Statistical Analysis of Performance Indicators in UK Higher Education

submitted by

Mark John Gittoes

for the degree of Ph.D.

of the

University of Bath

2001

Summary

In this thesis I examine a variety of statistical issues surrounding the construction and interpretation of *performance indicators*, which are intended to be measures of how effectively institutions such as schools, hospitals, and universities are carrying out their public mandate. A number of fundamental questions are raised and examined, principally using a UK Higher Education case-study involving a binary outcome variable:

- the analysis of large sparse contingency tables in an observational study involving many *potential confounding factors* (PCFs);
- the statistical formalisation of a non-model-based *input-output* quality assessment approach using indirect standardisation;
- the generation of reliable estimators of the difference $\hat{D} = \hat{O} - \hat{E}$ between observed and expected institutional success rates, based on a significant number of PCFs;
- the creation of a valid repeated-sampling standard error for $\hat{D}$, and the demonstration of its validity via a large number of calibration simulations;
- the demonstration of the practical equivalence between the indirect standardisation approach and a particular form of fixed-effects multilevel model;
- an examination of the link between fixed- and random-effects quality-assessment methodologies;
- the presentation of results for a variety of artificial and real worlds exhibiting wide variation in the sparseness of the data;
- the development of a reporting system for institutions that identifies areas of both excellence and concern (differential effectiveness);
- the creation of a longitudinal model structure for examining performance indicators across time;
- the development of a studentized method for assessing institutions in such a way that their observed outcomes do not affect their expected results; and
- the comparison of several IO analyses with *gold standards* based on the direct measurement of the quality of institutional processes.

"Am I a professor? Goodness. I expect I was hopeless, was I?"
- Gilderoy Lockhart

"Well ... when we were in our first year, Harry - young, carefree, and innocent - well more innocent than we are now anyway."
- George Weasley

"A computer is a stupid machine with the ability to do incredibly smart things, while computer programmers are smart people with the ability to do incredibly stupid things. They are, in short, a dangerously perfect match."
- Bill Bryson

"Judge a man by his questions rather than his answers."
- Voltaire

Acknowledgements

Firstly I would like to thank my parents for their continuing encouragement of my academic studies and for giving me enough sense to choose Professor David Draper as my supervisor. In most acknowledgements, it would usual to write something like "David's guidance throughout my PhD was invaluable and I would not have completed this PhD without his help". That is obviously true but I wish I could provide the words to show David's true worth. I can't except to say that if I needed advice on anything, David would be a first port of call. My thanks also to Andrea, David's wife, for her help during the comings and goings of my American adventure.

I would like to thank the HEFCE team (Ali, Judy, John and Mark) for their experience, knowledge and for supplying the data to make this project possible.

Thanks to all those who survived working with in 4.19, especially Vicky and Emily (who combined this with living with me as well) and Jon (for pointing out when I was being statistically and sportingly stupid). I would also like to thank those who shared a house with me over the last three years but I'd like to especially thank a few of these fine people: Marianne, my sister, for just being about when I needed her; Aaron for drinking more than me; and Paul, Gordon and Matt for supporting worse football teams than me.

I would like to thank: Anyone who improved my football playing skills over the last three years which helped me to keep sane, especially David Hobson (for his silly hats and general advice), Jock (for not being English), Mark Penney (for believing we could restart the Maths football team), Pete (for scoring goals), Alf (for organising the league), Marie (for being mad enough to support the team), and Rich and Andy (for being the other Maths captains); Beth and Madeline for putting up with my mad rantings for so long; Mark Willis for his computing advice; The gang: Bob (for being madder than me), James (for providing entertainment), Marie (for being a Swedish blonde), Jezza (for being Jezza), Matt P. (for showing me you can finish in three years), Matt L. (for fantasy formula one and keeping my younger sister sane), Sarah (for being social everything), Joerg (for a Germanic point of view), Fas (for his programming advice) and the rest (you know who you are); and Hayley, Marky F., Rich R., Liz H. and Claire H. for that moral support stuff.

Special thanks to the Department of Mathematical Sciences especially the members of the Statistics Group, and to the EPSRC for their financial support. Finally thanks to anyone I've forgotten who helped me in anyway with my PhD and to the Romans for building Bath.

Preface

All singular subjects in this thesis are female by convention.

It is HEFCE policy not to publish information about named institutions unless the institutions have first had the opportunity to see the figures. As HEFCE supplied this data, all institutions in this document have been anonymised.

Those who are acquainted with the multilevel modelling package `MLwiN` should note that the convention in this thesis regarding levels 1 and 2 of two-level models is opposite to that in `MLwiN`: here level 1 (indexed by i) represents universities and level 2 (indexed by j) represents students nested within universities, as in Scheffé (1959).

Contents

List of Tables

List of Figures

Chapter 1

Problem introduction

1.1 Motivation of problem

Over the last decade, there has been increasing pressure placed on the UK Government to assess institutional performance within the public sector. This has heralded a dramatic increase in the use of league tables. The application of performance indicators in the public sector is now widespread. They are also used on a smaller scale in private companies. The Government's drive to improve monitoring and quality assessment in all institutions under its control has rapidly increased their importance in recent years.

In March 2000 the British Medical Association published BMA (2000), which discussed a number of issues relating to hospital league tables, clinical indicators and other associated issues. They state that "... crude league tables are misleading and it is time for a complete change in methods of analysis of data". The report recommends that health trusts should be legally required to collect, maintain and provide data of sufficient quality to enable accurate performance indicators to be measured. They also indicate that comparative data for institutions should be adjusted for case mix and relative risk using the most up to date and rigorous methodologies. With this in mind, along with many others factors, the NHS executive have recently published their second set of NHS performance indicators (NHS (2000)). The report is based on high level performance and clinical indicators, which provide information on individual NHS hospital Trusts and Health Authorities. The focus of these NHS indicators is on the quality of services as well as the efficiency of service delivery to their patients and the public.

One of the early drivers for performance indicators was to establish fair comparisons in school achievements. Goldstein (1997) warns about the use of crude league-tables in assessing schools, but also highlights the problems involved the more complex "value-added" tables. The two principal problems highlighted are: the inability of the tables to indicate schools that are "differentially effective" (i.e., perform better with certain types of students); and the scores attached to each school for quality assessment in value-added (or crude) tables typically have a large margin of error associated with them. Goldstein (2000) indicates that a higher degree of sophistication is required within educational league-tables and that a less political ideal is

required in school assessment. Goldstein suggests that political interest is driving misleading assessments and may be doing fundamental harm to education.

Further Education(FE) performance indicators have also been published by the Further Education Funding Council (FEFC) (FEFC (2000)) which provides information on 426 FE and sixth form colleges in England and 216 LEA maintained and independent external institutions. These FE indicators have three main purposes:

- to enable institutions to compare themselves to their peers;
- to provide information to the FE fund-holders; and
- to enable institutions to monitor changes in performances at each institution over time.

Goldstein and Spiegelhalter (1996) make an extensive overview of the use of performance indicators in both health and educational applications. They also examined a potential framework for comparing institutions when using outcome data. The article concludes that there will always be limitations in making comparisons between institutions and any results produced should be treated as "suggestive rather than definitive". Goldstein and Spiegelhalter (1996) is still considered as one of the most extensive papers on performance indicators when taken with its associated discussion.

1.2 Performance indicators in higher education

The Government have extended their assessments to try and make universities accountable for their actions and results. Following the recommendation of the National Committee of Inquiry into Higher Education, the Government approached the funding councils to develop suitable indicators and benchmarks of performance in the HE sector. The Performance Indicators Steering Group (PISG) was set-up to take forward these recommendations.

Very few countries have implemented HE performance indicators on a country-wide scale. Australia have published one document on the characteristics of its HE institutions (HEDA (1998)). They had four principal aims they hoped the publication would cover: add to the information available to prospective HE students and those that aid the student's HE decisions; for institutions to compare their performance to their peers; to illustrate the diversity in HE establishments; and finally, to contribute to public accountability.

The UK indicators would focus on the requirements of the Government, funding councils and management of institutions, although any performance indicator (PI) would have interest to a wider audience. The funding councils, including the Higher Education Funding Council for England(HEFCE), have published a number of reports regarding these performance indicators, including HEFCE (1999a) and HEFCE (1999b).

HEFCE (1999b) states:

"Recognising the diversity of higher education (HE), the purpose of performance indicators is to:

- provide better and more reliable information on the nature and performance of the UK higher education sector as a whole;
- influence policy developments; and
- contribute to the public accountability of higher education."

These reports permit the Government and HE institutions to identify good practice and help disseminate it throughout the sector.

In HE, performance indicators are outcomes that are perceived to be a measure of the quality with which a university is carrying out its public mandate. The PISG identified three major groups of performance indicators for HE:

- Access - These indicators provide information regarding how accessible universities are to under-represented groups.

  The PISG decided on three access indicators: the percentage of students who attended a school or college in the state sector; the percentage whose parent's occupation is classed as skilled manual, semi-skilled or unskilled; and the percentage whose home area is known to have a low proportion of 18 and 19 year olds attending HE.

- Learning and Teaching - The group identified a number of indicators that would help to assess learning and teaching in universities.

  One of these indicators concentrates on student non-continuation. This establishes the percentage of students who are absent from the HE system (apart from those graduating) in October 1997, having started at at a new university in October 1996. This variable would give an idea of how well a university is doing at holding on to its students. There are two principal reasons that students should be absent from university a year after entry. The student could have been asked to leave by the university for failing to meet the standards of that institution. Or the student may have decided that the HE life was not for her, due to financial difficulties or emotional problems for example. A number of papers have examined the reasons why students drop-out of HE. Brunsden and Davies (2000) updates, expands and extends the work of Tinto (1975), who completed some of the early work on student drop-out. Brunsden and Davies (2000) highlights the effects of a student's academic and social integration into university life on her probability of drop-out.

- Research - How the university performed in research relative to the resources it had at its disposal.

  The measures used here are the number of PhDs awarded and the amount of research grants and contracts obtained.

The Select Committee on Education and Employment (SCEE), working on behalf of the Government, have very recently reviewed the importance and use of performance indicators in

HE (SCEE (2001)). They repeat previous concerns about access to HE for under-represented groups but do indicate that some progress has already been made in improving access. The SCEE note that the HEFCE access performance indicators, which are based on the work completed in this thesis, have helped to highlight problem areas in the HE institutions. They also state that "higher education institutions should consider these performance indicators regularly ... as a means of examining their own performance and setting new targets".

I will focus on only one of these indicators, student non-continuation, which falls into the "Learning and Teaching Outcomes" measures. For the majority of my analysis, I will use data, provided by HEFCE, for students starting in October 1996. Data from students starting in October 1997 was also made available to me which I also examined but to a lesser extent.

