LR (P-value)	Comment
Grade 8	0.3056	0.106	0.0544	Reduces hazard
Grade 9	1.6621	0.174	0.1829	Increases hazard
Grade 10	1.0518	0.890	0.8900	Increases hazard
Grade 11	0.8877	0.907	0.9053	Reduces hazard

The hazard ratio of 0.3056 for grade 8 implies that being in grade 8 reduces the hazards of offences by 31%. The hazard ratio of 1.6621 implies that being in grade 9 increases the hazards of offences by 66%. The hazard ratio of 1.0518 implies that being in grade 10 increases the hazards of offences by only 5%. The hazard ratio of 0.8877 means that being in grade 11 reduces the hazards of offences by 11%. We notice that none of the p-values for all grades is less than 0.05 which means that the null hypotheses that the hazard ratios are equal to 1, are not rejected at the 5% level of significance. In general grade is not a significant factor to learners to committing a second offence. The non-significant LR P-values imply that the presence of the variable grade is not useful in predicting learner offence.

4.3.3 Examining the Continuous Variables

Table 4.23 gives the results obtained by fitting the Cox proportional hazards model with each of the continuous covariates. The results show the effect of each variable while holding other variables fixed.

Table 4.23: Learner performance LR test

Variable	$\exp(\beta)$	P-value	LR (P-value)
English	0.9613	0.002*	0.0022*
Mathematics	0.9338	0.002*	0.0009*
Life Orientation	0.9674	0.002*	0.0017*

We observe that the hazard ratios for all the three variables are less than 1, which means that exposure to these variables is protective. English has a hazard ratio of 0.9613 implying that an increase by 1 mark obtained in English decreases the hazard by 4%. For Mathematics, the hazard ratio is 0.9338 which means that an increase of 1 mark in Mathematics decreases the hazard by 7%. Similarly, the hazard ratio of 0.9674 for Life Orientation means that an increase by 1 mark in Life Orientation decreases the hazard by 3%. We further observe the small p-values (less than 0.05) in testing for the effect of each of these three variables, meaning that at the 5% level of significance, the chi-square likelihood ratio test rejects the null hypotheses that the learner performance has no effect on survival time to leaner offence. Thus, better scores reduce a learner's chance of a second offence. We observe the significant LR P-values meaning that it is useful to include learner performance in each of the subjects (English, Mathematics, and Life Orientation) as the predictor variable.

4.3.4 Examining the Interaction effect of Variables

As mentioned in Section 3.4 (Model building) we are interested in the interaction effect of Grade and Parental involvement, and also the interaction effect of Grade and Suspected substance abuse. The results of the Cox proportional hazards model that was fitted are given in Table 4.24.

Table 4.24: LR test for Interaction effect

Variables	$\exp(\beta)$	P-value	LR (P-value)
Grade and Parental involvement	2.4481	0.149	0.1202
Grade and Suspected substance abuse	4.8977	0.111	0.0000 *

Since the P-value (0.149) is greater than 0.05 we cannot reject the null hypothesis (at the 5% level) that there is no interaction effect of Grade and Parental involvement. That means there is no significant difference in survival time of learners to offend again due to interaction between Grade and Parental Involvement. By observing the P-value (0.111) we also fail to reject the null hypothesis (at the 5% level) that there is no interaction effect of Grade and Suspected Substance abuse. That is, also the interaction between Grade and Suspected Substance abuse has no significant effect on survival time of learners to commit a second offence. We observe that for the interaction between Grade and Suspected substance abuse the LR P-value (0.0000) is significant, implying that it is useful to include this interaction in the model to predict learner offence.

4.3.5 Examining the Cox Proportional Hazards Model with all the Covariates without Interactions

The Cox proportional hazards model was fitted with all the variables included without interactions in 4.3.4. The results obtained are given in Table 4.25.

Table 4.25: LR test for combined effects without interaction

Variable	Coefficient (β)	P-value
Gender	35.8164	1.000
Grade	-0.6395	0.073
Repeating	0.6946	0.355
Home location	-2.6469	0.009*
Hostel residence	-0.8755	0.247
Parent involvement	-1.2483	0.049*
Suspected substance abuse	2.3692	0.001*
English	0.0554	0.026*
Mathematics	-0.1105	0.006*
Life Orientation	-0.0257	0.130
LR		0.000*

With all the covariates included, the analyst can have an overview of the effects of coefficients combined. The hypotheses tested are stated in terms of the coefficients rather than hazard ratios. With the coefficient vector $\boldsymbol{\beta}_X$, the null hypotheses, H_0 : $\boldsymbol{\beta}_X$ = **0**, state that the coefficients are equal to zero which is the same as saying that the hazard ratios are equal to one. That is $\boldsymbol{\beta}_X$= **0** $\Leftrightarrow \exp(\boldsymbol{\beta}_X)$ = **1** as indicated in Section 2.2.1. From the p-values we observe that the significant variables at 5% level are Home location, Parent Involvement, Suspected substance abuse, English, and Mathematics. Among the significant variables we observe that Home location, Parent involvement, and Mathematics have their coefficients less than 0. This has the same interpretation as hazard ratios less than 1. Thus, exposure to these variables decreases hazards of a second offence. Suspected substance abuse and English have their coefficients greater than zero implying that their hazard ratios are greater than 1. Thus, exposure to these variables increases hazards of second offence. We observe a significant LR P-value (0.000) meaning that there are useful predictor variables included in this model.

4.3.6 Examining the Cox Proportional Hazards Model with all the Covariates and Interactions

The Cox proportional hazards model was fitted again with all the variables including interactions in 4.3.4. The results obtained are given in Table 4.26.

Table 4.26: LR test for combined effects including interaction

Variable	Coefficient (β)	P-value
Gender	38.23602	-
Grade	-1.417874	0.106
Repeating	0.4377734	0.570
Home location	-2.380192	0.020*
Hostel residence	-0.9702021	0.202
Parent involvement	-13.70303	0.108
Suspected substance abuse	-3.329293	0.701
English	0.0565412	0.037*
Mathematics	-0.1163634	0.005*
Life Orientation	-0.0267153	0.126
Grade and Parent Involvement	1.293354	0.132
Grade and Suspected substance abuse	0.6807169	0.497
LR		0.000*

From the p-values we observe that the significant variables at 5% level are Home location, English, and Mathematics. Unlike in the model without interactions in the previous section, Parental involvement and suspected substance abuse are not significant at 5% level. They are not significant even at 10% level. For Home location, English, and Mathematics, we observe similar results from the previous model. Home location and Mathematics have their coefficients less than 0 meaning that exposure to these variables decreases hazards of a second offence. English has a coefficient greater than 0 meaning that exposure to this variable increases hazard of a second offence. For Home location, we have $\exp(-2.3802) = 0.0925$, meaning that local

learners have a lower hazard of repeat offences than non-local ones. For English, we have $\exp(0.0565) = 1.0581$, meaning that an increase of 1 mark in English increases the hazards of a second offence by 5.8%. For Mathematics, we have $\exp(-0.1164) = 0.8901$, meaning that an increase of 1 mark in Mathematics decreases the hazards of a second offence by 11%. The interaction of Grade and Parent Involvement, and also the interaction of Grade and Suspected substance abuse were included in the model. For these interactions we see the p-values of 0.132 and 0.497, respectively, implying that we cannot reject the null hypothesis (at the 5% level of significance) that there is no interaction effect. LR P-value (0.0000) is significant, indicating the presence of useful predictors in the model fitted.

4.4 Applying Parametric Hazards Models to Learner Offence Data

In the previous sections we have seen that the nonparametric model made no assumption about the hazard. That makes comparing learners within risk sets difficult. As indicated in Section 2.2.1 Nonparametric and Semi parametric models compare learners at the times when failures (offending again) occur. Parametric models do not base their results on such comparisons but depict what occurs over the whole interval, given what is known about the learner during the current time. In this section we apply the Exponential, Weibull, and Gompertz distributions to regress survival time on various groups of learners.

4.4.1 Regressing survival time on the Parental Involvement group.

The three parametric hazards models that were fitted to the data are Exponential, Weibull, and Gompertz, with their corresponding outputs given in appendices G.1, G.2, and G.3. The results are summarised in Table 4.27.

Table 4.27: Parametric regression: Parent involvement LR test

Distribution	$\exp(\beta)$	Constant	P-value	LR (P-value)
Exponential	0.4083	0.0104	0.095	0.0639
Weibull	0.4190	0.0516	0.105	0.0728
Gompertz	0.4140	0.0293	0.101	0.0686

The three models fitted indicate non-significant results at the 5% level. That is the null hypothesis that the group effect is zero, is not rejected at 0.05 significance level for all the distributions. We conclude that whether the parent was involved or not, there was no significant difference on survival time of learners to commit a second offence. The hazard ratios are less than 1 indicating that parental involvement reduces the hazard, but not significantly as the test indicated. Similar results of non-significance about parental involvement were also found when applying the Cox proportional hazards model and when testing the equality of survivor functions. We observe that in all the three distributions the LR P-value is greater than 0.05 implying that the model fits the data.

4.4.2 Regressing survival time on Suspected Substance Abuse membership

The results of the three models that were fitted to the data are given in Table 4.28.

Table 4.28: Parametric regression: Suspected substance abuse LR test

Distribution	$\exp(\beta)$	Constant	P-value	LR (P-value)
Exponential	10.9729	0.0016	0.000*	0.0000*
Weibull	10.1322	0.0072	0.000*	0.0000*
Gompertz	10.0230	0.0045	0.000*	0.0000*

By observing the smallest p-value (0.000) the three models give high significant results. That is the null hypothesis of no group effect is rejected at the 5% level of significance, in favour of the alternative hypothesis that there is a group effect. In other words belonging to a group of Suspected Substance Abusers has significant effect on survival time of learners to offend again. We see the hazard ratios are greater than 10 meaning that the hazard faced by the group of learners who are suspected of substance abuse is 10 times the hazard of the group not suspected of substance abuse. Thus, exposure to substance abuse is not safe because it increases hazard. Suspected substance abuse was also found to be significant when applying

the Cox Proportional Hazards Model. In all the three distributions the LR P-value (0.0000) < 0.05 meaning that the model is not fitting the data.

4.4.3 Regressing survival time on Repeaters

In analysing the learner groups according to (Repeating), again the three parametric hazards models fitted t were Exponential, Weibull, and Gompertz. The results are given in Table 4.29.

Table 4.29: Parametric regression: Repeating LR test

Distribution	$\exp(\beta)$	Constant	P-value	LR (P-value)
Exponential	6.2077	0.0021	0.003*	0.0002*
Weibull	5.3879	0.0103	0.006*	0.0007*
Gompertz	4.9514	0.0066	0.009*	0.0013*

The p-values are less than 0.05, indicating that the null hypothesis of no group effect is rejected. The two groups of the variable (Repeating) differ significantly at the 5% level. That is there is a difference in survival times to offend again between a learner belonging to group of repeaters and leaners belonging to a group of non-repeaters. We also found significant results with regard to Repeating when applying the Cox proportional hazards models and when testing the equality of survivor functions. We also see from the high hazard ratios 6.2077, 5.3879, and 4.9514 that repeating is not safe since it increases the hazard. The LR P-value < 0.05 in all the three distributions meaning that the model does not fit the data.

4.4.4 Regressing survival time on the variable (Home location)

For the variable (Home location), we compare local learners, i.e. those who live in the area where the school is located and those coming from other areas or far from school. The three parametric hazards models were fitted and the results are given in Table 4.30.

Table 4.30: Parametric regression: Home location LR test

Distribution	$\exp(\beta)$	Constant	P-value	LR (P-value)
Exponential	0.4492	0.0187	0.274	0.3280
Weibull	0.5970	0.0707	0.483	0.5136
Gompertz	0.7034	0.0343	0.632	0.6483

The p-values are much higher than 0.05, showing insignificant results for all distributions. The null hypothesis that the group effect is zero, is not rejected. That is, there is no difference in survival time to commit a second offence due to learners' proximity to school. Similar results of non-significance were found when applying the Cox proportional hazards model and when testing the equality of survivor functions based on the variable (Home location). The hazard ratios are less than 1 indicating that the hazard faced by local leaners is lower than the one faced by non-local learners, although not significantly so. For all the three distributions LR P-value > 0.05 implying that the model fits the data.

4.4.5 Regressing survival time on Hostel residence membership

Again the effect of learner's home on learner offence is tested based on hostel residents versus non-hostel residents. Table 4.31 provides the results of the three models that were fitted.

Table 4.31: Parametric regression: Hostel residence LR test

Distribution	$\exp(\beta)$	Constant	P-value	LR (P-value)
Exponential	0.9929	0.0087	0.992	0.9922
Weibull	1.1099	0.0440	0.887	0.8885
Gompertz	1.2753	0.0244	0.740	0.7484

The p-values reported by these models are higher than 0.05. The null hypothesis of no group effect is not rejected at the 5% level of significance. That is, there is no difference in survival time to offend again between learners who live at the hostel and those who are not hostel

residents. We also see that the hazard ratios are very close to 1. That is, exposure of learners to hostel residence seems not to change the hazard to second offence. We can say that the place where the learners come from, is not an issue of major concern as the Cox proportional hazards model and the test for equality of survivor functions also reported non-significant results based on the variables (Home location and Hostel residence). For all the three distributions we see LR P-value > 0.05 implying that the model fits the data.

