School of Mathematics and Statistics

MSc STATISTICS
SEPTEMBER 2010

Forecasting the Demand for National Qualifications of the Scottish Qualifications Authority

By

Gillian Louise Boyle

Abstract

The Scottish Qualifications Authority (SQA) currently uses an *ad hoc* method to predict the number of entries for each of its examinations for the coming year, based on a weighted average of the numbers of entries in the previous three years. Data for the years 2004-9 were analysed with the aim of providing a more accurate method of prediction or, failing that, statistical justification for the prediction method currently used. The best method of prediction identified by this work would then be used by the SQA for planning purposes such as future resourcing and funding.

Standard multiple regression models are explored for predicting the number of entries from explanatory variables such as year (a simple linear trend) and total school roll. If the numbers of entries for the same subject in successive years are autocorrelated, then this information might be used to improve the model predictions. So the Durbin-Watson test is applied to determine whether first-order autocorrelation is present and, where there is evidence that it is, an autoregressive term is added to the models. The mean square error of prediction is used to compare the performance of different models, including the simple weighted average model currently favoured by the SQA.

The modelling has to be carried out with just 5 data points, for the years 2004 - 2008. Data from before 2004 can not be used because it was noticed that the entry profile for the intermediate courses had changed and is no longer useful when trying to model the courses in their current structure. In an attempt to overcome the deficiencies of standard parametric analysis when there are so few data points, bootstrapping is applied. Bias-corrected and accelerated bias-corrected percentile confidence limits are calculated for the Durbin-Watson statistic.

Data for seven school subjects are explored in detail: Accounting and Finance, Art, English, Mathematics, Physics, Psychology and Spanish. It is concluded that the current SQA method produces a prediction value that is always close to the actual number of entries but which can sometimes underestimate it. The prediction which produces a value close to the actual number of entries, but which never underestimates it, is the number of subjects enrolled by the schools and colleges.

Acknowledgements

The first and most important thank you must go to my supervisor Prof. John H. McColl, firstly for convincing me to do an MSc in the beginning but also for his guidance, support and patience. It has been both a pleasure and a privilege to work with him. My second thank you is to the SQA for providing me with the data to work with and also the funding to do so. I must mention specifically Susan Kirk and Marie McGhee whose advice and support have been invaluable throughout. I would also like to acknowledge the encouragement and help from my friends and family, whether it was in the form of an ear to vent my problems to, even when they didn't quite understand, or a laugh and a joke to see the day through with a smile, it was more than I could have hoped for. My last thank you goes to my parents. They have always been that voice in the back of my mind telling me to focus, work to the best of my ability and most importantly to never give up and for that I am eternally grateful.

Contents

Chapter 1

Introduction

1.1 The Scottish School System

In the current education system in Scotland, children usually attend primary school for seven years, from ages four and a half or five years to eleven and a half or twelve years. After that they attend a secondary school for between four and six years. For the first two years at secondary school, all pupils study a wide range of subjects. At the end of second year each pupil chooses a selection of, most commonly eight, subjects to study for usually a further two years. At the end of this period pupils generally sit examinations set by the Scottish Qualifications Authority. The most common type of examination sat is Standard Grade although recently there has been a change and it is becoming more common for pupils to sit Intermediate level National Courses at this stage. At the end of fourth year and beyond, if pupils choose to stay at school, they sit National Courses examinations at varying levels to gain National Qualifications (NQs). It is also possible to leave school and achieve the same qualifications at another educational institution. The level of National Course that is sat is usually determined by the pupil's performance in earlier examinations. After secondary school pupils can then elect to attend college or university. Entry to either is usually related to the performance of the pupil in Standard Grades and, most importantly, National Courses.

This thesis investigates how to predict the number of pupils who will be presented for every subject for all levels of NQ approximately 12 months in advance of the April/May examination diet.

1.2 The Scottish Qualifications Authority

The Scottish Qualifications Authority (SQA) is the main examination board within Scotland. The SQA is:

"an executive non-departmental public body (NDPB) sponsored by the Scottish

Government Schools Directorate."

The main aim of the SQA, as stated on their website is:

"to manage the qualifications system below degree level to allow students to fulfil their potential to participate in the economy, society and communities of Scotland."

The SQA have three main types of qualifications:

- Units
- Courses
- Group Awards

This report will only touch on Courses and further information is available on the other qualifications from the SQA. Courses can be split into Standard Grades, National Courses and Skills for Work. Skills for Work are practical courses with no final examination and are not included in this report. Standard Grades and National Courses are combined under the heading NQs.

As previously mentioned, it is common practice for Standard Grades to be taken over two years, typically in third and fourth year at secondary school. There are typically three levels of attainment possible for Standard Grade, each of which is associated with two grades:

- Foundation – Grade 5 or 6
- General – Grade 3 or 4
- Credit – Grade 1 or 2

Grade 1 is the highest possible award at the level of Standard Grade and Grade 6 the lowest, though it is also possible for Grade 7 or "No Award" to be recorded when a pupil's level of attainment falls below that required for Grade 6. "No Award" is also used if a pupil fails to attend for the final examination in the subject.

In almost all subjects, examinations are set at all three levels and pupils sit two of these: General and either Foundation or Credit. Based on their performance, pupils who sit the Foundation and General examinations can be awarded any of Grades 6 up to 3 but only pupils who sit General and Credit can be awarded Grades 1 and 2. However, this system does not apply to every Standard Grade subject. For example,

all candidates for Standard Grade Art sit the same assessment, though they are awarded Grades from 6 up to 1. Physics has no Foundation level examination, so all pupils sit the General and Credit papers but are again awarded Grades from 6 up to 1.

If a high enough level of attainment is reached and the pupil chooses to do so, the next step is National Courses. National Courses are split into 6 different levels:

- Access 2
- Access 3
- Intermediate 1
- Intermediate 2
- Higher
- Advanced Higher

Only the highest four levels of National Courses will be included in this report. Access courses are not included in the analysis as they are benchmarked against Standard Grade levels and do not include an external examination.

Qualifications at the four higher levels consist of three units, each assessed internally by a unit test, with a final external assessment – typically an examination – that tests material from the whole course. In order to achieve a qualification, a pupil must pass all the unit tests but the tests are not themselves graded, nor do they contribute to the grading of the award, which is determined wholly by the external assessment. This is graded A – D, with A being the best grade. A D is usually viewed as a narrow fail rather than a pass. "No award" is recorded for those whose level of attainment falls below that required for a D (including failing a unit test) or who fail to present themselves for the external assessment.

There are some National Courses, for example Skills for Work and Project-based National Courses, which have no examination. There is equivalence between Credit level Standard Grade and Intermediate 2 and also between General level Standard Grade and Intermediate 1. In recent years schools have exploited this equivalence by presenting pupils for Intermediate examinations instead of Standard Grades.

1.3 Data

1.3.1 Data Received

Data were received from the SQA in the form of numerous Excel worksheets. The

worksheets received were National Course Entries for 2004 – 2008, National Course Results for 2002 – 2008, Standard Grade Entries for 2004 – 2008 and Standard Grade Results for 2002 – 2008.

The Entries and Results worksheets both contained a number of "entries" for that year, split by subject and also by gender and stage of those entered, but the numbers in the two sets of files were not generally in agreement with each other. Each entries worksheet is a cut taken from the SQA's live database at the beginning of May and reflects the actual number of learners expected at that stage to sit an examination. The results files record the number of candidates who, when certification is completed in the August following the examination diet, are deemed to have entered for the examination. The causes of the differences between the two sets of figures are well understood by the SQA; for example, candidates who have a certified illness during the examination diet might not be included in the results file but candidates who attend the examination though not listed in advance by their school or college will be given a result. In this thesis, on the advice of the SQA, the "entries" figures from the entries files have been used in all calculations, except when calculating an award or pass rate; this is because the data is available much earlier than the results files but award or pass rate is lagged so the results files is available from the previous year.

Also received were data on the number of centres (i.e. schools or colleges) who had students wishing to sit each level of examination, the number of candidates enrolled by schools and colleges to sit each examination (i.e. the number of candidates at each school or college who, at the start of the year, wished to be considered to sit that level of examination), and the predicted entries for 2004 (using the actual entries from 2001, 2002 and 2003). The number of candidates enrolled to sit each examination, known as the School Prediction from here on, would be the most accurate "entries" data to use for prediction, however they are received too late for the SQA to use them for planning purposes.

In general, the number of passes and entries for the same subject at the same level can be used to calculate a pass rate. However, a particular problem arises with Standard Grades. It is not possible to calculate the exact Credit Pass Rate (for example), since the number of Standard Grade entries is not split by level so the exact number of Credit entries is not known. Therefore, as an alternative measure used for some

purposes in this thesis, a Credit Award Rate was calculated as the total number of Credit passes divided by the total number of Standard Grade entries. Clearly, this measure is not intended to be comparable to a pass rate in a National Course but it does allow some comparisons between Standard Grades in different subjects or different years.

National Ratings were also provided for every subject in every year. The National Rating is an index calculated by the SQA to provide a measure of the difficulty of a subject relative to other subjects at the same level. For example, a subject that has a National Rating of -0.50 is a subject in which candidates scored half a grade lower than the average of their other subjects. The SQA has a 'tolerance zone' which means that subjects should not have a National Rating outside the interval -0.5 to +0.5. It is expected that certain subjects, creative subjects such as music, will have National Ratings outside the 'tolerance zone'. This is accepted as it is thought that only musical students would choose that subject. Subjects with a low uptake may also have unusual National Ratings. The National Ratings are for internal use only, although more information is available in the statistics section of the SQA website (accessed via www.sqa.org.uk/sqa/CCC_FirstPage.jsp)

The data received contained information for all subjects available at each available level. It was decided to focus initially on only a few subjects and then try to extend any inferences made to the wider population of subjects. The SQA nominated the following seven subjects for initial investigation:

- English
- Mathematics
- Physics
- Spanish
- Art
- Accounting and Finance
- Psychology

English and Mathematics were chosen as they are the two most common subjects. In most schools it is compulsory for almost all pupils to take these subjects up to the end of fourth year and many students continue with English and Mathematics to National

Course level. The remainder of the subjects were chosen to give the widest possible range; a science, a language, a practical subject, a subject with a small number of entries (Accounting and Finance) and a subject predominately taken by college students (Psychology). As these specific subjects were chosen by the SQA and they represent a broad range of subjects the results for all 7 subjects are discussed in this thesis.

It was believed that, particularly for compulsory subjects, school roll would be a remarkably good predictor of entries. Data for the total size of the school roll in Scotland was found on the Scottish government website.

1.3.2 Problems with Data

Some problems arose with the data which had to be resolved before analysis could start.

The results data is available from 2002; however the entry data is not available until 2004. As the data is only available until 2008 this means that for each subject there are only 5 data points that can be used to model the number of entries. This poses a great number of problems which will be addressed throughout the thesis. It is not possible to create longer data series as, in 2004, it was noticed that the entry profile was changing for Intermediate Courses. To use entry data from before 2004 would create more problems than solutions. There is a continuing programme which looks at revisions to courses and this led to some other problems discussed in a later paragraph.

As discussed in Section 1.3.1, the number of entries given on the entries worksheet is different from the number of entries given on the results worksheet. It was decided that for modelling purposes the number of entries on the entries worksheet would be used. The main reason for making this decision was that in future planning the

number of entries on the entries sheet is available months before the number of entries on the results sheet. When trying to calculate an award or pass rate, the number of entries on the results worksheet was used. This was to provide as accurate an award or pass rate as possible.

There were some changes to Accounting and Finance over the time period. Standard Grade Accounting and Finance remained unchanged but there was a change in the National Courses. For 2004 all levels of National Course are entitled Accounting and Finance, in 2005 Accounting and Finance only accounts for the Advanced Higher entries and all other National Courses are classed as Accounting. After 2005 all National Courses were labelled as Accounting. This happened as there was a revision of the National Course in 2004, except for Advanced Higher, which took place in 2005. This was more than a simple name change; the whole syllabus for each course was changed. As the two courses did not run at the same time for any level, the only way to avoid losing information was to treat the courses as equivalent in terms of entries. Any pupil wanting an Accounting qualification would have had to sit whatever course was available at the time so it is believed that, by and large, pupils would have sat the examination whatever the course content. In any case, detailed information about the syllabus content is not information that is widely available to pupils before they undertake the course.

A similar situation affected Psychology. Standard Grade Psychology does not exist, therefore no problem was observed with that. In 2005 Psychology (new), the revised Psychology course, ran alongside the original Psychology course. In the years after 2005 only the revised course, Psychology (New), was available. It was decided that, as a student could only take one or other of these courses, the number of entries for the two could simply be added together in 2005 to give an approximate number of entries to model. There may be more issues with combining these subjects when comparing pass rates.

When looking at the issues that affect Standard Grade English, it was found that there are an additional three variations of the subject; English and Communication, English – Spoken, English – Alternative Communication. According to the SQA, English and Communication has to be treated as English and the data can be summed together. However, English – Spoken and English – Alternative Communication should not be taken into account as they are taken by very few people with the number of entries usually estimated as 10.

1.4 Aims

The ability to provide accurate estimates for the number of entries for NQs in both the long and the short term would be greatly beneficial to the SQA. This would enable them to know in advance how many papers to print and deal with other such operational planning issues, for example timetabling and financial forecasts, but also would have an impact on long term policy and course development.

The current method of producing NQ projections is both qualitative and quantitative. A weighted average of the previous 3 years' entries is used as the basis of the predictions. Ultimately, though, the final decision is a somewhat subjective judgement made by the individual SQA Qualification Manager who decides whether they feel the prediction reflects what they know about their particular subject. National school roll figures are also taken into account but are only used as an indicator and not used in any calculations. There are no more details to define the current method any more clearly.

Forecasts are available for all qualifications except Access. The main purpose of these forecasts are to allow for operational planning for the academic year ahead and are very rarely used for beyond that year.

The purpose of this project is to review the current system and either provide statistical reasoning for it or provide a more accurate forecasting tool. It is also of

interest to try to include Access courses in the forecasts and, if possible, produce long term forecasts.

In the remainder of this thesis Chapter 2 outlines some of the theory that will be used in the thesis and details why it is needed. Chapter 3 contains the exploratory analysis of the data, by time series plots for each subject and level and provisional regression models. After the subjective impressions of Chapter 3, Chapter 4 displays the results of autocorrelation analysis, testing if any relationship between the number of entries in different years can be exploited to achieve a more accurate prediction, and examines the predictive power of the model which best fits the data. Once the 'best' model for each level and subject has been found, Chapter 5 goes on to display the predictions produced for 2009 and then compares them with the total number of students at all schools and colleges enrolled to sit the examination, the prediction calculated by the SQA and the actual number of entries. Finally, Chapter 6 provides a summary of results throughout the thesis, discusses limitations and problems posed by the study and discusses possible future work.

Chapter 2
Literature and Methods

2.1 Bootstrapping

Since the time series provided by the SQA generally consisted of just 5 data points, standard parametric methods of analysis seemed unlikely to give useful results. Alternative approaches based on bootstrapping were therefore investigated. Bootstrapping, so called first by Efron (1979), is a resampling method using only the sample provided and no external information.

According to Manly (1997) the main idea behind bootstrapping is that, if the only information available regarding the population is a random sample, then the best way to estimate the distribution of the population is to use the distribution in the random sample. In order to simulate a resample of the population, the easiest way is to resample, with replacement, the random sample. The unknown distribution of the real population is modelled by the observed values in the sample, each with probability 1/n, where n is number of values in the original sample. Bootstrapped samples can be taken numerous times from the one sample and this simulates sampling from the population numerous times.

Bootstrapping can be extended to calculate valid confidence limits for population parameters. There are various methods for producing confidence limits from bootstrap samples, such as the standard bootstrap confidence limits, simple percentile confidence limits, bias-corrected percentile confidence limits and accelerated bias-corrected percentile limits.

An estimate for the population parameters is calculated from each bootstrapped sample. An average of all these estimates will give an overall bootstrapped estimate. The bootstrapped estimate is then used to derive an estimate of the standard error and confidence intervals for the population parameters.

The standard bootstrap 100(1-α) % confidence interval is defined as:

$$\text{Estimate} \pm z_{1-\alpha/2} \quad \text{(Bootstrap estimated standard error)} \tag{2.1}$$

Here, $z_{1-\alpha/2}$ is the upper percentile of the standard normal distribution. According to Manly (1997) the standard bootstrap confidence interval has been shown to be unsatisfactory in general, tending to be too narrow to achieve the nominal coverage. This interval is constructed on the assumption that the sampling distribution of the estimator is approximately normal, where this is not the case it will cause coverage problems.

The simple percentile method is described by Manly (1997) as follows:

> *"the 100(1 – α) % limits for a parameter are just the two values that contain the central 100(1 – α) % of the estimates obtained from bootstrapping the original sample"*

If the bootstrap sampling distribution of the estimate is skewed, the simple percentile method, like the standard bootstrap method, can be biased and therefore methods that eliminate this bias need to be developed. One method is bias-corrected percentile confidence limits, detailed in the Durbin Watson secton (2.2) below. The final method mentioned above is the accelerated bias-corrected percentile method, which is called accelerated because, according to Efron and Tibshirani (1993), it captures the rate at which the standard error of the estimate changes with respect to the true value of the parameter. This method requires a less restrictive assumption than the bias-corrected percentile method which effectively sets the acceleration constant to zero.

These methods, although thoroughly explained generically in Manly (1997), will be discussed in more detail in the next section in one context for which they were used in this study, which is investigating first-order autocorrelation in the SQA time series.

2.2 The Durbin Watson Statistic

It seemed possible that the number of entries for a particular SQA examination in successive years might be autocorrelated, in which case these relationships might be exploited when predicting entries one year ahead. The first step when examining any relationship between variables is simple linear regression. A simple linear regression equation is of the form

$$y_t = \alpha + \beta x_t + \varepsilon \tag{2.2}$$

where α is the intercept, β is the slope of the line and ε_t is the error term, $\varepsilon_t \sim N(0,\sigma^2)$. In this thesis, t will usually denote time (years). The standard assumptions that this model are based on are that the errors are independently and normally distributed, the variance is constant across time and that the relationship between y and x is linear.

The next step is to look at the simplest form of autocorrelation in a regression, which is first order autocorrelation or first order autoregression (denoted AR(1)). This is defined by

$$y_t = \mathbf{X}_t \beta + u_t \tag{2.3}$$

where y_t is the response at time t, $\mathbf{X_t}$ a vector of the values of the explanatory variables at time t, β a vector of unknown parameters and if u follows an AR(1) process then the error terms are related by $u_t = \rho u_{t-1} + e_t$, where $|\rho| < 1, e_t \sim N(0, \sigma^2)$ and e_t are uncorrelated.

The Durbin-Watson (D-W) test (Durbin and Watson, 1951) is commonly used to test for first-order autocorrelation in the residuals from regression analysis (see MacKinnon, 2002).