When looking at student non-continuation (or drop-out), consider a student who enters a course at an institution in a particular year. In the following year that student may continue at the same institution, transfer to another institution, or be absent from HE completely. A few students might graduate in this year, like those on a HND course, but they are counted as continuing in HE. For the purposes of this study, I will create a dichotomous variable by grouping students who continue at an institution or transfer to another HE institution as progressing, and those students not in HE in the next year as non-progressing. Other groupings are possible, i.e., coding "transferring to another university" as non-progressing, but for the sake of this analysis I will use the groupings as stated. This does not mean that I consider this to be the correct way of grouping the data, it's just the way the PISG decided to group it.

Due to the way in which HEFCE collect student records, students who drop out very early in the academic year are not included in this data. This omission will obviously affect any results given but it only relates to the initial data and not the methods involved in developing quality indicators. The principal focus of this thesis is the methods involved in calculating performance indicators rather than the results that specific datasets provide.

1.3 Statistical base

The outcome in this problem is student non-continuation or drop-out. To create an optimistic outcome variable (Y) for students, we will consider student progression as a positive outcome. The tables published by HEFCE (HEFCE (1999a)) also use student progression rather than student drop-out. This means the binary outcome variable has the form:

$$Y_i = \begin{cases} 0 & \text{if student } i \text{ does not continue in HE for a second year} \\ 1 & \text{if student } i \text{ does progress} \end{cases}$$

We have a variable that is the underlying quality of the university, which obviously affects the outcome variable Y. Let the quality of the university be defined as S. This is an unobserved or unmeasured variable and is called the supposedly casual factor (SCF).

In standard modelling, the variable whose effect on the outcome factor is being modelled, is called the treatment or risk factor (Anderson et al. (1980)). The main difference between these terms is that the treatment is something that can be applied specifically to affect the outcome

variable Y, whereas the risk factor is accidental or uncontrollable. In our modelling, the SCF is an unobserved risk factor.

So in our statistical set-up, we have a large observational study (Rubin (1974)) where students choose their university using a decision mechanism that could be confounded with the quality of the university, the SCF. A observational study is one in which the observations are selected by some means not chosen by the investigator (Cochran (1983)). In any observational study where a treatment effect is being estimated, it is critical to identify any bias in the analysis. Biases in observational studies are usually caused by confounding factors.

A confounding factor, defined by Anderson et al. (1980), is any variable that has the following properties:

1. is statistically associated with the risk factor, or in our case the SCF;

2. directly affects the outcome Y.

Thus in our modelling, we must consider all potential confounding factors (PCFs) as possible biases on the SCF. These PCFs are variables that could potentially be correlated with student progression at a university and also with university quality. It is unlikely that all the PCFs will be identified, but it is critical to discover as many PCFs as possible. PCFs not taken into account may cause biased and flawed results.

1.4 Multilevel data

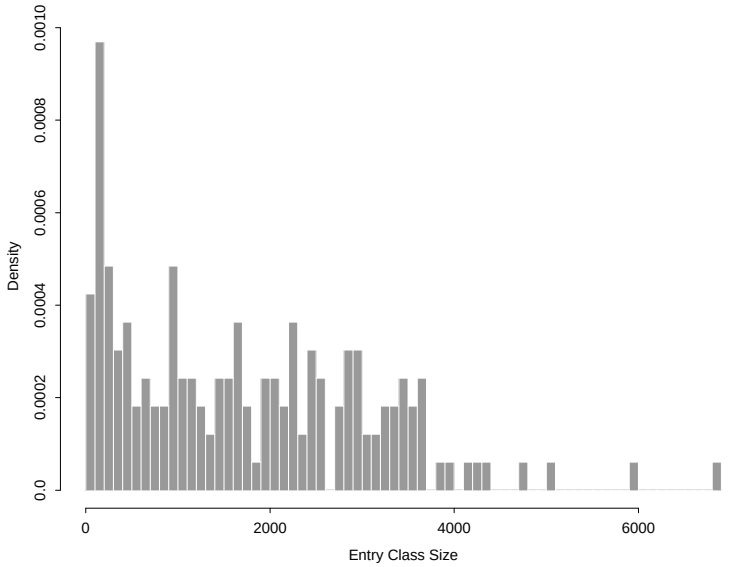

Figure 1-1: Histogram of numbers of entering students in 1996–97.

The data I examine has a multi-level structure with 284,399 students nested within 165 universities. Multilevel data has a hierarchical structure and is commonly found in both observational and designed experiments. It is important not to ignore this clustering structure in

datasets of this type as individuals in the same group tend to be similar in their performance. This was shown by Aitkin et al. (1981), where the statistical effects of ignoring a multilevel structure in educational data are highlighted. Wilkinson et al. (1999) is one of many examples that examine and describe how individuals in the same clusters have a large influence on her peers.

There have been a number of publications focusing on the analysis of multilevel models with regard to quality assessment. Goldstein (1995) provides a good overview of multilevel models in both general and educational datasets and examines the effects of model-based approaches on assessment, with a particular interest in random-effects modelling. Raudenbush and Willms (1991) is a series of papers from the international community that aimed to show that multilevel analyses were no longer restricted to being extremely technical and that a variety of more practical applications were taking place. Yang and Woodhouse (2001) examine a large multilevel dataset, with nearly 700,000 students in 2,794 institutions. They analyse how quality assessment adjustment methods affect the apparent effects of gender and institutional effectiveness based on student A-level outcomes.

In the multilevel quality assessment literature, there seems to be an overwhelming interest in effect sizes, e.g., how much effect does being a male rather than female have on attainment, and not so much targeting of institutional effects. In these multilevel articles, the writers seem to shy away from describing methods for specific institutions with unusual performance. This is a worrying feature but with the increased interest in league tables and other such assessment exercises, this trend should die out.

There is a large variation in the entry class size in our data, ranging from Institution 33 accepting 55 students to Institution 1 receiving nearly 7,000. Figure 1-1 shows how the numbers of students vary at each of the 165 UK universities in the 1996/1997 data.

1.5 Progression rates

The overall progression rate for students in the 1996/1997 data is 90.1%. The lowest progression rate is recorded at the Institution 33, with only 80.0% of their students progressing, and the highest rate, 99.0%, is achieved at Institution 14.

Figure 1-2 shows that the majority of universities have progression rates between 87% and 97%. The rate distribution has a bimodal flavour to it with peaks at 89.0% and 94.5%. Figure 1-3 examines the effect of university size (based on number of students entering) on a university's progression rate. There is an indication that smaller universities have higher progression rates but this effect diminishes when a university has more than 2,000 students.

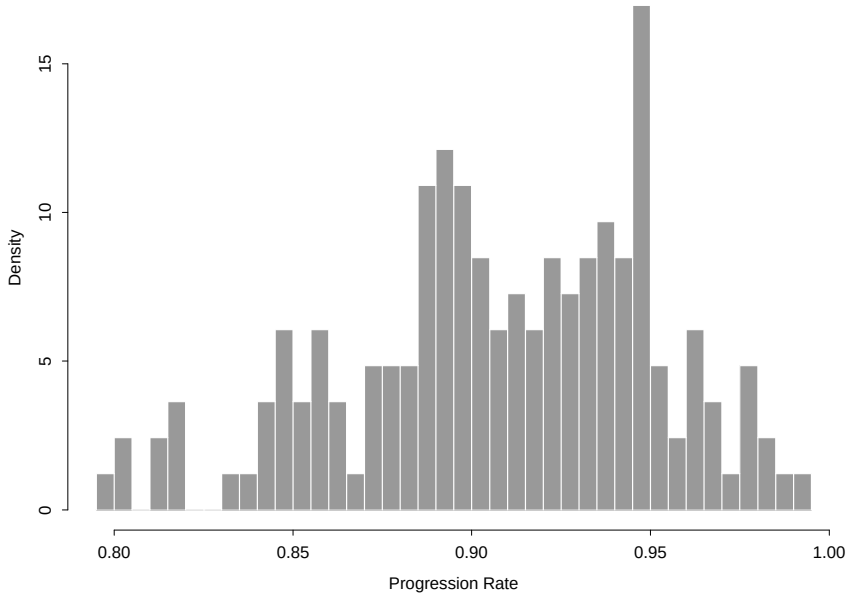

Figure 1-2: Histogram showing the distribution of progression rates for universities.

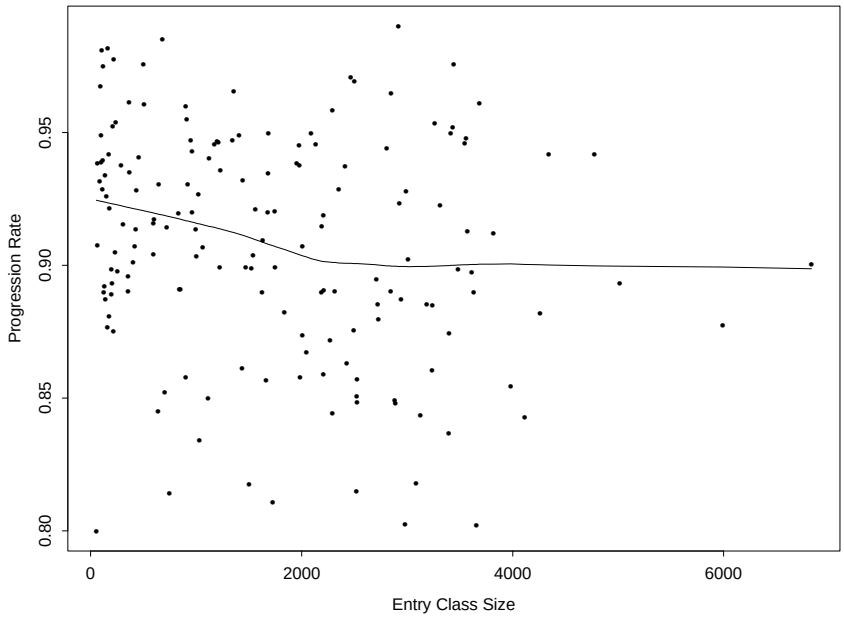

Figure 1-3: University size against progression rate.

1.6 Potential confounding factors (PCFs)

Introduction to PCFs

By linking the UCAS and Higher Education Statistics Agency (HESA) databases, HEFCE was able to provide the following eight PCFs for possible inclusion in the analysis. All these PCFs are at the student level. The appropriateness of these PCFs is not the main focus of debate for this thesis, but they were chosen based upon data availability, previous HEFCE work on progression and suitability. The data only pertain to full-time students with valid progression information. A full-time student is defined as someone who is studying full-time at an institution provided that the course is expected to last at least 24 weeks. There may be students whose course length is unknown within the data and these are classed as full-time, a definitional decision made by HEFCE.

The following sections define each of the eight PCFs and highlight their progression distributions.

Student school type

This variable is taken straight from UCAS records. It measures where the student has applied from, prior to university. It has three levels:

- The student attended a *state school* as her latest school education.
  This includes students from sixth form colleges and FE colleges.
- The student attended an *independent school* as her latest school education.
- School information is *unknown*.
  This could be down to a variety of reasons. For example the person might be a mature student, which may mean she did not apply directly from a school or FE college, or the student did not come through UCAS.