4.4.6 Regressing survival time on Gender.

Table 4.32 gives the results of testing the effect of gender on learner offence by fitting the Exponential model, Weibull model, and Gompertz model.

Table 4.32: Parametric regression: Gender LR test

Distribution	$\exp(\beta)$	Constant	P-value	LR (P-value)
Exponential	4185539	$2.67e^{-09}$	0.983	0.0001*
Weibull	6875576	$7.94e^{-09}$	0.986	0.0001*
Gompertz	4204962	$7.45e^{-09}$	0.983	0.0001*

From the p-values we observe that the null hypothesis that there is no gender effect, is not rejected at the 5% level of significance. That is according to these models there is no difference in survival times to offend again between a boy learner and a girl learner. This is due to the fact that there were no girl who committed a second offence. However, we observe a very big hazard ratio for gender as it was also the case when applying the Cox proportional hazards model. It means the hazard faced by male learners is highly greater than the hazard faced by female learners. From the tests for equality of survivor functions, we have found the variable (Gender) to have significant effect on survival time of learners to offend again. For all the three distributions the model does not fit the data since LR P-value is less than 0.05.

4.4.7 Regressing survival time on (Grade) membership

The variable (Grade) has four categories without numerical or ordinal value. It indicates whether the learner belongs to grade 8, 9, 10, or 11. The results of the Exponential model, Weibull model, and Gompertz model that were fitted are provided in Table 4.33.

Table 4.33: Parametric regression: Grade LR test

Distribution	$\exp(\beta)$	Constant	P-value	LR (P-value)
Exponential	1.3305	0.0006	0.211	0.2020
Weibull	1.2537	0.0051	0.323	0.3158
Gompertz	1.1897	0.0048	0.444	0.4390

Looking at the p-values we see that the effect of learner grade on learner offence is not significant at 0.05 level of significance. That means there is no difference in survival times of learners to offend again due to learner belonging to grade 8, 9, 10, or 11. The hazard ratios from all the three models are very close to one. That is, there is no significant difference in the hazards of a second offence faced by learners due to movement from one grade to the other. The LR P-value is greater than 0.05 in all the distributions meaning that the model is fitting the data.

4.4.8 Regressing survival time on Learner Performance

The Exponential model, Weibull model, and Gompertz model were fitted to analyse the effect of learner performance in each of the three subjects on learner offence. The results are given in Table 4.34.

Table 4.34: Parametric regression: Learner performance LR test

Variable	Distribution	$\exp(\beta)$	Constant	P-value	LR (P-value)
English	Exponential	0.9573	0.0431	0.001*	0.0007*
	Weibull	0.9594	0.1838	0.001*	0.0012*
	Gompertz	0.9586	0.1104	0.001*	0.0011*
Mathematics	Exponential	0.9298	0.0326	0.001*	0.0004*
	Weibull	0.9324	0.1437	0.002*	0.0007*
	Gompertz	0.9316	0.0847	0.002*	0.0006*
Life Orientation	Exponential	0.9622	0.0367	0.001*	0.0005*
	Weibull	0.9655	0.1503	0.001*	0.0010*
	Gompertz	0.9661	0.0862	0.001*	0.0009*

We observe that the hazard ratios for all the three variables are less than 1, which means that exposure to these variables is protective. English has a hazard ratio of 0.96 implying that an increase by 1 mark obtained in English decreases the hazard by 4%. For Maths, the hazard ratio is 0.93 which means that an increase of 1 mark in Mathematics decreases the hazard by 7%. The hazard ratio of 0.97 for Life Orientation means that an increase by 1 mark in this subject decreases the hazard by 3%. Thus, good marks reduce hazards to second offence. We further observe p-values less than 0.05 when testing for the effect of learner performance on offences, meaning that at the 5% level of significance, the chi-square likelihood ratio test rejects the null hypotheses that the effect of learner performance on learner offence is zero. Based on the three subjects, we conclude that there is a difference in survival time of learners to offend again due to learner performance. The higher the learner marks the less likely they are to offend again. Although we expected Life Orientation to be the most significant variable it turns out Mathematics plays a bigger role in reducing offence. For all the three variables we see the LR P-values < 0.05 for all distributions, implying that the model does not fit the data.

4.5 Fitting Parametric Models with all the Covariates Included

We further fit a model with all the predictors to capture their interaction. As mentioned in the model building section, there is also a specific interest on the interaction between Grade and Parental involvement, and the interaction between Grade and Suspected substance abuse. The results of the three models that were fitted are given in Table 4.35 and the corresponding Stata outputs are provided in Appendices P1, P2, and P3.

Table 4.35: Parametric regression: Combined effect LR test

	Weibull		Exponential		Gompertz	
Variable	$\exp(\beta)$	P-value	$\exp(\beta)$	P-value	$\exp(\beta)$	P-value
Constant	0.0727185	0.999	0.2477727	0.999	0.0107822	0.998
Gender	3587648	0.992	2876512	0.991	5171484	0.992
Grade	0.1953836	0.069	0.1703106	0.048*	0.2252984	0.093
Repeating	1.655032	0.512	1.592268	0.547	1.697999	0.492
Home location	0.1070689	0.019*	0.0623134	0.003*	0.1061301	0.016*
Hostel residence	0.2745176	0.095	0.2169307	0.047*	0.3465424	0.166
Parent involvement	2.19e-07	0.061	7.47e-08	0.048*	3.84e-07	0.078
Substance abuse	0.008467	0.586	0.0049791	0.546	0.0376602	0.707
English	1.054369	0.043*	1.063945	0.020*	1.054193	0.041*
Mathematics	0.88434	0.003*	0.8661981	0.001*	0.8916483	0.004*
Life Orientation	0.9735348	0.126	0.9703673	0.092	0.9738822	0.119
Grade and Parent Involvement	4.389278	0.072	4.85688	0.059	4.088064	0.095
Grade and Substance abuse	2.33234	0.403	2.527121	0.361	1.968912	0.502
LR		0.000		0.000		0.000

We observe that Home location, English, and Mathematics are significant at 5% level for all distributions. For the Exponential model three more variables are significant at 5% level, namely, Grade, Hostel residence, and Parent involvement. For both Weibull and Gompertz

distributions, the variables Grade and Parent involvement are only significant at 10% level. Among the significant variables only English has a hazard ratio greater than 1 implying that exposure to this variable increases hazards of a second offence. The variables Home location, Mathematics, Grade, and Parent involvement have their hazard ratios less than 1. Thus, exposure to these variables decreases hazards of a second offence. We cannot reject the null hypothesis (at the 5% level) that there is no interaction effect. That is none of the interactions is significant at the 5% level. It means that there no significant difference in survival time of learners to offend again due to interaction between Grade and Parental Involvement, and there no significant difference in survival time of learners to offend again due to interaction between Grade and Suspected substance abuse. Similar results about interaction were found when applying the Cox proportional hazards model. The LR P-value is less than 0.05 for all the distributions meaning that the model does not fit the data.

4.6 Modelling the Number of Repeat Offences

4.6.1 Summary Statistics

Table 4.36 provides summary statistics for the number of repeat offences.

Table 4.36: Summary Statistics for repeat offences

Variable	Obs	Mean	Std. Dev.	Min	Max
Repeat Offences	83	1.4699	2.8212	0	13

The mean of 1.4699 indicates that on average learners repeat an offence at least once but less than twice. However, there was a learner who committed offences repeatedly up to 13 times. A minimum value of 0 indicates that there were learners who had not committed an offence again since the first one. The bar graphs in section 4.6.2 to 4.6.8 are used to analyse the number of repeat offences in relation to the categorical variables.

4.6.2 Examining the Number of Repeat Offences in relation to Parental Involvement

The number of repeat offences in relation to parental involvement are represented graphically in Figure 4.16.

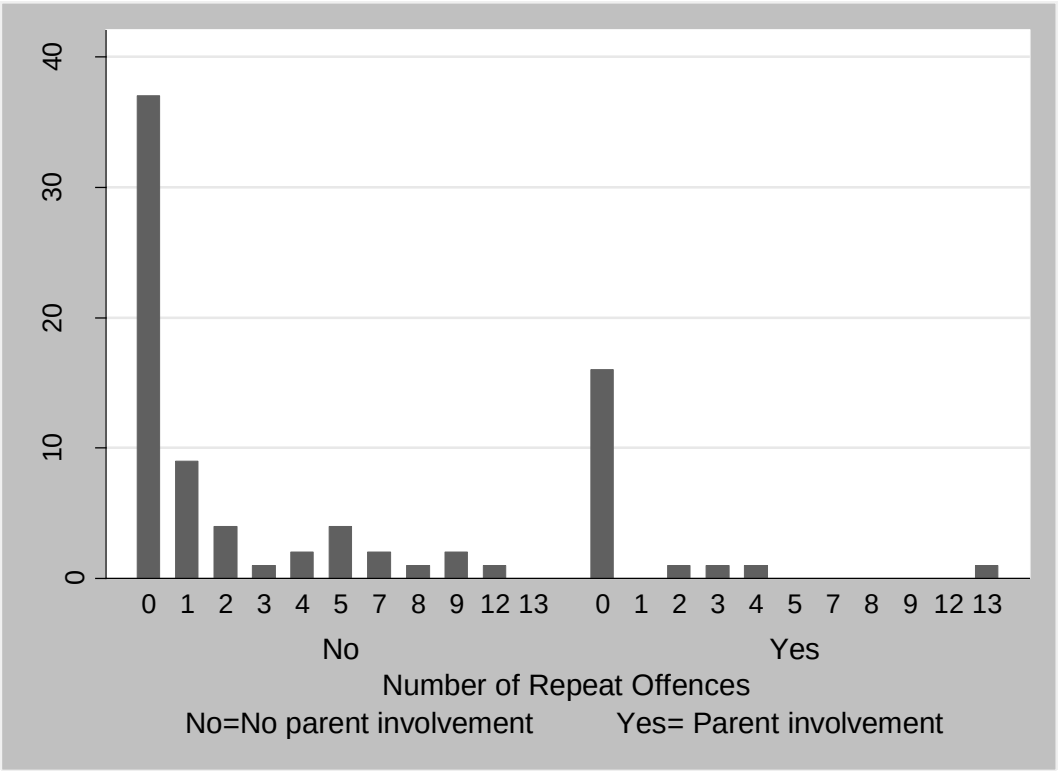


Figure 4.16: Repeat offences and Parental involvement

From the learners whose parents were involved in the handling of the first offence we observe that one learner committed 13 repeat offences, one committed 4 repeat offences, one committed 3 repeat offences, one committed 2 repeat offences, and about 16 learners never committed a repeat offence. From the learners whose parents were not involved in the handling of the first offence we observe that there were more of the learners who committed the repeat offences than those whose parents were involved. We also see from the learners whose parents were not involved in handling the first offence that more than 35 learners did not commit a repeat offence. In general the long bars in Figure 4.16 indicate that there were many learners who did not commit a repeat offence whether the parent was involved or not in the handling of the first offence.

4.6.3 Examining the Number of Repeat Offences in relation to Suspected Substance Abuse

The number of repeat offences in relation to suspected substance abuse are represented graphically in Figure 4.17.

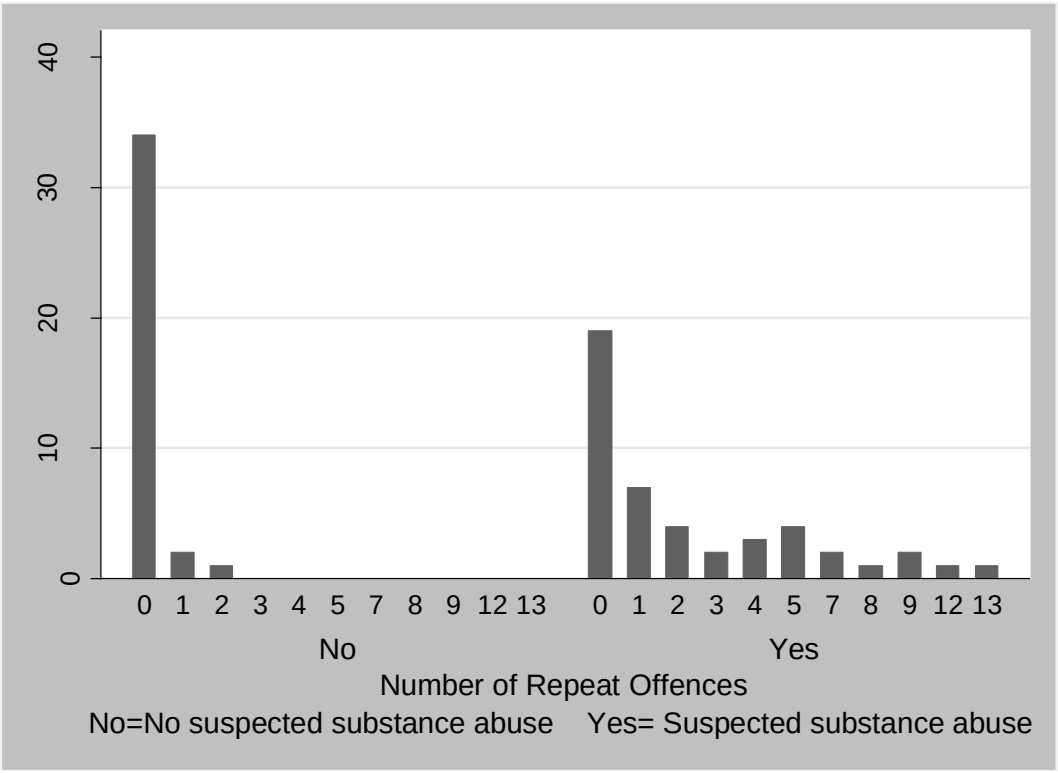


Figure 4.17: Repeat Offences and Suspected substance abuse

In Figure 4.17 we see a high frequency of learners who did not commit a repeat offence from both the learners suspected of substance abuse and those not suspected. From the learners who are not suspected of substance abuse, 33 learners did not commit a repeat offence, 2 learners committed 1 repeat offence, and only 1 learner committed 2 repeat offences. From those learners who are suspected of substance abuse, 19 learners did not commit a repeat offence, 7 learners committed 1 repeat offence, 4 learners committed 2 repeat offences, 2 learners committed 3 repeat offences, 3 learners committed 4 repeat offences, 4 learners committed 5 repeat offences, 2 learners committed 7 repeat offences, 1 learner committed 8

repeat offences, 2 learners committed 9 repeat offences, 1 learner committed 12 repeat offences, and 1 learner committed 13 repeat offences. Thus the learners who are suspected of substance abuse committed much more of the repeat offences than those not suspected.