If $\hat{u}_t$ denotes the t^{th} residual from ordinary least-squares (OLS) regression, as defined in equation 2.3, then the D-W statistic is

$$d = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2} \tag{2.4}$$

When testing autocorrelation the null hypothesis tested is $H_0 : \rho=0$ against $H_1 : \rho > 0$. The D-W statistic testing this hypothesis is defined by Jeong and Chung, 2001:

$$d = \frac{\hat{u}' A \hat{u}}{\hat{u}' \hat{u}} = \frac{u' MAMu}{u' Mu} \cong 2(1 - \hat{\rho}) \tag{2.5}$$

where $\hat{\rho}$ is the sample autocorrelation of the OLS residuals, $\hat{u} \equiv (I - X(X'X)^{-1}X')u \equiv Mu$ and

$$A \equiv \begin{vmatrix} 1 & -1 & 0 & . & . & . & . & 0 \\ -1 & 2 & -1 & & & & & 0 \\ 0 & -1 & 2 & & & & & 0 \\ . & & & . & & & & . \\ . & & & & . & & & . \\ . & & & & & . & & . \\ . & & & & & & 2 & -1 \\ 0 & 0 & 0 & . & . & . & -1 & 1 \end{vmatrix}$$

The D-W statistic produced must lie between 0 and 4. A value of d close to 2 suggests that there is no autocorrelation in the residuals. If the value of d produced is a very small value, lower that 2, it suggests that consecutive error terms are very close to each other numerically, this is also known as positive autocorrelation. If the value is larger than 2 it suggests the opposite is the case, that error terms which are close to each other in time are numerically very different, or negative autocorrelation.

The D-W test has proved popular, and useful in certain circumstances, but there are also limitations with the test. Two main limitations (Jeong and Chung, 2001) are that the distribution of the test statistic is not mathematically tractable and that it is dependent on the design matrix X. As a result, it is very difficult to construct unique critical values for the test statistic.

A solution to this, first suggested by Durbin and Watson (1951) themselves, then expanded on by many authors including Jeong and Chung (2001), is to calculate lower and upper bounds for the D-W statistic, d_L and d_U, whose distributions do not depend on the design matrix X. The null hypothesis, H_0, will be rejected if $d < d_L$, and will not be rejected if $d > d_U$. Here H_0: $\rho = 0$ and H_1: $\rho > 0$. This produces an 'indeterminate' range (d_L , d_U) and the result will be inconclusive if the test statistic, d, falls within this range. The existence of this range also reduces the power of the test.

The bounds for the statistic can often be far apart, and especially for small sample sizes the 'indeterminate' range is large relative to the possible range of d. The 'Theoretical Range of d Statistic' (Savin and White, 1977) for a sample size of 5 is (0.3820, 3.6180) so the D-W test seems unlikely to be useful for analysing the SQA data.

A more powerful procedure to assess the D-W statistic and a way of avoiding the intractability due to its dependency on the design matrix would be to use simulation methods, such as Monte Carlo and bootstrapping, the methodology for which was detailed in MacKinnon (2002.

- Estimate the parameters of a linear regression model where number of entries for a particular examination is the response and year is the explanatory variable. Evaluate the D-W statistic, d, from this model.
- The number of bootstrap simulations is then set, often B=99.
- 99 bootstrap samples are then generated by using the standard normal distribution and drawing vectors of errors $\mathbf{u}_j^*$.
- The vector of errors is then regressed on year to produce 99 simulated residual vectors $\hat{\mathbf{u}}_j^*$.
- For each of the 99 simulations, d_j^* is computed from the bootstrapped residuals $\hat{\mathbf{u}}_j^*$ using equation (2.1)
- The p-value to test positive serial correlation is

$$p^*(d) = \frac{1}{99}\sum_{j=1}^{99} I(d_j^* \leq d) \tag{2.6}$$

  where I (.) is an indicator function, which equals 1 if the statement inside the bracket is true and 0 if it is not.

- The null hypothesis is rejected, and hence there is evidence of positive serial autocorrelation, when p^* is less than α, the significance level.

Extending the sort of bootstrapping work described by MacKinnon (2002), the paper by Jeong and Chung (2001) suggests several bootstrapping methods to more accurately test for autocorrelation. The purpose of all these methods remains

"to eliminate the indeterminate range and improve the power"

The paper compares the behaviour of several different bootstrap tests on simulated datasets. Five different sample sizes were used; n = 10, 20, 50, 100 and 200, and

several different numbers regressors; k = 3, 5, 10 and 20. The paper found that the original D-W test has small sample properties that are unsatisfactory and therefore supported the proposal that another method needs to be used. The power when (n-k) is small, which is clearly the case in the SQA dataset, is unacceptable. Three methods examined by Jeong and Chung, 'Bootstrapped D-W test', 'Bootstrapped ρ (B – ρ)' and 'Bootstrapped- ρ BC $_a$ (BC- ρ) test', are detailed below.

The first alternative method taken from the paper is called the 'Bootstrapped D-W (BDW) test', which is similar to the method already mentioned by MacKinnon and therefore is not examined in any more detail here. The paper did find that this method has fairly accurate empirical size, even in small samples.

The second method discussed by Jeong and Chung is the 'Bootstrapped ρ (B – ρ) test'. The concept behind this method is that to test the null hypothesis it would be possible to obtain an OLS estimator of $\hat{\rho}$ by regressing $\hat{u}_j$ on $\hat{u}_{j-1}$ with the intercept set to zero. The advantages of bootstrapping $\hat{\rho}$ rather than the usual D-W statistic, as discussed by Jeong and Chung, are

- The D-W statistic is an indirect estimate of the parameter in the null hypothesis, while $\hat{\rho}$ is a direct estimate.
- Because the null value of ρ under H_0 is known and therefore a more sophisticated bootstrapped test procedure can be used.

Jeong and Chung believe that this test has worse small sample properties than the previous test. It has been argued that the reason for this is that the empirical distribution tends to be skewed and have fat tails. To correct for this the method used is an 'accelerated bias corrected percentile method'. This method uses both a variance-stabilizing transformation and a skewness-reducing transformation that has

been shown to be accurate in small samples. This method was denoted in Jeong and Chung by the 'Bootstrapped- ρ BC $_a$ (BC- ρ) test'.

The two methods selected from Jeong and Chung (2001) to be used on the SQA data were the 'Bootstrapped ρ (B – ρ) test' and the 'Bootstrapped- ρ BC $_a$ (BC- ρ) test'. The 'Bootstrapped- ρ BC $_a$ (BC- ρ) test' was chosen because it had been found in the paper to perform the best in very small sample examples and the 'Bootstrapped ρ(B – ρ) test' was an appropriate place to start and then extend the method to the more complex but perhaps more suitable 'Bootstrapped- ρ BC $_a$ (BC- ρ) test'.

The methods above can be extended to calculate intervals. The interval for a bias-corrected method, 'Bootstrapped- ρ BC $_a$ (BC- ρ) test', may be obtained as follow (Manly, 1997). The first step is to calculate z_0, the value from the standard normal distribution that is exceeded with probability p, where p is the proportion of times that the bootstrapped estimate of ρ is greater than the original estimate of ρ from the data. The upper and lower confidence limits are then the values that just exceed the proportions of the ordered bootstrapped estimates of ρ calculated by $\phi(2z_0 \pm z_{\alpha/2})$, which is the proportion of the standard normal distribution that is less than the formula contained within the bracket.

This can be developed into an accelerated bias-corrected method, 'Bootstrapped- ρ BC $_a$ (BC- ρ) test', as follows. The initial steps are the same as for the bias-corrected method produce bootstrapped values for ρ and calculate z_0 in the same way. The next step is to calculate the constant a by:

$$a \approx \sum_{i=1}^{n} (\hat{\rho}_{\cdot} - \hat{\rho}_{-i})^3 \Big/ [6\{\sum_{i=1}^{n} (\hat{\rho}_{\cdot} - \hat{\rho}_{-i})^2\}^{1.5}] \qquad (2.7)$$

where $\hat{\rho}_{-i}$ is the estimate of ρ with the i[th] observation removed and $\hat{\rho}_{\cdot}$ is the average of all $\hat{\rho}_{-i}$. The confidence limits are then set to INVCDF$\{\phi(z_L)\}$ and INVCDF $\{\phi(z_U)\}$ where z_L and z_U are as follows

$$z_L = (z_0 - z_{\alpha/2}) \Big/ \{1 - a(z_0 - z_{\alpha/2})\} + z_0 \,,\; z_U = (z_0 + z_{\alpha/2}) \Big/ \{1 - a(z_0 + z_{\alpha/2})\} + z_0 \,. \quad (2.8)$$

Initial results showed that the method from MacKinnon (2002) produced plausible results. The confidence interval produced by the 'Bootstrapped- ρ BC $_a$ (BC- ρ) test', shown in chapter 4, did not appear to be plausible and therefore an adaptation of the method was needed.

Jeong and Chung (2001) use OLS to calculate ρ and the residuals, which are then bootstrapped. Theoretically, the value of ρ should lie between -1 and +1. It was discovered that, when using OLS to estimate ρ, this was not always the case. An alternative way of estimating ρ was used. This was to estimate ρ as the correlation between the residuals and the lagged residuals from OLS estimation of the regression model. The output confirms that this is a more appropriate way to estimate ρ as the values produced are within the required range. The output for both methods is contained in the thesis for comparison.

2.3 Mean Squared Error of Prediction

Having a small number of data points available for regression analysis means that parameter estimates will lack precision. It also limits the number of possible explanatory variables that may be included in a model as there are a very limited number of degrees of freedom available. With a small number of data points there is a

possibility of over-fitting in the model. This means that any estimates produced would be extremely closely related to the data and the resulting model would be a poor predictor of future observations.

There is also a problem with identifying the best possible model from the choice of models available. Possible criterion for selecting the 'best' model are adjusted R squared and S. Adjusted R squared is the amount of variability in the response that is explained by the model, adjusted for the number of explanatory variables in the model; a higher value of adjusted R squared means that the model explains a greater amount of the variability. S is a measure of the predictive error and smaller values indicate a more accurate prediction. These values are very closely tied to the data and, due to the small number of data points, may be unreliable. Therefore another method of finding the best predictive model needs to be found. The mean squared error of prediction (MSEP) or prediction error is a measure of the accuracy of prediction. It is the average squared difference between the quantity of interest and the prediction given by the model. The smaller the value of MSEP, the better the prediction in any given context. It is also possible that the prediction may be biased, so it is of interest to estimate the amount of bias attached also.

According to Wallach and Goffinet (1989), the MSEP of a model is defined by:

$$\text{MSEP}(\hat{p}) = \text{E}\left[\left(y - f\left(X, \hat{p}\right)^2 \mid \hat{p}\right)\right] \tag{2.9}$$

Where E is the expectation (over the population), y is the quantity to be predicted (in this case the number of entries for an examination) and $f\left(X, \hat{p}\right)$ is the prediction

given by the model f. X denotes the explanatory variables in the model and $\hat{p}$, the estimated parameters.

MSEP is a measure over the entire population, not just the sample of interest, and cannot typically be directly measured. It therefore needs to be estimated from the sample data, denoted $\hat{MSEP}$.

In the SQA dataset, as there are only five data points, there is not enough data to be able to use independent data to test the predictive performance of any model. This could create a situation where the predictions would be very heavily biased towards the data. In this instance a resampling method needs to be used.

If the observations are independent of the test sample, as well as being independent themselves, then MSEP ($\hat{p}$) can be estimated by

$$\hat{MSEP}_1(\hat{p}) = \frac{1}{N}\sum_{i=1}^{N} ERR2_i \tag{2.10}$$

where

$$ERR2_i = (y_i - f(X_i, \hat{p}))^2 \tag{2.11}$$

That is the error is the squared difference between the actual quantity and the prediction from the model. In this case $\hat{MSEP}_1(\hat{p})$ will underestimate $MSEP(\hat{p})$ and therefore the formula will need to include a correction term to allow for this. The formula becomes

$$MSEP(\hat{p}) = \hat{MSEP}_1(\hat{p}) + OP \tag{2.12}$$

The final term is denoted OP for optimistic, as this is a measure of how overly optimistic one is in the predictive performance of the model by allowing $\hat{MSEP}_1(\hat{p})$ to estimate $MSEP(\hat{p})$. The only problem then becomes how to estimate OP. An estimate of OP, $\hat{OP}$ can be calculated using bootstrapping, subtracting $\hat{MSEP}_1(\hat{p})$ of the bootstrapped sample from $\hat{MSEP}_1(\hat{p})$ of the original data. $\hat{OP}$ is calculated for each bootstrapped sample and then the overall $\hat{OP}$ is calculated by summing all $\hat{OP}$ values for each bootstrapped sample and dividing by the number of bootstrapped samples.

Chapter 3
Initial Impressions

The first step is to gain an initial impression of the data. The data needs to be explored in a number of ways and any possible trend through time or any relationships between number of entries (the response) and possible explanatory variables examined. One of the explanatory variables that is expected to be important in modelling time trends in the number of entries is gender. The simplest and most common way of beginning this investigation is to look at plots of the number of entries over time. There were six plots produced for each subject, where possible, and these were

- Time Series Plot of Standard Grade Entries
- Time Series Plot of Intermediate 1 Entries
- Time Series Plot of Intermediate 2 Entries
- Time Series Plot of Higher Entries
- Time Series Plot of Advance Higher Entries
- Time Series Plot of Total Credit and Intermediate 2 Passes and Higher Entries.

As the aim of these plots is to gain subjective impressions not all of them have been included in the thesis. Some of the plots were also very repetitive, showed very little or were inconclusive.

To explore any relationships with possible explanatory variables, regression models were then fitted to the data. These models shall be explained in further detail later in the chapter.

3.1 Time Series Plots

As the data is collected over time, one of the easiest ways to display and examine the data is using a time series plot. This simply plots the time values, or in this case year

on the x-axis and the data values on the y-axis. The only plots included in the main body of the report are the time series plot of Standard Grade entries and the time series plot of Total Credit Passes and Intermediate 2 Passes and Higher Entries. The purpose of the Standard Grade plots is to visually examine any trend across time for all possible levels of Standard Grade. For the plot containing both the number of passes and entries the aim is to see if there is any possible relationship between the number of Higher entries and the total number of Credit and Intermediate 2 passes the previous year. For most subjects (e.g. Mathematics, English, French), pupils will only proceed to Higher if they have already passed the same subject at Credit or Intermediate 2 level. In a few subjects (e.g. Accounting and Finance) it is more common for students to begin their study of the subject in fifth or sixth year at Higher level ('Crash Higher'). If there is a relationship between passes at the lower level and entries at Higher, any trend that is visible in the total number of Credit and Intermediate 2 passes might appear in the Higher entries the following year.

3.1.1 Accounting and Finance

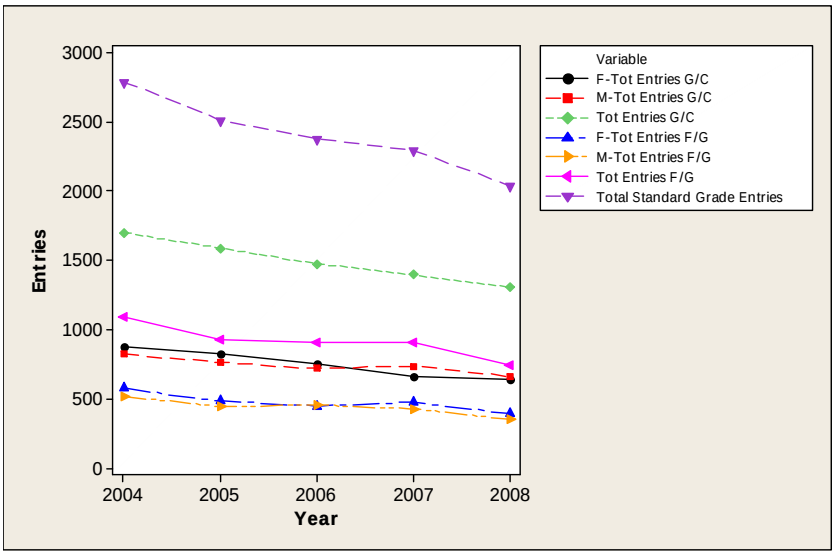


Figure 3.1.1 – The number of Standard Grade Entries 2004 – 2008 for Accounting and Finance

It can be seen that over the 5 year time period there is a decreasing trend in the total number of Standard Grade entries. The total number of Standard Grade entries is predominately made up of General/Credit entries. When this is split by gender it can

be seen that the number of General/Credit entries for males and females follow a similar pattern of decrease as the total number of entries. It is also seen that although the numbers are similar, only in 2007 is the number of General/Credit entries for males greater than females. For Foundation/General entries a slight decrease is also visible. The same gender pattern can also be seen, however the number of entries for males only exceeds that of females in 2006.

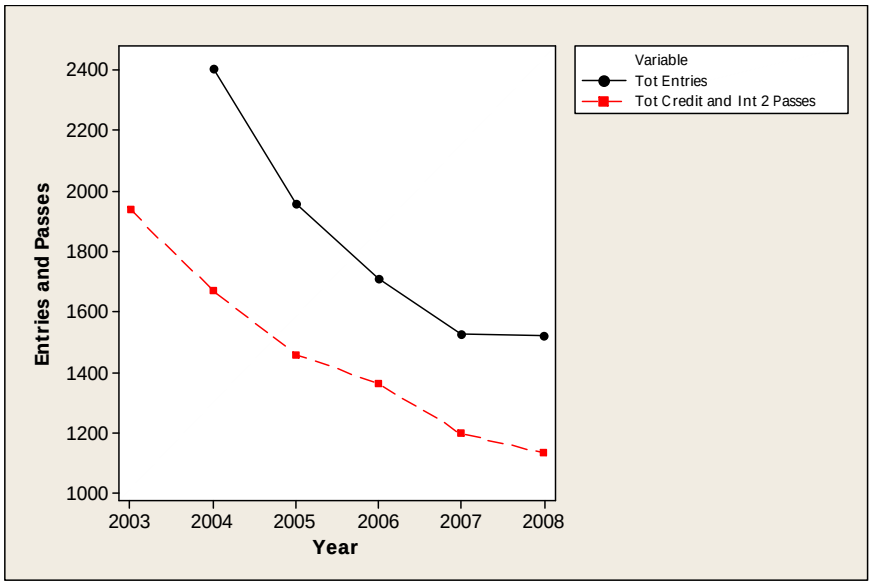


Figure 3.1.2 – The Total Number of Higher Entries 2004 – 2008 and Total Number of Credit and Intermediate 2 Passes 2003 – 2008 for Accounting and Finance

From Figure 3.1.2 it appears that there may be some relationship between the total number of Credit and Intermediate 2 passes and the Higher entries. The decrease in the number of passes is also evident in the number of entries. However the number of Higher entries does appear to be levelling off between 2007 and 2008 and this is not shown in the number of passes at Credit and Intermediate 2 the previous year. The number of Higher entries is also greater than the combined number of Credit and Intermediate 2 passes, this is not the case for all subjects.

It may also be the case that the subject is experiencing a fall in the number of entries across all levels. It is possible that this decline in popularity is related to the change in the name and syllabus of this subject in 2004, but no evidence is available to allow this hypothesis to be tested.