School Type	Freq.	Percent	Progression Rate
State	146295	51.4	.925
Independent	34159	12.0	.938
Unknown	103945	36.6	.856
Total	284399	100.0	.901

The table shows that, based on the marginal distribution, pupils from independent schools have the highest progression rate and pupils whose schools are unknown have the lowest continuation rates. This seems sensible as, historically, independent school students usually have higher entry qualifications (a good guide for progression) and those students who have no school information are, in general, mature students (who have a lower retention rate). Over a third of the dataset's population have an unknown school type.

Student age

This is a binary variable, with the following coding:

- *0 - Young Students*,
  who are those who enter an institution when they are aged less than 21; specifically they must be under 21 on the 30th September of the academic year in which they first enter the institution. For this data, i.e., students entering in 1996-1997, young students are those born after 30 September 1975.
- *1 - Mature Students*,
  who are 21 or over on the 30th of September in the academic year 1996-1997.

Age	Freq.	Percent	Progression Rate
Young	202494	71.2	.924
Mature	81905	28.8	.847
Total	284399	100.0	.901

The retention rates are as expected, with mature students finding it more difficult to remain in HE. There is a large difference between the two progression rates: 7.7 percentage points.

Student gender

Another binary variable with:

- *0 - Female Student*
- *1 - Male Student*

Gender	Freq.	Percent	Progression Rate
Female	145659	51.2	.917
Male	138740	48.8	.885
Total	284399	100.0	.901

Unusually there are more females entering HE in this year than males. The feminine progression rate is higher than the masculine one, with a marginal difference of 3.2 percentage points.

Student qualifications

This is a 21 level categorical variable. Most of the categories below are self explanatory, but there are a few points to note. The grouping has been chosen so that as far as possible the students within each group are relatively homogeneous.

- The A-level pts categories include Scottish Higher points, i.e., six Scottish Higher points count as six A-level points. This isn't a particularly efficient grouping but is sufficient for this study.
- The student qualifications are based upon data collected from UCAS and the universities themselves. This can cause problems as some universities define student qualifications differently to other universities.
- A pts 4 means 1-4 A-level pts and A pts 8 means 5-8 A-level pts. A pts 10 means 9-10 A-level pts. The higher A-level points use the logical pattern, i.e., the 2 pt range.
- When a student's qualification is completely missing, she goes into the unknown section and when the student is known to have A-level points but not how many, she is classified as A pts Unknown.

Entry Qual.	Freq.	Percent	Progression Rate
None	5234	1.8	.787
Others	8211	2.9	.793
Unknown	8031	2.8	.832
BTEC/ONC	15308	5.4	.849
GNVQ3+	8015	2.8	.856
HE	26493	9.3	.852
Access/Foundation	21906	7.7	.864
A pts Unknown	13611	4.8	.871
A pts 4	8353	2.9	.869
A pts 8	17678	6.2	.889
A pts 10	12138	4.3	.903
A pts 12	13434	4.7	.905
A pts 14	14539	5.1	.915
A pts 16	15112	5.3	.926
A pts 18	15375	5.4	.942
A pts 20	15371	5.4	.945
A pts 22	13763	4.8	.952
A pts 24	13275	4.7	.961
A pts 26	12289	4.3	.967
A pts 28	10891	3.8	.972
A pts 30	15372	5.4	.984
Total	284399	100.0	.901

Some interesting features of the marginal qualifications distribution are: there seems to be a very uniform distribution of students in the categories from A pts 10 to A pts 30; nearly 50,000 students already hold a HE qualification or have entered from access courses; students with no qualifications have the lowest retention rates (as expected); and considering only those students with A-levels, there is a strictly increasing distribution of progression rates from A pts 4 to A pts 30 (86.9% for students holding 1-4 A-level pts rising to 98.4% for students with 29-30 pts).

Parental occupation

This variable is taken from UCAS. It is based upon parental occupation for the main household earner, which is classified using the Standard Occupational Classification. The social classifications used are:

1. Social Class I - Professional;
2. Social Class II - Intermediate;
3. Social Class III N - Skilled non-manual;
4. Social Class III M - Skilled manual;
5. Social Class IV - Semi-skilled;
6. Social Class V - Unskilled.

These are combined as follows to create a three level variable:

- *Lower Class*:
  students whose parent's occupation is grouped within social classes III M, IV or V;
- *Higher Class*:
  students whose parent's occupation is grouped within social classes I, II or III N;
- *Unknown*:
  this information is not known about a student's parents.

Social Class	Freq.	Percent	Progression Rate
Lower	50480	17.8	.905
Higher	144250	50.7	.929
Unknown	89669	31.5	.854
Total	284399	100.0	.901

Half the student population are from the "higher" social classes and these students have the highest HE retention rates. This is partially down to her financial stability. Students with an unknown social class make up nearly a third of the data and these individuals have the highest non-continuation rates: 14.6% of them fail to progress into her second year of HE.

Year of (program of) HE study

All students within the data are defined as students who are recorded as commencing a programme of study at an institution during the academic year of interest (1996-1997). While most students go into her *first year* of a program of study, some will start on her *second, or later*, year of programme, e.g., students who transferred from another institution, or those who have gained additional credits at other institutions. Therefore this is a two level variable.

Study Year	Freq.	Percent	Progression Rate
2+ or Unknown	32612	11.5	.878
First Year	251787	88.5	.904
Total	284399	100.0	.901

Around 11% of students are defined as having an unknown year of study, or are in her second year or later. These students have the lower retention rate. This is to be expected as a proportion of them are students who have already decided to move from one university to other.

Low higher education participation

Enumeration districts across the country were clustered into 160 groups, which were then used for the analysis. Each student was allocated to an enumeration district, and so to one of these clusters, by her post code.Then the proportion of young people (aged under 21) in each of these areas who entered HE in academic years 1995-1996, 1996-1997 and 1997-1998 have been calculated. Areas for which this participation rate was less than two thirds of the UK average over the whole period have been defined as low participation neighbourhoods. This variable was developed from a study made by HEFCE.

Therefore students can either:

- come from a *low* participation post code area;
- come from a *non-low* participation post code area; or
- have *unknown* participation information. This is usually either because the information on where she lives is unreliable or the area in which she lives has not be classified.

Participation Status	Freq.	Percent	Progression Rate
Low	37955	13.4	.875
Non-Low	234878	82.6	.908
Unknown	11566	4.1	.846
Total	284399	100.0	.901

Over four fifths of the dataset come from areas of higher participation rates, with only a small 4.1% having an unknown participation status. Students from areas of low participation find it more difficult to remain at university than those from a higher participation area. This is probably down to the fact that low participation rate areas are more likely to be socially and financially deprived regions of the country, making student retention at university more difficult.

Student subject of study

This is taken from the subject of student qualification aim on entry.

The full subject categories are:

- Engineering and Technology;
- Mathematical sciences and Computer science;
- Architecture, Building, and Planning;
- Combined subjects - any combined studies programmes whose subjects lie completely within one category have been included in that category;
- Business and administrative studies and Librarianship & information science;
- Agriculture and related subjects;
- Social studies and Law;
- Subjects allied to medicine;
- Creative arts and Design;
- Education;
- Biological sciences, and Physical sciences;
- Languages and Humanities; and
- Medicine, Dentistry, and Veterinary science.

Subject	Freq.	Percent	Progression Rate
Engineering	22638	8.0	.864
Maths & Comp	20569	7.2	.877
Architecture	7151	2.5	.877
Combined	34748	12.2	.889
Business	36610	12.9	.891
Agriculture	2511	0.9	.903
SocSt + Law	33567	11.8	.903
Allied to M	15240	5.4	.904
Art + Design	23317	8.2	.906
Education	14543	5.1	.913
Biol + Phys	36498	12.8	.916
Lang + Hum	30199	10.6	.930
Medicine	6808	2.4	.980
Total	284399	100.0	.901

The largest subject category is the business, with agriculture being the subject area with the least amount of students. Medicine has the smallest drop-out rate, with only 2% of its students deciding to drop-out of HE. There are a number of potential reasons for this ranging from medicine students being highly motivated to the high qualifications required to read medicine. Engineering and the Mathematical Sciences have the greatest problem with retention rates, with only around 87% of their students progressing.

Chapter 2

The HEFCE approach

2.1 HEFCE's adjustment method

HEFCE's method for inferring the supposedly causal factor (SCF), university quality, is a version of input-output (Draper (1995)) or league-table analysis (Goldstein and Spiegelhalter (1996)). These approaches use an adjustment method to establish quality, i.e., you were given X, you provided Y and you should have produced Z given your inputs X. The difference between Y and Z is the starting point for quality assessment. Both Draper and Goldstein point out that the source of the data in a quality analysis is as important as the methods for carrying out the assessment. The idea behind the HEFCE approach is as follows: Picture a black box with a number of inputs and a single output. You are not allowed to open the black box and so you are unable to measure or define directly measure what is happening within the box. Therefore you must use only your inputs and output to define the processes occurring within the box. In a perfect world, we would like to directly the process within the box but this is typically difficult and/or expensive. This approach is discussed further in Draper (1996).

In our specific problem, the black box is the institution of interest, i.e., a university. I am not trying to formally measure the processes within the university. That means that no direct measure of, for example, teaching quality or support services for students is being taken. All that is available are the inputs, i.e., the student and general university characteristics, and the output, in our case a single binary outcome, student progression. The approach means that the quality of the university, i.e., the process going on in the box, is inferred indirectly.

The HEFCE method involves using a massive cross tabulation to calculate the inferred quality of a university. Picture taking a single PCF and cross tabulating it with another PCF. This would create a table of size c_1 by c_2, where c_i is the number of levels in PCF i. For example, if the binary indicator for student age is crossed with the binary indicator for student gender, then a two by two table would be created.

Now cross this two dimensional table with a third PCF to create a three-dimensional "shoe-box" shape, where the shoe box is split into $c_1c_2c_3$ cells, where c_3 is the number of levels of the

third PCF. Continue this process until you have included all the necessary W PCFs into the cross tabulation. A W-dimensional shape is created which is separated into a large number of cells, $C = c_1 c_2 ... c_W$.

Each student falls into one and only one of these cells. Each cell identifies individuals with a specific set of characteristics, for instance a young male student from a poor background with good entry qualifications who is studying mathematics. Some of these C cells will contain no individuals, meaning that students with certain characteristics do not exist in the dataset. A good example in our data might be students with very poor entry qualifications studying medicine.

Rather than picturing these cells within a q-dimensional shape, consider placing each cell in a long line of C cells and then removing those cells with no students in. This creates a single line of M cells where $M \leq C$. This process so far has taken no account of the university or multilevel structure of the data. I now complete a further cross-tabulation of these M cells against the N institutions in the analysis. A N by M table is created, where a single cell contains students at a certain university with a unique set of characteristics. Some of the cells in this new table will be empty as a selection of the universities will not have students with specific characteristics that other universities will have. For example, an arts university will not contain any students studying biology, whereas the majority of other HE establishments will have biological students.