4.6.4 Examining the Number of Repeat Offences in relation to Repeating a Grade

In Figure 4.18 we observe from the learners who were not repeating a grade that 21 learners never committed a repeat offence, 1 learner committed 1 repeat offence, 1 learner committed 2 repeat offences, and 1 learner committed 9 repeat offences.

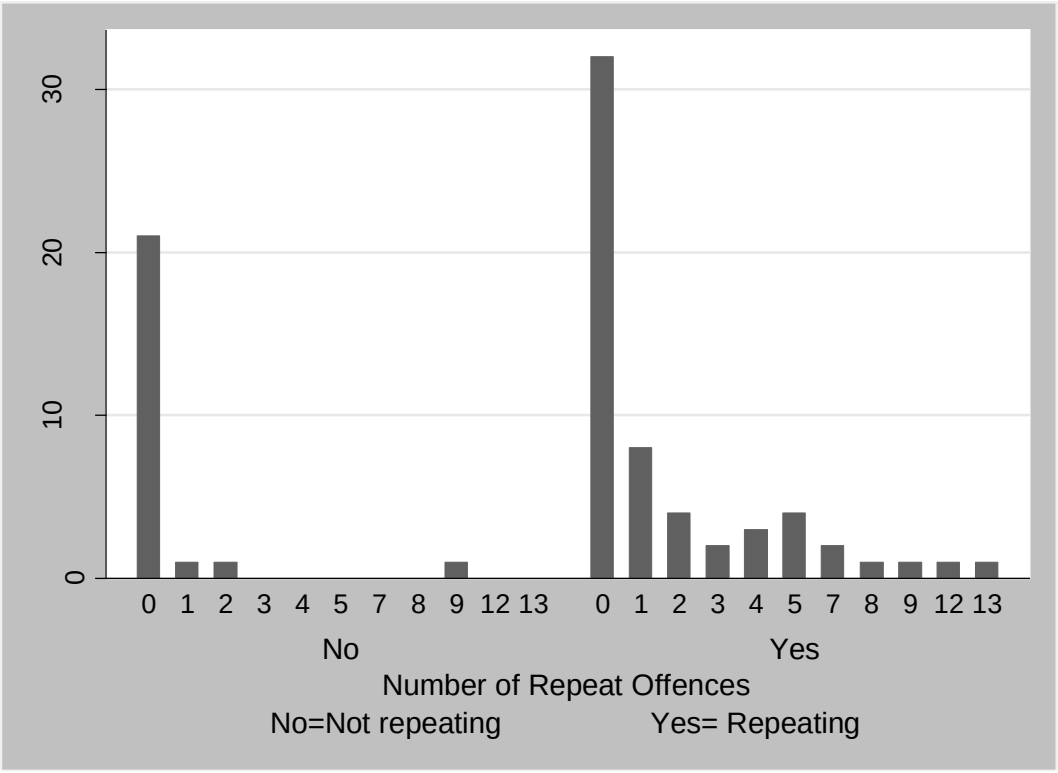


Figure 4.18: Repeat Offences and Grade Repeating

By examining learners who were repeating a grade, we see that 32 learners never committed a repeat offence, 8 learners committed 1 repeat offence, 4 learners committed 2 repeat offences, 2 learners committed 3 repeat offences, 3 learners committed 4 repeat offences, 4 learners committed 5 repeat offences, 2 learners committed 7 repeat offences, 1 learner committed 8 repeat offences, 1 learner committed 9 repeat offences, 1 learner committed 12 repeat

offences, and 1 learner committed 13 repeat offences. We can in general conclude that repeating a grade increases the number of repeat offences, however, there was one none repeating learner who committed 9 repeat offences. We have also observed that whether repeating a grade or not, there is a high frequency of learners who did not commit a repeat offence.

4.6.5 Examining the Number of Repeat Offences in relation to Gender

The graph in Figure 4.19 shows that there was no girl who committed a repeat offence. We also see high frequency (38) of boys who never committed a repeat offence.

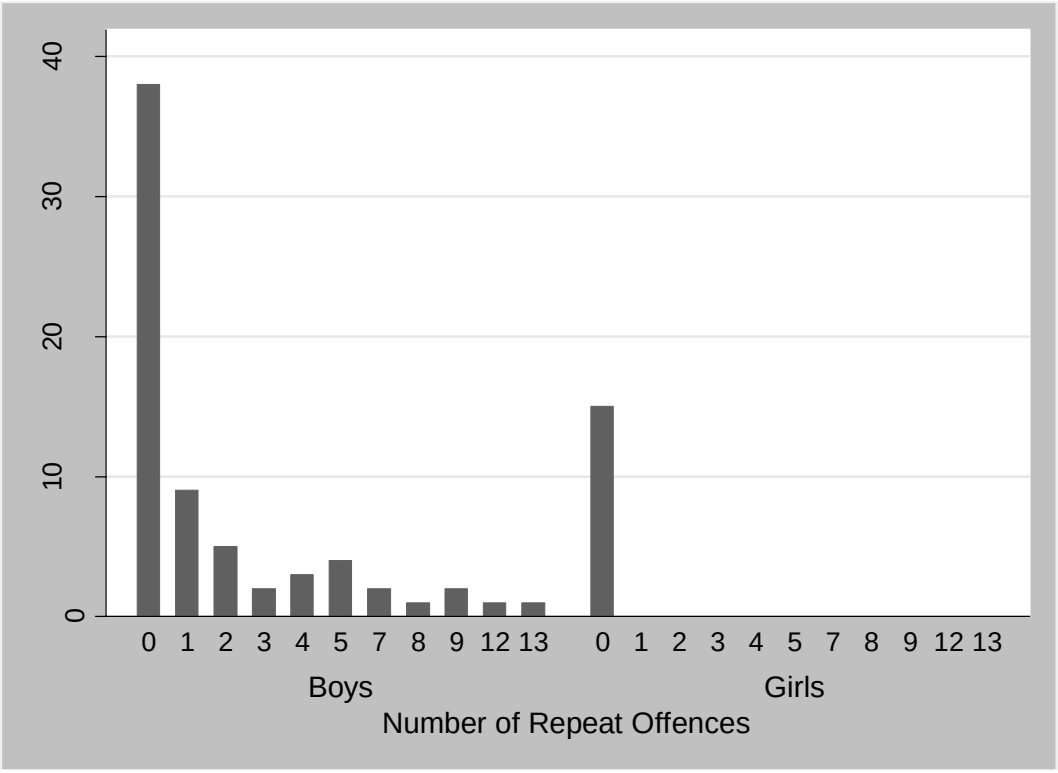


Figure 4.19: Repeat Offences and Gender

We observe that 9 boys committed 1 repeat offence, 5 boys committed 2 repeat offences, 2 boys committed 3 repeat offences, 3 boys committed 4 repeat offences, 4 boys committed 5 repeat offences, 2 boys committed 7 repeat offences, 1 boy committed 8 repeat offences, 2 boys committed 9 repeat offences, 1 boy committed 12 repeat offences, and 1 boy committed

13 repeat offences. Since there were very few girls in this study we may be wrong to suggest that being a girl learner decreases the number of repeat offences, based only on the bar graph. More analysis Gender is required.

4.6.6 Examining the Number of Repeat Offences in relation to Home Location

From the non-local learners in Figure 4.20 we see that 3 learners did not commit a repeat offence, 1 learner committed 4 repeat offences, and 1 learner committed 7 repeat offences.

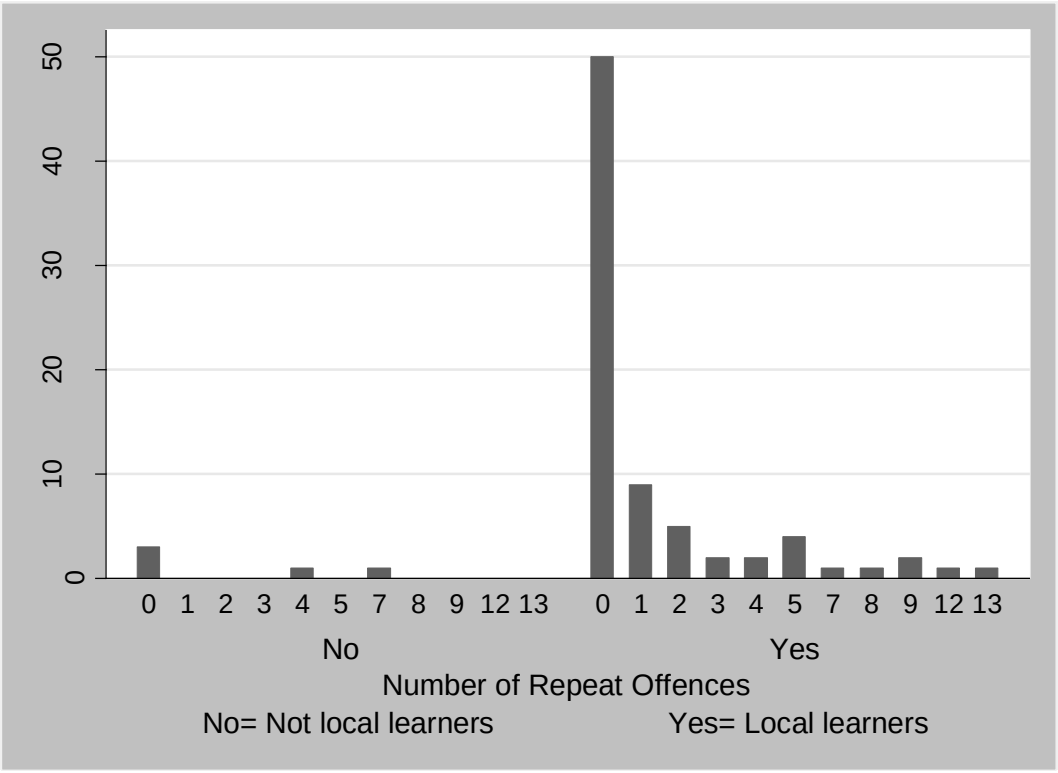


Figure 4.20: Repeat Offences and Home location

From the local learners we observe that 50 learners did not commit a repeat offence, 9 learners committed 1 repeat offence, 5 learners committed 2 repeat offences, 2 learners committed 3 repeat offences, 2 learners committed 4 repeat offences, 4 learners committed 5 repeat offences, 1 learner committed 7 repeat offences, 1 learner committed 8 repeat offences, 2 learners committed 9 repeat offences, 1 learner committed 12 repeat offences, and 1 learner

committed 13 repeat offences. We generally conclude that being a local learner a student is generally more likely to repeat an offence.

4.6.7 Examining the Number of Repeat Offences in relation to Hostel Residence

From the few learners who are hostel residents we see that in Figure 4.21 five learners committed a first offence and only one repeated an offence once, and one other learner committed 9 repeat offences.