3.1.2 Art

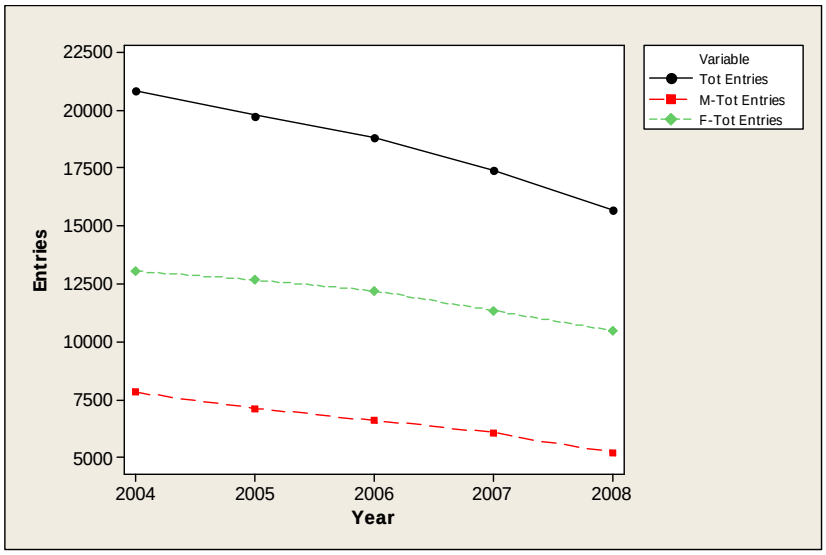


Figure 3.2.1 – The Number of Standard Grade Entries 2004 - 2008 for Art

There is also a noticeable downward trend in Standard Grade Art entries. Art cannot be split by level as all students sit the same examination regardless of level and are graded accordingly. There is a clear gender difference with females having a greater number of entries than males. However, the downward trend is visible in both genders.

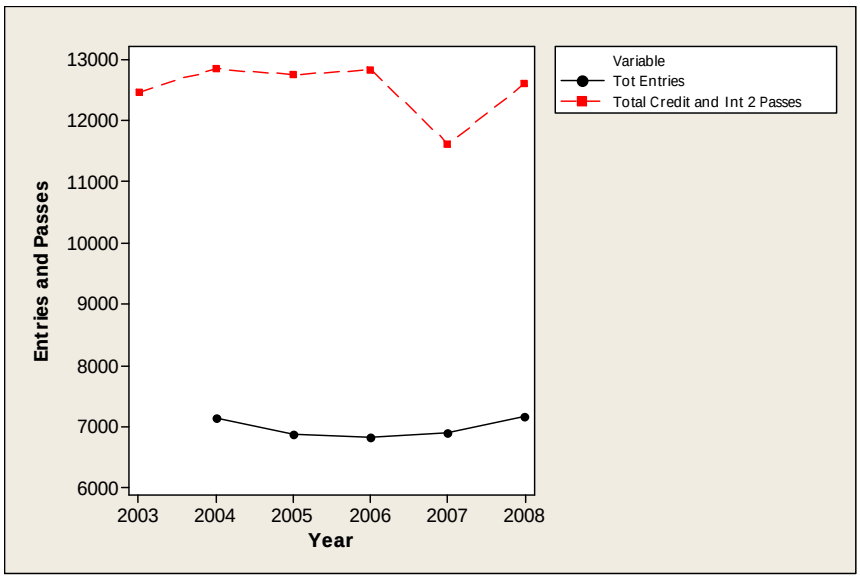


Figure 3.2.2 – The Total Number of Higher Entries 2004 - 2008 and Total Number of Credit and Intermediate 2 Passes 2003 - 2008 for Art

Figure 3.2.2 does not immediately indicate relationship between the total number of Credit and Intermediate 2 passes and Higher entries. There is a sharp decrease in the number of passes in 2007 that is not reflected in the 2008 Higher entries. It can be seen that the total number of Higher entries is substantially less than the total number of passes from Credit and Intermediate 2 for all years. This might be typical of patterns of uptake in 'practical' subjects (such as Art, Music, Physical Education), since the options made available to pupils force them to take a wide range of subjects at Standard Grade, not all of which can be continued at a higher level.

3.1.3 English

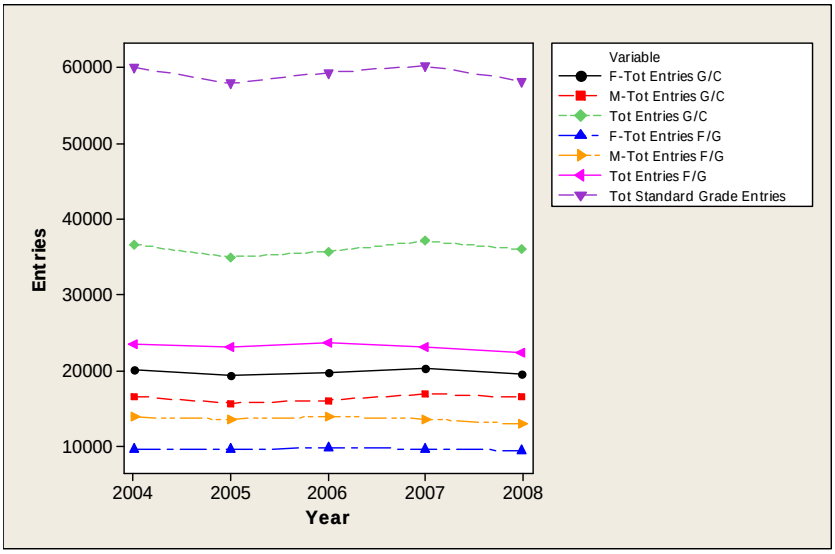


Figure 3.3.1 – The Number of Standard Grade Entries 2004 - 2008 for English

For Standard Grade English it appears that there is very little change over time. Again most of the Standard Grade entries are accounted for by the General/Credit entries. Although it appears that the number of entries are constant over time, gender differences can be seen. For General/Credit females have a higher number of entries than males and for Foundation/General this effect is reversed. This seems to support anecdotal evidence that girls develop better language skills than boys at an earlier age.

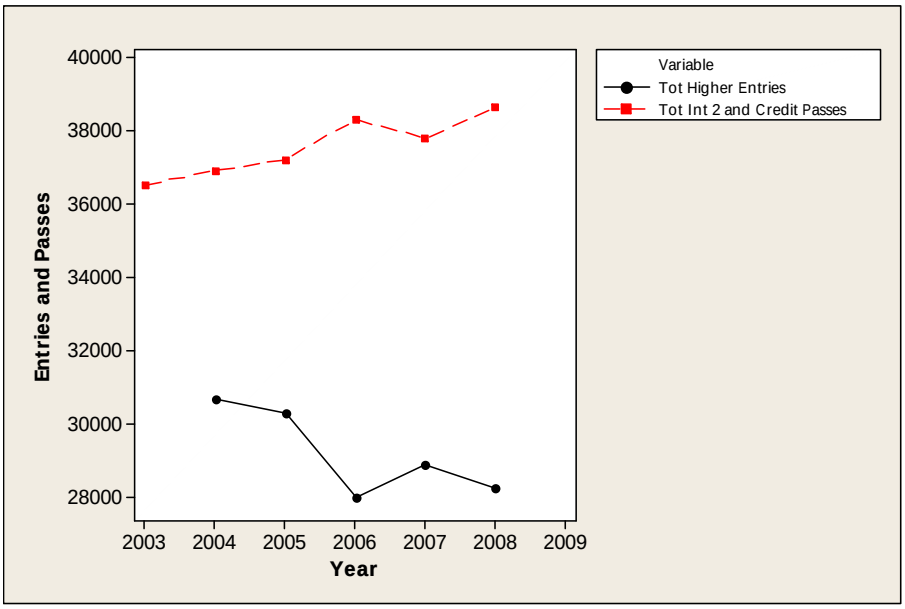


Figure 3.3.2 – The Number of Total Higher Entries 2004 - 2008 and Total Number of Credit and Intermediate 2 Passes 2003 - 2008 for English

There is no obvious relationship between lagged passes at lower levels and Higher entries for English. At the same time as the total number of passes has been gradually increasing, the number of entries has been slightly decreasing. There is a sharp decrease in the number of Higher entries in 2006 that does not reflect the previous year's passes at Credit and Intermediate 2.

3.1.4 Mathematics

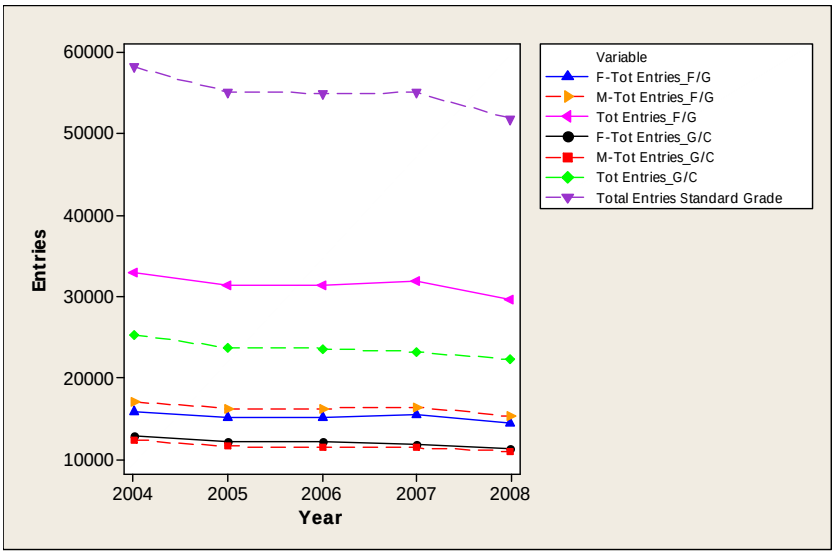


Figure 3.4.1 – The Number of Standard Grade Entries 2004 – 2008 for Mathematics

Mathematics is the only one of the subjects studied here in which the number of Foundation/General entries is greater than the number of General/Credit entries. It does appear that there is a decrease in the number of Standard Grade entries across the time period. On the other hand, the number of Intermediate 1 entries (not plotted here) increased by over 7000 from 2004 to 2008 and the number of Intermediate 2 entries increased by over 6500 from 2004 to 2008. These increases in the number of entries at Intermediate more than account for the decrease in Foundation/General and General/Credit.

The number of male entries is greater than female for Foundation/General and less than female for General/Credit, the same pattern as for English.

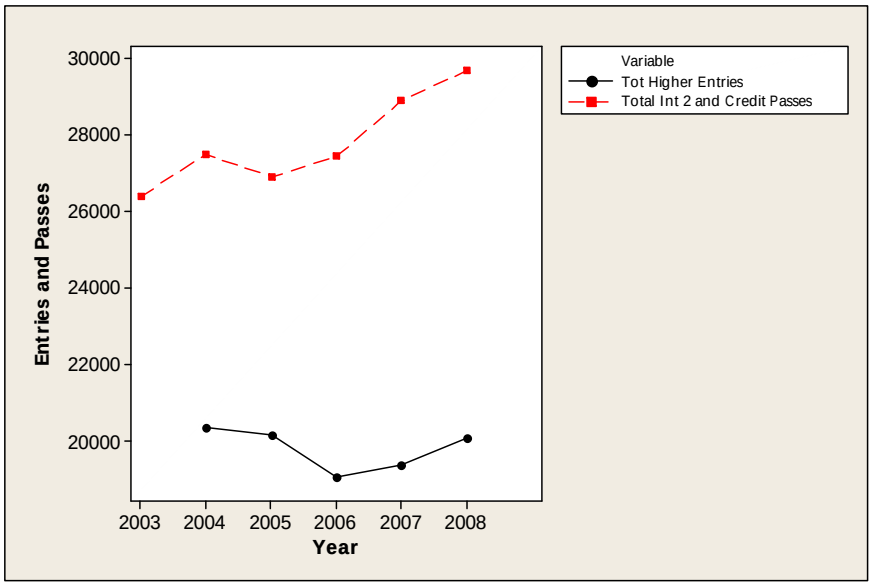


Figure 3.4.2 – The Total Number of Higher Entries 2004 – 2008 and Total Number of Credit and Intermediate 2 Passes 2003 – 2008 for Mathematics

As the number of Credit and Intermediate 2 passes increases in this time period, this is not reflected in the number of Higher entries which decrease and then increase again.

3.1.5 Physics

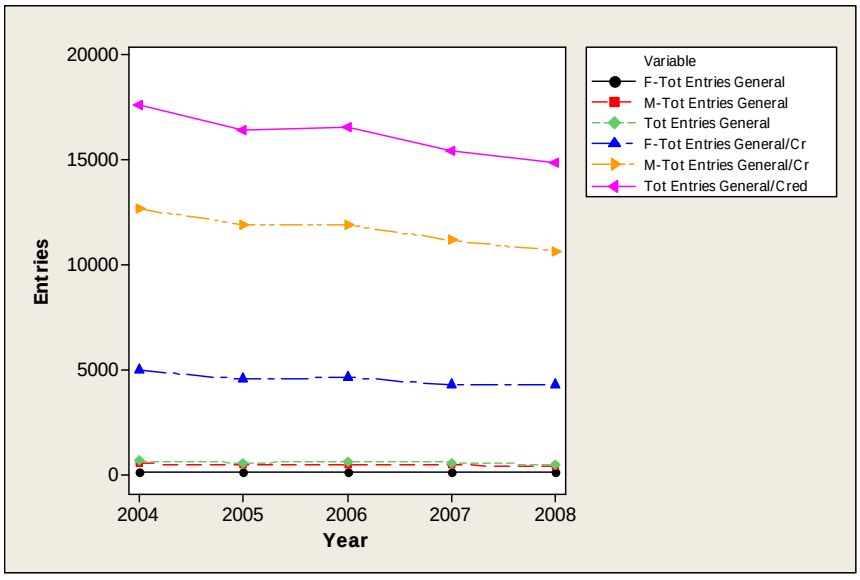


Figure 3.5.1 – The Number of Standard Grade Entries 2004 - 2008 for Physics

All Physics Standard Grade entries are classed as General/Credit or General as there is no Foundation level for discrete sciences. The total number of General/Credit entries again account for most of the total number of Standard Grade entries. It does appear

that across the time period there is a slight decrease in the number of entries.

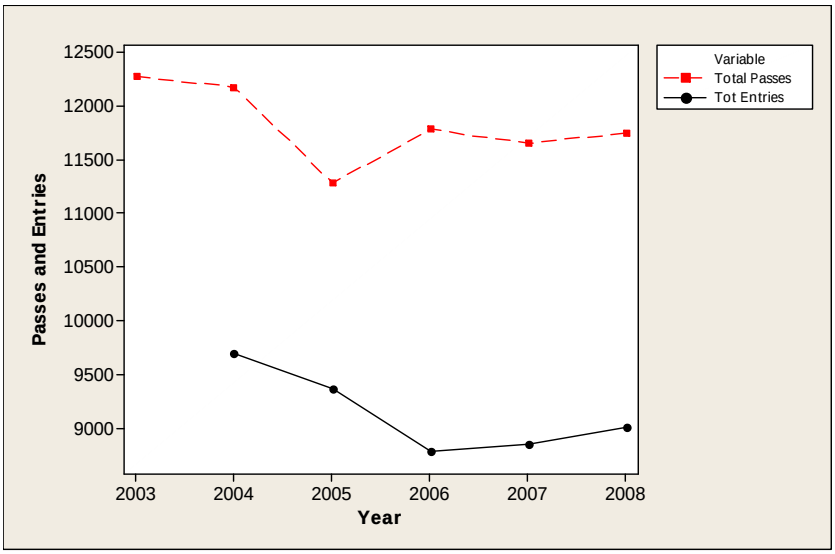


Figure 3.5.2 – The Total Number of Higher Entries 2004 - 2008 and Total Number of Credit and Intermediate 2 Passes 2003 – 2008 for Physics

Figure 3.5.2 shows that the total number of Standard Grade and Intermediate 2 passes appears to have an effect on the number of Higher entries. There is a sharp decrease in the number of passes in 2005 that is reflected by a decrease, not as dramatic, in the number of Higher entries a year later.

3.1.6 Psychology

For Psychology it is not possible to show the plot of Standard Grade entries as Standard Grade Psychology is not an option in schools. Therefore the time series plot involves only Intermediate 2 entries.

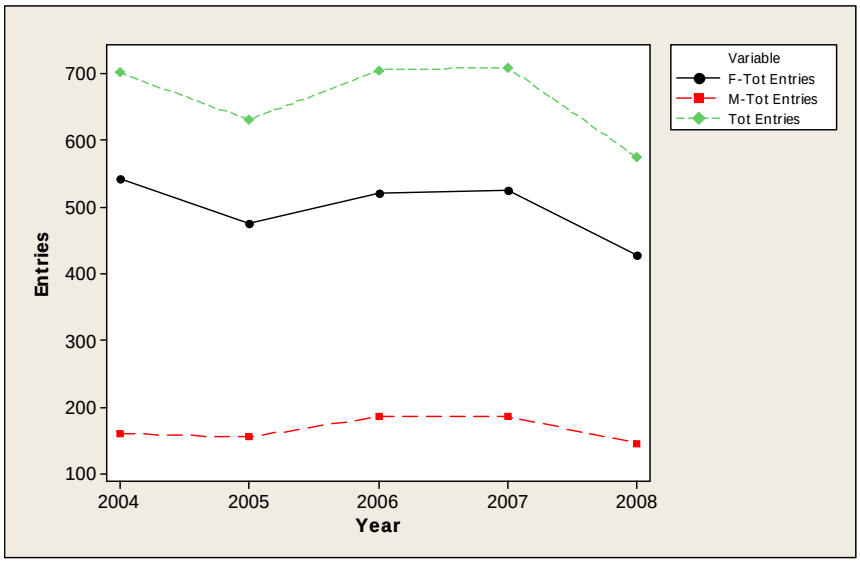


Figure 3.6.1 – The Number of Intermediate 2 Entries 2004 - 2008 for Psychology

It appears that there is no general trend across all the years for Intermediate 2 Psychology. The gender difference is still visible, even though there is a small number of total entries for each year.

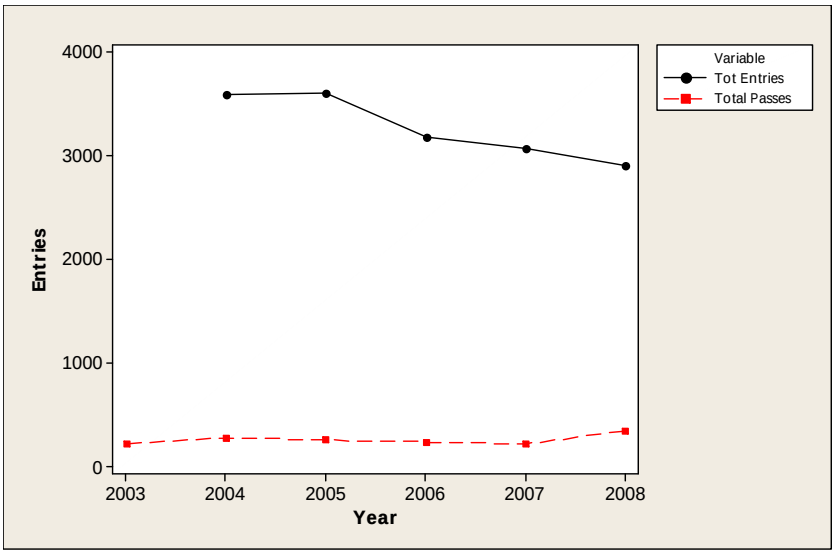


Figure 3.6.2 – The Total Number of Higher Entries 2004 - 2008 and Total Number of Intermediate 2 Passes 2003 - 2008 for Psychology

Due to the small number of people sitting Intermediate 2 Psychology it is very difficult to distinguish if there is any relationship with Higher entries. From the limited data it does not appear that there is a relationship. It appears that the number

of passes has a very slight increase, whereas the number of entries is decreasing.

