Metodološki zvezki,

# A Longitudinal Study of Student Performance in English using Repeated Measures, Multilevel and Logistic Regression Models

Liberato Camilleri[1] and Georgiana Xuereb[2]

**Abstract**

This paper presents three statistical models that analyze longitudinal data on student performance in English. A random sample, comprising male and female students who attend either a state or a private school, was selected to investigate gender and school bias in this subject. The English annual marks attained by each student were recorded during the last three years in primary schools. In the first approach, we present a repeated measures analysis of variance that captures the correlation between the repeated measures. Several tests are carried out to check for within subjects and between subjects effects; equality of covariance matrices and sphericity. In the second approach, we fit a two-level random coefficient model to examine the effect of time on student performance in English. This model allows the student-specific coefficients describing individual trajectories to vary randomly. In the third approach, we fit a Logistic regression model to estimate the probability of passing the Eleven-Plus examination that students sit for when they terminate Primary education.

## 1 Introduction

The Junior Lyceum eleven-plus examination, in Malta, is an important benchmark of academic performance as students proceed from Primary to Secondary school. This examination assesses students on five subjects, including English and Mathematics. To attend a Junior Lyceum a student has to pass in all five subjects. These Junior Lyceums are rated amongst the best schools on the island where students receive a sound education in all aspects of their development. A prevailing outcome of this examination is that females outperform their male counterpart in the

[1] University of Malta, Department of Statistics, Malta; liberato.camilleri@um.edu.mt
[2] University of Malta, Department of Statistics, Malta; georgiana.xuereb@um.edu.mt

languages, while almost matching them in Mathematics. Although students sit in the same classroom, read the same textbooks and listen to the same teacher, the percentage of successful females in English has exceeded the percentage of successful male students by more than 5% for several years. Another notable contrast is the discrepancy in performance between students attending private schools and those attending state schools. Each year, approximately 75% of the students attending private schools do get a pass in English. This contrasts considerably with a 50% pass rate for students attending state schools. This performance discrepancy is more attributed to diverse socio-economic family status rather than different teaching methods. It is known that parents with a high educational background and high socio-economic status are more likely to send their children to private schools rather than state schools. These parents, besides affording to pay for the school fees, are more likely to prioritize their children's educational upbringing. A study conducted in 2006 revealed that 38.6% of the parents whose children attended private schools had a professional occupation. This contrasted considerably with 8.4% of the parents whose children attended state schools.

# 2 A Repeated Measures Model

Suppose we have $N$ individuals and for the $i^{th}$ individual we have $n_i$ observations. Let $\mathbf{y}_i$ denote the vector of $n_i$ responses for the $i^{th}$ individual and let $\mathbf{y}$ denote the vector of responses for all individuals. A Normal linear regression model for $\mathbf{y}$ is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \mathbf{V}) \tag{2.1}$$

$\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is a vector of fixed effects parameters, $\boldsymbol{\varepsilon}$ is a vector of error terms and $\mathbf{V}$ is the variance-covariance matrix. $\mathbf{V}$ has a block diagonal form if we assume that the responses for different individuals are independent. The probability density function for $\mathbf{y}_i$ is given by:

$$f(\mathbf{y}_i) = |2\pi\mathbf{V}_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\} \tag{2.2}$$

Suppose that $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)'$ is a random sample from $N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ then the likelihood function is given by:

$$l(\mathbf{y}) = \sum_{i=1}^{N} \log f(\mathbf{y}_i) = \sum_{i=1}^{N} \log|2\pi\mathbf{V}_i| - \frac{1}{2}\left\{\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\} \tag{2.3}$$

$\boldsymbol{\beta}$ can be estimated by maximizing the log-likelihood function given that the elements of **V** are known. Thus, the score equations become

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \mathbf{X}_i^{'} \mathbf{V}_i^{-1} \left( \mathbf{y}_i - \mathbf{X}_i^{'} \boldsymbol{\beta} \right) = \mathbf{X}^{'} \mathbf{V}^{-1} \left( \mathbf{y} - \mathbf{X}^{'} \boldsymbol{\beta} \right) = \mathbf{0} \tag{2.4}$$

The maximum likelihood estimator of $\boldsymbol{\beta}$ is given by:

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^{N} \mathbf{X}_i^{'} \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^{N} \mathbf{X}_i^{'} \mathbf{V}_i^{-1} \mathbf{y}_i \right) = \left( \mathbf{X}^{'} \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^{'} \mathbf{V}^{-1} \mathbf{y} \tag{2.5}$$

The sampling distribution of $\hat{\boldsymbol{\beta}}$ has an asymptotic multivariate normal distribution where

$$\hat{\boldsymbol{\beta}} \sim N\left[ \boldsymbol{\beta},\ \left( \mathbf{X}^{'} \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \right] \tag{2.6}$$

In practice, $\mathbf{V}_i$ is unknown and has to be estimated from the data by an iterative process. There are several forms for the matrix $\mathbf{V}_i$.