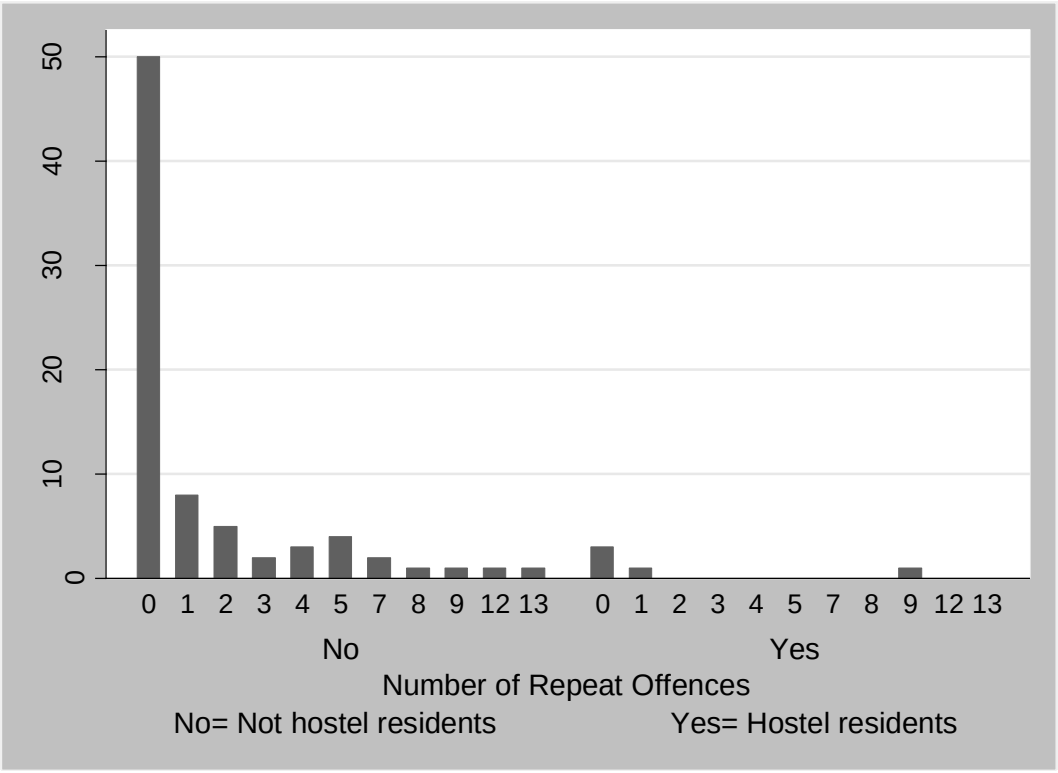


Figure 4.21: Repeat Offences and Hostel residence

For the learners not residing at the hostel, we see 50 learners who never repeated an offence, 8 learners committed 1 repeat offence, 5 learners committed 2 repeat offences, 2 learners committed 3 repeat offences, 3 learners committed 4 repeat offences, 4 learners committed 5 repeat offences, 2 learners committed 7 repeat offences, 1 learner committed 8 repeat offences, 1 learner committed 9 repeat offences, 1 learner committed 12 repeat offences, and

1 learner committed 13 repeat offences. We cannot conclude that residing at the hostel increases or decreases the number repeat offences since there are only 5 hostel residents.

4.6.8 Examining the Number of Repeat Offences in relation to Grade

The number of repeat offences in relation to learner grade are represented graphically in Figure 4.22.

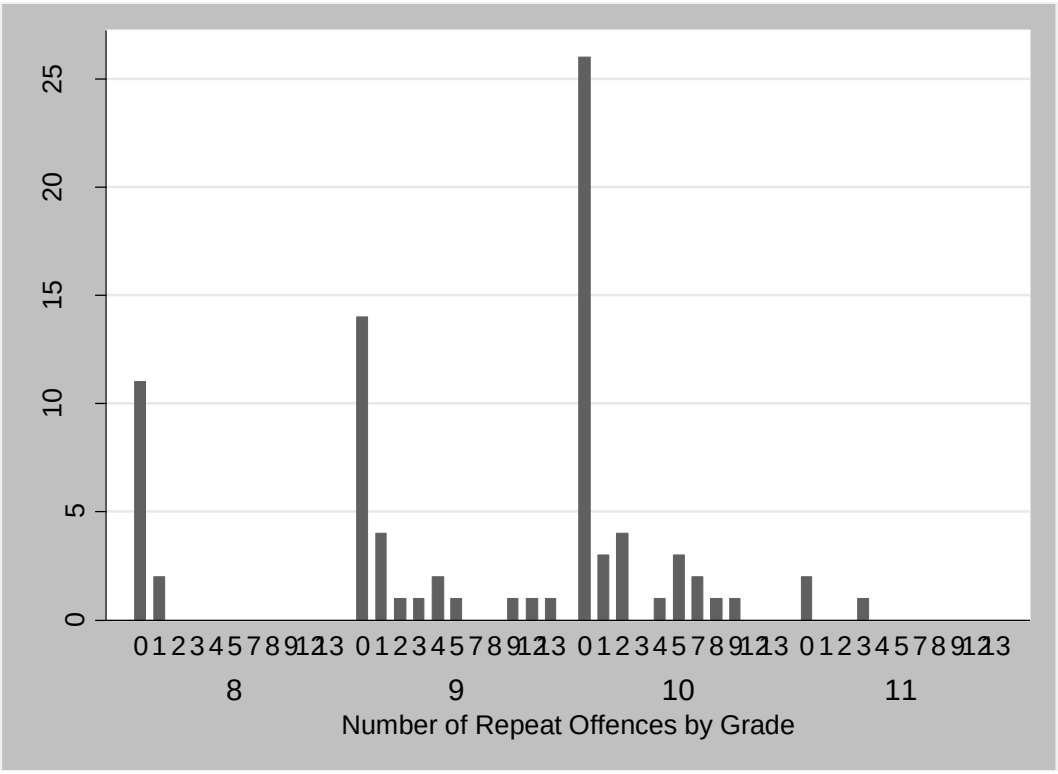


Figure 4.22: Repeat Offences and Grade

We see grade 8, 9, 10, and 11, all having high frequencies of learners who did not commit a repeat offence. There was no learner in grade 8 who committed more than one repeat offence. We see the majority of learners who committed more than one repeat offence are in grades 9 and 10.

4.6.9 Exploring Models for Count Data

As indicated in Section 3.4.2, the Poisson Model, Negative Binomial Model and Zero Inflated Models are applied to investigate the effects of variables on the number of repeat offences committed by the learners. The results of fitting these models are summarised in Table 4.37 and Table 4.38 obtained from Appendices Q, R, S1, S2, and S3 which provide more details.

Table 4.37: Comparison of Count Models (Substance abuse and Maths in the Zero part)

		POISSON		NEGATIVE BINOMIAL		ZIP	
Count part	**Variable**	β	**P-value**	β	**P-value**	β	**P-value**
	Parentinvo	- 0.1655	0.509	- 0.3212	0.551	0.0491	0.855
	Substance	2.7103	0.000*	2.8343	0.000*	2.3254	0.003*
	Repeat	0.7165	0.066	0.6801	0.235	0.5284	0.126
	Gender	16.4837	0.988	16.8883	0.990	16.1603	0.985
	Homeloc	- 0.3945	0.277	- 0.5022	0.593	0.3314	0.398
	Hostelres	- 0.2297	0.496	- 0.5201	0.489	-0.1960	0.613
	Grade	- 0.4806	0.010*	- 0.5402	0.121	-0.4353	0.081
	English	0.01947	0.089	0.0075	0.754	0.0064	0.591
	Maths	- 0.0201	0.219	- 0.0198	0.533	0.0459	0.030*
	LO	- 0.0296	0.000*	- 0.0201	0.242	-0.0304	0.001*
	Constant	-13.1562	0.990	-12.8289	0.993	-13.6482	0.988
	Pseudo R2	0.3834		0.1900			
	LR(P-value)	0.0000		0.0000		0.0001	
Zero part	**Variable**					β	**P-value**
	Substance					-0.7471	0.603
	Maths					0.1228	0.012*
	Constant					-2.2471	0.222
	Vuong test						0.0234

By observing the count part in Table 4.37 all the three models indicate the variable (Substance abuse) to have significant effect (at the 5% level) on the number of repeat offences committed by the learners. A positive coefficient (β) for (Substance) means that being suspected of substance abuse increases the number of repeat offences. The Poisson model and the Zero-Inflated Poisson (ZIP) model also show that LO (Life Orientation) has a significant effect (at the 5% level) on the number of repeat offences committed by the learners. A negative coefficient

(β) for Life Orientation means that getting higher marks in this subject decreases the number of repeat offences. The variable (Grade) was found to be significant (at the 5% level) only by Poisson model. A lower pseudo R square (38.34%) from the Poisson model and the LR (P-value) < 0.05, imply that the model fitting was not good. To test for overdispersion on Poisson, we have $\frac{157.25}{82-11} = 2.21$ which is greater than 2, implying that there is overdispersion. We try the Negative Binomial model to deal with overdispersion. We have $\frac{46.54}{82-11} = 0.6555$. Since $0.5 < 0.6555 < 2$, overdispersion is handled by the Negative Binomial model. However the model is not fitting the data as we see the lower pseudo R-square (19%) and LR (P-value) = 0.000.

Because we noticed from Figure 4.16 to 4.22 that we had an excess of zeros in repeat offences as most of the learners did not have a second offence, the Zero Inflated Poisson (ZIP) model was fitted. Appendix S1 provides the results of the initial ZIP model that was fitted with all the predictor variables included on the inflate part. The results suggest that Maths and Substance abuse play a role in the inflation of zeros since they were the only variables that were significant (at the 10% level) both on the Offences part and the inflate part. The ZIP model in Table 4.37 was fitted with only Maths and Substance abuse on the inflate part and the results show that Substance abuse was not significant (p-value = 0.603) on the inflate part. The ZIP model having only Substance abuse on the inflate part was also fitted, and again it was not significant (p-value = 0.432) as the results show in Appendix S4. Substance abuse was dropped when fitting the ZIP model that produced the results in Table 4.38.

Table 4.38: Comparison of Count Models (Maths in the Zero part)

		POISSON		NEGATIVE BINOMIAL		ZIP	
Count part	**Variable**	***β***	**P-value**	***β***	**P-value**	***β***	**P-value**
	Parentinvo	- 0.1655	0.509	- 0.3212	0.551	0.0416	0.876
	Substance	2.7103	0.000*	2.8343	0.000*	2.5657	0.000*
	Repeat	0.7165	0.066	0.6801	0.235	0.5383	0.117
	Gender	16.4837	0.988	16.8883	0.990	18.0126	0.994
	Homeloc	- 0.3945	0.277	- 0.5022	0.593	0.3135	0.421
	Hostelres	- 0.2297	0.496	- 0.5201	0.489	-0.1918	0.620
	Grade	- 0.4806	0.010*	- 0.5402	0.121	-0.4561	0.061
	English	0.01947	0.089	0.0075	0.754	0.0069	0.562
	Maths	- 0.0201	0.219	- 0.0198	0.533	0.0442	0.035*
	LO	- 0.0296	0.000*	- 0.0201	0.242	-0.0305	0.001*
	Constant	-13.1562	0.990	-12.8289	0.993	-15.5205	0.994
	Pseudo R2	0.3834		0.1900			
	LR(P-value)	0.0000		0.0000		0.0000	
Zero part	**Variable**					***β***	**P-value**
	Maths					0.1282	0.008*
	Constant					-3.0527	0.004
	Vuong test						0.0235

The results of Table 4.38 suggest that Maths plays a role in the inflation of zeros. We have $\exp(0.1282) = 1.14$, meaning that a higher Maths score increases the likelihood of learners not to commit a second offence by 14%, and this is statistically significant (p=0.008). The estimate (2.5657) of the variable Substance abuse in the ZIP model gives $\exp(2.5657) = 13.01$. It means that it is statistically significant (p-value=0.000) that among those learners who are at risk of committing a second offence, being suspected of substance abuse increases the expected rate of committing a second offence by more than 100%, while holding other variables constant. An increase in Mathematics mark slightly increases the mean number of second offences by $\exp(0.0442) = 1.05$, i.e. by 5%. From the estimate (-0.0305) of the variable LO in the ZIP model, we have $\exp(-0.0305) = 0.97$. It means that among those learners who are at risk of committing a second offence, a higher LO score decreases the likelihood of committing a second offence by 3%, while holding other variables constant.

4.7 Summary

In this Chapter, the survival analysis methods were applied to study the time to a second offence committed by learners. We have seen from the Kaplan-Meier survival estimate that, in general a learner who survive 18 weeks before committing a second offence is less likely to commit a second offence. The Nelson-Aalen estimate has shown that the hazards of a second offence decrease between 10 and 30 weeks from the first offence, but increase after 30 weeks. The Kaplan-Meier survival estimate was also used to compare groups of learners based on the categorical variables. It was shown that learners whose parents were involved in handling of offence cases had a better survival experience to second offence than the group of learners without parental involvement. The learners who are not suspected of substance abuse had far much better survival experience than those suspected of substance abuse. Although there were very few girls in the study, the boys were at risk of committing a second offence until after 54 weeks from the first offence. There was a lower survival experience to a second offence for repeaters than for non-repeaters. Most of the learners are local with very few none local learners, and that made it difficult to compare the Kaplan-Meier survival plots. It was also shown from the plots of hostel residents versus non-hostel residents that the survivor probabilities for learners to commit a second offence do not differ according to learners' place of residence. We observed that grade 8 had a better survival experience as compared to other grades, while grade 9 had the worst survival experience to a second offence. For the continuous variables, the summary statistics had shown the worst performance of learners (1^{st} offenders) in Mathematics with an average score of 19.8. In Life Orientation and English, the learners obtained average scores of 40.4 and 38.8, respectively.