3.1.7 Spanish

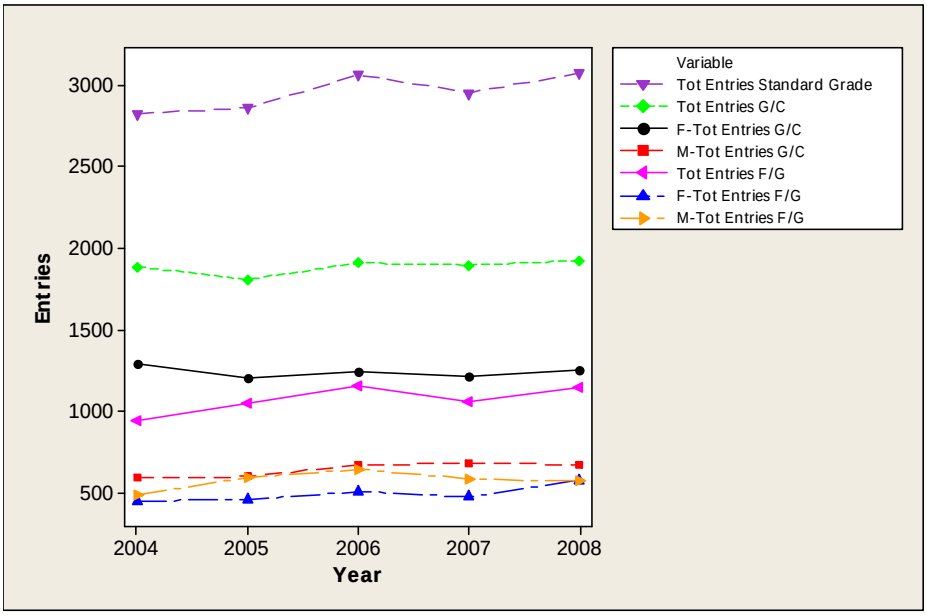


Figure 3.7.1 – The Number of Standard Grade Entries 2004 - 2008 for Spanish

It appears that over the time period there is a slight increase in the number of entries. The total number of General/Credit entries appears constant and the increase can be accounted for by Foundation/General. There is also a clear gender difference visible, with female General/Credit having a greater number of entries than any other.

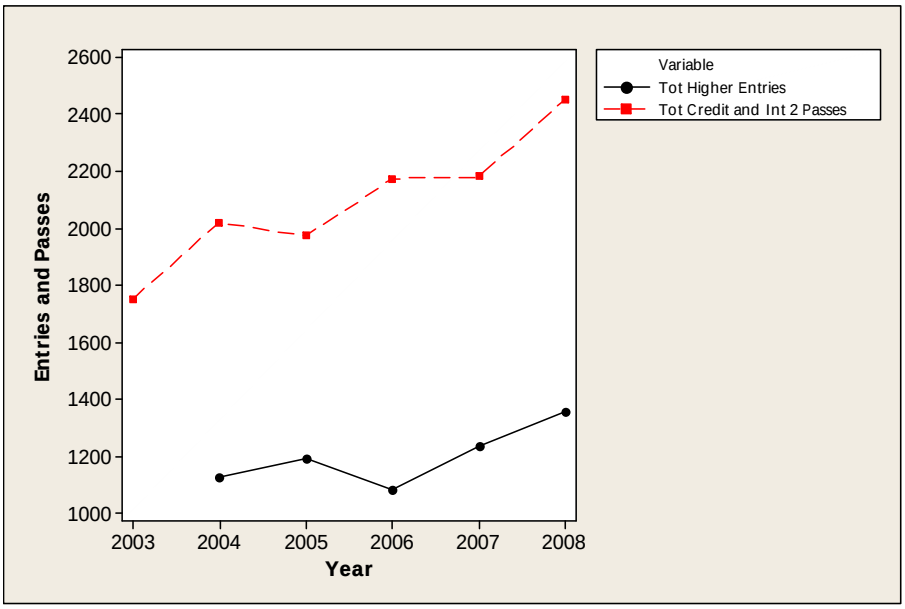


Figure 3.7.2 – The Total Number of Higher Entries 2004 – 2008 and Total Number of Credit and Intermediate 2 Passes 2003 - 2008 for Spanish

For Spanish it also appears that the total Credit and Intermediate 2 passes may be related to the number of Higher entries. The pattern that appears in the total number of Credit and Intermediate 2 passes is reflected in the number of Higher entries. The increase in the number of Higher entries might also be due to an increase in the subjects popularity across all levels and not the increase in the pass rate.

3.1.8 Overall

There is trend seen in some of the subjects. It is difficult to generalise this trend as it unique for each subject. In Art, English, Physics, Psychology and Spanish there is a visible gender difference which differs in direction and severity. In certain subjects there is a decrease in the number of Standard Grade entries which can occasionally be accounted for by an increase in the number of Intermediate 2 entries though this has not been explored further here. It is also the case that for Spanish, Physics and Accounting and Finance it is possible that there is a relationship between the total number of Standard Grade and Intermediate 2 passes for the previous year and the number of Higher entries. From the perspective of the whole examination system, a decrease in one subject may be caused by an increase in another. For example if there is a decrease in the number of entries for Spanish, the decrease may be accounted for

by an increase in another language. This interrelationship between subjects is not explored in this thesis.

3.2 Higher Entries Regression Models

Regression models were fitted using some of the possible explanatory variables. The response variable used in this first set of models was number of predicted entries at Higher level. The possible explanatory variables were chosen based on the data available. When year is mentioned it is the year in which the academic year started, that is the 2003-2004 academic year is simply labelled as 2003. The first explanatory variable was the combined number of passes at Credit and at Intermediate 2, lagged at year n-1. This was chosen as in most cases students progress from Credit or Intermediate 2 into Higher the following year; and they have been combined as they are equivalent and recent years has seen Intermediate 2 substituted in favour of Credit. Another possible explanatory variable is S5 School Roll, as it is believed that if there is a visible difference in the 5th year school roll then this will be related to the number of students sitting Higher examinations. In these models, only 4 data points are available, since school rolls were not yet available for 2008. These models are very provisional.

When examining the number of Higher entries four possible models were considered. The first of these models was the model containing year only, this model was chosen to see if any variation in the number of entries could be described adequately by a simple trend. The second model was the model containing lagged Credit and Intermediate 2 passes, in which the Higher in a subject is essentially taken by a proportion of those who are qualified to take it. The third model contained year and lagged Credit and Intermediate 2 passes. Finally the model containing 5th Year school roll only seeks to explain number of entries as a proportion of all pupils.

Another problem that arises in model selection is collinearity. Collinearity is when there is a strong correlation between two explanatory variables. For example in the SQA data Year and School roll could be strongly correlated, as could Year and National Rating. Collinearity does not reduce the predictive power of the model and as that is the main aim here it is not a major problem. It does mean that when

examining models with one explanatory variable, the models that contain variables which are collinear may produce very similar predictive performance and therefore be equally as good and difficult to choose between.

More complicated models were not investigated as there are so few data points that increasing the complexity of the model may lead to oversaturation. The previous years entries were not used here as a possible explanatory variable as any possible relationship between the previous years entries will hopefully be exploited using autoregression in Chapter 4.

There are underlying assumptions made when using linear regression for prediction purposes. These are a linear relationship between the independent and dependent variables, that the errors are independent, homoscedasticity or constant variances across time and that the variables are normally distributed. These assumptions can be checked by examining plots of the residuals.

The plots were examined but not included in the thesis. Due to the limited amount of data the residual plots were non informative. The QQ plots appeared to support the assumptions. A few of the residual versus fitted values plots suggest that a quadratic term may be needed in the model. However this is not possible as this would over complicate the model.

The model labelled ‘best’ out of the four possible models, is that which has the largest R-squared and the smallest residual standard deviation (s), the value quoted in the table is that of the adjusted R-squared which penalises the R-squared value depending on the number of parameters in the model. This criterion is chosen because of the lack of power to detect significant effects in models due to the small sample size. This means that the ‘best’ model may include variables that are deemed non-significant. This is because the ‘best’ model is for forecasting and not for examining relationships. The output shown for each subject is the ‘best’ model. The table contains the value for the coefficient, the standard error of the coefficient in brackets, the S value and the adjusted R-squared.

Subject	Model	S	R-Sq (Adj) %
Accounting	130 (260) + 0.821 (0.317) Credit Passes + 2.50 (0.821) Intermediate 2 Passes	79.49	95.4
Art	10737 (1046) – 0.276 (0.08) Credit Passes – 0.372 (0.106) Intermediate 2 Passes	82.14	73.7
English	121117 (27808) + 7290 (1808) Year – 1.36 (0.599) Credit Passes – 8.52 (2.039) Intermediate 2 Passes	293.62	94.1
Mathematics	-15541 (12483) + 1.56 (0.54) Credit Passes + 0.82 (0.314) Intermediate 2 Passes	338.59	62.6
Physics	3595 (1057) – 220 (32.33) Year + 0.507 (0.083) Credit Passes + 1.05 (0.108) Intermediate 2 Passes	48.55	98.4
Psychology	4414 (209.9) – 191 (34.06) Year	107.69	88.4
Spanish	999 (812.9) – 0.148 (0.604) Credit Passes + 0.598 (0.262) Intermediate 2 Passes	79.94	44.6

Table 1.1 – 'Best' model output for Higher subjects

3.2.2 Overall

Qualitatively different models might be required for different subjects. School Roll in 5th Year appears to always have a negative effect for all subjects. However this effect is not significant. The sign and size of effects of lagged Credit and Intermediate 2 pass variables are inconsistent and difficult to understand.

3.3 5th Year Higher Entries Regression Models

It is common policy that the Higher examinations students sit in 5th year are continuations of subjects that the students have previously sat at Credit or Intermediate 2 the previous year, most 'Crash Highers' are sat in 6th year. For this reason it was believed that modelling only 5th Year Higher entries using the lagged Credit and Intermediate 2 passes may produce more accurate predictions than

modelling all Higher entries. The only subject for which there was an improvement from the previous 'best' model for the total 5th year entries is Spanish. An improvement is determined by an increase in R-Squared (adjusted) value. Although this is not commonly done, as comparing R-Squared (adjusted) values for different datasets is not good practice, it has been used here as the datasets are not greatly dissimilar. The value for the 'best' model increases from 44.6% to 50.5%. The 'best' model for 5th Year Higher entries includes only the 4th Year Intermediate 2 Passes.

3.4 General/Credit Entries Regression Models

The next step of interest was to see if General/Credit could be modelled better than all Standard Grade entries. In terms of planning purposes, the two levels of Standard Grade are unique examinations which require different papers and different examinations; therefore it is of interest to model them separately. The next response variable used is the number of General/Credit Entries. The explanatory variables used to model General/Credit entries were Year, S4 School Roll and Credit Award Rate, lagged at year n-1. The Credit Award Rate has previously been explained in section 1.3.1. As previously mentioned there is no Standard Grade Psychology and therefore there is no General/Credit regression model for that subject.

Subject	Model	S	R-Sq (Adj) %
Accounting	2240 (105.5) – 101 (3.73) Year – 295 (194.5) Credit Award Rate	10.46	99.6
Art	15610 (6687) – 1232 (76.47) Year + 21680 (13731) Credit Award Rate	223.57	98.8
English	67129 (27139) – 0.488 (0.427) S4 School Roll	903.95	9.3
Mathematics	27545 (764.4) – 663 (124.0) Year	392.13	87.3
Physics	29932 (6375) – 579 (90.97) Year – 17481 (11225) Credit Award Rate	256.13	94.1
Spanish	2790 (296.7) – 1922 (629.1) Credit Award Rate	26.11	67.6

Table 1.2 – 'Best' model output for General/Credit entries

3.4.2 Overall

Again it is the case that different subjects require different models. As with the Higher entries even subjects which need the same variables in the model have different signs and size of the effects. Also with Higher, the 'best' model may include non significant terms, as deemed by p-values, but have significant t-ratios. The subjects whose General/Credit model has a higher R-Squared value than their Higher model are Accounting and Finance, Art, Mathematics and Spanish.

3.5 National Ratings Regression Models

As previously mentioned the National Ratings are an index used to provide a measure of how difficult a subject is relative to other subjects at the same level. Two models containing National Rating were looked at for each subject, these models were lagged National Rating for Higher on its own and then with Year. This was looking at the Higher entries. The only improvement on the Adjusted R-Squared from the total Higher Entries model is for Spanish.

Subject	Model	S	R-Sq (Adj) %
Spanish	233 (391.5) + 78.3(25.04) Year + 1000 (566.2) National Rating	62.478	66.1

Table 1.3 – Model Improved by inclusion of National Rating, Higher Level

It can be seen that the S value has decreased from 79.94 to 62.48 and the adjusted R-squared has increased from 44.6% to 66.1%. Both of these suggest that the model including year and national rating is a better model for predicting than the previous 'best' model for Higher entries.

3.5.2 Overall

When looking at the models with National Rating as explanatory variables it is clear that the original models shown in Table 1.1 are better predictive models, with the

exception of Spanish.

Chapter 4
Further Modelling

In Chapter 3 simple modelling was used to examine relationship between the possible explanatory variables and the number of entries, with the focus on Higher entries. In Chapter 4 more complex modelling is examined, trying to exploit any relationship with the previous year's entries, tested using the extensions of the Durbin-Watson statistic. The predictive ability of the model is assessed using the root mean squared error of prediction.

4.1 Autocorrelation

When testing for autocorrelation, as previously mentioned, three methods were looked at. For each level of examination the output produced is a p-value for the Durbin-Watson Monte Carlo [D-W MC] and ρ (B – ρ) methods (testing Ho: ρ=0), the lower and upper confidence limits for ρ produced by the accelerated bias-corrected percentile [ab – c] method,. At each level two tables were produced, one using the OLS estimate of ρ and the other using the correlation estimate of ρ, to allow for direct comparison between the two methods (see page 20). The p-values highlighted in bold are significant at 5% and the intervals highlighted bold do not contain zero. The order of the output in the tables below (2.1 – 2.6) is the estimate of the autocorrelation, ρ, either by OLS or correlation, the p-value for D-W MC method, the p-value for the ρ (B – ρ) method and 95% confidence interval for ρ from the ab – c method. A positive autocorrelation suggests a proportional increase in the number of entries from the previous year and a negative autocorrelation suggests a proportional decrease in the number of entries from the previous year. If there is significant autocorrelation then the relationship with the previous years entries can be used to provide a more accurate prediction.

4.1.1 Foundation/General

Subject	$\hat{\rho}$ (OLS)	Durbin-Watson Monte Carlo [p-value]	ρ (B – ρ) [p-value]	Bounds for 95% CI (ab – c method)	
Accounting	-0.434	0.485	0.081	**-1.200**	**-0.198**
English	-0.202	0.172	0.071	**-0.698**	**-0.113**
Mathematics	-0.501	0.576	**0.040**	**-1.122**	**-0.492**
Physics	-0.405	0.424	0.051	**-0.754**	**-0.342**
Spanish	-0.255	0.444	0.222	-1.648	0.314

Table 2.1.1 – OLS estimate output for Foundation/General

Subject	$\hat{\rho}$ (corr)	Durbin-Watson Monte Carlo [p-value]	ρ (B – ρ) [p-value]	Bounds for 95% CI (ab – c method)	
Accounting	-0.427	0.485	0.162	**-0.769**	**-0.228**
English	-0.206	0.172	0.465	-0.969	0.994
Mathematics	-0.456	0.576	0.182	**-0.854**	**-0.170**
Physics	-0.365	0.424	0.242	-0.973	0.520
Spanish	-0.294	0.444	0.404	-0.947	0.509

Table 2.1.2 – Alternative estimate output for Foundation/General

In the first table it can be seen that there is a discrepancy in the significance level of the p-values produced by the first two methods. The range of values for ρ, seen in both the bounds for the CI, are out with the required range of [-1, 1] for an autocorrelation.

In the second table the p-values produced by the first two methods consistently yield the same qualitative conclusions that there is no significant autocorrelation (relative to a significance level of 5%) though the p-values themselves can be very different. The CIs are quite wide suggesting difficulty in accurately calculating the true value of autocorrelation. For Accounting and Mathematics the confidence interval suggests that ρ is negative and significantly different from zero. In general, the range of values generated by bootstrapping are more acceptable than when using OLS to estimate ρ. The point estimates produced are all negative suggesting negative autocorrelation in the time series.

4.1.2 General/Credit

Subject	$\hat{\rho}$ (OLS)	Durbin-Watson Monte Carlo [p-value]	ρ (B – ρ) [p-value]	Bounds for 95% CI (ab – c method)	
Accounting	-0.011	0.111	0.343	-1.628	1.089
Art	0.022	**0.010**	0.364	-1.103	0.989
English	-0.369	0.455	0.121	-1.268	-0.006
Mathematics	-0.417	0.556	0.061	-1.719	-0.223
Physics	-0.746	0.960	0.101	-1.080	-0.515
Spanish	-0.691	0.798	0.141	-0.893	-0.358

Table 2.2.1 – OLS estimate output for General/Credit

Subject	$\hat{\rho}$ (corr)	Durbin-Watson Monte Carlo [p-value]	ρ (B – ρ) [p-value]	Bounds for 95% CI (ab – c method)	
Accounting	-0.063	0.111	0.505	-0.927	0.713
Art	-0.061	**0.010**	0.455	-0.875	0.927
English	-0.400	0.455	0.172	**-0.855**	**-0.139**
Mathematics	-0.485	0.556	0.192	**-0.794**	**-0.169**
Physics	-0.778	0.960	0.111	**-0.924**	**-0.542**
Spanish	-0.814	0.798	0.071	**-0.974**	**-0.702**

Table 2.2.2 – Alternative estimate output for General/Credit

All of the subjects are for General/Credit level with the exception of Art which is all Standard Grade entries. In the first table only Art has a significant p-value when using the D-W MC method. However there are some large differences in the values of p-value produced by the two methods. As with Foundation/General the range of values for ρ are out with the required range from -1 to 1.

In the second table Art is again the only subject with a significant p-value. In this level only Accounting and Art have a bootstrap confidence interval that includes zero and therefore suggests that ρ is not significantly different from zero. The ranges of the non-significant intervals are extremely wide and almost include the whole range of possible values for ρ. As with the previous level the minimum and maximum values are within the expected range.

Predominately the results produced for the other levels have similar conclusions. For this reason only the results for the alternative estimate of ρ are shown below.