The simplest structure for $\mathbf{V}_i$ is given by (2.7). This pattern assumes that the correlation is constant regardless of the distance between the measurements. This equicorrelation matrix is said to have compound symmetry when $\rho = \sigma_a^2 \big/ \left( \sigma_a^2 + \sigma_b^2 \right)$ and $\sigma^2 = \sigma_a^2 + \sigma_b^2$.

$$\mathbf{V}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix} \tag{2.7}$$

The autoregressive structure given by (2.8) has homogeneous variances and correlations that decrease exponentially with distance.

$$\mathbf{V}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \rho \\ \rho^{n-1} & \rho^{n-2} & \cdots & \rho & 1 \end{bmatrix} \tag{2.8}$$

The Toeplitz structure given by (2.9) assumes that any pair of elements equidistant from the diagonal has the same correlation. The first-order autoregressive structure is a special case of the Toeplitz.

$$\mathbf{V}_i = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} \tag{2.9}$$

If $\mathbf{V}_i$ is assumed to be an unstructured correlation matrix then all the correlation terms may be different. The merit about this structure is that it involves no assumptions about the correlations. The demerit is that a large number of parameters have to be estimated when the order of $\mathbf{V}_i$ is large.

There are a number of criteria for selecting a covariance structure that trade-off between a simple structure and a complex one. Simple structures tend to yield incorrect estimates of the sampling variability that leads to misleading inferences, whereas, complex structures sacrifice power and efficiency. The information criteria that are used for the selection of the covariance structure are based on the bias-corrected log-likelihood and a penalty term. This penalty term measures the model complexity.

$$C = -2\log L(\mathbf{\Psi}) + dc \tag{2.10}$$

where $d$ is the number of estimated parameters and $c$ is the penalty constant. The well-known Akaike information criterion (AIC) arises when $c = 2$. The Bayesian information criterion (BIC) arises when $c = \log(N)$ where $N$ is the number of respondents. For $N$ > 8, BIC penalizes complex models more heavily than AIC. These criteria are functions of the log likelihood and can be compared across different models. Hence, one can fit several models, having different covariance structures, and compute their respective information criterion. The model with the smallest information criterion has the optimal covariance structure.

# 3 Results of the Repeated-Measures Model

The data set comprise the English annual examination marks of male and female students attending private or state schools. These marks were recorded during the last three years in Primary school. One of the aims of this study is to analyze this longitudinal data set through a Repeated Measures analysis to reveal correlations between successive assessments. The within-subject variable comprise the English marks during year 4, 5 and 6; whereas the between-subjects variables are the student gender and school type each having two levels. Several hypothesis tests are carried out to test for within subjects and between subjects effects; equality of covariance matrices and sphericity. Moreover, we examine the gender-school effect on student performance in English and fit models that predict the mean English annual marks.

An assumption that needs to be satisfied when conducting an ANOVA with a repeated measures factor is sphericity. Sphericity refers to the equality of the variances of the differences between levels of the repeated measures factor. Mauchly's test of sphericity verifies the null hypothesis that the variance-covariance matrix of the orthonormalized-transformed dependent variables is proportional to the identity matrix. The sphericity assumption is satisfied when the p-value exceeds 0.05 implying that the variances for each set of difference scores are equal.

**Table 1:** Mauchly's test of Sphericity

| Within Subject Effect | Mauchly's W | Chi-Square | df | P-value |
|---|---|---|---|---|
| English Marks | 0.998 | 0.603 | 2 | 0.740 |

The *p*-value, displayed in Table 1, exceeds the 0.05 level of significance implying that the variance-covariance structure is circular in form.

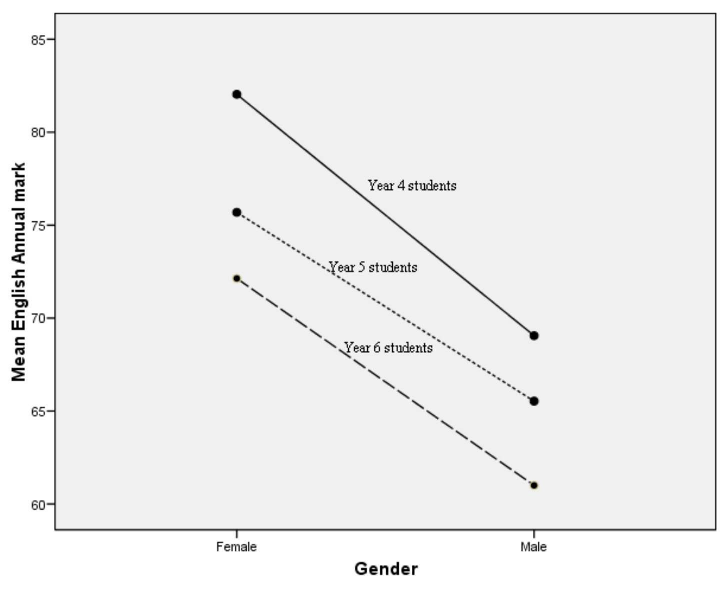


**Figure 1:** Line graphs displaying variations in English mean marks by Year and Gender

Figure 1 reveals a discrepancy in English language performance between males and females but exhibits no interaction effects between gender and English marks. Females perform better than their male counterparts, however, this performance deteriorates for both gender groups as students progress from year 4 to year 6.