The Log-rank, Wilcoxon and Peto-Peto tests for equality of survivor functions between groups of learners were performed. It was shown that the variables (suspected substance abuse, gender, repeating a grade) had a significant effect on survival time of learners to commit a second offence, while the variables (parental involvement, home location, hostel residence,

learner grade) were found not to be significant. The Cox proportional hazards model was also applied to learner offence data. When applied individually, the variables (suspected substance abuse, repeating a grade, English, Mathematics, and Life Orientation) were found to have a significant effect on the hazards of a second offence. However, when all the predictor variables and interaction between some variables were included in the model, the significant variables were Home location, English, and Mathematics. Home location and Mathematics had their coefficients less than 0, meaning that exposure to these variables decreases hazards of a second offence. English had a coefficient greater than 0 meaning that exposure to this variable increases hazards of a second offence. For Home location, we had $\exp(-2.3802) = 0.0925$, meaning that local learners face the hazards of a second offence 9% greater than the non-local learners. For English, we had $\exp(0.0565) = 1.0581$, meaning that an increase of 1 mark in English increases the hazards of a second offence by 5.8%. For Mathematics, we had $\exp(-0.1164) = 0.8901$, meaning that an increase of 1 mark in Mathematics decreases the hazards of a second offence by 11%.

When fitting the parametric hazards models, again the variables were first considered individually, and the variables (suspected substance abuse, repeating a grade, English, Mathematics, and Life Orientation) were found to have a significant effect on the hazards of a second offence, as was the case with the Cox proportional hazard model. However, when all the predictor variables and interaction between some variables were included in the model, we observed that Home location, English, and Mathematics were significant at 5% level for all the three distributions. From two of the three parametric hazards models, Home location had a hazard ratio of 0.11, implying that the hazard of a second offence faced by local learners is 0.11 times the hazard faced by non-local learners. English had a hazard ratio of about 1.05 from all the three parametric hazards models, implying that an increase of 1 mark in English increases the hazards of a second offence by 5%. Mathematics had a hazard ratio of about 0.88 from all the three models, implying that an increase of 1 mark in Mathematics decreases the hazards of a second offence by 12%.

The methods of count data were applied to study the number of repeat offences. Since the bar graphs had shown an excess of zeros in the second offences, the Zero Inflated Poisson (ZIP) model was fitted. We have seen that Mathematics contributes significantly (p-value=0.008) to more zeros, meaning that it reduces the chances of a second offence. We observed from the results of the ZIP model that once the initial offence occurs, the variables (Suspected substance abuse and Mathematics) influence the number of second offences. Suspected substance abuse increases the mean number of second offences by 1201%. An increase in Mathematics mark slightly increases the mean number of second offences by 5%.

Chapter 5

Summary and Conclusions

5.0 Introduction

The study aimed to model learner lack of discipline at schools. All the data about the dependent variables and related covariates used in the study of learner offence were obtained from a public school, which is a governmental educational institution. The data were obtained from school records for the period July 2013 to June 2015. All learners were eligible for inclusion in the study because the sample comes from the whole learner population enrolled at the school during that period. The dataset used is provided in Appendix A and the results of analyses in subsequent appendices mainly produced by STATA 14 (2015) statistical software package. All the variables were described in Chapter 1.

5.1 Summary

The covariates used in the analysis consist of 7 categorical variables and 3 continuous variables. The literature review in Chapter 2 and methodology in Chapter 3 discuss the statistical techniques applicable for this study, whereas the detailed analysis followed in Chapter 4. The analytical methods used in this study complement each other.

The use of survival analysis methods was supported by observation of the random variable *T* that takes time measures (in weeks) between first and second learner offence. The failure event is when the learner make an offence again. One limitation for using survival analysis in this study was lack of left-censoring. There were no delayed entries since all subjects (learners) entered the study at time 0, but not all the learners had the first offence at the beginning of the study period. To allow for left censoring, learners might have entered the study at different

times since the survival time recorded is the time between the first and the second offence for an individual learner. Survival analysis reported 30 failures out of 83 observations in the data.

Non-parametric analysis using Kaplan-Meier survival estimate produced the following results about the potential covariates: The learners whose parents were involved in their offence cases had a better survival experience than those whose parents were not involved. The learners who were not suspected of substance abuse had far much better survival experience than those suspected of substance abuse. There was a better survival experience for girls than for boys. There was a better survival experience for non-repeaters than for repeaters. Grade 9 learners had the worst survival experience amongst all the grades involved in the study. The results were further investigated by testing the hypotheses about the equality of survivor functions across two or more groups of each of the categorical variables. According to these tests (e.g. the log-rank test), the survivor probabilities of the learners whose parents were involved in the learner offence cases and those whose parents were not involved, are equal. The contradiction with the result of Kaplan-Meier survival estimate might be due to the fact that 76% of the learners in the dataset had no parental involvement and only 24% had parental involvement. The tests also reported equal survival probabilities between two or more groups of leaners for the variables: home location, hostel residence, and grade. The tests found the significant difference in the survivor probabilities of two groups of learners for the following variables: suspected substance abuse, gender, and repeating.

Semi-parametric analysis was done using the Cox proportional hazards regression model. The model was fitted firstly using each of the covariates individually. The variables that indicated significant effect on survival time for learner to offend again, were suspected substance abuse, repeating a grade, and learner performance in English, Life Orientation, and Mathematics. However, only the variables (Home location, English, and Mathematics) were significant when including interaction in the combined model. There was no significant difference in survival time of learners to offend again due to interaction between Grade and Parental Involvement, and also interaction between Grade and Suspected substance abuse. Similar results about the

significant variables were also found when fitting the parametric hazards regression models (exponential, Weibull, and Gompertz). We have also seen from both the Cox proportional hazards model and parametric regression hazards models that the combined model with all the variables included, does not fit the data. The Cox proportional hazards model and parametric regression hazards model that fitted the data was when investigating the effects of the following variables individually: parental involvement, home location, hostel residence, learner grade, and parent involvement.

The number of offences committed by learners after their first offence were examined by fitting the Poisson regression model, Negative Binomial regression model and Zero-Inflated Poisson (ZIP) model. The LR P-values from the count part and the Vuong P-value from the zero part suggest that the models do not fit the data. However, the results from these models indicated the variables (Suspected substance abuse and Mathematics) to have a significant effect on the number of repeat offences. We have seen that the effect of the variable (Suspected substance abuse) is to increase the odds of repeating an offence, and that an increase in Mathematics mark slightly increases the mean number of second offences. From the zero part of the ZIP model, we have observed that Mathematics plays a significant role in the inflation of zeros. A higher Mathematics score increases the likelihood of learners not to commit a second offence.

5.2 Conclusion

Based on the results of the models that were applied, this study found the variables (Home location, English, Mathematics, and Suspected substance abuse) to be significant. When dealing with learners committing an offence at school, educators should pay careful attention to local learners because local learners face the hazards of a second offence more than the non-local learners. They should also pay attention to learner performance, especially in Mathematics because a higher score in Mathematics decreases the hazards of a second offence. We have seen from the combined models that an increase of 1 mark in English increases the hazards of a second offence by about 5%. It means that, unlike Mathematics, learners might still commit the

second offence even if they get higher marks in English. We have also seen that when English was applied independently, an increase of 1 mark decreases the hazards of a second offence by 4%. We may suspect that English is an easier subject for any learner to be able to score higher marks even if the learner had a history of lack of discipline. Educators should monitor learners who are suspected of substance abuse and make their parents aware since the study had shown that being suspected of substance abuse increases the likelihood of repeat offences.

5.3 Recommendations

Although parental involvement in dealing with learner offence case was found to be not significant, it is important to involve parents in school matters. Without parental involvement teachers are left frustrated by learners' behaviour and end up risking their jobs by applying corporal punishment which is not allowed according to existing policies in the Department of Education. When developing policy on discipline and the code of conduct for learners, educators should involve the parents in order to have a working plan that will monitor the learners' movements all the time whether they are at school or at home. That will assist in earlier detection of any suspicions of substance abuse that may lead learners to commit offences. Schools and government should educate parents and learners about the problems associated with substance abuse.

When arranging meetings for parents, the schools are advised to have additional meetings meant specifically for parents from the local area where lack of discipline is the issue for discussion. The code of conduct may include a clause that put harsh measures against leaners from the local home location who commit offence again. We should expect local learners to be a good example to learners coming from other areas so that a good reputation of the school in that community is maintained.

Parents should be encouraged to check the homework books on daily basis and assist children to get more information on what is being taught. Although all the subjects are important, learners should be encouraged to spend more time on studying Mathematics. The department

or the schools may review the teaching time and allocate more time to Mathematics. That will assist learners to achieve good marks in Mathematics assessment tasks, and reduce their chances of committing offences.

Learners' lack of discipline is an ongoing challenge and concern. Studying learners' lack of discipline and misconduct using statistical techniques can be done in more other schools, locally and globally for comparability. This will make the findings more informative and the recommendations applicable to schools in general.

REFERENCES

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics,* No 6: pp. 701-726.

Baumrind, D. (1996). *Effects of authoritative parental control on child behaviour*, vol. 37, no 4, pp. 887-907.

Beck, A.J. and Shipley, B.E. (1987). *Recidivism of young parolees*, Bureau of Justice Statistics, U.S. Department of Justice, 11 pages.

Breslow, N.E. (1970). A generalized Kruskal-Wallis test for comparing *k* samples subject to unequal patterns of censorship. *Biometrika*, No 57: pp. 579-594.

Cameron, A.C. and Trivedi P.K. (2013). *Regression Analysis of Count Data,* 2nd ed., Cambridge University Press, Vol. 53, New York, USA, 567 pages.

Cleves, M.A., Gould, W.W. and Gutierrez, R.G. (2004). *An introduction to survival analysis using Stata,* revised edition, Stata Corporation, College Station, Texas, 308 pages.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*. 34: pp. 187-220.

Dirks, M. (2012). *Educators and learners' perceptions and experiences regarding the effectiveness of school rules in the Fezile Dabi District*, unpublished M Ed dissertation, North-West University. URI: http://hdl.handle.net/10394/10281.

Dishion, T.J. and MacMahon, R.J. (1988). Parental monitoring and the prevention of child and adolescent problem behaviour: A conceptual and empirical formulation. *Clinical Child and Family Psychology Review,* Vol. 1, No. 1, pp. 43-62.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. and Tatham, R.L. (2005). *Multivariate Data Analysis,* 6th ed., Prentice Hall, New Jersey, 899 pages.

Hall, D.B. (2000). Zero-inflated Poisson and Binomial regression with random effects: A case study. *Biometrics*, Vol. 56, No. 4, pp. 1030-1039.

Harrisunker, N. (2014). *The perceptions of teachers, pupils, and parents regarding discipline in newly-integrated Lenasia schools*, unpublished M Ed thesis, University of the Witwatersrand, URI: http://hdl.handle.net/10539/14201.

James, M.T. (2014). *Parental involvement as a strategic tool to improve the culture of teaching and learning in the township schools*, unpublished M. Ed dissertation, University of South Africa, URI: http://hdl.handle.net/10500/13354.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data,* 2nd ed., New York, John Wiley & Sons, 447 pages.

Keating, J.B. (2011). Security in the workplace of the foundation phase educator: an education law perspective, unpublished M Ed thesis, North West University, URI: http://hdl.handle.net/10394/4697.

Kleinbaum, D.G. and Klein, M. (2012). *Statistics for Biology and Health, Survival Analysis: A Self Learning Text,* 3rd ed., New York, Springer Science + Business Media, 524 pages.

Lambert, D. (1992). Zero-inflated Poisson Regression with Application to Defects in Manufacturing, *Technometrics*, Vol. 34, No. 1, pp. 1-14.

Lambert, N.M. and McCombs, B.L. (1998). *How learners learn: Reforming schools through learner-centred education,* Washington DC, US: American Psychological Association, 540 pages.

Long, J.S. and Jeremy, F. (2006). *Regression Models for Categorical Dependent Variables Using Stata*, second edition, Stata Corporation, College Station, Texas, 526 pages.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, Vol 22: pp. 719-748.

Mestry, R., van der Merwe, M. and Squelch, J. (2006). Bystander behaviour of school children observing bullying, *SA-eDUC Journal*, Vol. 3, No. 2, pp. 46-59.

Mullahy, J. (1997). Heterogeneity, Excess Zeros, and the Structure of Count Data Models. *Journal of Applied Econometrics,* Vol. 12, No.3, pp. 337-350.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics,* No 14: pp. 940-965.

Ntuli, L.T. (2013). *Managing discipline in a post-corporal punishment era environment at secondary schools in the Sekhukhune school district, Limpopo*, unpublished M Ed dissertation, University of South Africa, URI: http://hdl.handle.net/10500/9982.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society*, Vol 135: pp. 185-206.

Republic of South Africa (2000). *The South African Council for Educators Act,* Act No. 31 of 2000, Government Printer Pretoria.

Republic of South Africa (1998). *The Employment of Educators Act,* Act No. 76 of 1998, Government Printer Pretoria.