4.1.3 Intermediate 1

Subject	$\hat{\rho}$ (corr)	Durbin-Watson Monte Carlo [p-value]	ρ (B – ρ) [p-value]	Bounds for 95% CI (ab – c method)	
Accounting	-0.621	0.222	0.111	**-0.988**	**-0.280**
Art	-0.057	**0.020**	0.576	-0.891	0.863
English	-0.454	0.657	0.152	**-0.998**	**-0.149**
Mathematics	-0.090	0.051	0.515	-0.766	0.925
Physics	-0.162	0.061	0.364	-0.852	0.585
Psychology	-0.324	0.374	0.465	-0.956	0.489
Spanish	-0.544	0.818	0.343	-0.959	0.725

Table 2.3 - Alternative estimate output for Intermediate 1

4.1.4 Intermediate 2

Subject	$\hat{\rho}$ (corr)	Durbin-Watson Monte Carlo [p-value]	ρ (B – ρ) [p-value]	Bounds for 95% CI (ab – c method)	
Accounting	-0.216	0.051	0.374	-0.867	0.626
Art	-0.480	0.717	0.242	-0.922	0.030
English	-0.576	0.647	0.111	**-0.940**	**-0.368**
Mathematics	-0.745	0.899	0.202	**-0.946**	**-0.070**
Physics	-0.376	0.485	0.283	-0.908	0.007
Psychology	-0.331	0.253	0.303	-0.936	0.715
Spanish	-0.101	**0.040**	0.444	-0.913	0.697

Table 2.4 –Alternative estimate output for Intermediate 2

4.1.5 Higher

Subject	$\hat{\rho}$ (corr)	Durbin-Watson Monte Carlo [p-value]	ρ (B – ρ) [p-value]	Bounds for 95% CI (ab – c method)	
Accounting	-0.218	0.061	0.343	-0.735	0.816
Art	-0.170	**0.030**	0.333	-0.814	0.880
English	-0.402	0.616	0.384	-0.963	0.858
Mathematics	-0.086	0.131	0.495	-0.653	0.969
Physics	-0.038	0.061	0.455	-0.907	0.950
Psychology	-0.676	0.849	0.172	**-0.970**	**-0.242**
Spanish	-0.201	0.182	0.374	-0.977	0.897

Table 2.5 –Alternative estimate output for Higher

4.1.6 Advanced Higher

Subject	$\hat{\rho}$ (corr)	Durbin-Watson Monte Carlo [p-value]	ρ (B – ρ) [p-value]	Bounds for 95% CI (ab – c method)	
Accounting	-0.469	0.546	0.121	**-0.967**	**-0.209**
Art	-0.873	0.778	0.303	-0.997	0.657
English	-0.198	**0.010**	0.404	-0.799	0.917
Mathematics	-0.984	0.970	**0.000**		
Physics	-0.586	0.748	0.253	-0.945	0.069
Spanish	-0.666	0.667	0.141	**-0.975**	**-0.276**

Table 2.6 – Alternative estimate output for Advanced Higher

The bounds for the 95% CI for the ab – c method does not have values for Mathematics as the model failed to converge.

4.1.7 Conclusions for Autocorrelation

There are some discrepancies between the p-values produced by D-W MC and ρ (B – ρ), using either OLS or correlation. Even in the instances when the conclusion drawn from the p-values is the same, the numerical difference between the two is often very large. When using OLS the range for ρ is vastly different from what is expected. The most severe example being Intermediate 2 Mathematics which has a range of (-1.65, 7.74). The range of the bootstrapped values using the ab – c method are more accurate when using the alternative method for ρ, correlation. All the point estimates produced when using the alternative method are negative. However when using OLS the point estimates for Standard Grade Art and for Intermediate 1 Art are positive. Not all models have significant intervals when using the ab – c method. It appears that more of the models in the above tables have non-significant intervals rather than significant ones. The significant intervals do not correspond to significant p-values from either of the other methods and those p-values that are significant do not always correspond to a significant interval. Perhaps there is a suggestion that there are consistently negative autocorrelations but the sample size is so small that there is little power to detect that.

4.2 Root Mean Squared Error of Prediction

In Chapter 3, initial impressions were gained by using a limited number of explanatory variables and focusing on specific areas of interest, for example 5th Year Higher Entries and General/Credit Entries. This section looks at all the available explanatory variables for each subject and each level and assesses how accurate the models are in producing predictions. The prediction is assessed and compared across subjects by calculating the root mean squared error of prediction (RMSEP), see section 2.3. This not only allows an assessment of how well the model predicts the number of entries but also provides an estimate of the amount of bias attached.

When looking at RMSEP, the lower the value the better the predictive ability of the model. Only the value for the 'best' model has been contained in the main body of the report.

The full table with all possible models is shown for Intermediate 2 Accounting only. This subject was chosen as it contained the best example of model progression. The possible explanatory variables change depending on the level, for example the variables available for General/Credit are Year, S4 School Roll and National Rating. Only a total of 3 variables were fitted at the one time as fitting any more variables may lead to oversaturation given the limited amount of data.

Single Variable Models	**RMSEP**
Year	72.1193
S5 School Roll	103.7508
S4 School Roll	83.1675
Standard Grade General Passes	62.7476
National Rating	75.9896
Two Variable Models	
Year + S5 School Roll	95.7125
Year + S4 School Roll	88.8691
Year + Standard Grade General Passes	51.7119
Year + National Rating	81.6633
S5 School Roll + S4 School Roll	99.5699
S5 School Roll + Standard Grade General Passes	85.8242

S5 School Roll + National Rating	92.0249
S4 School Roll + Standard Grade General Passes	91.4396
S4 School Roll + National Rating	89.1210
Standard Grade General Passes + National Rating	77.4493
Three Variable Models	
Year + S5 School Roll + S4 School Roll	100.7406
Year + S5 School Roll + Standard Grade General Passes	102.2358
Year + S5 School Roll + National Rating	103.5716
Year + S4 School Roll +Standard Grade General Passes	101.0609
Year + S4 School Roll + National Rating	103.8840
Year + Standard Grade General Passes + National Rating	88.6803
S5 School Roll + S4 School Roll + Standard Grade General Passes	106.5835
S5 School Roll + S4 School Roll + National Rating	107.2550
S5 School Roll + Standard Grade General Passes + National Rating	101.1763
S4 School Roll + Standard Grade General Passes + National Rating	109.4803

Table 2.7 – Root mean squared error of prediction for all possible models for Intermediate 2 Accounting and Finance

Starting with models including only one explanatory variable. It can be seen that the best predictive model, the one with the lowest RMSEP, is the model with Standard Grade General Passes. The next step is to look at the values for all models with two explanatory variables. It is only worth using a more complex model if the RMSEP is lower than the best model with one explanatory variable. In this instance it can be seen that the model with Year and Standard Grade General Passes has a lower

RMSEP than the model with only Standard Grade General Passes. The same is then done for models with 3 explanatory variables. For Accounting it can be seen that none of the RMSEP values for models with 3 explanatory variables is lower than the best model with only two variables. Therefore the best predictive model for Intermediate 2 Accounting is the model including Year and Standard Grade General Passes. This exact procedure is followed for every qualification.

The following tables contain the best model, the value of RMSEP, the average number of entries for that qualification and the ratio of RMSEP divided by the average number of entries. As the number of entries in a typical year varies greatly from subject to subject, this ratio was calculated in order to allow for a more direct comparison of the predictive performance of the best model for different subjects at the same level.

4.2.2 Foundation/General

Subject	'Best' Model	RMSEP (R)	Average Entries (E)	Ratio (R)/(E)
Accounting	Year	104.2	1089	0.096
English	S4 School Roll	432.4	23413	0.018
Mathematics	S4 School Roll	812.6	32878	0.025
Physics	S4 School Roll	53.6	642	0.083
Spanish	National Rating	85.4	939	0.091

Table 2.8.1 – Root mean squared error of prediction ratio for Foundation/General

It can be seen that for 3 of the 5 subjects, English, Mathematics and Physics, the best model for producing predictions contains only S4 School Roll, for Accounting it is Year and for Spanish it is National Rating. When looking at the ratios it can be seen

that the best predictive model, defined by the lowest ratio value, is for English and the worst is for Accounting.

The other levels, contained in the tables below, produce similar results and conclusions.

4.2.3 General/Credit

Subject	'Best' Model	RMSEP(R)	Average Entries (E)	Ratio (R)/(E)
Accounting	Year	115.2	1698	0.068
Art	Year	1636.7	20865	0.078
English	National Rating	968.4	36644	0.026
Mathematics	Year	825.8	25228	0.033
Physics	National Rating	836.6	17596	0.048
Spanish	Year	49.3	1885	0.026

Table 2.8.2 – Root mean squared error of prediction ratio for General/Credit

4.2.4 Intermediate 1

Subject	'Best' Model	RMSEP(R)	Average Entries (E)	Ratio (R)/(E)
Accounting	Standard Grade Foundation Passes	43.8	230	0.190
Art	Year	646.3	859	0.752
English	Year	787.1	5371	0.147
Mathematics	Year	2217.6	6881	0.322
Physics	Year	496.7	1164	0.404
Psychology	National Rating	54.4	83	0.655
Spanish	National Rating	78.6	706	0.111

Table 2.8.3 – Root mean squared error of prediction ratio for Intermediate 1

4.2.5 Intermediate 2

Subject	'Best' Model	RMSEP(R)	Average Entries (E)	Ratio (R)/(E)
Accounting	Year + Standard Grade General Passes	51.7	594	0.087
Art	Year	776.0	3432	0.226
English	Year	1591.8	16185	0.098
Mathematics	Year	1873.7	14733	0.127
Physics	Year	503.9	2392	0.211
Psychology	S5 School Roll	40.1	702	0.057
Spanish	S5 School Roll	152.5	777	0.196

Table 2.8.4 – Root mean squared error of prediction ratio for Intermediate 2

4.2.6 Higher

Subject	'Best' Model	RMSEP(R)	Average Entries (E)	Ratio (R)/(E)
Accounting	Intermediate 2 Passes	318.3	2403	0.132
Art	Year + Intermediate 2 Passes	152.3	7135	0.021
English	Standard Grade Passes	1105.2	30667	0.036
Mathematics	Standard Grade Passes + Intermediate 2 Passes	578.7	20372	0.028
Physics	Year	366.8	9691	0.038
Psychology	Year	257.1	3587	0.072
Spanish	S5 School Roll	92.7	1124	0.082

Table 2.8.5 – Root mean squared error of prediction ratio for Higher

4.2.7 Advanced Higher

Subject	'Best' Model	RMSEP(R)	Average Entries (E)	Ratio (R)/(E)
Accounting	Year	36.6	172	0.213
Art	S6 School Roll	40.1	1468	0.027
English	Higher Passes	64.1	1764	0.036
Mathematics	Year	158.9	2475	0.064
Physics	Higher Passes	22.1	1474	0.015
Spanish	S6 School Roll	18.6	152	0.123

Table 2.8.6 – Root mean squared error of prediction ratio for Advanced Higher

4.2.8 Conclusions for RMSEP

When looking at the best model for all qualifications, as depicted by RMSEP, the initial impression that different models will be required for different subjects and different levels is confirmed. The predictive ability of the best model across the various levels and across subjects within levels is extremely diverse. The value of the ratio also differs. The worst model, defined by the largest ratio value is either for Accounting or Art. The subjects with best predictive model at each level are English, Spanish twice, Psychology, Art and Physics.

4.3 Comparison of Number of Simulations

As the number of simulations was arbitrarily chosen to be 99 this needs to be tested to see if increasing the number of simulations will have an effect on any conclusions reached. This was tested for both autocorrelation and RMSEP. A selection of subjects at each level were chosen, the number of simulations was increased to 999 and the output recorded and compared to the previous values.

When looking at autocorrelation, the values differed slightly but the conclusions reached by each method remained the same.

When looking at the RMSEP all the same models were selected with the exception of Foundation/General Accounting. The best model when using 999 simulations is S4 School Roll and is Year when using 99 simulations. Both these models are close in value and the reason for the difference when increasing the number of simulations could be due to collinearity.

4.4 Fitting an Autoregressive Model

As some of the intervals for the ab – c method appear to be significant and therefore suggest that there may be autocorrelation, it was decided that the next logical step was to fit an autoregressive model. The best way to account for any trend across the time period was to include year as a covariate. An AR(1) model was chosen as, see section 2.2, due to the limited number of data points, it was not feasible to fit a higher order autoregressive term.

The tables for each subject are shown below. The first table contains the coefficient for the AR(1) section, the intercept and coefficient for the year covariate (x1) and the RMSEP, calculated the same as previously. The second table contains the values produced for the original model containing only year. The values for RMSEP allow for direct comparison between the two models. It is to be remembered that the lower the RMSEP the better the predictive ability of the model. It was also hoped that the coefficient for AR(1) would be similar to the value of ρ produce by the year only model.

4.4.1 Foundation/General

Subject	AR	Intercept	x1	RMSEP
Accounting	-0.626	1250.0	-57.3	323.428
English	-0.221	24482.1	-224.9	644.742
Maths	-0.880	33880.8	-412.0	1469.265
Physics	-0.716	719.8	-26.0	60.634
Spanish	-0.230	940.5	24.1	119.470

Table 2.9.1 – AR output from Foundation/General

Subject	ρ	Intercept	x1	RMSEP
Accounting	-0.427	1347.2	-72.8	113.219
English	-0.206	24388.8	-216.6	569.919
Maths	-0.456	35072.6	-614.7	1125.365
Physics	-0.365	776.4	-35.7	69.750
Spanish	-0.294	816.4	42.2	93.420

Table 2.9.2 – Year model output from Foundation/General

The above tables are used to compare the AR model and the model containing year only for Foundation/General Entries. It can be seen that the coefficient for the autoregressive term is more negative than ρ for all subjects, with the exception of Spanish. When looking at the RMSEP, to determine if the AR model has better predictability, Physics is the only subject for which this is the case. This is not consistent with previous tables in which Accounting and Mathematics had significant ab – c 95% Confidence Intervals, indicating that an autoregressive model would be better.

4.4.2 General/Credit

Subject	AR	Intercept	x1	RMSEP
Accounting	0.213	2029.6	-91.3	147.426
Art	0.855	43734.5	-2698.3	1453.714
English	-0.460	33719.0	367.7	1098.262
Maths	-0.466	26729.8	-539.2	900.275
Physics	-0.746	19772.3	-604.1	1046.663
Spanish	-0.657	1723.2	25.7	48.335

Table 2.10.1 – AR output from General/Credit

Subject	ρ	Intercept	x1	RMSEP
Accounting	-0.063	2083.0	-98.8	114.352
Art	-0.061	26145.6	-1273.6	1554.924
English	-0.400	35604.6	79.5	986.816
Maths	-0.485	27544.6	-662.6	880.119
Physics	-0.778	20033.6	-643.9	867.484
Spanish	-0.814	1783.6	16.8	47.513

Table 2.10.2 – Year model output from General/Credit

For General/Credit Entries the autoregressive term is different from ρ for Accounting, Art, English and Spanish, for Maths and Physics the coefficient for ρ is larger than the coefficient for AR. When comparing predictability, Art is the only subject for which the AR model is an improvement. This is not consistent with the previous table in which Accounting and Art were the only subjects that did not have a significant ab – c 95% Confidence Intervals, indicating that an AR model would be better for the remaining subjects.

4.4.3 Intermediate 1

Subject	AR	Intercept	x1	RMSEP
Accounting	-0.342	301.0	-26.3	37.731
Art	0.311	-1677.1	573.9	731.052
English	-0.926	630.9	2609.2	918.428
Maths	0.729	7405.9	869.3	1978.629
Physics	0.192	621.6	254.5	465.795
Psychology	-0.263	152.7	1.6	148.904
Spanish	-0.535	649.7	19.6	125.176

Table 2.11.1 – AR output from Intermediate 1

Subject	ρ	Intercept	x1	RMSEP
Accounting	-0.621	332.6	-30.5	45.558
Art	-0.0573	-1282.2	519.7	664.784
English	-0.454	3076.8	550.5	754.589
Maths	-0.090	577.2	1618.9	1792.377
Physics	-0.162	-115.6	357.7	495.542
Psychology	-0.324	58.4	15.4	64.421
Spanish	-0.544	629.2	22.5	79.331

Table 2.11.2 – Year model output from Intermediate 1

4.4.4 Intermediate 2

Subject	AR	Intercept	x1	RMSEP
Accounting	0.201	442.9	-5.2	141.656
Art	-0.579	579.6	673.4	601.067
English	-0.536	11407.7	1240.7	1828.450
Maths	-1.039	8356.8	1560.9	2128.580
Physics	-0.378	603.7	387.4	577.991
Psychology	-0.851	696.1	-4.3	75.668
Spanish	0.501	-442.0	200.1	426.457

Table 2.12.1 – AR output from Intermediate 2

Subject	ρ	Intercept	x1	RMSEP
Accounting	-0.216	711.4	-42.8	78.645
Art	-0.480	798.6	638.9	822.972
English	-0.576	11182.4	1275	1665.216
Maths	-0.745	8769.2	1486.6	2026.292
Physics	-0.376	907.4	341.8	464.219
Psychology	-0.331	769.8	-17.5	63.369
Spanish	-0.101	284.2	106.2	156.861

Table 2.12.2 – Year model output from Intermediate 2

4.4.5 Higher

Subject	AR	Intercept	x1	RMSEP
Accounting	0.814	-4248.8	456.8	292.6
Art	0.836	-4316.8	973.9	143.9
English	-0.391	32721.0	-595.7	1066.8
Maths	0.397	17544.5	291.0	660.4
Physics	0.593	7108.5	205.9	352.0
Psychology	-0.679	4557.1	-213.3	287.6
Spanish	-0.657	1723.2	25.7	46.0

Table 2.13.1 – AR output from Higher

Subject	ρ	Intercept	x1	RMSEP
Accounting	-0.218	3140.4	-219.6	361.255
Art	-0.170	6923.0	9.4	197.609
English	-0.402	32972.8	-627	1170.235
Maths	-0.086	20603.8	-131.7	626.394
Physics	-0.038	10274.6	-189.7	392.176
Psychology	-0.676	4414.0	-191.1	250.179
Spanish	-0.201	891.0	51.1	110.511

Table 2.13.2 – Year model output from Higher

4.4.6 Advanced Higher

Subject	AR	Intercept	x1	RMSEP
Accounting	-0.941	306.9	-31.3	334.402
Art	-1.537	1469.3	-2.5	55.921
English	0.354	1151.8	74.4	77.655
Maths	-0.974	2051.2	86.2	176.346
Physics	-0.743	1555.8	-17.4	30.245
Spanish	-1.434	165.5	-1.3	52.353

Table 2.14.1 – AR output from Advanced Higher

Subject	ρ	Intercept	x1	RMSEP
Accounting	-0.469	287.4	-27.9	39.751
Art	-0.873	1414.8	8.3	61.125
English	-0.198	1678.8	2.7	74.474
Maths	-0.984	2071.4	85.4	171.770
Physics	-0.586	1536.0	-14.1	24.710
Spanish	-0.666	130.8	5.1	27.335

Table 2.14.2 – Year model output from Advanced Higher

4.4.7 Conclusions

There are differences between the coefficients for ρ and the AR term for different models for across the various levels. For most levels there are only one or two subjects for which the AR model is an improvement. The exception to this is Higher level entries for which 5 of the 7 subjects, Maths and Psychology being the remaining subjects, have an improvement in predictability when using the AR model. The conclusions produced when looking at the AR models are different from those produced ab – c method.

4.5 Common Model

It was of interest to try and find one model for groups of subjects, to see if a single model could be used to predict the number of entries for a common group of subjects. It was decided to group subjects by type and to see if the 'best' model for each subject was the same. This was done in an exploratory way for Higher level subjects and for Science and Languages.