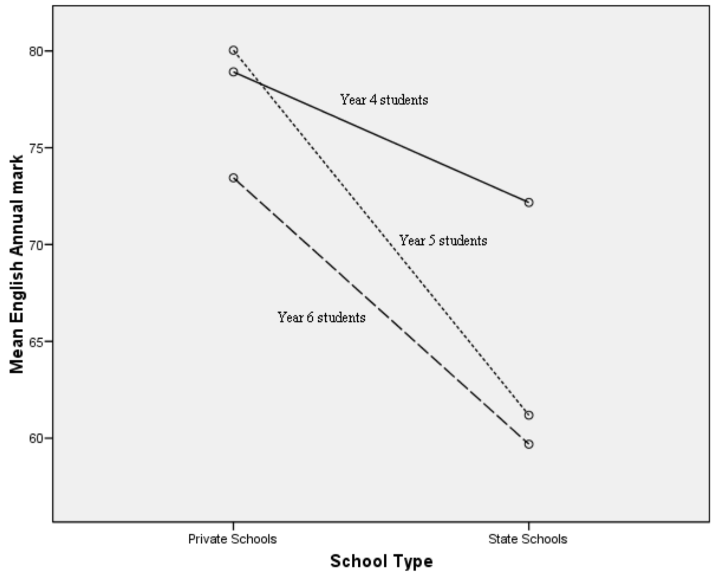


**Figure 2:** Line graphs displaying variations in English mean marks by Year and School

Figure 2 reveals a discrepancy in English language performance between students attending private and state schools. It also exhibits an interaction effects between school type and English marks. Students attending private schools achieve better results; however, the rate of deterioration in English language performance is more conspicuous for students attending state schools.

**Table 2:** Tests for Between-Subjects Effects

| Term | Sum of Squares | df | Mean Square | F | P-value |
|---|---|---|---|---|---|
| Gender | 7312.968 | 1 | 7312.968 | 10.856 | 0.001 |
| School | 9641.267 | 1 | 9641.267 | 14.312 | 0.000 |
| Gender*School | 724.218 | 1 | 724.218 | 1.075 | 0.301 |
| Error | 175146.590 | 260 | 673.641 | | |

Table 2 displays the analysis of variance for tests of between-subjects effects. It exhibits significant gender and school main effects.

**Table 3:** Tests for Within-Subjects Effects

| Term | Sum of Squares | df | Mean Square | F | P-value |
|---|---|---|---|---|---|
| Marks | 3022.268 | 2 | 1511.134 | 25.211 | 0.000 |
| Marks*Gender | 76.886 | 2 | 38.443 | 0.641 | 0.527 |
| Marks*School | 1379.831 | 2 | 689.916 | 11.51 | 0.000 |
| Marks*Gender*School | 130.548 | 2 | 65.274 | 1.089 | 0.337 |
| Error | 31168.187 | 520 | 59.939 | | |

Table 3 displays the analysis of variance for tests of within-subjects effects when the sphericity assumption is satisfied. The *F*-statistics for the English main effect and the English-school interaction effect are both significant. The non-parallel, well-separated line graphs displayed in figure 2 complement this result. Moreover, the *F*-statistic for the English-gender interaction effect is not significant and which complements the parallel line graphs displayed in figure 1.

**Table 4:** Correlation Table

| Year | Pearson Correlation | P-value |
|---|---:|---:|
| Year 4 – Year 5 | 0.811 | 0.000 |
| Year 4 – Year 6 | 0.768 | 0.000 |
| Year 5 – Year 6 | 0.822 | 0.000 |

Table 4 displays the correlation matrix for the three sets of English annual marks. As expected, the correlations between repeated measures are positive and decrease gradually with an increase in time separation.

**Table 5:** Summary of Information Criteria

| Covariance Structure | Deviance | No of parameters | AIC | BIC |
|---|---:|---:|---:|---:|
| Unstructured | 6115.0 | 6 | 6127.0 | 6148.4 |
| Compound Symmetry | 6122.5 | 2 | 6126.5 | 6133.6 |
| Toeplitz | 6121.2 | 3 | 6127.2 | 6137.9 |
| Autoregressive | 6155.1 | 2 | 6159.1 | 6166.2 |

The Akaike and Bayesian information criteria, displayed in table 5, identify the optimal covariance structure. Both criteria elicit the compound symmetry form as the optimal structure.