Republic of South Africa (1996). *The National Education Policy Act,* Act No. 27 of 1996, Government Printer Pretoria.

Republic of South Africa (1996). *The South African Schools Act,* Act No. 84 of 1996, Government Printer Pretoria.

Singh, G.D. (2012). *Managing learner aggression in rural secondary schools in the Empangeni District of Kwa-Zulu Natal*, unpublished M Ed dissertation, University of South Africa, URI: http://hdl.handle.net/10500/7041.

Stata Corp. (2015). *Stata Statistical Software: Release 14.* College Station, Texas 77845, USA.

Wolhuter, C.C. and Russo, C. (2013). Dealing with incidents of serious disciplinary problems amongst learners: A comparative study between South Africa and selected countries, *Koers – Bulletin for Christian Scholarship* 78 (3), pp. 1-11.

APPENDIX A

The learner offence dataset shows 12 variables and 83 observations.

Observations are for those subjects (learners) who had cases of offence during the period starting from July 2013 to December 2014 with follow up study until 30 June 2015. The variable time represents survival time in weeks until the learner offends again or the length of follow up for censored observations. The marks recorded in the three subjects are those obtained by a learner at the end of the year in which an offence was committed. There is a missing value in Mathematics for observation with identity 20. The last column provides the number of offences committed by learners after the first offence.

Id	Time	Cense	Gender	Grade	Repeat	Homeloc	Hostelres	Parentinvo	Substance	English	Maths	LO	No of repeat offences
1	98	0	1	10	1	1	0	0	0	60	18	35	0
2	98	0	1	10	1	1	0	0	1	40	15	42	0
3	92	0	1	9	1	1	1	0	1	28	11	35	0
4	92	0	1	8	0	1	1	0	0	27	10	32	0
5	92	0	1	8	0	1	0	0	1	40	40	16	0
6	50	1	1	9	0	1	0	0	1	43	27	39	1
7	4	1	1	9	1	1	0	0	1	24	11	28	5
8	47	1	1	9	0	1	0	0	1	45	22	52	9
9	92	0	1	9	0	1	0	1	0	40	5	40	0
10	92	0	1	9	0	1	0	0	0	10	10	19	0
11	41	1	1	10	1	1	0	0	1	23	13	40	1
12	92	0	1	9	0	1	0	0	0	31	40	47	0
13	91	0	1	10	0	1	0	1	0	76	26	59	0
14	87	0	0	8	0	1	0	1	0	55	25	81	0
15	86	0	1	9	1	1	0	0	1	18	13	26	0
16	84	0	0	8	0	1	0	1	0	43	15	71	0
17	82	0	1	8	0	1	0	1	1	54	24	50	0
18	80	0	1	8	1	1	0	0	0	54	31	33	0
19	78	0	0	9	1	1	0	0	0	67	40	61	0
20	67	0	1	10	1	1	0	0	1	46		42	0
21	67	0	1	10	1	1	0	0	0	25	16	30	0
22	17	1	1	10	1	1	0	0	1	7	2	16	2
23	67	0	1	10	0	1	0	0	0	33	8	26	0
24	3	1	1	9	1	1	0	1	1	37	17	22	13
25	10	1	1	10	1	1	0	0	1	19	11	3	5
26	67	0	1	10	1	1	0	0	1	38	30	40	0
27	54	1	1	10	1	1	0	0	1	44	17	40	2
28	19	1	1	9	1	1	0	0	1	29	8	43	4
29	6	1	1	9	1	1	0	0	1	11	6	37	1

30	18	1	1	10	1	1	0	1	1	35	10	20	2
31	58	0	1	10	1	1	0	0	1	36	23	36	0
32	33	1	1	10	1	1	0	0	1	31	13	8	5
33	13	1	1	10	1	0	0	1	1	12	17	5	4
34	55	0	0	8	1	1	0	1	1	31	15	33	0
35	10	1	1	10	1	1	0	0	1	30	16	24	8
36	53	0	1	10	1	1	0	0	1	18	12	21	0
37	11	1	1	9	1	1	0	0	1	35	30	48	12
38	51	0	0	10	1	1	0	0	1	38	25	24	0
39	51	0	0	10	1	1	0	0	1	54	30	49	0
40	51	0	0	10	0	1	0	0	0	40	33	42	0
41	51	0	1	10	0	1	0	1	1	67	33	57	0
42	6	1	1	10	1	1	1	0	1	5	5	5	9
43	50	0	1	10	1	1	0	0	1	43	35	31	0
44	45	0	1	8	0	1	0	1	0	33	20	50	0
45	4	1	1	10	1	1	0	0	1	24	9	23	1
46	4	1	1	9	1	1	0	0	1	41	26	49	4
47	4	1	1	9	1	1	0	0	1	20	12	45	3
48	44	0	0	9	1	1	0	0	0	40	10	64	0
49	1	1	1	10	1	0	0	0	1	44	20	19	7
50	43	0	1	9	1	1	0	1	1	19	4	53	0
51	3	1	1	10	0	1	0	0	1	44	22	31	2
52	8	1	1	9	1	1	0	0	1	36	16	44	1
53	42	0	0	10	1	1	0	1	0	59	32	64	0
54	42	0	0	10	1	1	0	1	0	32	16	52	0
55	42	0	1	10	1	1	0	0	1	55	22	62	0
56	42	0	1	10	1	1	0	0	1	33	20	41	0
57	1	1	1	10	1	1	0	0	1	42	6	43	5
58	3	1	1	10	1	1	0	0	1	30	11	13	7
59	42	0	1	11	0	1	0	1	0	40	20	56	0
60	40	0	1	10	1	1	0	0	0	52	42	72	0
61	40	0	1	10	0	1	0	0	0	49	30	4	0
62	40	0	1	9	1	1	0	0	0	40	15	53	0
63	2	1	1	9	1	1	1	0	1	49	27	48	1
64	10	1	1	8	1	1	0	0	0	41	14	39	1
65	39	0	1	10	1	1	1	0	1	49	19	53	0
66	38	0	1	9	1	1	0	0	0	36	14	64	0
67	19	1	1	8	1	1	0	0	0	50	12	44	0
68	2	1	1	10	1	1	0	0	1	34	9	31	0
69	38	0	1	11	0	1	0	0	0	44	32	47	0
70	37	0	0	10	1	1	0	0	0	63	22	59	0

71	37	0	0	10	1	1	0	0	0	54	21	63	0
72	38	0	1	10	1	1	0	0	0	53	23	42	0
73	5	1	1	9	1	1	0	0	0	30	24	45	2
74	25	1	1	11	1	1	0	1	1	50	16	58	3
75	37	0	1	8	0	1	0	0	0	40	19	47	0
76	37	0	1	8	0	1	0	0	0	56	42	54	0
77	33	0	1	10	1	1	0	0	0	30	18	31	0
78	33	0	1	9	1	1	0	1	1	40	17	40	0
79	33	0	0	8	1	1	0	1	0	30	7	8	0
80	31	0	0	9	0	0	0	1	0	40	13	75	0
81	31	0	0	9	0	0	0	0	0	51	27	51	0
82	31	0	1	9	0	0	0	0	0	45	43	48	0
83	30	0	1	10	1	1	0	1	0	62	40	60	0

APPENDIX B.1

```
. summarize Time, detail

                            Time
-------------------------------------------------------------
      Percentiles      Smallest
 1%             1              1
 5%             3              1
10%             4              2       Obs                  83
25%            17              2       Sum of Wgt.          83

50%            40                      Mean           41.46988
                        Largest         Std. Dev.      28.76242
75%            55             92
90%            91             92       Variance       827.2765
95%            92             98       Skewness       .3875198
99%            98             98       Kurtosis       2.205557
```

APPENDIX B.2

```
. stdes

        failure _d:  Cense
  analysis time _t:  Time
```

Category	total	per subject: mean	min	median
no. of subjects	83			
no. of records	83	1	1	1
(first) entry time		0	0	0
(final) exit time		41.46988	1	40
subjects with gap	0			
time on gap if gap	0			
time at risk	3442	41.46988	1	40
failures	30	.3614458	0	0

APPENDIX C

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf. Int.]	
1	83	2	0	0.9759	0.0168	0.9071	0.9939
2	81	2	0	0.9518	0.0235	0.8767	0.9816
3	79	3	0	0.9157	0.0305	0.8312	0.9589
4	76	4	0	0.8675	0.0372	0.7735	0.9243
5	72	1	0	0.8554	0.0386	0.7595	0.9152
6	71	2	0	0.8313	0.0411	0.7319	0.8964
8	69	1	0	0.8193	0.0422	0.7183	0.8869
10	68	3	0	0.7831	0.0452	0.6781	0.8574

11	**65**	**1**	**0**	**0.7711**	**0.0461**	**0.6649**	**0.8474**
13	**64**	**1**	**0**	**0.7590**	**0.0469**	**0.6518**	**0.8373**
17	**63**	**1**	**0**	**0.7470**	**0.0477**	**0.6389**	**0.8270**
18	**62**	**1**	**0**	**0.7349**	**0.0484**	**0.6259**	**0.8167**
19	**61**	**2**	**0**	**0.7108**	**0.0498**	**0.6004**	**0.7959**
25	**59**	**1**	**0**	**0.6988**	**0.0504**	**0.5877**	**0.7853**
30	**58**	**0**	**1**	**0.6988**	**0.0504**	**0.5877**	**0.7853**
31	**57**	**0**	**3**	**0.6988**	**0.0504**	**0.5877**	**0.7853**
33	**54**	**1**	**3**	**0.6859**	**0.0511**	**0.5739**	**0.7741**
37	**50**	**0**	**4**	**0.6859**	**0.0511**	**0.5739**	**0.7741**
38	**46**	**0**	**3**	**0.6859**	**0.0511**	**0.5739**	**0.7741**
39	**43**	**0**	**1**	**0.6859**	**0.0511**	**0.5739**	**0.7741**
40	**42**	**0**	**3**	**0.6859**	**0.0511**	**0.5739**	**0.7741**
41	**39**	**1**	**0**	**0.6683**	**0.0527**	**0.5535**	**0.7598**
42	**38**	**0**	**5**	**0.6683**	**0.0527**	**0.5535**	**0.7598**
43	**33**	**0**	**1**	**0.6683**	**0.0527**	**0.5535**	**0.7598**
44	**32**	**0**	**1**	**0.6683**	**0.0527**	**0.5535**	**0.7598**
45	**31**	**0**	**1**	**0.6683**	**0.0527**	**0.5535**	**0.7598**
47	**30**	**1**	**0**	**0.6460**	**0.0554**	**0.5262**	**0.7428**
50	**29**	**1**	**1**	**0.6237**	**0.0578**	**0.4997**	**0.7253**
51	**27**	**0**	**4**	**0.6237**	**0.0578**	**0.4997**	**0.7253**
53	**23**	**0**	**1**	**0.6237**	**0.0578**	**0.4997**	**0.7253**
54	**22**	**1**	**0**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
55	**21**	**0**	**1**	**0.5954**	**0.0618**	**0.4642**	**0.7044**

58	**20**	**0**	**1**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
67	**19**	**0**	**4**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
78	**15**	**0**	**1**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
80	**14**	**0**	**1**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
82	**13**	**0**	**1**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
84	**12**	**0**	**1**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
86	**11**	**0**	**1**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
87	**10**	**0**	**1**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
91	**9**	**0**	**1**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
92	**8**	**0**	**6**	**0.5954**	**0.0618**	**0.4642**	**0.7044**
98	**2**	**0**	**2**	**0.5954**	**0.0618**	**0.4642**	**0.7044**

APPENDIX D1

```
Cox regression -- Breslow method for ties

No. of subjects =           83                 Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                               LR chi2(1)       =        3.11
Log likelihood  =   -122.61236                 Prob > chi2      =      0.0778
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Parentinvo	.4244778	.2281097	-1.59	0.111	.1480569	1.216974

APPENDIX D2

Cox regression -- Breslow method for ties

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	22.78
Log likelihood =	-112.7774	Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Substance	9.32069	5.68026	3.66	0.000	2.82296	30.77453

APPENDIX D3

Cox regression -- Breslow method for ties

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	9.82
Log likelihood =	-119.25664	Prob > chi2 =	0.0017

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Repeater	4.800603	2.928533	2.57	0.010	1.452224	15.86931

APPENDIX D4

Cox regression -- Breslow method for ties

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	0.29
Log likelihood =	-124.02408	Prob > chi2 =	0.5924

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Homeloc	.6581636	.4853457	-0.57	0.571	.1551091	2.792739

APPENDIX D5

```
Cox regression -- Breslow method for ties

No. of subjects =           83                  Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                                LR chi2(1)       =        0.09
Log likelihood  =   -124.12385                  Prob > chi2      =      0.7678
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Hostelres	1.250183	.9159962	0.30	0.761	.2973769	5.255816

APPENDIX D6

```
Cox regression -- Breslow method for ties

No. of subjects =           83                  Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                                LR chi2(1)       =        3.70
Log likelihood  =   -122.31714                  Prob > chi2      =      0.0544
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Grade8	.3056149	.2240139	-1.62	0.106	.0726522	1.285583

APPENDIX D7