After these tabulations, the following grids can be defined:

	PCF Categories				Weighted
University	1	2	...	M	Row Mean
1	$\hat{p}_{11}$	$\hat{p}_{12}$	...	$\hat{p}_{1M}$	$\hat{p}_{1\cdot}$
2	$\hat{p}_{21}$	$\hat{p}_{22}$	...	$\hat{p}_{2M}$	$\hat{p}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N	$\hat{p}_{N1}$	$\hat{p}_{N2}$	...	$\hat{p}_{NM}$	$\hat{p}_{N\cdot}$
Weighted Column Mean	$\hat{p}_{\cdot 1}$	$\hat{p}_{\cdot 2}$	...	$\hat{p}_{\cdot M}$	$\hat{p}_{\cdot\cdot}$

Table 2.1: Proportions grid for PCF by university breakdown.

	PCF Categories				Row
University	1	2	...	M	Sum
1	n_{11}	n_{12}	...	n_{1M}	n_{1+}
2	n_{21}	n_{22}	...	n_{2M}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N	n_{N1}	n_{N2}	...	n_{NM}	n_{N+}
Column Sum	n_{+1}	n_{+2}	...	n_{+M}	n_{++}

Table 2.2: Numbers grid for PCF by university breakdown.

Cell ij contains n_{ij} individuals from institution i with PCF category j characteristics. The proportion of these individuals whose progression outcome is successful is $\hat{p}_{ij}$. Each weighted row mean, $\hat{p}_{i.}$, is the observed progression rate for university i and can be written in terms of the $\hat{p}_{ij}$ and n_{ij}:

$$\hat{p}_{i.} = \frac{\sum_{k=1}^{M} n_{ik}\hat{p}_{ik}}{\sum_{k=1}^{M} n_{ik}} = n_{i+}^{-1} \sum_{k=1}^{M} n_{ik}\hat{p}_{ik} \tag{2.1}$$

The weighted column mean, $\hat{p}_{.j}$ is the observed national progression rate for individuals in PCF category j and has the following form:

$$\hat{p}_{.j} = \frac{\sum_{k=1}^{N} n_{kj}\hat{p}_{kj}}{\sum_{k=1}^{N} n_{kj}} = n_{+j}^{-1} \sum_{k=1}^{N} n_{kj}\hat{p}_{kj} \tag{2.2}$$

$\hat{p}_{..}$ is the overall success rate for all students and can be calculated as a weighted mean of the rows or the columns. n_{+j} is the number of students from the whole population that fall into PCF category j:

$$n_{+j} = \sum_{i=1}^{N} n_{ij}$$

and n_{i+} is the entry class size to university i:

$$n_{i+} = \sum_{j=1}^{M} n_{ij}$$

The HEFCE method then creates an institutional progression benchmark using a standardisation approach, which is covered in the forthcoming sections.

2.2 Example datasets

Motivation for example datasets

For illustration I introduce four different datasets derived from the original 1996/1997 student information provided by HEFCE. These different sets are used throughout the thesis to provide a numerical and practical guide to the theoretical results. In all cases, universities will never lose any of their students but the dataset will be reduced in size by either losing whole universities or by not taking certain PCFs into account. The following subsections describe each of these "worlds" in descending size.

Big World: the original dataset

The Big World contains all of the students and adjusts for all eight PCFs available. This means the dataset contains $n_{++} = 284,399$ individuals in $N = 165$ universities. Adjusting for the

eight PCFs produces $M = 17,799$ different PCF classes. This is the complete data matrix for the HE case study.

HEFCE Publication World

The tables published by HEFCE (HEFCE (1999a), HEFCE (2000)) only adjust for two PCFs: student qualifications and subject of study. This adjustment produces $M = 272$ PCF categories. The published tables are based on all the students in the 1996/1997 data, so there are again $n_{++} = 284,399$ individuals divided amongst $N = 165$ universities.

Medium World

The Medium World is made up of $N = 10$ universities, which were chosen at random. They are: Institution 1 - Institution 10. Class sizes for these universities range from 116 (Institution 33) to 6831 (Institution 1).

Four PCFs are adjusted for in the Medium World. They are: age; gender; low HE participation; and school type. There are $n_{++} = 23318$ students, who fall into $M = 36$ unique PCF characteristic categories.

Small World

Consider a Small World with only $N = 5$ universities and $M = 4$ PCF categories, defined by cross-tabulating two PCFs only: age against gender. This set-up allows us to picture Tables 2.1 and 2.2 in practical terms, as in Tables 2.3 and 2.4.

We can see that the Institution 33 has no young students registering and, in fact, only has 55 new students in total. As Institution 33 has the lowest observed progression rate in the original dataset, it also has the smallest rate in this reduced world (80%). These universities were chosen to form a good variety from the original 165 universities. Institution 118 and Institution 14 have high progression rates; Institution 1 has a near average rate; and Institution 42 and Institution 33 have some of the lowest observed progression rates. The overall Small World progression rate is larger than that in the Big World (92.4% vs 90.1%).

$\hat{p}_{ij}$	PCF Categories				
	Young		Mature		Weighted
University	Male	Female	Male	Female	Mean
33	—	—	.800	.800	.800
42	.838	.889	.819	.878	.858
1	.897	.923	.868	.905	.900
118	.941	.972	.884	.887	.947
14	.993	.992	.958	.969	.990
Weighted Mean	.933	.947	.870	.903	.924

Table 2.3: Proportions grid for the Small World.

n_{ij}	PCF Categories				
	Young		Mature		Row
University	Male	Female	Male	Female	Sum
33	0	0	5	50	55
42	198	271	227	205	901
1	2133	2099	1443	1156	6831
118	712	501	69	62	1344
14	1467	1176	144	130	2917
Column Sum	4510	4047	1888	1603	12048

Table 2.4: Numbers grid for the Small World.

2.3 Standardisation

Introduction to standardisation

All standardisation methods have a counter-factual (Holland (1986)) flavour to them. The idea behind counterfactuals is the "what if ..." question. We know the outcome under certain conditions but what if the conditions had been something else, what would the outcome have been? This counterfactual approach is regularly used in clinical trials. For example, the factual information might be "Mr Blunkett, given Drug A for 10 days, showed signs of mental instability after 30 days". The counterfactual question to this might be "Would Mr Blunkett have showed signs of mental instability after 30 days if he had received Drug B instead?". It is important to note, and the clinical trial example highlights this, that there isn't just one potential counterfactual and it is important to identify which counterfactual you are interested in.

In the HE analysis, the factual data is the observed progression rate (O_i) at university i. The counterfactual data of interest is what the expected progression rate (E_i) would have been at university i, if .. . One of the key issues is that this counterfactual argument assumes that all other factors, except for those mentioned in the "what if..." statement, would remain the same.

The observed progression rate for university i, O_i, is simply $\hat{p}_{i.}$ in my notation. The expected progress rate is dependent on the counterfactual question being asked. There are two principal counterfactuals in input-output (IO) analysis, each corresponding to the two main types of standardisation (Anderson et al. (1980)): direct and indirect. Bishop et al. (1974) concluded that indirect standardisation was preferred for estimating rates but further work is required to discover which works better under these quality assessment conditions.

Direct standardisation

The easiest way to explain the different types of standardisation is use to use some example data. In Anderson et al. (1980), they use work from Herring (1936) which is based on breast cancer death rates among female aged 25 or older. The outcome variable is death due to breast

cancer and the risk factor (equivalent to our SCF) is martial status (ever married or not). They identify age as one potential confounding factor.

Anderson et al. (1980) state that the direct standardisation approach in the breast cancer problem is to ask "What would the cancer rates have been for ever-married and not ever-married women if the age distribution for both married and single women had been the same as in some standard population but the age-specific rates were the same as the observed"? The counterfactual nature of the problem is apparent - this happened, but what would have happened if ...?

So what is the equivalent question in the HE analysis? There is not a direct equivalence between the two analysis as our risk factor is unobserved and they have an observed risk factor. The key issue here is that the cancer rates are kept as they are in the age factor, their PCF. So to be consistent with the direct standardisation method, we should keep our PCF category rates as they are in the data. They means we have to adjust the distribution of students across these PCF categories in each university, to some standard distribution across the PCF categories. One natural standard distribution is to consider the population as one huge university and use the PCF distribution for the whole population.

We are asking the question "What would the observed overall progression rate have been at this university if its progression rates in the PCF categories had been what they were, but its distribution of students across PCF categories had instead matched the national distribution"? This is direct standardisation to the national cohort, as the standard distribution is the national cohort's distribution. This is equivalent to imagining that the government decided to send everyone to the university in question, rather than the university just receiving its own students.

In formulaic terms, the direct standardisation expected value for university i, $\hat{E}_i$, is calculated by using a similar method to calculating the observed rate for the university i, $p_{i.}$, but the $\frac{n_{ik}}{n_{i+}}$s in Equation 2.1 change to the population distribution $\frac{n_{+k}}{n_{++}}$s for the PCF classes. For ease, we will use $\hat{O}_i$ as the observed progression rate for university i (same as $\hat{p}_{i.}$). The p_{ik} remain unchanged.

$$\text{Direct: } \hat{p}_{i.} = \quad \hat{O}_i = \quad n_{i+}^{-1} \sum_{k=1}^{M} n_{ik} \hat{p}_{ik}$$

is compared with

$$\hat{E}_i = \quad n_{++}^{-1} \sum_{k=1}^{M} n_{+k} \hat{p}_{ik} \tag{2.3}$$

Indirect standardisation

An alternative to direct is indirect standardisation. In the Anderson et al. (1980) cancer example, an indirect standardisation approach asks "What would the cancer rate have been amongst single women if the age distributions for single and married women were the same as

observed, but the age-specific cancer rates had been the same as in some standard population?". The standard population chosen in the cancer case was the other risk group, married women, although the whole population could have also been adjusted over.

So in the cancer case, the age distribution of single women remained as observed but the age-specific rates for single women changed to some other rates based on some standard population. This method is the exact opposite of the direct method where the distribution changed and the rates remained the same. So now, in my HE example, we compute $\hat{E}_i$ by holding the $\frac{n_{ik}}{n_{i+}}$'s constant in Equation 2.1 and changing the p_{ik} to $p_{.k}$, i.e., the PCF category rate for the whole population in the study.

$$\text{Indirect: } \hat{p}_{i.} = \quad \hat{O}_i = \quad n_{i+}^{-1} \sum_{k=1}^{M} n_{ik} \hat{p}_{ik}$$

is compared with

$$\hat{E}_i = \quad n_{i+}^{-1} \sum_{k=1}^{M} n_{ik} \hat{p}_{.k} \tag{2.4}$$

So we are asking "What would the observed overall progression rate have been at university i, it its distribution of students across the PCF categories had been what it was but its progression rates in the PCF categories were replaced by the national rates?". This is like asking what would the university progression rate have been if the university had performed as the whole university population did considering the university's students. This standardisation can be called indirect standardisation to the university cohort, i.e we only use the university's distribution of students to base the expected rate on. HEFCE's initial method was based upon indirect standardisation to the university cohort and that will be the principal focus here.