**Table 6:** Parameter Estimates for the Repeated Measures Model

| Effect | Gender | School | Year | Estimate | St Error | t-value | P-value |
|---|---|---|---|---:|---:|---:|---:|
| Intercept | | | | 55.8272 | 1.6283 | 34.29 | 0.0000 |
| Gender | Female | | | 7.8283 | 1.9930 | 3.93 | 0.0001 |
| Gender | Male | | | 0 | . | . | . |
| School | | Private | | 9.5213 | 5.5284 | 1.72 | 0.0862 |
| School | | State | | 0 | . | . | . |
| Year | | | Year 4 | 11.8902 | 0.6847 | 17.36 | 0.0000 |
| Year | | | Year 5 | 1.9394 | 0.6847 | 2.83 | 0.0048 |
| Year | | | Year 6 | 0 | . | . | . |
| Gender*School | Female | Private | | 7.1895 | 6.9340 | 1.04 | 0.3008 |
| Gender*School | Female | State | | 0 | . | . | . |
| Gender*School | Male | Private | | 0 | . | . | . |
| Gender*School | Male | State | | 0 | . | . | . |

Table 6 displays the parameter estimates for the Repeated Measures model. The coefficient of Gender indicates that the English annual mark for female students is, on average, 7.83 higher than for male students. Moreover, the coefficient of School indicates that students attending private schools score, on average, 9.52 more marks than students attending state schools. The mean English annual mark decreases by almost 12 points as the students progress from year 4 to year 6.

# 4 A Multilevel Model

An alternative approach for analyzing repeated measures data is to use hierarchical models. Multilevel models are hierarchical linear mixed models or random coefficient models that provide an extremely flexible approach to the analysis of longitudinal data by dropping the assumption of independence between the responses. These models facilitate the analysis of hierarchical data when observations are nested within higher levels of classification. Linear mixed models are linear in the parameters and the independent variables involve a mix of fixed and random effects. In general, linear mixed models for Normal responses can be expressed in the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{4.1}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects whereas, $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ are vectors of random effects. In addition, $\mathbf{y}$ is a vector of responses whereas, $\mathbf{X}$ and $\mathbf{Z}$ are both design matrices. The columns of $\mathbf{X}$ comprise the main effects and interactions that are included in the fixed part of the model whereas the columns of $\mathbf{Z}$ comprise the main effects and interactions that are included in the random part of the model. $\mathbf{X}\boldsymbol{\beta}$ is the fixed component and $\mathbf{Z}\boldsymbol{\eta}$ is the random component of the model. Both $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ are assumed to be independent and normally distributed random variables.

$$\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right) \text{ and } \boldsymbol{\eta} \sim N\left(\mathbf{0}, \boldsymbol{\Psi}\right) \tag{4.2}$$

The mean vector and the variance-covariance matrix for $\mathbf{y}$ are:

$$E\left(\mathbf{y}\right) = \mathbf{X}\boldsymbol{\beta} \text{ and } \operatorname{Var}\left(\mathbf{y}\right) = \mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}' + \boldsymbol{\Sigma} \tag{4.3}$$

Linear regression models are special cases of mixed models with $\mathbf{Z} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$. The parameters of interest are the elements of $\boldsymbol{\beta}$ and the variance and covariance elements in $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$. For Normal models, these can be estimated using either maximum likelihood or REML estimation.

It is often convenient to specify a linear mixed model in terms of an explicitly defined hierarchy of simpler models, which correspond to the levels of a clustered or longitudinal data set. When linear mixed models are specified in this way, they are referred to as hierarchical linear models or multilevel models. Let $i$ be the indicator for the level-1 units within the level-2 units, let $j$ be the indicator for the level-2 units and let $M$ be the number of random effects. A two-level random coefficient models is of the form:

$$y_{ij} = \underset{\text{Fixed part}}{\mathbf{X}_{ij}^{'}\boldsymbol{\beta}} + \underbrace{\sum_{m=1}^{M} \eta_{mj}^{(2)} z_{mij}^{(2)} + \varepsilon_{ij}}_{\text{Random part}} \tag{4.4}$$

$\eta_{mj}^{(2)}$ is a random coefficient at level 2 that allows the effects of $z_{mij}^{(2)}$ to vary between clusters $j$. $\boldsymbol{\eta}$ is a vector of random coefficients and has a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Psi}$

# 5 Results of the Multilevel Model

Growth curve or random coefficient models are used for the analysis of longitudinal data when the analyst is interested in modeling the effects of time at level 1 on a continuous dependent variable, and wishes to investigate the amount of between-subject variance in the effects of the variables across level 2 units. In this application, we fit a random coefficient model that examines the effect of age on student performance in English and simultaneously allows the student-specific coefficients describing individual trajectories to vary randomly. This hierarchical model has two levels of nesting, reflecting the contribution of the age level (level 1) and the student level (level 2). The student level 2 variables are gender and the type of school.