Subject	'Best' Model	RMSEP	Average Entries (b)	Ratio (a)/(b)
Biology	Standard Grade Passes + Intermediate 2 Passes	104.024	9181	0.011
Biotechnology	Intermediate 2 Passes + National Rating	9.003	53	0.170
Chemistry	Intermediate 2 Passes	202.433	9633	0.021
Human Biology	Intermediate 2 Passes	80.736	3829	0.021
Physics	Year	366.843	9691	0.038

Table 2.15.1 – 'Best' model for Sciences

For four out of the five subjects the lagged Intermediate 2 Passes is included in the 'best' model. However only two of the subjects, Chemistry and Human Biology, have the same model. There is a wide range of predictability across the models. From looking at the ratios, the lowest ratio having the best predictive ability, the model for Biology is the best predicatively and Biotechnology is the worst.

Subject	Best' Model	RMSEP (a)	Average Entries (b)	Ratio (a)/(b)
Classical Greek	Year	4.108	13	0.316
English	Standard Grade Passes	1105.176	30667	0.036
French	National Rating	163.419	4701	0.035
Gaelic (learners)	Year	15.583	145	0.107
Gaidhlig	National Rating	13.846	92	0.151
German	Standard Grade Passes	171.441	1827	0.094
Italian	Standard Grade Passes	32.354	274	0.118
Latin	Year + Intermediate 2 Passes	17.715	248	0.071
Russian	Intermediate 2 Passes	3.200	17	0.188
Spanish	S5 School Roll	92.678	1124	0.082

Table 2.15.2 – 'Best' model for Languages

For Languages, there is no common model or even a common term. Year is used in three of the ten models, and is the closest to a common term. The model with the best predictability is National Rating only for French Entries and the worst is the model with Year only trying to model the number of entries for Classical Greek. Classical Greek only has an average of 13 entries across the five years. This makes it incredibly difficult to predict the number of entries with values so small.

4.5.2 Conclusions

From examining the number of Higher Entries for Sciences and Languages, it can be seen that when looking for a common model, grouping subjects by type does not produce acceptable results. There is a wide range for ratio values showing that the ability to predict the number of entries for a single type of subjects varies immensely. If it is possible to group subjects, to predict the number of entries using a single model for a group of subjects, grouping sciences and languages into two unique groups is not an appropriate way to do it.

4.6 Gender Difference

The next area of interest was in fitting models for boys and girls separately and then for the total number of entries still split by gender. This was done for General/Credit Entries and Higher Entries only. When the results were examined it was seen that there was spurious positive correlation when the two genders are treated together. For this reason it was decided that the results from the split gender analysis would not be used.

Chapter 5
Predictions

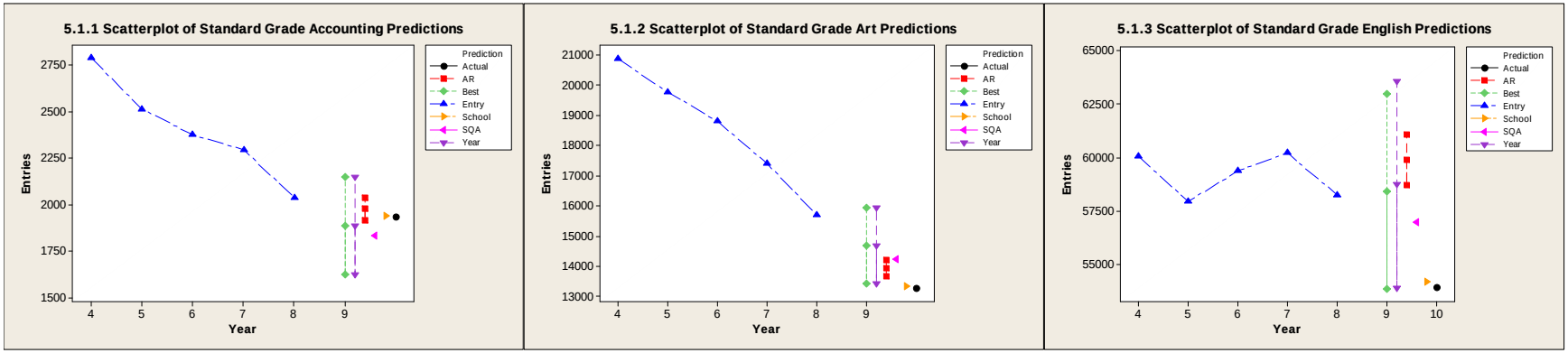
This chapter contains the predictions, both point estimates and, where possible, 95% confidence intervals, of the number of entries for each subject at each level for the 2009 entries. Predictions were produced by the 'Best' model (the model with the largest adjusted R-squared value and smallest s value), the Year model (the model containing only year as an explanatory variable) and the AR model. The point estimates and confidence intervals produced by these models were then compared with the actual number of entries. The predicted entries produced by the SQA and also the number of subjects enrolled by the schools and colleges were also included in the comparison. If the value of the prediction is lower than the actual entry this could cause more severe problems than if the prediction was greater, for example not printing enough examination papers or assigning enough resources. If the prediction is considerably greater than the actual value the obvious problem would be greater cost and a waste of resources. The information that will be presented is number of entries and predicted number of entries as previously mentioned both in a table and graphically. The graph also includes the entries from the previous 5 years. The remaining table that will be presented for each level contains the percentage error for each subject for the 'Best' model, year model, AR model and the SQA prediction. As a lower prediction is deemed worse than a higher prediction a large negative percentage error is the worst possible outcome. The best outcome would be a small positive percentage error, as close to zero as possible.

5.1 Standard Grade

The data received from the SQA containing the SQA predictions and the actual number of entries only contained values for the total number of Standard Grade entries and not individually for Foundation/General and General/Credit. For this reason, predictions for the total number of Standard Grade entries were calculated and displayed in table 3.1.1 below.

				Best			Year			AR		
	Actual	SQA	School	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
Accounting	1 932	1 830	1 937	1 886	1 622	2 150	1 886	1 622	2 150	1 976	1 915	2 038
Art	13 280	14 220	13 319	14 683	13 421	15 945	14 683	13 421	15 945	13 921	13 655	14 188
English	53 927	57 000	54 171	58 424	53 845	63 003	58 760	53 912	63 607	59 895	58 717	61 073
Maths	46 779	49 490	47 334	52 747	47 117	58 378	51 122	47 090	55 153	53 035	52 083	53 987
Physics	14 780	14 610	14 801	18 341	16 786	19 896	14 694	13 243	16 144	14 905	14 618	15 192
Spanish	3 299	3 130	3 311	3 137	2 739	3 535	3 131	2 773	3 489	3 112	3 030	3 193

Table 3.1.1 – Standard Grade Predictions



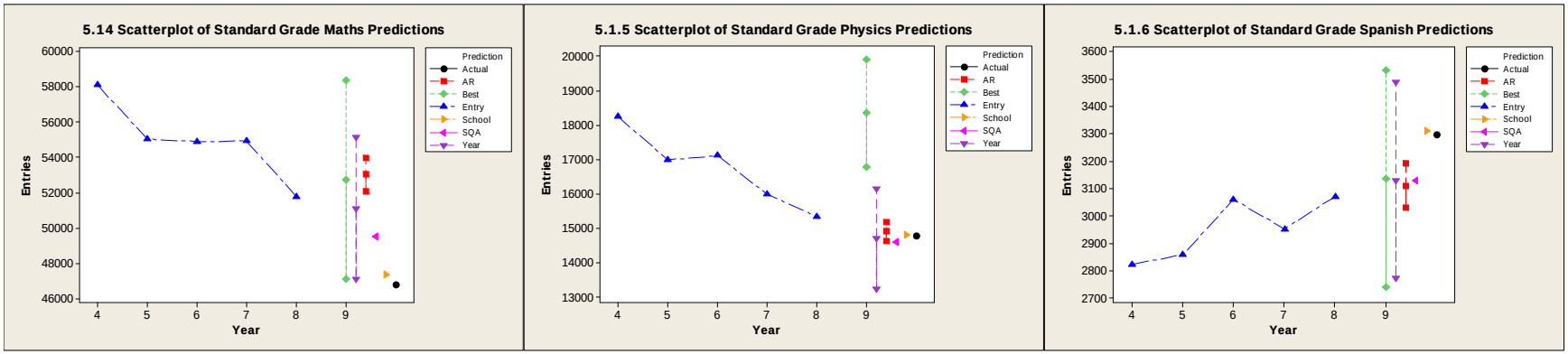


Figure 5.1.1 – 5.1.6 Standard Grade Predictions by subject

Subject	SQA	School	Best Model	Year Model	AR Model
Accounting	-5.280	0.259	-2.391	-2.391	2.301
Art	7.078	0.294	10.566	10.566	4.830
English	5.698	0.452	8.340	8.961	11.067
Maths	5.795	1.186	12.758	9.283	13.373
Physics	-1.150	0.142	24.091	-0.585	0.845
Spanish	-5.123	0.364	-4.910	-5.092	-5.683

Table 3.1.2 –Standard Grade percentage error

From the Figures 5.1.1 – 5.1.6 and Table 3.1.2 it can be seen that the results differ for the subjects. The one consistent conclusion that can be reached is that the prediction produced by the schools is always the closest to the actual number of entries and is also always greater than the actual number of entries. Looking at Art and Mathematics we can see that none of the intervals contain the actual number of entries, they are all greater. The prediction for these two subjects by the SQA is above the actual number of entries. For the remaining subjects there is no clear conclusion. The SQA prediction is only greater than the number of entries for English, for Accounting, Physics and Spanish the SQA prediction is below the number of entries and this would be costly. For Accounting all the intervals include the number of entries, with the interval for the AR model being closest and therefore less costly. For English the interval for the 'Best' model and year model only just include the value and upper values of the interval and then greatly above the number of entries. For physics both the year and AR model contain the value of interest, with the interval for AR being the narrowest and therefore closest to the actual number of entries. The only subject with an interval completely below the actual number of entries and therefore the worst possible scenario is the AR interval for Spanish. The other intervals do include the number of entries.

The percentage error is reasonably small for all subjects with none of the subjects having a percentage greater than 25 and Physics the only subject greater than 20%. The school and college value has a positive value of percentage error for all subjects, it also produces smallest value for all subjects. When comparing only the three methods that produce confidence intervals, the AR model has the lowest percentage error for half of the subjects, Accounting, Art and Physics. For English and Maths the SQA has the smallest positive percentage error, although the percentage error for the AR method is still relatively small and positive. Spanish is the only subject where all 4 methods have a negative percentage error and in this instance the 'Best' model produces the best percentage error.

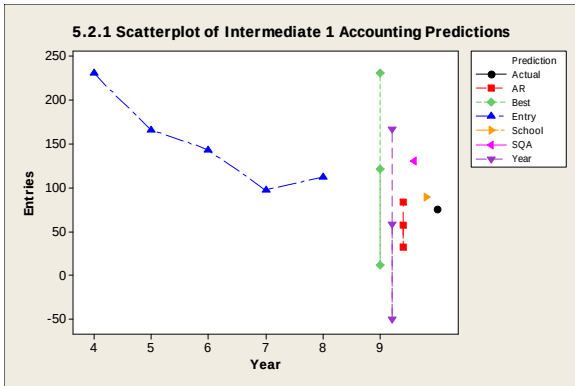
From the Standard Grade tables it can be seen that the value produced by the schools and colleges is always the closest to the actual number of entries and is always above the number of entries. The AR method consistently produces the narrowest interval and if this could

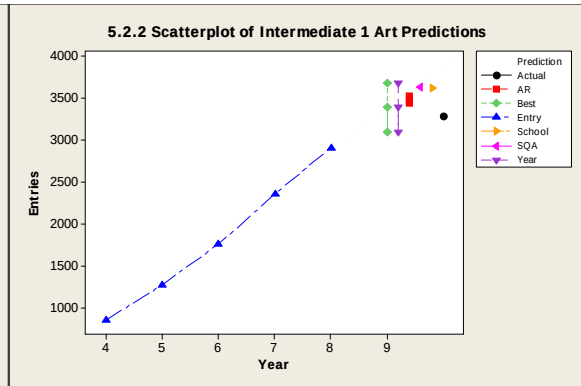
frequently contain the actual number of entries, then would be the most effective method to use. The school method consistently produces the smallest, positive percentage error for all subjects. The results produced by the other prediction methods differ per subjects and therefore it is very difficult to pick one method over the others. The method of prediction that produces the smallest number of negative percentage errors and the lowest values of percentage errors is the AR method. The hardest subject to predict, in terms of percentage error, would be Spanish, only the school method produces a non-negative prediction error.

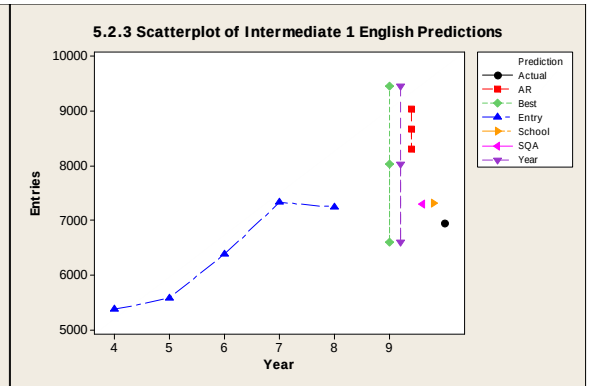
5.2 Intermediate 1

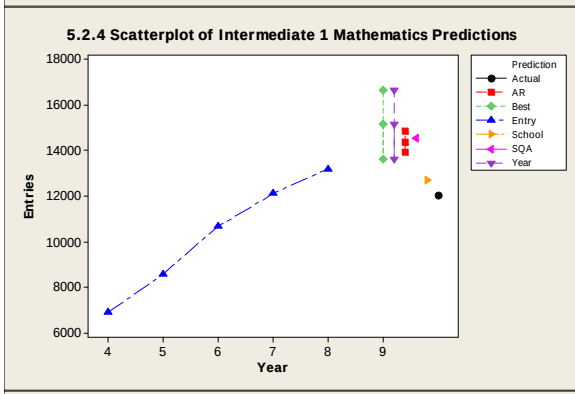
				Best			Year			AR		
	Actual	SQA	School	estimate	lower	upper	estimate	lower	upper	estimate	lower	upper
Accounting	76	130	89	121	11	232	58	-50	167	57	31	84
Art	3 287	3 640	3 623	3 395	3 100	3 690	3 395	3 100	3 690	3 488	3 447	3 529
English	6 955	7 300	7 318	8 031	6 602	9 460	8 031	6 602	9 460	8 667	8 303	9 032
Maths	12 061	14 530	12 716	15 147	13 629	16 666	15 147	13 629	16 666	14 388	13 933	14 844
Physics	2 557	2 880	2 733	3 104	2 445	3 763	3 104	2 445	3 763	2 914	2 866	2 962
Psychology	117	220	130	163	-57	384	197	-78	473	166	86	245
Spanish	805	810	823	726	458	993	832	501	1 162	829	728	931

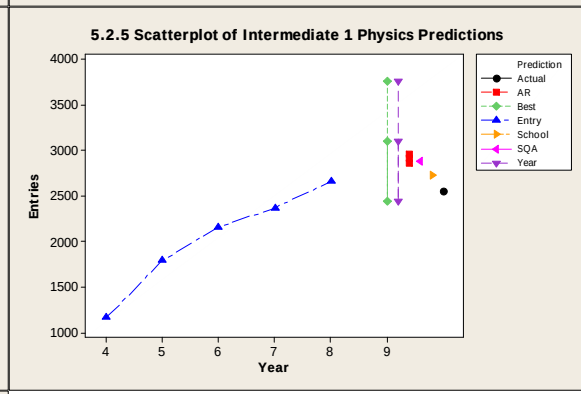
Table 3.2.1 – Intermediate 1 Predictions

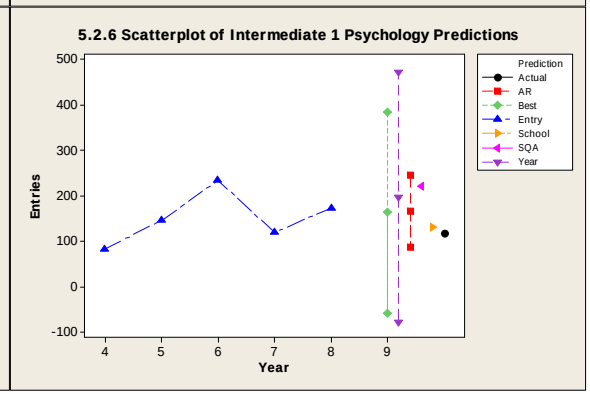

5.2.1 Scatterplot of Intermediate 1 Accounting Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.2.2 Scatterplot of Intermediate 1 Art Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.2.3 Scatterplot of Intermediate 1 English Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.2.4 Scatterplot of Intermediate 1 Mathematics Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.2.5 Scatterplot of Intermediate 1 Physics Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.2.6 Scatterplot of Intermediate 1 Psychology Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year

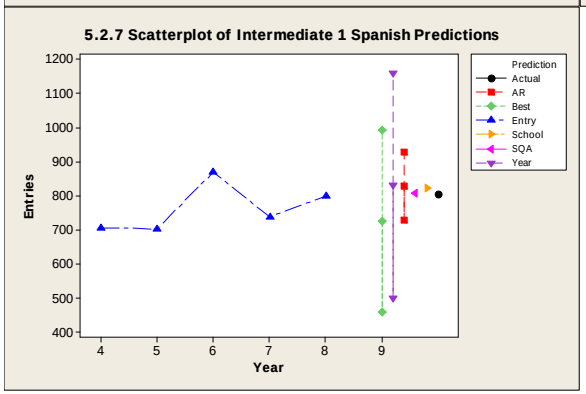

5.2.7 Scatterplot of Intermediate 1 Spanish Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year

Figure 5.2.1 – 5.2.7 Intermediate 1 Predictions by subject

Subject	SQA	School	Best Model	Year Model	AR Model
Accounting	71.053	17.105	59.665	-23.553	-24.366
Art	10.739	10.222	3.289	3.289	6.123
English	4.960	5.219	15.475	15.475	24.620
Maths	20.471	5.431	25.589	25.589	19.296
Physics	12.632	6.883	21.381	21.381	13.947
Psychology	88.034	11.111	39.172	68.376	41.583
Spanish	0.621	2.236	-9.852	3.317	3.040

Table 3.2.2 –Intermediate 1 percentage error

The Intermediate 1 subjects can be separated into 3 groups: those where the number of entries falls into none of the intervals (Mathematics); those where the number of entries falls into all of the intervals (Accounting, Psychology and Spanish) and those where the number of entries falls into the 'Best' and year models only (Art, English and Physics). As with the standard grade entries, the school and college enrolment is consistently the closest to the actual number of entries and is also always above the number of entries. The SQA prediction is also always greater than the number of entries however it is not always as close as the school and college value.