Indirect standardisation to the university cohort formulation

In the notation used in Equation 2.1, the observed progression rate at university i, $\hat{O}_i$, is a weighted average of the form

$$\hat{O}_i = \hat{p}_{i.} = \frac{\sum_{j=1}^{M} n_{ij} \hat{p}_{ij}}{\sum_{j=1}^{M} n_{ij}} = n_{i+}^{-1} \sum_{j=1}^{M} n_{ij} \hat{p}_{ij}, \tag{2.5}$$

and the HEFCE expected progression rate at university i, $\hat{E}_i$, based on indirect standardisation (to the university cohort), is also a weighted average:

$$\hat{E}_i = n_{i+}^{-1} \sum_{j=1}^{M} n_{ij} \hat{p}_{.j}. \tag{2.6}$$

HEFCE compute the difference of these two values to give an idea of the "quality" of a university with regard to its progression rate

$$\hat{D}_i = \hat{O}_i - \hat{E}_i \tag{2.7}$$

So the following equations apply:

$$\begin{aligned}
\hat{D}_i = \hat{O}_i - \hat{E}_i &= \left(\frac{1}{n_{i+}}\sum_{j=1}^{M} n_{ij}\hat{p}_{ij}\right) - \left(\frac{1}{n_{i+}}\sum_{j=1}^{M} n_{ij}\hat{p}_{.j}\right) \\
&= \frac{1}{n_{i+}}\sum_{j=1}^{M}\left\{n_{ij}\left(\hat{p}_{ij} - \hat{p}_{.j}\right)\right\}
\end{aligned}$$

but

$$\hat{p}_{.j} = \frac{1}{n_{+j}}\sum_{k=1}^{N} n_{kj}\hat{p}_{kj}$$

implies that:

$$\begin{aligned}
\hat{D}_i &= \frac{1}{n_{i+}}\sum_{j=1}^{M}\left\{n_{ij}\left(\hat{p}_{ij} - \frac{1}{n_{+j}}\sum_{k=1}^{N} n_{kj}\hat{p}_{kj}\right)\right\} \\
&= \sum_{j=1}^{M}\frac{1}{n_{i+}}\left\{n_{ij}\hat{p}_{ij} - \frac{n_{ij}}{n_{+j}}\sum_{k=1}^{N} n_{kj}\hat{p}_{kj}\right\} \\
&= \sum_{j=1}^{M}\left\{\frac{n_{ij}}{n_{i+}}\hat{p}_{ij} - \sum_{k=1}^{N}\frac{n_{ij}n_{kj}}{n_{+j}n_{i+}}\hat{p}_{kj}\right\}
\end{aligned} \tag{2.8}$$

Now consider Equation 2.8 in terms of a summation over all the j and k terms. The inner bracket is a summation over k, except there is one additional term $\frac{n_{ij}}{n_{i+}}\hat{p}_{ij}$ for each j term. I shall include this additional term into the summation over k by including it when $i = k$.
So when $i = k$, the term for the k summation is:

$$\begin{aligned}
k \text{ summation term} &= \frac{n_{kj}}{n_{i+}}\hat{p}_{kj} - \frac{n_{kj}}{n_{+j}n_{i+}}n_{kj}\hat{p}_{kj} \\
&= \frac{n_{kj}}{n_{i+}}\left\{1 - \frac{n_{kj}}{n_{+j}}\right\}\hat{p}_{kj}
\end{aligned}$$

When $i \neq j$, the $n_{ij}\hat{p}_{ij}$ term is not required and the k summation term takes the following form:

$$k \text{ summation term} = -\frac{n_{ij}n_{kj}}{n_{+j}n_{i+}}\hat{p}_{kj}$$

So the $\hat{D}_i$ can be written as a weighted sum of all the NM cells in the "university by PCF" grid:

$$\hat{D}_i = \sum_{j=1}^{M}\sum_{k=1}^{N} \lambda_{ikj}\hat{p}_{kj}$$

where

$$\lambda_{ikj} = \left\{ \begin{array}{cc} \frac{n_{ij}}{n_{i+}}\left(1-\frac{n_{ij}}{n_{+j}}\right) & \text{for } i=k \\ -\frac{n_{ij}n_{kj}}{n_{i+}n_{+j}} & i \neq k \end{array} \right\} \tag{2.9}$$

2.4 Comments on the HEFCE method

HEFCE have already published a number of documents trying to establish a record of the UK universities with regard to progression rates (HEFCE (1999a), HEFCE (2000)). In this section, I describe the original process that HEFCE used to produce these initial and raw results.

In the original results, HEFCE used only two PCFs when it created its grids of PCF categories crossed with the universities. They were unable to use anymore PCFs as their method was computationally intensive and very slow. The two PCFs they used were student qualifications and the general subject to be studied by the student. This created 272 PCF categories, i.e., there were 272 different types of students dependent on these two factors. However, HEFCE did split these results into three different sections: young entrants, mature entrants, and both together.

The major problems for HEFCE at this early stage were:

- They were unable to add more PCFs to the analysis and there were obviously other PCFs that could be included, e.g., a student's social background.
- There was no way of calibrating the quality differences, $\hat{D}_i$, discovered in each university. For example, a difference of -5.0 percentage points was discovered at Falmouth College of Arts. Is 5.0% a large difference or a small difference? How likely is it that an equivalent difference would be discovered in the forthcoming years?

To get an idea of whether the quality difference was large for a university, HEFCE decided to calculate the a z-score for each university, $\hat{z}_i$, where:

$$\hat{z}_i = \frac{\hat{D}_i}{\hat{\mathrm{SE}}(\hat{D}_i)}$$

in which

$$\hat{\mathrm{SE}}(\hat{D}_i) = \sqrt{\hat{\mathrm{Var}}(\hat{D}_i)} \tag{2.10}$$

This z-score can be used as a basis for identifying "good" and "bad" universities. Essentially the z-score acts as a marker to say whether the $\hat{D}_i$ is statistically different from zero. The

z-scores are assumed in the original HEFCE method to come from a standard normal, i.e.,

$$z_i \sim N(0,1) \tag{2.11}$$

Therefore, under standard assumptions, if I chose 1.96 as the z-score cut-off point for identifying unusual universities, 5% of the time a $\hat{D}_i$ flagged as being significantly different from zero will be incorrectly categorised as being unusual. In HEFCE's first report, universities with $|\hat{z}_i| \geq 3$ and $|\hat{D}_i| \geq 0.03$ were flagged as being unusual and were marked with an asterisk(*).

2.5 Standard errors for the $\hat{D}_i$

Previous work

A major issue for all quality assessment methods of this type is how to place a standard error on the quality term for the institutions in the a analysis. There seems to have been very little work completed on this problem in the literature. Some work has tried to produce a reasonable e SE estimate for the ratio, $R_i = \frac{O_i}{E_i}$ including Hosmer and Lemeshow (1995), who used bootstrap and log transformation methods to achieve approximate SE estimates. Burgess et al. (2000) uses a Gamma distribution to examine the SEs of the R_i. Daley et al. (1988) completed some of the early work k on establishing a SE for the difference between observed and expected rates. DeLong et al. (1997) mentions that a difference estimator, i.e., $D_i = O_i - E_i$, can be used for institutional assessment but then ignore it and suggest SEs for the ratio based on Hosmer and Lemeshow (1995) and Daley et al. (1988). In both Hosmer and Lemeshow (1995) and Daley et al. (1988) there is a reliance on external information, i.e., a "test" dataset is required to provide sensible SEs for $\frac{O_i}{E_i}$. Further work on the ratio approach is examined in Section 7.2. Smith (1994) examines three potential SE forms for $\hat{D}_i$ based solely upon the observed or expected progression rate at each institution and a method that is a less-rigorous version of the local estimation approach described later in this section. Thomas et al. (1994) combined a Bayesian framework and the Δ-method to provide SE estimates for their version of our $\hat{D}_i$. They conclude that Daley et al. (1988) SEs were incorrect and produced misleading significance levels that exaggerates the evidence about the extreme nature of exceptional hospitals.

A superpopulation approach

Given that we have all the information on the universities and their students in any one year, there can be no perceived error in the calculated $\hat{D}_i$, as both the observed and expected rates for university i are fact. In order to provide sensible standard errors for the $\hat{D}_i$, we need to regard this year's results as a random sample from a super population (Cochran (1977)). The general idea behind this super population argument is that if "history was rerun" how much would the results differ? Or if a university received similar students in another year and did not vary in its progression performance, how variable would the results be?

As was seen in Section 2.3, the $\hat{D}_i$'s estimated for each university in the indirect standardisation case can be written as a weighted sum of NM cells in the grid formulation. This produces Equation 2.8:

$$\hat{D}_i = \sum_{j=1}^{M}\sum_{k=1}^{N} \lambda_{ikj}\hat{p}_{kj}$$

where

$$\lambda_{ikj} = \left\{ \begin{array}{cc} \frac{n_{ij}}{n_{i+}}\left(1-\frac{n_{ij}}{n_{+j}}\right) & \text{for } i=k \\ -\frac{n_{ij}n_{kj}}{n_{i+}n_{+j}} & i \neq k \end{array} \right\}$$

This construction of the $\hat{D}_i$'s can help us to discover a potential form for the variance of these estimators:

$$\text{Var}(\hat{D}_i) = \text{Var}\left(\sum_{j=1}^{M}\sum_{k=1}^{N} \lambda_{ikj}\hat{p}_{kj}\right) \tag{2.12}$$

Under superpopulation sampling the $\hat{p}_{kj}$ are independent, so

$$\begin{aligned} \text{Var}(\hat{D}_i) &= \text{Var}\left(\sum_{j=1}^{M}\sum_{k=1}^{N} \lambda_{ikj}\hat{p}_{kj}\right) \\ &= \sum_{j=1}^{M}\sum_{k=1}^{N} \text{Var}\left(\lambda_{ikj}\hat{p}_{kj}\right) \\ &= \sum_{j=1}^{M}\sum_{k=1}^{N} \lambda_{ikj}^2 \text{Var}\left(\hat{p}_{kj}\right) \end{aligned}$$

where

$$\lambda_{ikj} = \left\{ \begin{array}{cc} \frac{n_{ij}}{n_{i+}}\left(1-\frac{n_{ij}}{n_{+j}}\right) & \text{for } i=k \\ -\frac{n_{ij}n_{kj}}{n_{i+}n_{+j}} & i \neq k \end{array} \right\} \tag{2.13}$$

This gives us a formula that can be used to calculate the variance of the $\hat{D}_i$ as long as we can find a valid expression for the $\text{Var}(\hat{p}_{kj})$. To find the standard error of the $\hat{D}_i$, we can simply use:

$$\hat{\text{SE}}(\hat{D}_i) = \sqrt{\hat{\text{Var}}(\hat{D}_i)} \tag{2.14}$$

Local variance estimation

We need to derive an expression for $\text{Var}(\hat{p}_{kj})$. Under repeated superpopulation sampling $\hat{p}_{kj}$,

$$\text{Var}(\hat{p}_{kj}) = \frac{\hat{p}_{kj}(1-\hat{p}_{kj})}{n_{kj}}. \tag{2.15}$$

One possibility to estimate this is *local variance estimation*:

$$\hat{\text{SE}}(\hat{p}_{kj}) = \sqrt{\frac{\hat{p}_{kj}(1-\hat{p}_{kj})}{n_{kj}}} \tag{2.16}$$

With this approach I am estimating the underlying cell progression rate (p_{kj}) by the observed cell progression rate ($\hat{p}_{kj}$) from the data. I am also making the assumptions that: the true p_{kj} isn't too close to zero or one; and that n_{kj} isn't particularly small. Obviously as the number of PCFs included in the study increases, each cell in the "PCF by university" grid has fewer and fewer students in it, causing problems with this estimation of the $\text{Var}(\hat{p}_{kj})$. I also cannot be sure that the true cell progression rate isn't particularly close to 0 or 1. These potentially invalid assumptions can and will cause problems with estimation of $\text{Var}(\hat{p}_{kj})$, as we will see later.