The **level-1 model** is:

$$y_{ij} = b_{0j} + b_{1j}\mathrm{A}_{ij} + \varepsilon_{ij} \tag{5.1}$$

$A_{ij} = 0$ (Year 4 student), $A_{ij} = 1$ (Year 5 student) and $A_{ij} = 2$ (Year 6 student)

The **level-2 model** is:

$$\begin{aligned} b_{0j} &= \beta_0 + \beta_1 \mathrm{G}_j + \beta_2 \mathrm{S}_j + \eta_{0j} \\ b_{1j} &= \beta_3 + \beta_4 \mathrm{G}_j + \beta_5 \mathrm{S}_j + \eta_{1j} \end{aligned} \tag{5.2}$$

$G_j = 0$ (Male student) and $G_j = 1$ (Female student)

$S_j = 0$ (State school) and $S_j = 1$ (Private school)

The combined model is given by:

$$y_{ij} = \underbrace{\beta_0 + \beta_1 G_j + \beta_2 S_j + \beta_3 A_{ij} + \beta_4 G_j.A_{ij} + \beta_5 S_j.A_{ij}}_{\text{Fixed Part}} + \underbrace{\eta_{0j} + \eta_{1j} A_{ij} + \varepsilon_{ij}}_{\text{Random Part}} \quad (5.3)$$

The parameters $\beta_0$ through $\beta_5$ represent the fixed effects associated with the intercept, the main effects and interaction terms in the model. $\beta_0$ is the mean predicted English mark for a male student in year 4 attending a state school. The fixed effects $\beta_1$ and $\beta_2$ respectively represent the difference in the intercept for the levels of gender and the levels of school. The fixed effect $\beta_3$ represents the rate of change of the mean English mark with age. The fixed effects $\beta_4$ and $\beta_5$ respectively represent the differences in the linear effect of age between the levels of gender and the levels of school. The terms $\eta_{0j}$ and $\eta_{1j}$ represent the random effects associated with the student-specific intercept and linear effect of age for student $j$. The term $\varepsilon_{ij}$ represents the residual associated with the observation at time $i$ on student $j$. It is assumed that these residuals are independent of the random effects.

$$\begin{pmatrix} \eta_{0j} \\ \eta_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{00} & \psi_{01} \\ \psi_{10} & \psi_{11} \end{pmatrix}\right] \text{ and } \varepsilon_{ij} \sim N\left(0, \sigma^2\right) \quad (5.4)$$

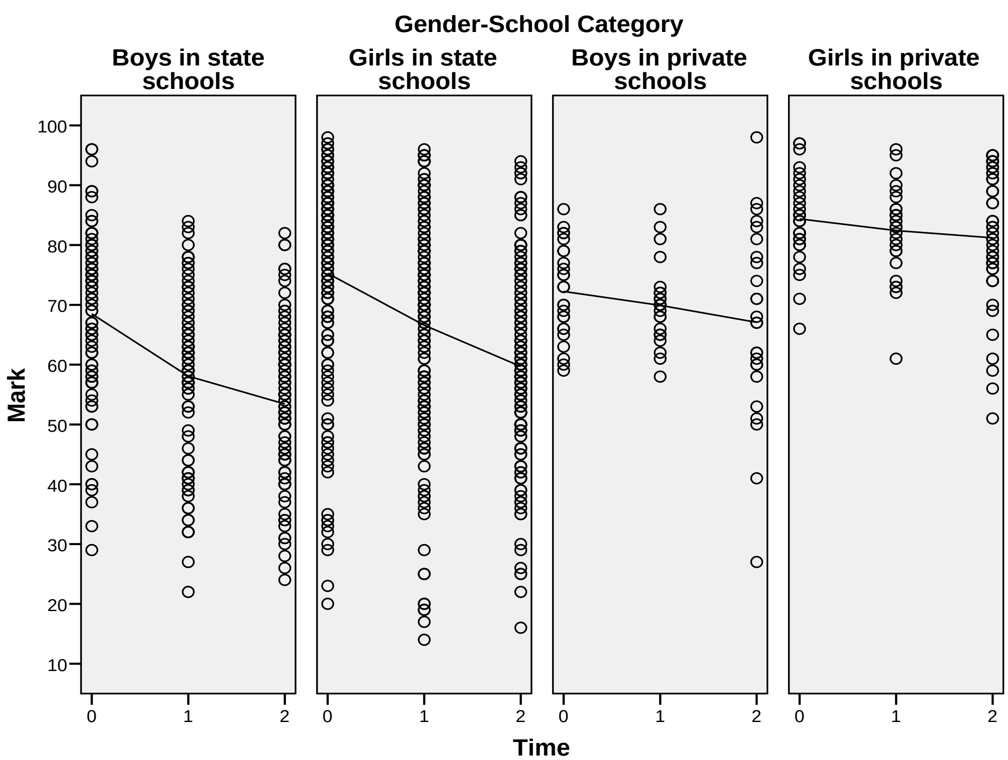


**Figure 3:** Mean English marks categorized by year, student gender and type of school

Figure 3 displays the observed English marks for individual students within each gender - school category. The plots show substantial variation in attainment between the students. Moreover, the between-student variability in the English marks tends to increase at each successive year of age, particularly for students attending private schools. The random coefficient model (5.3) accommodates these variations by the inclusion of the random effects $\eta_{0j}$ and $\eta_{1j}$. In addition, the marks of some students tend to decrease, as the students get older, whereas the marks for other students remain relatively constant. The between-student variability in the English marks differs considerably within each gender - school category. This is most evident when contrasting students attending state and private schools. The mean profiles displayed in Figure 3 show that the mean English mark generally decrease linearly with age.