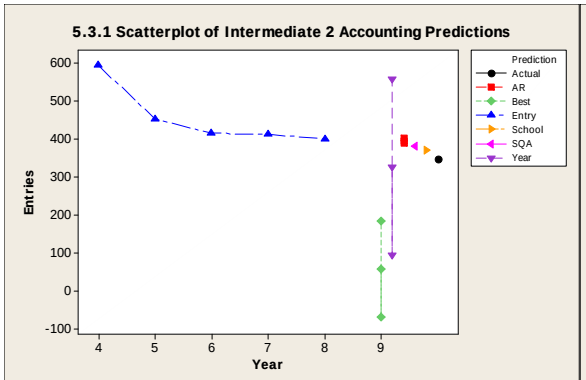
Although the prediction methods, and the lowest, positive percentage error, fall into 3 distinct groups, the pattern of the groups is not the same. The school and college enrolment and the SQA method are the only two methods which produce positive values for all subjects. The school and college value produces the lowest, positive percentage error for all subjects except Art, English and Spanish. The SQA produces the lowest, positive percentage error for the last two subjects and the 'Best' and Year models produce the smallest, positive percentage error for Art. There is a wide range of values of prediction error showing that it is easier to predict some subjects more than others. The only subjects which have any negative values of percentage error are Accounting and Spanish.

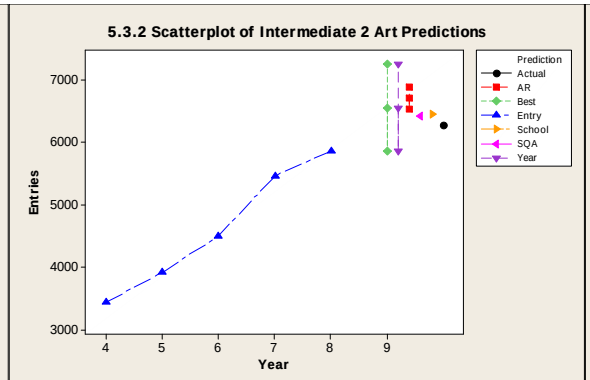
It again appears the school and college value obtains a consistent value above but still close to the actual number of entries, for four of the seven subjects. The SQA method also always produces a value greater than the number of entries but the value is not always as close as the one produced by the school method. Of the remaining prediction methods it appears that there is very little to separate the 'Best' and year models. The AR model still produces the narrowest of intervals but as can be seen these intervals rarely contains the actual number of entries. When taking into account the percentage prediction error the two best methods of prediction are the School and College value, followed by the SQA method.

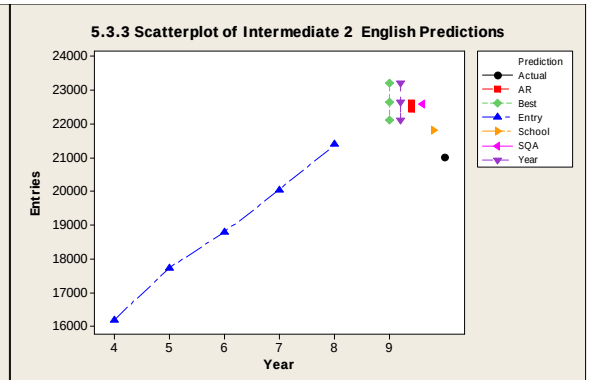
5.3 Intermediate 2

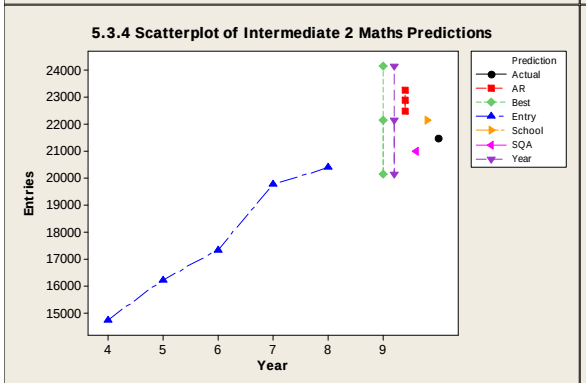
				Best			Year			AR		
	Actual	SQA	School	estimate	lower	upper	estimate	lower	upper	estimate	lower	upper
Accounting	348	380	370	58	-68	185	326	95	557	396	389	403
Art	6 264	6 410	6 452	6 549	5 854	7 243	6 549	5 854	7 243	6 704	6 527	6 881
English	21 025	22 600	21 821	22 657	22 108	23 207	22 657	22 108	23 207	22 540	22 460	22 621
Maths	21 485	21 000	22 167	22 149	20 150	24 147	22 149	20 150	24 147	22 881	22 490	23 271
Physics	3 796	3 800	3 923	3 984	3 187	4 781	3 984	3 187	4 781	4 122	3 919	4 324
Psychology	542	500	629	666	455	878	612	329	895	731	640	822
Spanish	1 224	1 370	1 254	926	293	1 558	1 240	931	1 549	1 372	1 359	1 385

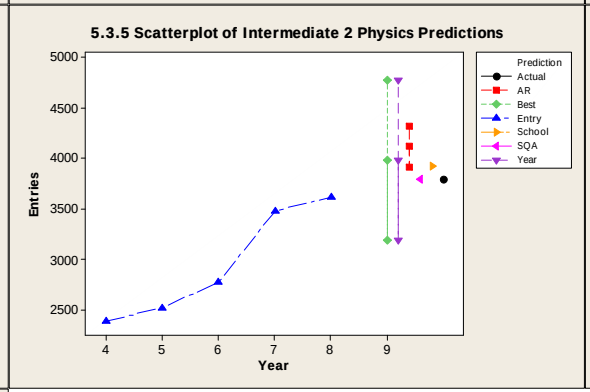
Table 3.3.1 – Intermediate 2 Predictions

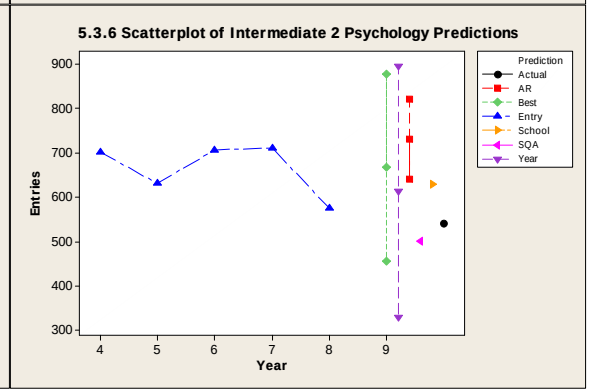

5.3.1 Scatterplot of Intermediate 2 Accounting Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.3.2 Scatterplot of Intermediate 2 Art Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.3.3 Scatterplot of Intermediate 2 English Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.3.4 Scatterplot of Intermediate 2 Maths Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.3.5 Scatterplot of Intermediate 2 Physics Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.3.6 Scatterplot of Intermediate 2 Psychology Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year

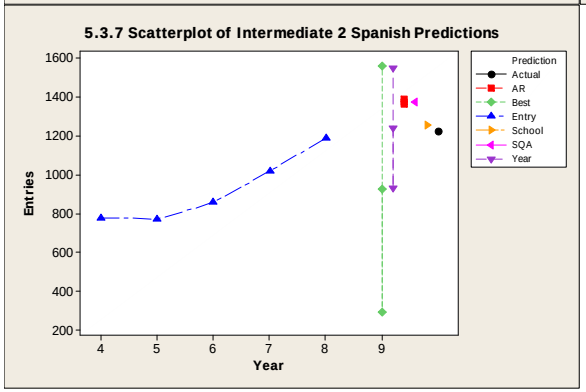

5.3.7 Scatterplot of Intermediate 2 Spanish Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year

Figure 5.3.1 – 5.3.7 Intermediate 2 Predictions by subject

Subject	SQA	School	Best Model	Year Model	AR Model
Accounting	9.195	6.322	-83.243	-6.264	13.671
Art	2.331	3.001	4.545	4.545	7.030
English	7.491	3.786	7.764	7.764	7.208
Maths	-2.257	3.174	3.089	3.089	6.496
Physics	0.105	3.346	4.942	4.942	8.577
Psychology	-7.749	16.052	22.927	12.970	34.882
Spanish	11.928	2.451	-24.379	1.307	12.106

Table 3.3.2 –Intermediate 2 percentage error

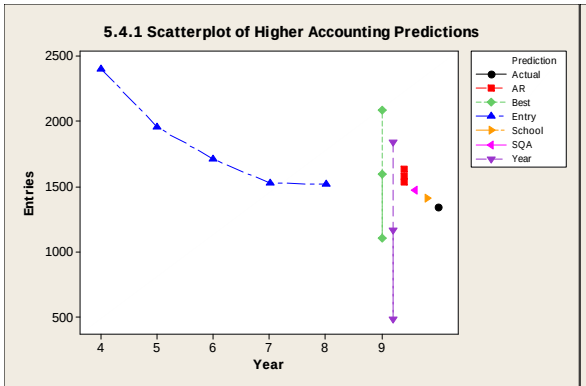
The value produced by the school and college enrolment is again always the closest and always above the actual number of entries. The prediction produced by the SQA method is not always positive and is below the actual number of entries for Mathematics and Psychology. For all the subjects at Intermediate 2 level, except Accounting and Finance and English, the intervals produced by the 'Best' method and the year method contain the actual number of entries. For Accounting and Finance the interval produced by the year model contains the actual number of entries, the interval produced by the 'Best' model falls completely below the number of entries. For the final subject, English, all of the intervals produced lie above the actual number of entries. The intervals produced by the AR method are still the narrowest of all the intervals but do not include the number of entries for any of the subjects and constantly lie completely above the number of entries. This leads to a positive percentage error for all subjects for the AR model, this is also the case for the school and college enrolment value. The values of percentage error produced by the AR method are reasonably small for all subjects with the exception of Psychology. The methods that produce the smallest, positive percentage error values for each individual subject are the School and college value for Accounting and Finance and English, the SQA method, for Art and Physics, the Year method, for Mathematics, Psychology and Spanish.

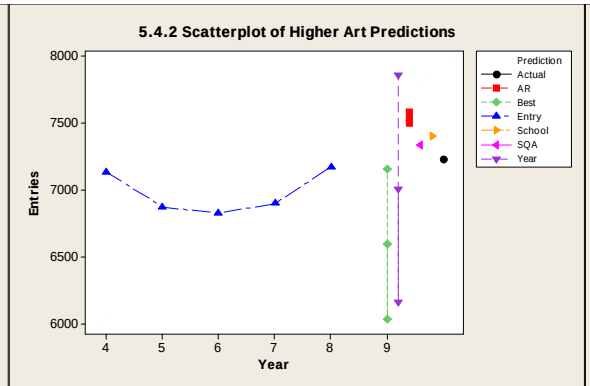
The only method that consistently produces a value that is both close and positive is the school and college value. For the AR method, although the interval never contains the actual number of entries, has a positive percentage error for every subject, as does the school and college value, but never produces the smallest percentage error.

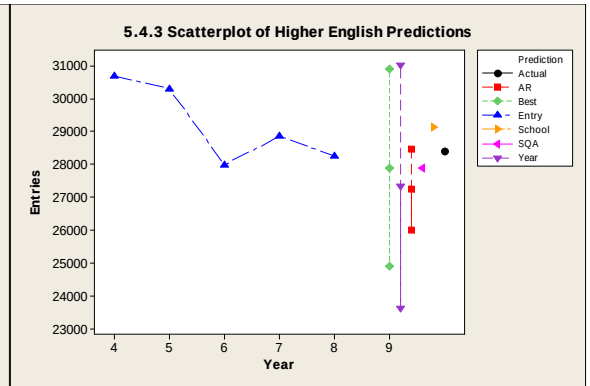
5.4 Higher

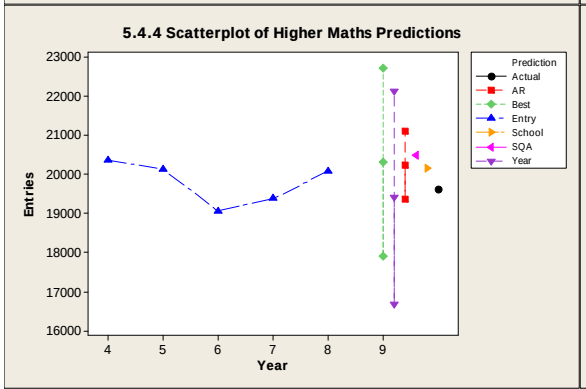
				Best			Year			AR		
	Actual	SQA	School	estimate	lower	upper	estimate	lower	upper	estimate	lower	upper
Accounting	1 344	1 470	1 416	1 598	1 106	2 090	1 164	485	1 843	1 583	1 531	1 634
Art	7 232	7 330	7 400	6 592	6 029	7 155	7 008	6 158	7 857	7 538	7 497	7 579
English	28 389	27 900	29 119	27 894	24 900	30 887	27 330	23 637	31 022	27 244	26 020	28 469
Maths	19 631	20 500	20 155	20 319	17 895	22 744	19 419	16 684	22 153	20 250	19 381	21 118
Physics	9 001	9 100	9 225	8 567	7 283	9 852	8 568	7 283	9 852	9 109	8 758	9 459
Psychology	2 762	2 590	3 102	2 694	2 197	3 191	2 694	2 197	3 191	2 601	2 506	2 696
Spanish	1 364	1 490	1 386	1 198	765	1 630	1 351	974	1 728	1 957	1 940	1 975

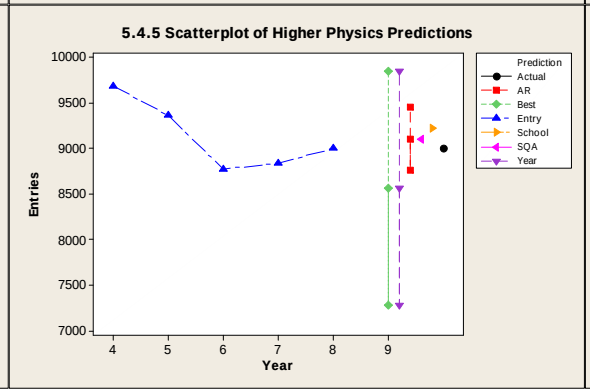
Table 3.4.1 – Higher Predictions

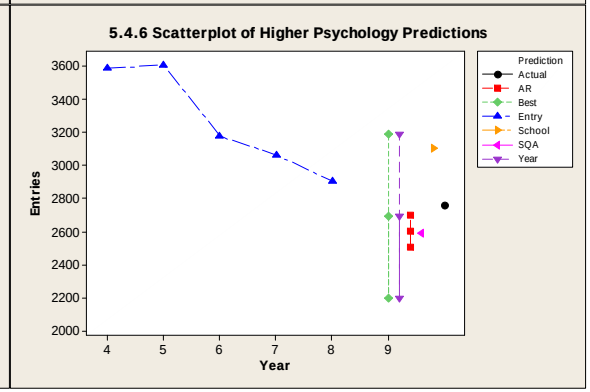

5.4.1 Scatterplot of Higher Accounting Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.4.2 Scatterplot of Higher Art Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.4.3 Scatterplot of Higher English Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.4.4 Scatterplot of Higher Maths Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.4.5 Scatterplot of Higher Physics Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year


5.4.6 Scatterplot of Higher Psychology Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year

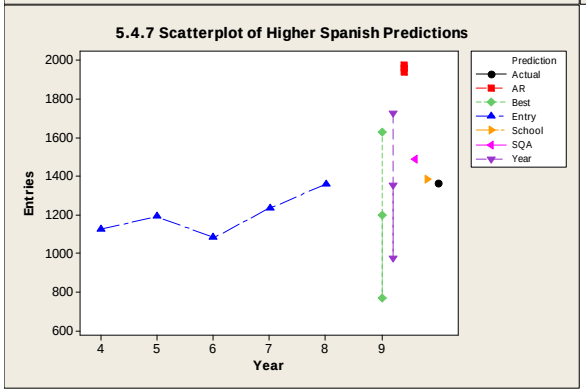

5.4.7 Scatterplot of Higher Spanish Predictions
Entries
Year
Prediction
Actual
AR
Best
Entry
School
SQA
Year

Figure 5.4.1 – 5.4.7 Higher Predictions by subject

Subject	SQA	School	Best Model	Year Model	AR Model
Accounting	9.375	5.357	18.906	-13.393	17.753
Art	1.355	2.323	-8.846	-3.103	4.233
English	-1.722	2.571	-1.745	-3.731	-4.032
Maths	4.427	2.669	3.506	-1.082	3.151
Physics	1.100	2.489	-4.818	-4.818	1.198
Psychology	-6.227	12.310	-2.458	-2.458	-5.818
Spanish	9.238	1.613	-12.205	-0.960	43.511

Table 3.4.2 –Higher percentage error

The value produced by the school and college is the only value which always produces a value higher than the number of estimates. For four of the seven subjects, Accounting and Finance, Art, Psychology and Spanish, the interval produced by the AR method does not include the value of the actual number of entries. The interval produced by the 'Best' model does not include the number of entries for Art. The year model contains the actual number of entries for all subjects.

When taking the percentage error of prediction into consideration it would appear that the only method that never produces a negative value is the school and college value, for psychology it is the only method which produces a positive value. For five subjects the value produced by the schools and colleges is the smallest, positive percentage error, for the remaining subjects, Art and Physics, the smallest, positive error is produced by the SQA method.

When looking at both the prediction values produced and the percentage prediction error it would appear that the only method which produces a prediction for every subject that is above the actual entry and is the closest value in five out of the seven subjects would be the schools and colleges enrolment value.