```
Cox regression -- Breslow method for ties

No. of subjects =           83                  Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                                LR chi2(1)       =        1.77
Log likelihood  =    -123.2804                  Prob > chi2      =      0.1829
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Grade9	1.662077	.6205188	1.36	0.174	.7995844	3.454922

APPENDIX D8

```
Cox regression -- Breslow method for ties

No. of subjects =           83                  Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                                LR chi2(1)       =        0.02
Log likelihood  =   -124.15786                  Prob > chi2      =      0.8900
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Grade10	1.051819	.3843511	0.14	0.890	.5139234	2.152699

APPENDIX D9

```
Cox regression -- Breslow method for ties

No. of subjects =           83                  Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                                LR chi2(1)       =        0.01
Log likelihood  =   -124.16034                  Prob > chi2      =      0.9053
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Grade11	.8877464	.9049548	-0.12	0.907	.1203891	6.546218

APPENDIX D.10

```
Cox regression -- Breslow method for ties

No. of subjects =           83                  Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                                LR chi2(1)       =       14.54
Log likelihood  =   -116.89652                  Prob > chi2      =      0.0001
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Gender	2.38e+15	4.19e+22	0.00	1.000	0	.

APPENDIX E.1

```
Cox regression -- Breslow method for ties

No. of subjects =           83                  Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                                LR chi2(1)       =        9.38
Log likelihood  =   -119.47793                  Prob > chi2      =      0.0022
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
English	.9613017	.0124748	-3.04	0.002	.9371598	.9860654

APPENDIX E.2

```
Cox regression -- Breslow method for ties

No. of subjects =           82                  Number of obs    =          82
No. of failures =           30
Time at risk    =         3375
                                                LR chi2(1)       =       10.96
Log likelihood  =   -118.17102                  Prob > chi2      =      0.0009
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Maths	.9337663	.0211191	-3.03	0.002	.8932777	.97609

APPENDIX E.3

```
Cox regression -- Breslow method for ties

No. of subjects =           83                  Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                                LR chi2(1)       =        9.89
Log likelihood  =   -119.22328                  Prob > chi2      =      0.0017
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
LO	.9673715	.0102278	-3.14	0.002	.9475318	.9876267

APPENDIX F.1

Cox regression -- Breslow method for ties

No. of subjects =	82	Number of obs =	82
No. of failures =	30		
Time at risk =	3375		
		LR chi2(10) =	51.68
Log likelihood =	-97.809795	Prob > chi2 =	0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Gender	35.8164	2.37e+07	0.00	1.000	-4.65e+07	4.65e+07
Grade	-.6395235	.3565957	-1.79	0.073	-1.338438	.0593912
Repeat	.6946347	.7506217	0.93	0.355	-.7765567	2.165826
Homeloc	-2.646942	1.014885	-2.61	0.009	-4.636079	-.6578038
Hostelres	-.8755268	.7560052	-1.16	0.247	-2.35727	.6062162
Parentinvo	-1.248254	.6343302	-1.97	0.049	-2.491518	-.0049892
Substance	2.369214	.7364418	3.22	0.001	.9258145	3.812613
English	.0553676	.0248681	2.23	0.026	.0066269	.1041082
Maths	-.1105124	.0398781	-2.77	0.006	-.188672	-.0323529
LO	-.0256985	.0169686	-1.51	0.130	-.0589563	.0075593

APPENDIX F2

Cox regression -- Breslow method for ties

No. of subjects =	82	Number of obs =	82
No. of failures =	30		
Time at risk =	3375		
		LR chi2(11) =	55.47
Log likelihood =	-95.917733	Prob > chi2 =	0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Gender	38.23602	.	.	.	.	.
Grade	-1.417874	.8774629	-1.62	0.106	-3.13767	.3019214
Repeat	.4377734	.7708686	0.57	0.570	-1.073101	1.948648
Homeloc	-2.380192	1.025545	-2.32	0.020	-4.390223	-.3701604
Hostelres	-.9702021	.7609451	-1.27	0.202	-2.461627	.5212228
Parentinvo	-13.70303	8.524299	-1.61	0.108	-30.41035	3.004287
Substance	-3.329293	8.677409	-0.38	0.701	-20.3367	13.67812
English	.0565412	.0270389	2.09	0.037	.003546	.1095364
Maths	-.1163634	.041288	-2.82	0.005	-.1972863	-.0354405
LO	-.0267153	.0174768	-1.53	0.126	-.0609692	.0075386
GradeParent	1.293354	.8586584	1.51	0.132	-.3895854	2.976294
GradeSub	.6807169	1.001069	0.68	0.497	-1.281343	2.642776

APPENDIX F.3

Cox regression -- Breslow method for ties

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(3) =	26.50
Log likelihood =	-110.91915	Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Grade	.2496424	.2381656	-1.45	0.146	.0384813	1.619524
Substance	7.56e-06	.0000638	-1.40	0.162	4.93e-13	115.7342
GradeSubst~e	4.897703	4.880375	1.59	0.111	.6947083	34.52887

APPENDIX F.4

Cox regression -- Breslow method for ties

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(3) =	5.83
Log likelihood =	-121.25232	Prob > chi2 =	0.1202

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Grade	.9594342	.2538404	-0.16	0.876	.571229	1.611462
Parentinvo	.0000789	.0004857	-1.53	0.125	4.51e-10	13.79631
GradeParent	2.448064	1.517956	1.44	0.149	.7261452	8.253196

APPENDIX G.1

Exponential regression -- log relative-hazard form

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	3.43
Log likelihood =	-107.39865	Prob > chi2 =	0.0639

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Parentinvo	.4082966	.2192904	-1.67	0.095	.1424975	1.169888
_cons	.0104	.0020396	-23.28	0.000	.0070811	.0152745

APPENDIX G.2

Weibull regression -- log relative-hazard form

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	3.22
Log likelihood =	-101.10096	Prob > chi2 =	0.0728

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Parentinvo	.4189547	.225058	-1.62	0.105	.1461877	1.200669
_cons	.0515893	.0216573	-7.06	0.000	.0226581	.1174616
/ln_p	-.5294901	.1663378	-3.18	0.001	-.8555062	-.2034739
p	.5889052	.0979572			.425068	.8158915
1/p	1.698066	.2824527			1.225653	2.352565

APPENDIX G.3

Gompertz regression -- log relative-hazard form

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	3.32
Log likelihood =	-97.598653	Prob > chi2 =	0.0686

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Parentinvo	.4140416	.2223837	-1.64	0.101	.1444973	1.186392
_cons	.0293235	.0081153	-12.75	0.000	.0170469	.0504414
/gamma	-.0493151	.0135702	-3.63	0.000	-.0759121	-.0227181

APPENDIX H.1

```
Exponential regression -- log relative-hazard form

No. of subjects =           83                 Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                               LR chi2(1)       =       27.13
Log likelihood  =    -95.54796                 Prob > chi2      =      0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Substance	10.97292	6.677907	3.94	0.000	3.328875	36.16987
_cons	.0015865	.0009159	-11.17	0.000	.0005117	.0049189

APPENDIX H.2

```
Weibull regression -- log relative-hazard form

No. of subjects =           83                 Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                               LR chi2(1)       =       24.89
Log likelihood  =    -90.26551                 Prob > chi2      =      0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Substance	10.13223	6.173275	3.80	0.000	3.069678	33.44394
_cons	.0072298	.005071	-7.03	0.000	.0018285	.0285868
/ln_p	-.4740418	.1610862	-2.94	0.003	-.789765	-.1583186
p	.6224812	.1002732			.4539514	.8535778
1/p	1.606474	.2587809			1.171539	2.202879

APPENDIX I.1

```
Exponential regression -- log relative-hazard form

No. of subjects =           83                 Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                               LR chi2(1)       =       14.20
Log likelihood  =   -102.01637                 Prob > chi2      =      0.0002
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Repeat	6.207659	3.777861	3.00	0.003	1.883229	20.46221
_cons	.0021352	.0012328	-10.65	0.000	.0006887	.0066204

APPENDIX I.2

```
Weibull regression -- log relative-hazard form

No. of subjects =           83                 Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                               LR chi2(1)       =       11.59
Log likelihood  =   -96.915901                 Prob > chi2      =      0.0007
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Repeat	5.387941	3.287897	2.76	0.006	1.629266	17.81778
_cons	.0102721	.0073355	-6.41	0.000	.0025339	.0416411
/ln_p	-.4759904	.1645985	-2.89	0.004	-.7985975	-.1533832
p	.6212695	.10226			.4499596	.8578009
1/p	1.609607	.264939			1.165772	2.222422

.

APPENDIX J.1

Exponential regression -- log relative-hazard form

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	0.96
Log likelihood =	-108.63714	Prob > chi2 =	0.3280

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Homeloc	.4491754	.3287627	-1.09	0.274	.1070041	1.885522
_cons	.0186916	.0132169	-5.63	0.000	.0046747	.0747372

APPENDIX J.2

Weibull regression -- log relative-hazard form

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	0.43
Log likelihood =	-102.49695	Prob > chi2 =	0.5136

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Homeloc	.5969936	.4387578	-0.70	0.483	.1413783	2.520905
_cons	.0707478	.0547318	-3.42	0.001	.0155313	.3222684
/ln_p	-.5280982	.1677123	-3.15	0.002	-.8568082	-.1993882
p	.5897255	.0989042			.4245149	.8192318
1/p	1.695704	.2843904			1.220656	2.35563

APPENDIX K.1

Exponential regression -- log relative-hazard form

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	0.00
Log likelihood =	-109.11539	Prob > chi2 =	0.9922

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Hostelres	.9928881	.7267197	-0.01	0.992	.2365291	4.167888
_cons	.00872	.0016479	-25.09	0.000	.0060208	.0126293

APPENDIX K.2

Weibull regression -- log relative-hazard form

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	0.02
Log likelihood =	-102.70047	Prob > chi2 =	0.8885

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Hostelres	1.109879	.8126698	0.14	0.887	.264249	4.661631
_cons	.0440129	.018361	-7.49	0.000	.0194304	.0996963
/ln_p	-.5372905	.1673734	-3.21	0.001	-.8653363	-.2092447
p	.5843293	.0978012			.42091	.8111967
1/p	1.711364	.2864367			1.232747	2.375805

APPENDIX L

```
Exponential regression -- log relative-hazard form

No. of subjects =           83                 Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                               LR chi2(1)       =       14.84
Log likelihood  =   -101.69778                 Prob > chi2      =      0.0001
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Gender	4185539	2.95e+09	0.02	0.983	0	.
_cons	2.67e-09	1.88e-06	-0.03	0.978	0	.

APPENDIX M.1

```
Exponential regression -- log relative-hazard form

No. of subjects =           83                 Number of obs    =          83
No. of failures =           30
Time at risk    =         3442
                                               LR chi2(1)       =        1.63
Log likelihood  =   -108.30136                 Prob > chi2      =      0.2020
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Grade	1.330526	.3038533	1.25	0.211	.8504204	2.081675
_cons	.0005941	.0012935	-3.41	0.001	8.33e-06	.0423744

APPENDIX M.2

Weibull regression -- log relative-hazard form

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	1.01
Log likelihood =	-102.20722	Prob > chi2 =	0.3158

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Grade	1.253741	.2870177	0.99	0.323	.8004666	1.963688
_cons	.005127	.0114667	-2.36	0.018	.000064	.4107604
/ln_p	-.5260757	.167631	-3.14	0.002	-.8546263	-.197525
p	.5909194	.0990564			.4254421	.8207596
1/p	1.692278	.2836782			1.218384	2.350496

APPENDIX N.1

Weibull regression -- log relative-hazard form

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	10.42
Log likelihood =	-97.500775	Prob > chi2 =	0.0012

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
English	.9594007	.0124201	-3.20	0.001	.9353641	.9840551
_cons	.1837829	.1045375	-2.98	0.003	.0602745	.5603723
/ln_p	-.4999299	.1634549	-3.06	0.002	-.8202957	-.1795641
p	.6065732	.0991474			.4403014	.8356344
1/p	1.648606	.2694728			1.196696	2.271171

APPENDIX N.2

Weibull regression -- log relative-hazard form

No. of subjects =	82	Number of obs =	82
No. of failures =	30		
Time at risk =	3375		
		LR chi2(1) =	11.55
Log likelihood =	-96.412403	Prob > chi2 =	0.0007

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Maths	.9323578	.0210704	-3.10	0.002	.8919617	.9745833
_cons	.1436989	.0750458	-3.71	0.000	.051632	.3999337
/ln_p	-.5010866	.1632231	-3.07	0.002	-.820998	-.1811753
p	.6058719	.0988923			.4399923	.8342891
1/p	1.650514	.269402			1.198625	2.272767

APPENDIX O

Weibull regression -- log relative-hazard form

No. of subjects =	83	Number of obs =	83
No. of failures =	30		
Time at risk =	3442		
		LR chi2(1) =	10.84
Log likelihood =	-97.290265	Prob > chi2 =	0.0010

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
LO	.9654565	.0103746	-3.27	0.001	.9453353	.986006
_cons	.1502531	.0775836	-3.67	0.000	.0546139	.4133744
/ln_p	-.4988162	.163704	-3.05	0.002	-.8196701	-.1779624
p	.6072491	.0994091			.440577	.8369739
1/p	1.646771	.2695829			1.19478	2.269751

APPENDIX P.1

Weibull regression -- log relative-hazard form

No. of subjects = 82 Number of obs = 82
No. of failures = 30
Time at risk = 3375
LR chi2(12) = 60.67
Log likelihood = -71.855419 Prob > chi2 = 0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Gender	3587648	5.18e+09	0.01	0.992	0	.
Grade	.1953836	.1753295	-1.82	0.069	.0336553	1.134286
Repeat	1.655032	1.271141	0.66	0.512	.3673195	7.457076
Homeloc	.1070689	.1017663	-2.35	0.019	.0166195	.689777
Hostelres	.2745176	.2124882	-1.67	0.095	.060215	1.251515
Parentinvo	2.19e-07	1.79e-06	-1.87	0.061	2.38e-14	2.004524
Substance	.008467	.0741342	-0.54	0.586	2.98e-10	240214.1
English	1.054369	.0276188	2.02	0.043	1.001603	1.109915
Maths	.88434	.0369243	-2.94	0.003	.8148517	.959754
LO	.9735348	.0170728	-1.53	0.126	.9406412	1.007579
GradeParent	4.389278	3.612261	1.80	0.072	.8747347	22.02469
GradeSub	2.33234	2.363464	0.84	0.403	.3200599	16.99622
_cons	.0727185	104.8953	-0.00	0.999	0	.
/ln_p	-.2511228	.1515923	-1.66	0.098	-.5482381	.0459926
p	.7779269	.1179277			.5779672	1.047067
1/p	1.285468	.194867			.955049	1.730202

.