**Table 7:** Estimated fixed effects for the Random Coefficient Model

| | Estimate | St Error | z | P-value | Lower Limit | Upper Limit |
|---|---|---|---|---|---|---|
| $\beta_0$ | 66.83926 | 1.474709 | 45.3 | 0.000 | 63.94888 | 69.72963 |
| $\beta_1$ | 8.536799 | 1.763042 | 4.84 | 0.000 | 5.081301 | 11.99230 |
| $\beta_2$ | 7.389658 | 2.158862 | 3.42 | 0.001 | 3.158365 | 11.62095 |
| $\beta_3$ | -7.695683 | 0.584691 | -13.2 | 0.000 | -8.841655 | -6.549711 |
| $\beta_4$ | -0.233419 | 0.704745 | -0.33 | 0.740 | -1.614695 | 1.147856 |
| $\beta_5$ | 6.011440 | 0.884542 | 6.80 | 0.000 | 4.277770 | 7.745110 |

**Table 8:** Variance-covariance parameter estimates

| | Estimate | St Error |
|---|---|---|
| $\sigma^2$ | 70.227 | 4.064 |
| $\psi_{00}$ | 167.86 | 17.96 |
| $\psi_{01}$ | -3.009 | 4.925 |
| $\psi_{11}$ | 0.054 | 0.173 |

Tables 7 displays the parameter estimates for the fixed effects and table 8 shows the variance-covariance parameter estimates. The level 1 variance (70.227) is a measure of the variation in the English marks throughout the years within students; whereas the level 2 variance (167.91) is interpreted as the variation in the English marks between students. The total variance is 238.14. The proportion of the total variance, which arises due to differences between student attainments, is 70.51%; whereas the remaining 29.49% of the total variance arises from year to year fluctuations of student performance in English.

# 6 A Logistic Regression Model

A further task is to examine a number of explanatory variables that predict the outcome of the Eleven-Plus English examination by fitting a Logistic regression model. A Logistic regression model is a probability model that relates a dichotomous response variable to a number of predictors. The model assumes a Bernoulli error distribution and a logit link function.

Let $Y$ be a discrete random variable with two possible outcomes

$$Y = \begin{cases} 1 & \text{if outcome is a success} \\ 0 & \text{if outcome is a failure} \end{cases} \tag{6.1}$$

$Y$ has a Bernoulli distribution with $P(Y=1)=p$. If there are $N$ independent random variables $Y_1, Y_2, \ldots, Y_n$ with $P(Y_i=1)=p_i$, then the likelihood function is given by:

$$L = \prod_{i=1}^{N} p_i^{y_i} \left(1-p_i\right)^{1-y_i} = \exp\left[\sum_{i=1}^{N} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{N} \log\left(1-p_i\right)\right] \tag{6.2}$$

The above distribution belongs to the exponential family, which is given by:

$$\exp\left[\sum_{i=1}^{N} y_i b\left(p_i\right) + \sum_{i=1}^{N} c\left(p_i\right) + \sum_{i=1}^{N} d\left(y_i\right)\right] \tag{6.3}$$

$b(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$ is the natural parameter and $c(p_i) = \log(1-p_i)$ is the cumulant function. The Logistic regression model that relates the proportion of successes $p_i$ to a number of predictors is given by:

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i\,'\boldsymbol{\beta} = \eta_i \tag{6.4}$$

$\boldsymbol{\beta}$ is a vector of parameters, $\mathbf{x}_i$ is a vector of explanatory variables and $\eta_i$ is the linear predictor. Since $Y_i$ has a Bernoulli distribution with parameter $p_i$ then

$$\mu_i = E\left(Y_i\right) = p_i = \frac{\exp\left(\eta_i\right)}{1+\exp\left(\eta_i\right)} \tag{6.5}$$

$$\operatorname{var}\left(Y_i\right) = p_i\left(1-p_i\right) = \frac{\exp\left(\eta_i\right)}{\left[1+\exp\left(\eta_i\right)\right]^2} \tag{6.6}$$

For generalized linear models, maximum likelihood estimates can be obtained by an iterative weighted least squares procedure using the iterative equation

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{z} \tag{6.7}$$

**X** is the design matrix and **W** is a diagonal matrix containing iterative weights $w_{ii}$.

$$w_{ii} = \frac{1}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{\exp(\eta_i)}{\left[1+\exp(\eta_i)\right]^2} = p_i(1-p_i) \tag{6.8}$$

The working variate **z** has elements

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) = \log\left(\frac{p_i}{1-p_i}\right) + \frac{y_i - p_i}{p_i(1-p_i)} \tag{6.9}$$

**W**, **z** and $\hat{\boldsymbol{\beta}}$ are updated at the each iteration using results of the previous iteration. $\hat{\boldsymbol{\beta}}^{(m)}$ is taken as the maximum likelihood estimate when the difference between the successive approximations $\hat{\boldsymbol{\beta}}^{(m-1)}$ and $\hat{\boldsymbol{\beta}}^{(m)}$ is sufficiently small.