5.5 Advanced Higher

				Best			Year			AR		
	Actual	SQA	School	estimate	lower	upper	estimate	lower	upper	estimate	lower	upper
Accounting	78	70	80	36	-23	97	36	-23	97	9	-5	25
Art	1 544	1 620	1 675	1 471	1 235	1 706	1 490	1 229	1 750	1 322	1 312	1 332
English	1 590	1 830	1 633	1 665	1 467	1 863	1 703	1 378	2 028	1 823	1 822	1 824
Maths	3 027	2 900	3 090	2 840	2 305	3 375	2 840	2 305	3 375	2 755	2 721	2 789
Physics	1 550	1 420	1 580	1 446	1 397	1 495	1 409	1 331	1 487	1 389	1 369	1 409
Spanish	196	200	201	165	54	276	177	57	297	97	76	118

Table 3.5.1 – Advanced Higher Predictions

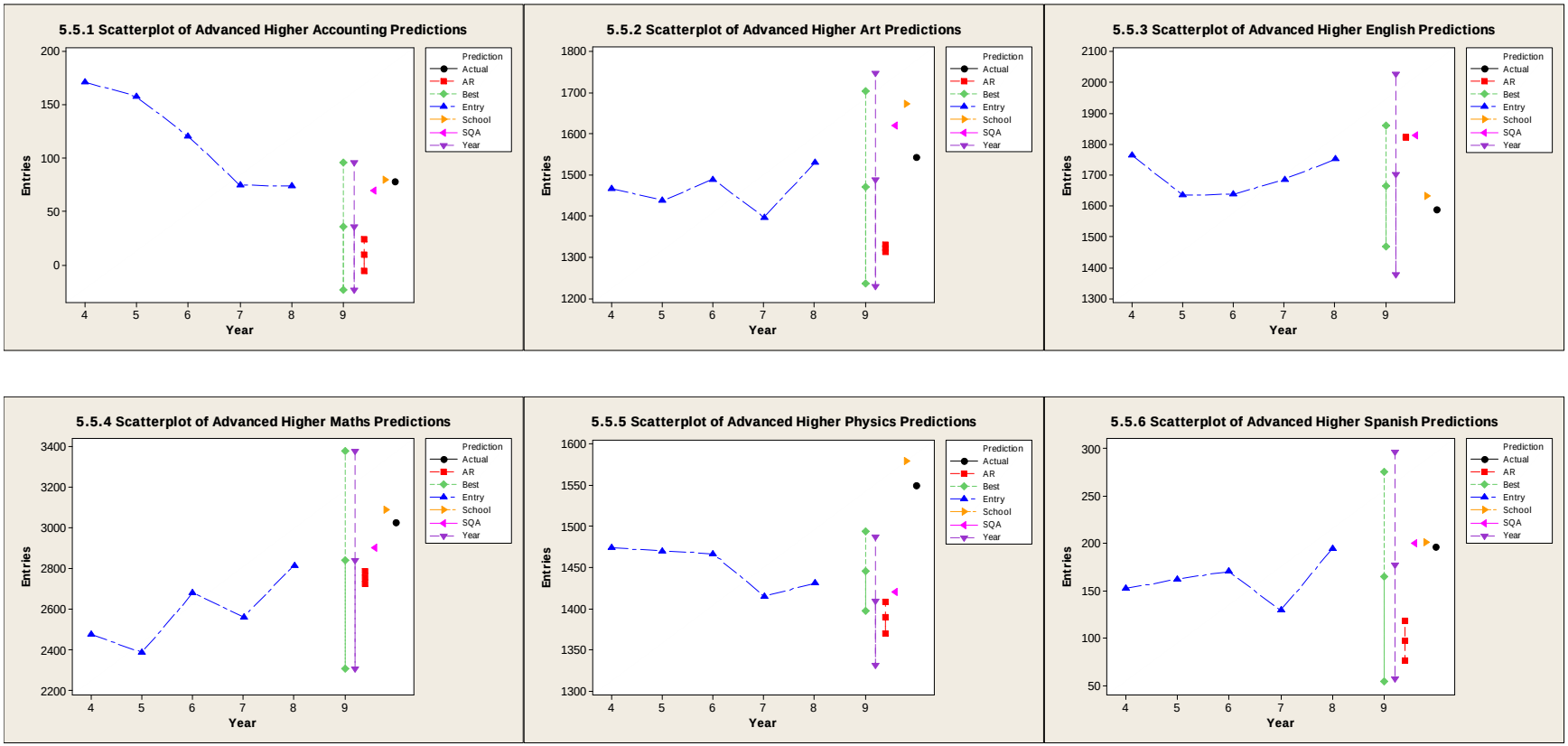


Figure 5.5.1 – 5.5.6 Advanced Higher Predictions by subject

Subject	SQA	School	Best Model	Year Model	AR Model
Accounting	-10.256	2.564	-53.462	-53.462	-87.900
Art	4.922	8.484	-4.758	-3.530	-14.383
English	15.094	2.704	4.706	7.113	14.663
Maths	-4.196	2.081	-6.178	-6.178	-8.986
Physics	-8.387	1.935	-6.734	-9.090	-10.386
Spanish	2.041	2.551	-15.903	-9.847	-50.416

Table 3.5.2 –Advanced Higher percentage error

As with all previous levels the value produced by the schools and colleges is the only value which produces a number greater than the actual number of entries for all subjects. The SQA method produces a prediction that is lower than the actual number of entries for three subjects, Accounting and Finance, Mathematics and Physics. For all subjects, with the exception of physics, the intervals produced by the year and 'best' methods include the actual number of entries. For Physics only the estimate produced by the Schools and colleges is greater than the actual number of entries. The AR method does not include the actual number of entries for any of the subjects; the entire interval is below the actual number of entries for all subjects except English.

The SQA method only produces the smallest, positive percentage prediction error for Art and Spanish, for the remaining subjects the School and colleges produces the smallest, positive percentage error. The school method is the only method which has a positive value of percentage error for all subjects. The estimate produced by the school and colleges method comes too late to be used by the SQA and therefore could not be used to make any form of planning decisions, either short or long term.

Chapter 6

Discussion and Conclusions

6.1 Summary

The main aim of this study was to take the current system used by the SQA for predicting the number of entries for NQs and either provide statistical reasoning to justify continuing in the same way, by showing that the prediction is accurate and that no better computational method can be found, or provide a more accurate forecasting tool. If this was possible then the next step was to extend the forecasting to long term forecasts to provide a more accurate tool for the SQA in order to make long term planning decisions.

It was very clear from the beginning that, due to the limitations posed by the data, the main limitation being the limited number of time points available, something which as previously mentioned could not be remedied, typical statistical methods could not be employed to test this and other solutions needed to be researched. The main resolution to this problem was to bootstrap the sample. There were various methods of bootstrapping employed to give a variety of results and these were bootstrap confidence limits; simple percentile confidence limits, bias-corrected percentile confidence limits and accelerated bias-corrected percentile limits.

6.1.1 Initial Impressions

Chapter 3 was used in several ways to explore the data and gain subjective impressions. The two methods used in this chapter to do that were time series plots of each level and also one of total Credit and Intermediate 2 passes summed together and Higher entries and also regression models. The purpose of the final plot was to examine if there was a relationship between the number of passes the previous year at the level below and the number of entries the next year for the level above.

The conclusions drawn from the time series plots are unique to each subject and level. In subjects where there is a trend decrease at Standard Grade this is occasionally countered by an increase in Intermediate 2, which has progressively more often been used as a replacement for General/Credit Standard Grade as a matter of school policy. Due to the fact that the number of subjects a student can select is fixed, it is believed, but it cannot be tested, that a decrease in one subject may be due to an increase in another subject, as pupils' interests and abilities alter from year to year.

Regression models were fitted for Higher entries, 5th Year Higher entries only, General/Credit entries and Higher entries modelled using lagged National Rating and lagged National Rating with year. The 'best' model for predicting the number of entries was assessed as having the largest R-squared and the smallest s (residual standard deviation). The regression models confirm the impression from the graphs that it is extremely difficult to make generalisations across subjects or even across levels. When focusing on 5th Year Higher entries the increasing in being able to accurately predict the number of entries was marginal, an increase in the adjusted R-squared value of 5.9%, from 44.6% to 50.5%, for Spanish only. By modelling only 5th Year Higher entries, in order to gain an accurate total number of entries for a subject at this level, all years which are capable of sitting that examination would need to be modelled separately and then summed together. The lack of increased predictability when modelling 5th Year Higher entries only suggests that the addition work required does not produce a valid increase in the accuracy of the result.

6.1.2 Results

Chapter 4 displays all the formal results obtained during this study. It was believed that it may be possible that the number of entries in successive years are related. If this is the case then the relationship can be exploited when predicting the number of entries for the coming year. A formal test to see if the values are related year on year is the Durbin-Watson test. Chapter 4 also contains the results from the Root Mean Squared Error of Prediction (RMSEP) for small sample sizes. This assesses how accurate the prediction produced by a model is.

The autocorrelation, tested by the Durbin-Watson statistic, was tested only on the model contain year as the explanatory variable for every subject as each level. It was also used to examine if there was any gender difference visible at General/Credit level and Higher Level. It was found that the OLS method of estimating ρ was unreliable and produced invalid

estimates outside the limits [-1,1], therefore the alternative method, estimating ρ as the correlation between the residuals and the lagged residuals from OLS regression model, was used to examine if any autocorrelation existed. The results produced by the OLS method are more often significant than those produced by the alternative, correlation approach.

The three methods used often produced different results and different conclusions. The p-values produced by D-W MC and ρ (B – ρ) can result in different conclusions or when the conclusions are the same the magnitude of the p-values can be extremely different. The ρ (B – ρ) method often produces a smaller p-value than the D-W MC method. The point estimates produced for ρ are all negative suggesting that the number of entries alternate between having a large number of entries followed by a smaller number of entries. Most of the subjects have non-significant 95% confidence intervals and those intervals correspond to significant p-values from the other methods. The subjects that do have significant 95% confidence intervals do not have p-values that are significant. This shows that there is no relationship between the p-values produced by the D-W MC and ρ (B – ρ) methods and the 95% confidence intervals from the ab – c method. The overall conclusion reached using the D-W is again that no common rule can applied to each level or each subject.

The D-W statistic was also used to examine if when splitting the entries by gender they were easier to model. This was not the case and no more information was gained and it was not easier to model. Still no common conclusion was reached.

The next step of chapter 4 was looking at the root mean squared error of prediction to select the model that will produce the most accurate prediction. A variety of models was examined, a simple approach just using the possible explanatory variables and selecting the ones which produce the lowest RMSEP, an AR model, which was used as the D-W statistic, suggested that some subjects may have autocorrelation and RMSEP was also used to see if a common model could be found for languages and sciences.

Using the RMSEP again confirmed that there appears to be no common way to model the number of entries for each subject and each level. It does not help us to provide a better forecasting tool for the SQA by using different models for each subject and each level. The more complex AR model only proved to produce more accurate predictions in a handful of cases. The AR models produce different conclusions from the ab – c method for the D-W

statistic. That is models which had a significant ab – c interval did not always produce a better value for RMSEP.

6.1.3 Prediction

The final chapter of results compares the actual number of entries in 2009 with the predictions for 2009 from the various methods; the SQA prediction, the school and colleges enrolment value, the 'Best' method, the year method and the AR method. The predictions were calculated and then the percentage error of prediction, how close the prediction is to the number of entries, was calculated.

The first clear result to be seen when looking at the predictions is that on average the best method of prediction is the school and college method. This is best in terms of the smallest, positive percentage prediction error. The school and college method never produces a prediction value below the actual number of entries. The SQA method produces a value that is consistently one of the closest values to the actual number of entries. Unlike the school and college method the SQA method does not always produce a value that is above the number of entries.

The remaining three methods produce confidence intervals. The purpose of the interval is not only to provide a point estimate but to show how confident you are that the prediction produced is close to the actual number of entries. The AR method consistently produces the narrowest interval. If the interval could be relied upon to always contain the value of the number of entries then this would be the best method to use. A narrow interval, which contains the number of entries, would mean that the upper limit of the confidence interval would always be above the actual number of entries, this would avoid the issue of not printing enough papers for an examination, the upper limit would also be close to the actual number of entries and therefore the amount of waste would be limited.

Looking at the 3 interval methods the Year method contains the actual number of entries in the interval more times than the other two methods. This does not account for the width of the intervals in any way, just the fact that of the three methods this is the one that is most reliable to contain the actual number of entries.

Comparing each method individually, then it can be seen, as mentioned previously, that the school and college enrolment value is the only method which has a positive percentage error of every subject at all levels. The next best method, in terms of prediction error is the SQA method, which has a negative value for 10 out of the 33 possible subject and level combinations. Of the remaining three methods the one with the greatest number of positive percentage error values is the AR method followed by the 'Best' method and the worst method, in terms of percentage error is the Year method, which has positive values for only 16 out of a possible 33 predictions. This is in contrast with the conclusion reached when taking into account whether or not the interval produced contains the number of entries.

It would appear that no method comes suitably close to the school method, with either a point estimate or a prediction interval. The school and college value appears to produce a result that is more often than not the closest value and always greater than the number of students who actually sat the examination.

6.2 Limitations of the Study and Further Work

There are several limitations to the study and to the conclusions drawn. The first and most obvious is the nature of the dataset available. The issues arise from the lack and volume of that data at the same time. The lack of data causes problems as the data is only available from 5 years and it is not sensible to go back any further because of major changes in the education system, the qualifications and the subjects. The changes meant that data before 2004 could not be used to model the current system or subjects as they are too dissimilar. The small amount of data means that the usual statistical methods cannot be used. It also means that any conclusions drawn may be inaccurate or highly correlated to the data used.

The issue that arises from the volume of data is the sheer number and variety of subjects. Certain subjects are only available at a particular level and therefore this limits the explanatory variables available for modelling. The time and resources needed to model every possible subject at every available level, as done for the 7 subjects in this thesis, would be huge and highly unlikely. A possible way around this would be to try and group subjects together and model groups of subjects. This causes a whole new set of problems like how to group the subjects. What makes subjects similar? Is it subjects with similar number of entries, subjects in a similar genre for example sciences or languages or use statistical methods to

decide which subjects are similar, again this raises more questions than answers. The limited amount of data would restrict any methods of trying to group subjects.

Another problem would be new subjects or subjects at a new level. There would be no data available on previous entries or results to predict the number of entries and an alternative method may need to be introduced if the model uses these variables. If a grouping method was used then the new subject or level would need to be assigned to a group. The lack of data would mean that it would be several years before a new subject or level could be included into any prediction tool.

Another issue that would cause the prediction tool to be required to be updated would be changes in the examination or school system. This may mean that previous year's data may not be able to be used and that the prediction model may not work if it relies on previous results or entries. As with the introduction of new subjects, a different method of prediction may need to be found.

If a model could be used to produce accurate predictions, either one general model or individual models for each subject or a group of subjects, then the model will need to be updated to include the new data available each year. This will mean that the coefficients in the model will need to be recalculated. As new data becomes available it will become easier to accurately model the predictions and this may have an effect on the model, variables which are used to predict the number of entries may not be as effective and variables which are not used may produce better predictions. The model, or models, will need to be re-evaluated with each new year of data.

In order to produce the most accurate predictions the whole system would need to be modelled. This is an impossible task. There are pressures and influences that cannot be modelled, influence from parents, siblings and friends and students preferences and abilities are just some of the external influences that cannot be measured. Other factors which can affect the subjects a student selects are the subjects the school offers and the school's choices form. If two subjects that a student wishes to choose are in the same column then they must decide on only one subject. It is also possible that the other subjects a student has selected will influence their remaining choices, for example if they have already selected a science they may not wish to choose another one. Future career or university or college entry requirements may also influence student's decisions. There are many more influences that are

not mentioned that cannot be measured and included in a prediction tool. It will never be possible to model the whole system and achieve predictions which will be completely accurate.

It was seen that the best prediction available is the prediction produced by the schools. This prediction could be used as a base, instead of working from previous results and past entries. This base may then be used to produce intervals and may provide a more accurate forecasting tool than those presented here.

6.3 Further Work

There were many issues that were required to be overcome when modelling the number of entries in order to predict future entries. Some of these issues were resolved, for example the limited data, others were not, modelling all subjects for instance. The main conclusion that can be drawn is that each subject and each level requires a different model to produce the most accurate predictions. This conclusion is linked extremely closely to the current data; new data may alter the conclusions in this thesis drastically.

Using a model to produce the predictions would require masses more resources, in both time and people and does not produce significantly better predictions than the current SQA method, for all the additional effort.

The most accurate predictions would be achieved by modelling the whole system as one, all the subjects together to model the trade off between subjects, modelling students preferences and any external influences. This is not possible and therefore there will always be a large amount of variability in any prediction produced.

Using each student's individual data from 4^{th} through 5^{th} and 6^{th} year to model predictions may be of interest. This may be useful as, previously mentions in section 3.1, most students will proceed to Higher if they have already passed the subject at Credit or Intermediate 2 and a student will only progress to Advanced Higher if they passed the same subject at Higher.

A limited amount of time was available to investigate whether it might be possible to model groups of subjects rather than many different subjects. Further methods of grouping the subjects together should be investigated to see if different common model can provide a more accurate prediction.

It would appear that the current method used by the SQA would be an adequate method for producing predictions, as the extra effect needed to find the best model for each subject at every level, go on to produce predictions and update the models every year does not produce predictions which are dramatically better than the SQA weighted average method.

Bibliography

Box, G. E. P. and G. M. Jenkins (1970), 'Time Series Analysis: Forecasting and Control', San Francisco, Holden-Day

Breidt, F. Jay, Richard A. Davis and William T. M. Dunsmuir (1995), 'Improved Bootstrap Prediction Intervals for Autoregressions', Journal of Time Series Analysis, Vol.16 No.2

Clements, Michael P. and Nick Taylor (2001), 'Bootstrapping Prediction Intervals for Autoregressive Models', International Journal of Forecasting, Vol. 17

Durbin, J. and G. S. Watson (1950), 'Testing for Serial Correlation in Least Squares Regression. I', Biometrika, Vol. 37

Durbin, J. and G. S. Watson (1951), 'Testing for Serial Correlation in Least Squares Regression. II', Biometrika, Vol. 38 No. ½

Durbin, J. and G. S. Watson (1971), 'Testing for Serial Correlation in Least Squares Regression. III', Biometrika, Vol. 58

Efron, B. (1979) 'Bootstrap Methods: Another Look at the Jackknife', The Annals of Statistics, Vol. 7 No. 1

Efron, Bradley (1983), 'Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation', Journal of the American Statistical Association, Vol. 78 No. 382

Efron, Bradley (1987), 'Better Bootstrap Confidence Intervals', Journal of American Statistical Association, Vol. 82 No. 397

Efron, Bradley and Gail Gong (1983), 'A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation', The American Statistician, Vol. 38 No. 1

Efron, Bradley and Robert Tibshirani (1993), 'An Introduction to the Bootstrap', Chapman & Hall

Granger, Clive W. J. (2001) 'Essays in Econometrics. Collected Papers of Clive W. J. Granger. Volume II: Causality, Integration and Cointegration, and Long Memory', Econometric Society Monographs

Harville, David A. and Daniel R. Jeske (1992), 'Mean Squared Error of Estimation or Prediction Under a General Linear Model', Journal of American Statistical Association, Vol. 87 No. 419

Jeong, Jinook and Seoung Chung (2001), 'Bootstrap Tests for Autocorrelation', Computation Statistics & Data Analysis, Vol. 38

MacKinnon, James G. (2002), 'Bootstrap Inference in Econometrics', Canadian Journal of Economics, Vol. 35 No. 4

Manly, Bryan F. J. (1997), 'Randomization, Bootstrap and Monte Carlo Methods in Biology', Chapman & Hall

Mevik, Bjorn-Helge and Henrik Rene Cederkvist (2004), 'Mean Squared Error of Prediction (MSEP) Estimates for Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR)', Journal of Chemometrics, Vol. 18

Orcutt, Guy H. and Herbert S. Winokur, Jr. (1969), 'First Order Autoregression: Inference, Estimation, and Prediction', Econometrica, Vol. 37 No. 1

Savin, N. E. and Kenneth J. White (1977), 'The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors', Econometrica, Vol. 45 No. 8

Sheiner, Lewis B. and Stuart L. Beal (1981), 'Some Suggestions for Measuring Predictive Performance', Journal of Pharmacokinetics and Biopharmaceutics, Vol. 9 No. 4

Stine, Robert A. (1987), 'Estimating Properties of Autoregressive Forecasts', Journal of American Statistical Association, Vol. 82 No. 400

Supit, I. (1997), 'Predicting National Wheat Yields Using a Crop Simulation and Trend Models', Agricultural and Forest Meteorology Vol. 88

Thombs, Lori A. and William R. Schucany (1990), 'Bootstrap Prediction Intervals for Autoregression', Journal of American Statistical Association, Vol. 85 No. 410

Wallach, D. and B. Goffinet (1989), 'Mean Squared Error of Prediction as a Criterion for Evaluating and Comparing System Models', Ecological Modelling, Vol. 44

Wallach, D. and B. Goffinet (1987), 'Mean Squared Error of Prediction in Models for Studying Ecological and Agronomic Systems', Biometrics, Vol. 43