APPENDIX P2

Exponential regression -- log relative-hazard form

No. of subjects =	82	Number of obs =	82
No. of failures =	30		
Time at risk =	3375		
		LR chi2(12) =	70.25
Log likelihood =	-73.400162	Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Gender	2876512	3.75e+09	0.01	0.991	0	.
Grade	.1703106	.1526505	-1.97	0.048	.0293971	.9866862
Repeat	1.592268	1.230807	0.60	0.547	.3499823	7.244131
Homeloc	.0623134	.059045	-2.93	0.003	.0097281	.3991508
Hostelres	.2169307	.1672631	-1.98	0.047	.0478638	.9831843
Parentinvo	7.47e-08	6.21e-07	-1.97	0.048	6.28e-15	.8895281
Substance	.0049791	.043681	-0.60	0.546	1.70e-10	146090.7
English	1.063945	.0282527	2.33	0.020	1.009988	1.120786
Maths	.8661981	.0367377	-3.39	0.001	.7971051	.94128
LO	.9703673	.0173239	-1.68	0.092	.9370002	1.004923
GradeParent	4.85688	4.059837	1.89	0.059	.9437199	24.99607
GradeSub	2.527121	2.563967	0.91	0.361	.3459506	18.46028
_cons	.2477727	323.098	-0.00	0.999	0	.

APPENDIX P3

Gompertz regression -- log relative-hazard form

No. of subjects = 82 Number of obs = 82
No. of failures = 30
Time at risk = 3375
LR chi2(12) = 59.15
Log likelihood = -69.195338 Prob > chi2 = 0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Gender	5171484	8.47e+09	0.01	0.992	0	.
Grade	.2252984	.2001049	-1.68	0.093	.039513	1.284624
Repeat	1.697999	1.306909	0.69	0.492	.3756541	7.675151
Homeloc	.1061301	.0992019	-2.40	0.016	.0169904	.6629388
Hostelres	.3465424	.2650403	-1.39	0.166	.0774009	1.551553
Parentinvo	3.84e-07	3.22e-06	-1.76	0.078	2.86e-14	5.164331
Substance	.0376602	.3285539	-0.38	0.707	1.41e-09	1004406
English	1.054193	.027161	2.05	0.041	1.00228	1.108794
Maths	.8916483	.0359419	-2.85	0.004	.8239145	.9649506
LO	.9738822	.0165512	-1.56	0.119	.9419768	1.006868
GradeParent	4.088064	3.446807	1.67	0.095	.7831347	21.34023
GradeSub	1.968912	1.985108	0.67	0.502	.2729122	14.20462
_cons	.0107822	17.65	-0.00	0.998	0	.
/gamma	-.0315595	.0125046	-2.52	0.012	-.056068	-.007051

APPENDIX Q

Poisson regression — Number of obs = 82; LR chi2(10) = 157.25; Prob > chi2 = 0.0000; Log likelihood = -126.43462; Pseudo R2 = 0.3834

Offences	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Parentinvo	-.165455	.2504684	-0.66	0.509	-.656364	.325454
Substance	2.710327	.5293215	5.12	0.000	1.672876	3.747778
Repeat	.7164742	.3894079	1.84	0.066	-.0467513	1.4797
Gender	16.48367	1101.215	0.01	0.988	-2141.858	2174.825
Homeloc	-.3945362	.3627748	-1.09	0.277	-1.105562	.3164894
Hostelres	-.2296889	.337293	-0.68	0.496	-.890771	.4313931
Grade	-.4806187	.1861726	-2.58	0.010	-.8455104	-.1157271
English	.0194696	.0114452	1.70	0.089	-.0029625	.0419018
Maths	-.0200823	.016354	-1.23	0.219	-.0521356	.0119709
LO	-.0295597	.0080312	-3.68	0.000	-.0453005	-.013819
_cons	-13.15619	1101.216	-0.01	0.990	-2171.501	2145.188

APPENDIX R

Negative binomial regression — Number of obs = 82; LR chi2(10) = 46.54; Prob > chi2 = 0.0000; Dispersion = mean; Log likelihood = -99.203408; Pseudo R2 = 0.1900

Offences	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Parentinvo	-.3212138	.5382178	-0.60	0.551	-1.376101	.7336737
Substance	2.834285	.6655997	4.26	0.000	1.529733	4.138836
Repeat	.6800813	.5727279	1.19	0.235	-.4424448	1.802607
Gender	16.88825	1374.602	0.01	0.990	-2677.283	2711.059
Homeloc	-.5021517	.9387556	-0.53	0.593	-2.342079	1.337776
Hostelres	-.5201096	.7522747	-0.69	0.489	-1.994541	.9543217
Grade	-.5402083	.347973	-1.55	0.121	-1.222223	.1418062
English	.0075458	.024033	0.31	0.754	-.039558	.0546497
Maths	-.0197949	.0317491	-0.62	0.533	-.082022	.0424323
LO	-.0200552	.0171365	-1.17	0.242	-.0536422	.0135318
_cons	-12.82889	1374.606	-0.01	0.993	-2707.008	2681.35

APPENDIX S1

Zero-inflated Poisson regression — Number of obs = 82; Nonzero obs = 30; Zero obs = 52

Inflation model = logit — LR chi2(10) = 25.08; Log likelihood = -97.89811 — Prob > chi2 = 0.0052

Offences	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Offences						
Parentinvo	.1642542	.2748695	0.60	0.550	-.37448	.7029885
Substance	1.848383	.7629644	2.42	0.015	.3530007	3.343766
Repeat	.3945428	.3588448	1.10	0.272	-.30878	1.097866
Gender	.694675	.	.	.	.	.
Homeloc	.3264927	.3995481	0.82	0.414	-.4566071	1.109593
Hostelres	-.0779708	.3669635	-0.21	0.832	-.797206	.6412643
Grade	-.3968524	.2502567	-1.59	0.113	-.8873466	.0936418
English	.0035358	.0120216	0.29	0.769	-.0200261	.0270976
Maths	.0418832	.021541	1.94	0.052	-.0003364	.0841028
LO	-.0267041	.0097114	-2.75	0.006	-.045738	-.0076701
_cons	2.082199	2.523395	0.83	0.409	-2.863564	7.027963
inflate						
Parentinvo	1.773836	1.130171	1.57	0.117	-.441258	3.988931
Substance	-2.327135	1.366688	-1.70	0.089	-5.005794	.3515235
Repeat	-.669535	1.039756	-0.64	0.520	-2.70742	1.36835
Gender	-15.21161	834.8933	-0.02	0.985	-1651.572	1621.149
Homeloc	3.439776	2.846658	1.21	0.227	-2.139571	9.019123
Hostelres	1.471009	1.218915	1.21	0.228	-.9180212	3.860039
Grade	.5946727	.588812	1.01	0.313	-.5593777	1.748723
English	-.0556506	.0520591	-1.07	0.285	-.1576846	.0463834
Maths	.1489674	.0638327	2.33	0.020	.0238575	.2740772
LO	.0262854	.0327497	0.80	0.422	-.0379028	.0904737
_cons	6.199595	834.9142	0.01	0.994	-1630.202	1642.601

Vuong test of zip vs. standard Poisson: z = 2.14 Pr>z = 0.0162

APPENDIX S2

Zero-inflated Poisson regression — Number of obs = 82; Nonzero obs = 30; Zero obs = 52

Inflation model = logit — LR chi2(10) = 36.38; Log likelihood = -100.5085 — Prob > chi2 = 0.0001

Offences	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Offences						
Parentinvo	.0490693	.2677751	0.18	0.855	-.4757603	.573899
Substance	2.32541	.7745063	3.00	0.003	.8074052	3.843414
Repeat	.5284225	.3449069	1.53	0.126	-.1475826	1.204428
Gender	16.16029	877.5416	0.02	0.985	-1703.79	1736.11
Homeloc	.3314495	.392333	0.84	0.398	-.437509	1.100408
Hostelres	-.1959541	.3879433	-0.51	0.613	-.9563089	.5644008
Grade	-.4353211	.2496165	-1.74	0.081	-.9245605	.0539183
English	.0064176	.0119455	0.54	0.591	-.0169952	.0298304
Maths	.0459252	.0212109	2.17	0.030	.0043525	.0874979
LO	-.0304492	.0093317	-3.26	0.001	-.048739	-.0121595
_cons	-13.64815	877.545	-0.02	0.988	-1733.605	1706.308
inflate						
Substance	-.7471311	1.436776	-0.52	0.603	-3.563159	2.068897
Maths	.1227515	.0486523	2.52	0.012	.0273948	.2181082
_cons	-2.247129	1.841398	-1.22	0.222	-5.856202	1.361944

Vuong test of zip vs. standard Poisson: z = 1.99 Pr>z = 0.0234

.

APPENDIX S3

Zero-inflated Poisson regression — Number of obs = 82; Nonzero obs = 30; Zero obs = 52

Inflation model = logit — LR chi2(10) = 58.19; Log likelihood = -100.6283 — Prob > chi2 = 0.0000

Offences	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Offences						
Parentinvo	.0416197	.266645	0.16	0.876	-.480995	.5642343
Substance	2.565656	.5889672	4.36	0.000	1.411302	3.72001
Repeat	.5382681	.3436882	1.57	0.117	-.1353484	1.211885
Gender	18.01263	2233.144	0.01	0.994	-4358.87	4394.895
Homeloc	.3135218	.3897999	0.80	0.421	-.450472	1.077516
Hostelres	-.1918055	.386566	-0.50	0.620	-.949461	.5658501
Grade	-.456111	.2438843	-1.87	0.061	-.9341155	.0218935
English	.0068963	.0119024	0.58	0.562	-.016432	.0302247
Maths	.0442013	.0209672	2.11	0.035	.0031063	.0852962
LO	-.0304577	.0093153	-3.27	0.001	-.0487152	-.0122001
_cons	-15.52051	2233.146	-0.01	0.994	-4392.406	4361.365
inflate						
Maths	.1281987	.0484088	2.65	0.008	.0333192	.2230782
_cons	-3.052728	1.061883	-2.87	0.004	-5.13398	-.9714766

Vuong test of zip vs. standard Poisson: z = 1.99 Pr>z = 0.0235

APPENDIX S4

Zero-inflated Poisson regression — Number of obs = 82; Nonzero obs = 30; Zero obs = 52

Inflation model = logit — LR chi2(10) = 35.18; Log likelihood = -104.9541 — Prob > chi2 = 0.0001

Offences	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Offences						
Parentinvo	.0856258	.2801923	0.31	0.760	-.463541	.6347926
Substance	2.430024	1.003688	2.42	0.015	.4628319	4.397217
Repeat	.4852762	.3656306	1.33	0.184	-.2313467	1.201899
Gender	15.5475	613.0347	0.03	0.980	-1185.978	1217.073
Homeloc	.248662	.3970288	0.63	0.531	-.5295001	1.026824
Hostelres	-.1160658	.3684159	-0.32	0.753	-.8381476	.6060161
Grade	-.5025152	.253125	-1.99	0.047	-.9986311	-.0063993
English	.0063213	.0119703	0.53	0.597	-.01714	.0297826
Maths	.0338892	.0218534	1.55	0.121	-.0089426	.076721
LO	-.0283123	.009676	-2.93	0.003	-.0472769	-.0093476
_cons	-12.25499	613.0399	-0.02	0.984	-1213.791	1189.281
inflate						
Substance	-1.195751	1.522819	-0.79	0.432	-4.180421	1.788919
_cons	.4995216	1.489687	0.34	0.737	-2.420211	3.419255

Vuong test of zip vs. standard Poisson: z = 1.64 Pr>z = 0.0506