# 7 Results of the Logistic Regression Model

The response variable for the Logistic regression model is the outcome of the Eleven-Plus English examination. The categories of this dichotomous variable are pass or fail. The three explanatory variables that were included to predict the Eleven-Plus English examination outcome were the Year 6 annual English mark ($x_1$), type of school attended by student ($x_2$) and student gender ($x_3$).

**Table 9:** Summary of the results using a forward procedure for model building

| Model | Deviance | Change in Deviance | Change in d.f | Pseudo R-square |
|---|---|---|---|---|
| $\beta_0$ | 334.797 | | | |
| $\beta_0 + \beta_1 x_1$ | 141.084 | 193.714 | 1 | 0.697 |
| $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ | 138.555 | 2.529 | 1 | 0.704 |
| $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ | 136.724 | 1.831 | 1 | 0.708 |

Table 9 displays the deviance and the Nagelkerke pseudo R-square value for each model fit. It is evident that $x_1$ contributes significantly in explaining variation in the responses; however, the contribution of $x_2$ and $x_3$ in improving the model fit is very small. The inclusion of these two predictors in the model fit results in a change in deviance that is less than $\chi^2_{0.05}(1) = 3.84$, implying that the parsimonious model is a one-predictor model.

**Table 10:** Parameter estimates, Odds Ratio and 95% Confidence Limits

| Term | Parameter Estimate | St. Error | Wald | df | P-value | Exp(B) | Lower Bound | Upper Bound |
|---|---|---|---|---|---|---|---|---|
| Intercept | -14.321 | 1.660 | 74.397 | 1 | 0.000 | | | |
| $x_1$ | 0.220 | 0.025 | 77.770 | 1 | 0.000 | 1.246 | 1.187 | 1.308 |

The odds ratio implies that the odds of passing the Eleven-plus English examination increases by 24.6% for every one-mark increment in the Year 6 English annual mark. Moreover, we are 95% confident that this odds ratio can vary between 18.7% and 30.8%.

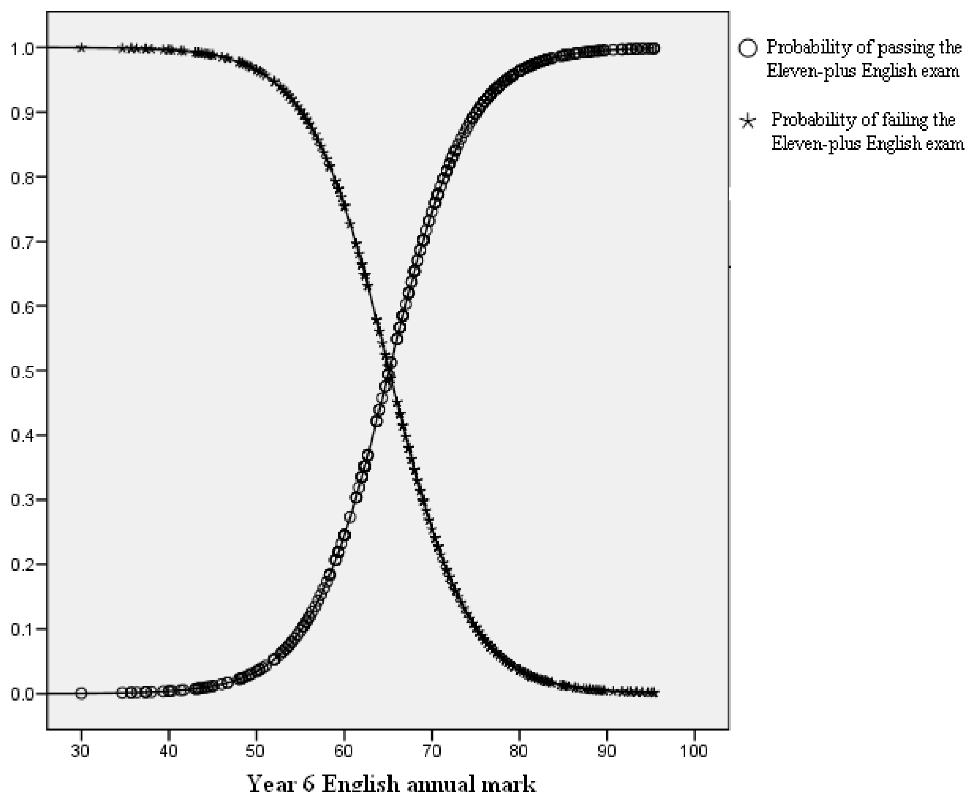


**Figure 4:** Probability curves for passing/failing the Eleven-Plus English examination

The probability curves show that students are more likely to pass, rather than fail, the Eleven-plus English examination if their Year 6 English annual mark exceeds 65. This indicates that Eleven-plus examinations are more stringent than the Year 6 annual examinations. To examine the predictive power of the logistic regression model the predicted outcome of the Eleven-plus examination was generated for each student using the estimated response probabilities. A predicted probability less than 0.5 correspond to an expected fail, whereas a predicted probability of at least 0.5 corresponds to an expected pass. Table 11 displays a crosstab of the classification by actual and predicted outcomes.