Results for local variance estimation: Small World

The local variance method was tested in the Small World, i.e., with five selected universities and four PCF categories. This meant calculating the quality difference D_i for each university ($i = 1, \ldots, 5$), the standard error relating to the D_i using the local variance method and thus, the associated z-score for each university. The results are given in Table 2.5.

University	n_{i+}	$\hat{O}_i$	$\hat{E}_i$	$\hat{D}_i$	$\widehat{SE}\left(\hat{D}_i\right)$	$\hat{z}_i$
33	55	.800	.900	−.100	.053	−1.89
42	901	.858	.914	−.056	.011	−5.18
1	6831	.900	.919	−.018	.002	−9.25
118	1344	.947	.933	+.014	.006	+2.39
14	2917	.990	.934	+.056	.003	+20.6

Table 2.5: Local estimation results in the Small World.

NB: By definition, $\sum_{i=1}^{N} n_{i+}\hat{D}_i = 0$.

If there were no university quality differences, we would expect the z-scores to be normally distributed around 0, with a standard deviation of 1. This implies that you would only expect |university z-score| ≥ 1.96 around 5% of the time using normal distributional assumptions, if underlying quality were the same. The local variance estimation results show that 80% (4 out of 5) of the universities would be tagged as unusual. Institution 42 and Institution 1 are unusually bad at progressing their students, compared to the performance of the whole Small World population. On the other side of the scale, Institution 118 and Institution 14 would be doing significantly better then expected.

80% seems very high and some of the z-scores look very large in comparision to the standard $N(0,1)$ distribution. One possibility is that the standard errors calculated using the local variance estimation method are too small. These z-scores may cast doubt on the effectiveness of the local variance approach. How can test this method to see if it is calibrated correctly? We need a model-based approach where a university's quality appears directly in the model definition. This will allow us to calibrate the results as we can develop models with "known truth".

Chapter 3

A model-based approach

3.1 Calibrating the model: introduction

The local variance method can be used to find z-scores associated with each university's $\hat{D}_i$. As we have seen in the previous section, this method produces z-scores that may be suspect. Therefore a calibration check method is required. A world needs to be created where it is known that there are no university quality differences, i.e., all universities perform equally with regard to student progression. This means developing a model (or models) where the quality of a university appears directly.

With university student progression, the supposedly casual factor S (university quality) is unobserved. We need a set of models where terms that stand for S appear within the model equations. That would allow us to fix the these university quality terms to whatever we want and then test our local variance method to see how well it is calibrated. Sections 3.2 - 3.4 describe a model-based approach that can be used to exactly reproduce the $\hat{D}_i$ found from a non-model-based approach. This model-based approach is then modified to help develop a model in which a university's quality appears directly in the model equation. This model is described in Section 3.5.

3.2 Regression formulation to calculate the $\hat{D}_i$

The original HEFCE method was very intuitive and has no statistical model base. The method is perfectly valid as it is both simply to understand and simple to use. However, it can and will fail to incorporate more complex methods and results. The $\hat{D}_i$ can be reproduced using a model-based method, which will give a more complex alternative point of view and thus increase the chances of completing more complex analysis on the data provided. This regression based method reproduces the HEFCE $\hat{D}_i$ exactly:

Step One

Fit a generalised linear model to the entire data set, ignoring the multilevel structure,

in which the progression status y_{ij} of student j at university i is the binary outcome. The predictors (xs) in the model are made from the W PCF variables in the analysis. If PCF variable k has c_k levels then it is converted into $(c_k - 1)$ indicator variables.

For example if we converted the state school PCF variable, which has three levels (state school attender, private school attender or not known) into indicator variables, we would create two variables. Variable one (x_{1ij}) indicates whether student j at university i attended private school ($x_{1ij} = 1$) or not ($x_{1ij} = 0$). Variable two (x_{2ij}) indicates whether a student had a "not known" school attendance data ($x_{2ij} = 1$) or not ($x_{2ij} = 0$). If we included another indicator variable for state school attendance or not, we would create a design matrix which would not have full rank. This process would be repeated for the other PCFs. This would create $C = (c_1 - 1) + (c_2 - 1) \ldots + (c_W - 1)$ predictors i.e., $x_1, \ldots, x_C$.

Step Two

We must now ensure that the model is fully saturated for all the predictors $(x_1, \ldots, x_C)$ included in the analysis. That means that every 2-way interaction term($x_p x_q$, with all $p < q$) must be included in the model, then every 3-way interaction ($x_p x_q x_r$, with all $p < q < r$) and continue this process until the C-way interaction ($x_1 x_2 \ldots x_C$) is introduced.

This seems a lot of terms but some of these terms can be eliminated because they are zero by design. The terms are those that include two (or more) indicator functions produced from the same PCF variable. This means that the model will only contain W-way interaction terms. So the model would has the following form:

$$(y_{ij} \mid p_{ij}) \overset{\text{indep}}{\sim} \text{Bernoulli}\,(p_{ij})\,, \tag{3.1}$$

$$\begin{aligned} F\,(p_{ij}) \quad = \quad & \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_W x_{Wij} \\ & + \text{all valid two-way interactions} \\ & + \text{all valid three-way interactions} \\ & + \ldots + \text{ the } W\text{-way interactions} \end{aligned} \tag{3.2}$$

where

- y_{ij} is whether student j at university i progressed into the 2nd year ($y_{ij} = 1$) or not ($y_{ij} = 0$);
- $F()$ is the link function;
- $x_{1ij} \ldots x_{(c_1-1)ij}$ are the indicator variables produced from PCF variable 1,
  $x_{(c_1)ij} \ldots x_{(c_1+c_2-2)ij}$ are the indicator variables produced from PCF variable 2,
  $\ldots$,
  and $x_{(C-c_W+1)ij} \ldots x_{Cij}$ are the indicator variables produced from PCF variable W.

Step Three

Use the model to obtain predicted values ($\hat{y}_{ij}$) for the outcome variable for each student.

Step Four

The difference between expected and observed progression rates at university i, $\hat{D}_i$, can be found using the following facts:

- $\hat{O}_i$ is just the mean of the student y values at university i; and
- $\hat{E}_i$ is the mean of the student $\hat{y}$ at university i.

3.3 Another regression model to calculate the $\hat{D}_i$

The regression model can be defined in another way that still produces the same $\hat{D}_i$:

Imagine creating one variable that identifies a student type, dependent on the W PCFs in the analysis. There would be $C = c_1 c_2 \ldots c_W$ potential student types, as defined in Section 2.1. Depending on the number of PCFs chosen and how many students are in the data set, some of the C student types may not exist. Concentrate on the M student types that do exist.

For example, if gender (0 - male, 1 - female) and age (0 - mature students, 1 - young students) were the two selected PCFs, there would be four potential student types: young males, young females, mature males and mature females. With only two small PCFs, it is highly likely that C equals M, i.e., all four student types appear in the selected data.

Then you can create $(M-1)$ indicator variables $(v_1, v_2, \ldots, v_{M-1})$ from this single variable, with each indicator showing whether an individual is a certain student type. As before, one student type does not have an indicator variable and acts as the baseline type. If student ij does not fall into any of the first $(M-1)$ categories, i.e., $v_{1ij}, \ldots, v_{(M-1)ij} = 0$, the student must be in the baseline category.

The new model has the following form and steps 2 and 3 can be carried out as before.

$$(y_{ij} \mid p_{ij}) \overset{\text{indep}}{\sim} \text{Bernoulli}\,(p_{ij})\,, \tag{3.3}$$

$$F\,(p_{ij}) = \beta_0 + \beta_1 v_{1ij} + \beta_2 v_{2ij} + \ldots + \beta_n v_{(M-1)ij} \tag{3.4}$$

where

- y_{ij} is whether student i in university j progressed into the 2nd year ($y_{ij} = 1$) or not ($y_{ij} = 0$);
- $F()$ is the link function;
- and v_{kij} identifies whether the student is type k.

3.4 The link function

In problems with a binary outcome, the response probability function, which is related to a linear combination of the predictors, is known as the link function. For the regression models given in the previous sections, the link function is $F()$. Collett (1991), along with many others, suggests a variety of different functions that could be used, ranging from complementary log-log to the logit functions. In all cases, $F()$ has to map from (0,1) to $(-\infty, +\infty)$. Collett (1991) favours using the logit function for the link but does state that it is important to consider other link functions as these may lead to simpler models or models that fit the data better. So it seems that the choice of link function can be critical when choosing models and an incorrect link can produce misleading results.

This is not the case in the models described in Sections 3.2 and 3.3. The key point for these models is that they are fully saturated and this means that the predicted outcome ($\hat{Y}_{ij}$) for each student is just the mean of the student type cell. Therefore the results that we are interested in do not change regardless of which link function is chosen. In models that are not fully saturated, the choice of link function is an issue and can affect any results for quality assessment. A proportion of our future models are not fully saturated and the effects of link choice on these methods are examined later (Section 8.2).

For computational efficiency, I use a linear link function for these fully saturated model because least-squares regression (produced by using a linear link) is much faster than any regression using a non-linear link (e.g., logistic regression).