**Table 11:** Summary of the results using a forward procedure for model building

| | | | Predicted Outcome | | Total |
|---|---|---|---|---|---|
| | | | Pass | Fail | |
| Actual Outcome | Pass | Count | 163 | 19 | 182 |
| | | Percentage | 51.9% | 6.1% | 58.0% |
| | Fail | Count | 24 | 108 | 132 |
| | | Percentage | 7.6% | 34.4% | 42.0% |
| Total | | Count | 187 | 127 | 314 |
| | | Percentage | 59.6% | 40.4% | 100% |

The entries in the leading diagonal of table 11 correspond to a correct classification. The Logistic regression model correctly classifies 86.3% of the students.

**Table 12:** Correct/Incorrect classification by Year 6 English mark categories

| | | | Classification | | Total |
|---|---|---|---|---|---|
| | | | Correct | Incorrect | |
| Year 6 English annual mark categories | Less than 50 | Count | 42 | 1 | 43 |
| | | Percentage | 97.7% | 2.3% | 100% |
| | 50 - 59 | Count | 42 | 6 | 48 |
| | | Percentage | 87.5% | 12.5% | 100% |
| | 60 - 69 | Count | 46 | 28 | 74 |
| | | Percentage | 62.2% | 37.8% | 100% |
| | 70 - 79 | Count | 70 | 8 | 78 |
| | | Percentage | 89.7% | 10.3% | 100% |
| | 80 - 100 | Count | 71 | 0 | 71 |
| | | Percentage | 100.0% | 0.0% | 100% |
| Total | | Count | 271 | 43 | 314 |
| | | Percentage | 86.3% | 13.7% | 100% |

The highest percentage of misclassified students correspond to that category of students whose Year 6 English annual mark is in the range 60 to 69. This range includes the cut-off mark (65) which determines whether a student is expected to pass or fail the Eleven-plus English examination. Of the misclassified students all the students, except one, got a pass in the Year 6 annual exam but failed the Eleven-plus exam.

# References

[1] Agresti, A. (2002): *Categorical Data Analysis*. Wiley Inter Science.

[2] Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1994): *Statistical modeling in GLIM*. Oxford Science Publications.

[3] Dobson, A.J. (2002): *An Introduction to Generalized Linear Models*. Chapman & Hall (CRC Press)

[4] Hand, D. and Crowder, M. (1996): *Practical Longitudinal Data Analysis*. Chapman & Hall, London.

[5] McCullagh, P. and Nelder, J.A. (1998): *Generalized Linear Models*. Chapman & Hall/CRC.

[6] Molenberghs, G. and Verbeke, G. (2005): *Models for Discrete Longitudinal Data*. Springer-Verlag, Berlin.

[7] Nelder, J.A. and Wedderburn, R.W.M. (1972): Generalized Linear Models, Journal of the Royal Statistical Society, A, 135, 370-384.

[8] Prigibon, D. (1981): Logistic Regression Diagnostics. *Annals of Statistics*, **9**, 705-724.

[9] Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2001): GLLAMM: A General Class of Multilevel Models & Stata Program, *Multilevel Modelling Newsletter*, **13**, 17-23.

[10] Raudenbush, S.W., Bryk, A.S. (2002): *Hierarchical Linear Models: Data Analysis Methods and Applications*. Sage Publications, Newbury Park, CA.

[11] Searle, S., Casella, G. and McCulloch, C.E. (1992): *Variance Components*. John Wiley & Sons, New York.

[12] Skrondal, A. and Rabe-Hesketh, S. (2004): *Generalized Latent Variable Modelling*. Chapman & Hall/CRC.

[13] Snijders, T.A.B. and Bosker, R.J. (1999): *Multilevel Analysis: An Introduction to Basic Multilevel Modeling*. Sage Publications, Newbury Park, CA.

[14] Verbeke, G and Molenbergh, G. (2000): *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, Berlin.

[15] West, B.T., Welch, K.B. and Galecki, A.T. (2007): *Linear Mixed Models – A Practical Guide using Statistical Software*. Chapman & Hall/CRC.

[16] Zeger, S.L., Liang, K.Y., Albert, P. (1988): Models for Longitudinal Data: A Generalized estimating equation approach. *Biometrics*, **44**, 1049-1060.