3.5 A fixed-effects model

Sections 3.2 and 3.3 described models that exactly reproduced the non-model-based $\hat{D}_i$'s. For calibration, we require a model where the terms for a university's perceived quality appear directly in the model. Therefore, based on the previous models, we require a prediction model that has terms that adjust for all the necessary PCFs and also contains quality terms. One potential approach is a fixed-effects model with the following structure, derived from the model in Section 3.3:

$$\begin{gathered} y_{ij} = \beta_0 + \textstyle\sum_{k=1}^{M} \beta_k \left(x_{ijk} - \bar{x}_k\right) + \alpha_i + e_{ij}, \\ e_{ij} \overset{\text{IID}}{\sim} N\left(0, \sigma_e^2\right), \quad \textstyle\sum_{i=1}^{N} n_{i+}\, \alpha_i = 0, \end{gathered} \tag{3.5}$$

where y_{ij} and x_{ijk} are the outcome and PCF "carrier" k for student j in university i and $\bar{x}_k$ is the grand mean of predictor k. To ensure linear independence within the model, I set β_1 to zero. Given that all the x_{ijk} are indicator functions (one or zero), the $\bar{x}_k$ terms are not strictly necessary. Dropping these additional terms gives us:

$$\begin{gathered} y_{ij} = \beta_0 + \textstyle\sum_{k=2}^{M} \beta_k x_{ijk} + \alpha_i + e_{ij}, \\ e_{ij} \overset{\text{IID}}{\sim} N\left(0, \sigma_e^2\right), \quad \textstyle\sum_{i=1}^{N} n_{i+}\, \alpha_i = 0, \end{gathered} \tag{3.6}$$

The two principal problems with fitting models of this type are: there can be a massive number of variables to adjust for; and the α side condition is not a standard restriction for

regression (more normally: $\alpha_1 = 0$, remaining α's unrestricted). Models such as this one are normally fitted using the EM algorithm (Dempster et al. (1977)). The EM algorithm uses an iterative maximum likelihood approach. The steps to fit this model are given below:

1. Obtain the estimates for the $\hat{\beta}$ from ordinary least squares, i.e., fit a standard fixed-effects model (Equation 3.6 without the additional α terms);
2. Calculate $\hat{\alpha}_i = y_{i\cdot} - \left[\hat{\beta}_0 + \sum_{j=1}^{n_{i+}} \sum_{k=2}^{M} \hat{\beta}_k x_{ijk}\right]$.
   NB $y_{i\cdot} = \hat{O}_i$ is the observed progression rate at university i;
3. Regress $(y_{ij} - \hat{\alpha}_i)$ on the PCF carriers $x_2, \ldots, x_M$ to get new $\hat{\beta}$ values; and
4. Repeat steps 2 and 3 until the β and α values have reached convergence;

Step 2 has the same flavour as the (observed - expected) equations in the original $\hat{D}_i$ calculations (i.e., Equation 2.7). This feature means that $\hat{\alpha}_i \doteq \hat{D}_i$ for all the universities. The difference between the $\hat{\alpha}$ and $\hat{D}$ is due to the fact that in the $\hat{D}_i$ set-up, the $\hat{\beta}$ are being regressed on the student progression outcome, y_{ij}. But in the $\hat{\alpha}$ part of the EM calculation, the $\hat{\beta}$'s are based on an adjusted outcome variable, $y_{ij} - \hat{\alpha}_i$. This means that the non-model-based $\hat{D}_i$ method is approximately equivalent to a fixed-effects multilevel model where the binary outcome is treated as continuous rather than a two level variable, i.e., the link function is linear rather than logistic (or probit).

3.6 Fixed-effects model vs non-model-based method

University (i)	n_i	$\hat{D}_i$	$\hat{\alpha}_i$
1	6831	0.0304	0.0325
2	3314	0.0262	0.0253
3	1113	−0.0306	−0.0314
4	2205	−0.0276	−0.0293
5	289	0.0681	0.0712
6	3238	−0.0126	−0.0131
7	2889	−0.0284	−0.0290
8	2292	−0.0235	−0.0250
9	1031	−0.0458	−0.0471
10	116	0.0441	0.0475

Table 3.1: A comparison of non-model-based vs fixed-effects quality assessment results.

In order to get an idea of how the fixed-effects method compares with the non-model-based method, the Medium World (described in Section 2.2) is examined. This consists of ten universities from the original 165 and looks at adjusting on four PCFs: gender; age; state school; and low participation. This gives us 36 PCF categories because all potential student types are present, i.e., M = two (gender levels) x two (age) x three (state school) x three (low participation) = 36.

The results for the inferred quality of a university using the non-model-based method (i.e., the $\hat{D}_i$'s) and by using the fixed-effects modelling ($\hat{\alpha}_i$) are given in Table 3.1.

There is close agreement between the two estimates ($\hat{D}_i$ and $\hat{\alpha}_i$) for each university. The largest differences between the two approaches occur in the smaller universities, e.g., Institution 10 (116 students), difference between parameters: 0.3%; Institution 5 (289 students), difference: 0.3%.

3.7 Standard errors for the fixed-effects model

We would like to obtain sensible standard errors (SEs) for the $\hat{\alpha}_i$ provided by the fixed-effects approach. This will help us to compare the non-model-based z-scores for university quality against a fixed-effects alternative. The $\hat{\text{SE}}(\hat{\alpha}_i)$ values can be estimated using the standard maximum likelihood technique involving the Fisher-Information matrix (Fisher (1973)). The method uses the following calculations for the fixed-effects model:

1. Calculate the log likelihood of the fixed-effects model: The model in Equation 3.4 can be rewritten as:

$$y_{ij} \sim N\left[\left(\beta_0 + \sum_{k=1}^{M} \beta_k x_{ijk} + \alpha_i\right), \sigma_e^2\right] \tag{3.7}$$

with the side condition $\sum_{i=1}^{N} n_{i+}\,\alpha_i = 0$. As before, y_{ij} and x_{ijk} are the outcome and PCF "carrier" k for student j in university i.

So the equivalent likelihood(L) and log-likelihood(LL) for the model is:

$$L = \prod_{i=1}^{N}\prod_{j=1}^{n_{i+}} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left[-\frac{1}{2\sigma_e^2}(y_{ij} - \beta_0 - \sum_{k=2}^{M} \beta_k x_{ijk} - \alpha_i)^2\right] \tag{3.8}$$

$$LL = \sum_{i=1}^{N}\sum_{j=1}^{n_{i+}} -\frac{1}{2}\ln(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{N}\sum_{j=1}^{n_{i+}}\left[y_{ij} - \beta_0 - \sum_{k=2}^{M} \beta_k x_{ijk} - \alpha_i)\right]^2 \tag{3.9}$$

But we need to include the side condition $\sum_{i=1}^{N} n_{i+}\,\alpha_i = 0$ so we restrict the final α, i.e., α_N. If we need:

$$\sum_{i=1}^{N} n_{i+}\,\alpha_i = 0 \tag{3.10}$$

then the restriction for α_N needs to be:

$$\alpha_N = -\frac{\sum_{i=1}^{N-1} n_{i+}\alpha_i}{n_{N+}} \tag{3.11}$$

The adjusted log-likelihood now looks like:

$$\begin{aligned} LL &= \sum_{i=1}^{N}\sum_{j=1}^{n_{i+}} -\frac{1}{2}\ln(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{N-1}\sum_{j=1}^{n_{i+}}\left[y_{ij} - \beta_0 - \sum_{k=2}^{M}\beta_k x_{ijk} - \alpha_i)\right]^2 \\ &\quad -\frac{1}{2\sigma_e^2}\sum_{j=1}^{n_{N+}}\left[y_{ij} - \beta_0 - \sum_{k=2}^{M}\beta_k x_{ijk} + \frac{\sum_{i=1}^{N-1} n_{i+}\alpha_i}{n_{N+}}\right]^2 \end{aligned}$$

2. Differentiate the log-likelihood with respect to each of the β's in turn (i.e., $\beta_0, \beta_2, \ldots, \beta_M$), then the α's present in the likelihood ($\alpha_1, \ldots, \alpha_{N-1}$) and finally the σ_e^2 variable. This process produces $(M + N)$ equations. Solving these $(M + N)$ equations simultaneously for $\beta_0, \beta_2, \ldots, \beta_p, \alpha_1, \ldots, \alpha_{N-1}, \sigma_e^2$ gives the maximum likelihood estimates (MLEs) for these variables (i.e., $\hat{\beta}_0, \hat{\beta}_2, \ldots, \hat{\beta}_M, \hat{\alpha}_1, \ldots, \hat{\alpha}_{N-1}, \hat{\sigma}_e^2$);

3. Create the Fisher Information matrix for this likelihood:

$$\begin{pmatrix} \frac{d^2LL}{d\beta_0\beta_0} & \frac{d^2LL}{d\beta_0\beta_2} & \cdots & \frac{d^2LL}{d\beta_0\beta_M} & \frac{d^2LL}{d\beta_0\alpha_1} & \cdots & \frac{d^2LL}{d\beta_0\alpha_{N-1}} & \frac{d^2LL}{d\beta_0\sigma_e^2} \\ \frac{d^2LL}{d\beta_2\beta_0} & \frac{d^2LL}{d\beta_2\beta_2} & & & & & & \vdots \\ \vdots & & \ddots & & & & & \vdots \\ \frac{d^2LL}{d\beta_M\beta_0} & & & \frac{d^2LL}{d\beta_M\beta_M} & & & & \vdots \\ \frac{d^2LL}{d\alpha_1\beta_0} & & & & \frac{d^2LL}{d\alpha_1\alpha_1} & & & \vdots \\ \vdots & & & & & \ddots & & \vdots \\ \frac{d^2LL}{d\alpha_{N-1}\beta_0} & & & & & & \frac{d^2LL}{d\alpha_{N-1}\alpha_{N-1}} & \vdots \\ \frac{d^2LL}{d\sigma_e^2\beta_0} & \frac{d^2LL}{d\sigma_e^2\beta_2} & \cdots & \frac{d^2LL}{d\sigma_e^2\beta_M} & \frac{dLL}{d^2\sigma_e^2\alpha_1} & \cdots & \frac{d^2LL}{d\sigma_e^2\alpha_{N-1}} & \frac{d^2LL}{d\sigma_e^2\sigma_e^2} \end{pmatrix} \tag{3.12}$$

4. Invert the Fisher Information Matrix and evaluate it at the MLEs;

5. The diagonal entries of this inverted matrix are the maximum likelihood estimates of the variance of the appropriate estimate for that row (or column) of the Fisher Information Matrix. So the $(M + 1)$ diagonal element is the variance for $\hat{\alpha}_1$. The $\hat{\text{SE}}$s for the $\hat{\alpha}$s are the square roots of the appropriate variances.

3.8 Using the fixed-effects approach with larger datasets

Introduction to the problem

So far only FE model results in relatively small models, in terms of parameters, have been examined. The Medium World had 46 parameters: ten α_i parameters, 35 β_i parameters and one β_0 parameter. The Small World had even fewer variables for regression. I would like to fit these fixed-effects models in the Big World (described in Section 2.2), i.e., 165 α_i parameters, 17,798 β_i parameters and one β_0 parameter, which gives around 18,000 regression parameters. The simple answer seem to be to just scale up the previous analysis and apply the appropriate calculations in the Big World. To fit a fixed-effects model we need to carry out two iterative EM steps (3.5):

1. $\hat{\alpha}_i = y_{i\cdot} - \left[\hat{\beta}_0 + \sum_{j=1}^{n_{i+}} \sum_{k=2}^{M} \hat{\beta}_k x_{ijk}\right]$;
2. Regress $(y_{ij} - \hat{\alpha}_i)$ on the PCF carriers $x_2, \ldots, x_M$ to get new $\hat{\beta}$ values;

where $y_{i\cdot} = \hat{O}_i$ is the observed progression rate at university i.

The first step for α_i is trivial but we now have to perform a regression with 17,799 β parameters on 284,399 students. Using the standard regression technique, the estimates for the β's ($\hat{\beta}$) can be calculated from $\left(X^T X\right)^{-1} X^T Y$. This method fails as X is a 284,399 by 17,799 matrix which means $X^T X$ is a 17,799 by 17,799 matrix and a matrix of this size is notoriously difficult to invert. The following subsection describes how to complete this analysis in the Big World, i.e., using a fully saturated approach. An alternative method is also described that is less restrictive than the EM approach because the adjustment process is not required to be fully saturated.

Implementing the EM algorithm in the Big World

Looking at each regression matrix in more detail helps to implement this large regression analysis. Each column of X, our design matrix, contains 284,399 rows (one for each student in the dataset) and 17,799 columns (a column of ones for β_0 and then indicator variables for membership of PCF groups $2 \ldots M$). Using these facts, it can be shown that $X^T X$ and $X^T Y$ have special forms:

$$X^T X = \begin{pmatrix} d_0 & d_1 & d_2 & \ldots & d_{M-1} \\ d_1 & d_1 & 0 & \ldots & 0 \\ d_2 & 0 & d_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ d_{M-1} & 0 & 0 & \ldots & d_{M-1} \end{pmatrix}, \tag{3.13}$$

where $d_0 = n_{++}$ is the total number of students and (for $j > 0$) $d_j = n_{+j}$ is the national number of students in PCF category j. The "missing" PCF group, the M^{th} in this case, is the baseline group to ensure there is no linear dependences in the regression analysis.

For X^TY, the j^{th} entry of the M x 1 vector is s_j, where:

- $s_0 = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{n_{ij}} y_{ijk}$; and
- $s_j = \sum_{i=1}^{N} \sum_{k=1}^{n_{ij}} y_{ijk}$ for $j \in (1, \ldots, M-1)$,
  which is the national sum of the y values in PCF category j.

X^TX can be symbolically inverted to produce:

$$\left(X^TX\right)^{-1} = \begin{pmatrix} \frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & -\frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & -\frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & \cdots & -\frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} \\ -\frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & \frac{d_1 + d_0 - \sum_{i=1}^{M-1} d_i}{d_1\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & \frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & \cdots & \frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} \\ -\frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & \frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & \frac{d_2 + d_0 - \sum_{i=1}^{M-1} d_i}{d_2\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & \cdots & \frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & \frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & \frac{1}{\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} & \cdots & \frac{d_{M-1} + d_0 - \sum_{i=1}^{M-1} d_i}{d_{M-1}\left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}} \end{pmatrix}, \tag{3.14}$$

or in a simplified form:

$$\left(X^TX\right)^{-1} = \begin{pmatrix} \frac{1}{b} & -\frac{1}{b} & -\frac{1}{b} & \cdots & -\frac{1}{b} \\ -\frac{1}{b} & \frac{d_1+b}{d_1 b} & \frac{1}{b} & \cdots & \frac{1}{b} \\ -\frac{1}{b} & \frac{1}{b} & \frac{d_2+b}{d_2 b} & \cdots & \frac{1}{b} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{b} & \frac{1}{b} & \frac{1}{b} & \cdots & \frac{d_{M-1}+b}{d_{M-1} b} \end{pmatrix}, \tag{3.15}$$

where $b = \left\{d_0 - \sum_{i=1}^{M-1} d_i\right\}$.

So to find the $\hat{\beta}$ parameter estimates, we need to calculate:

$$\begin{aligned} \hat{\beta} \quad &= \left(X^TX\right)^{-1} X^TY = \begin{pmatrix} \frac{1}{b} & -\frac{1}{b} & -\frac{1}{b} & \cdots & -\frac{1}{b} \\ -\frac{1}{b} & \frac{d_1+b}{d_1 b} & \frac{1}{b} & \cdots & \frac{1}{b} \\ -\frac{1}{b} & \frac{1}{b} & \frac{d_2+b}{d_2 b} & \cdots & \frac{1}{b} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{b} & \frac{1}{b} & \frac{1}{b} & \cdots & \frac{d_{M-1}+b}{d_{M-1} b} \end{pmatrix} \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ \cdots \\ s_{M-1} \end{pmatrix} \\ &= \frac{1}{b} \begin{pmatrix} s_0 - \sum_{j=1}^{M-1} s_j \\ -s_0 + \sum_{\substack{j=1, \\ j \neq 1}}^{M-1} s_j + \frac{d_1+b}{d_1} s_1 \\ -s_0 + \sum_{\substack{j=1, \\ j \neq 2}}^{M-1} s_j + \frac{d_2+b}{d_2} s_2 \\ \cdots \\ -s_0 + \sum_{\substack{j=1, \\ j \neq M-1}}^{M-1} s_j + \frac{d_{M-1}+b}{d_{M-1}} s_{M-1} \end{pmatrix} \end{aligned} \tag{3.16}$$

So the estimates can be viewed as:

$$\hat{\beta}_0 = \frac{\text{Number of Successes in the Baseline PCF Group}}{\text{Number in Baseline Group}} \tag{3.17}$$

and $\forall k \in (1, \ldots, M-1)$:

$$\hat{\beta}_k = \frac{\left\{ -\text{Successes}_{\{\text{Baseline \& PCF k}\}} + \left[\frac{N_{\{\text{Baseline}\}} + N_{\{\text{PCF k}\}}}{N_{\{\text{PCF k}\}}} \right] \text{Successes}_{\{\text{PCF k}\}} \right\}}{N_{\{\text{Baseline}\}}} \tag{3.18}$$

I have fit the fixed-effects multilevel model to the Big World using this technique and have discovered that the $\hat{\alpha}$ values closely match the non-model-based $\hat{D}$s. The next step would be to find the standard errors of these α_i terms but this can be very difficult when the problem is approached in this way. The forthcoming subsections suggest a more effective route for finding the $\hat{\alpha}$ parameter estimates and their SEs. The new approach is not restricted to dealing with fully saturated models.

Fitting an alternative side condition

Consider a standard FE model:

$$\begin{aligned} y_{ij} &= \beta_0 + \sum_{k=1}^{M} \beta_k x_{ijk} + \alpha_i + e_{ij} \\ e_{ij} &\overset{\text{IID}}{\sim} N\left(0, \sigma_e^2\right), \end{aligned} \tag{3.19}$$

with side conditions $\alpha_1 = 0, \beta_1 = 0$.

In this model, the standard side conditions are used for the $\hat{\alpha}_i$, i.e., fixing one as zero. This condition must be imposed as a model with all the $\hat{\alpha}_i$ unrestricted would create a linearly dependent set of parameters and an invalid model structure. These models can be fitted in most standard software packages, assuming that the software can deal with the model size.

The main problem with this α side condition is that the $\hat{\alpha}_i$'s now represent differences from a baseline university. We are more interested in university quality differences that are not relative to a certain university. The preference is for the university parameters to represent differences from the average university "quality". One way to impose this condition is to ensure that the weighted sum of the $\hat{\alpha}_i$'s is zero, with the weights depending on university numbers (as imposed in Section 3.5).

The side conditions for this multi-level model now look like this: $\beta_1 = 0$; and $\sum_{i=1}^{N} n_{i+}\, \alpha_i = 0$.

In models where there is a relatively small number of β parameters (≤ 300), statistical packages can be used to obtain the α parameters with the baseline side condition. When the model is fully saturated with regard to the PCFs, the student type can be identified using a

single variable. In this case, GLIM4's (GLIM (1993)) eliminate command can be used to fit the baseline model. This eliminate command can provide estimates for the $\hat{\alpha}_i$'s and their associated covariance matrix by essentially removing the student type from the regression. This means that, in the Big World, the software only needs to study a pseudo X matrix of around 165x165 rather than a matrix with over 18,000 rows and columns.

I would like to move from these baseline α estimates to the weighted estimates using a formulaic method, i.e., not use another iterative method to come up with these new weighted α estimates. The following algorithms describe how to obtain one type of α from the other.

Let α^w be the αs produced using the weighted side condition and α^b be the αs from a baseline side condition.

Algorithm 1: Weighted to Baseline

- Obtain the weighted α^w parameters from an appropriate statistical package using an appropriate iterative method;
- Select one of the α^w to be the baseline parameter, i.e., set α_1^w to be the baseline;
- Calculate the difference between α_1^w and each of the α^w in turn and set this difference to be the new baseline α (α_i^b), i.e., $\alpha_i^b = \alpha^w - \alpha_1^w$. Ensure that the sign of the new weighted α is correct: if $\alpha^w \geq \alpha_1^w$ then α_i^b should be positive and if $\alpha^w \leq \alpha_1^w$ then α_i^b should be negative.

- This method will automatically set the baseline α to zero as $\alpha_1^b = \alpha_1^w - \alpha_1^w = 0$.

Algorithm 2: Baseline to Weighted

- Obtain the baseline α^b estimates from a statistical package (small number of βs) or by some other appropriate method (large number of βs).
- One of the α^bs will be zero by default. Rearrange the α_i^b so α_1^b is the baseline α.
- Calculate the weighted mean ($\bar{b}$) of the α_i^b including α_1^b.
- Set $\alpha_1^w = -\bar{b}$.
- Calculate the remaining α^w using $\alpha^w = \alpha_1^w + \alpha_i^b$.

Obtaining standard errors for the weighted estimates

In most standard packages such as GLIM4, STATA (STATA (2001)) or S (Becker et al. (1988)), the baseline α^b estimates are given along with their associated co-variance matrix. For these performance indicator methods, we would like to obtain standard errors for the weighted α^w. Given that the α_i^w are a linear combination of the α_i^b, we can infer their standard errors of the α^w from the co-variance matrix of the α_i^b.

We know that for all $q \in 1, \ldots, U$:

$$\mathrm{SE}(\alpha_q^w) \quad = \quad \sqrt{\mathrm{Var}(\alpha_q^w)} \tag{3.20}$$

So for α_1^w:

$$\begin{aligned}
\mathrm{Var}\left[\alpha_1^w\right] &= \mathrm{Var}\left[-\frac{\sum_{k=2}^{U} n_k \alpha_k^b}{\sum_{k=1}^{U} n_k}\right] \\
&= \left(\frac{1}{\sum_{k=1}^{U} n_k}\right)^2 \mathrm{Var}\left[\sum_{k=2}^{U} n_k \alpha_k^b\right] \\
&= \left(\frac{1}{\sum_{k=1}^{U} n_k}\right)^2 \sum_{i=2}^{U}\sum_{j=2}^{U} n_i n_j \mathrm{Cov}(\alpha_i, \alpha_j)
\end{aligned} \tag{3.21}$$

For α_q^w, where q $\in 2, \ldots, U$:

$$\begin{aligned}
\mathrm{Var}\left[\alpha_q^w\right] &= \mathrm{Var}\left[-\frac{\sum_{k=2}^{U} n_k \alpha_k^b}{\sum_{k=1}^{U} n_k} + \alpha_q^b\right] \\
&= \mathrm{Var}\left[\frac{-\sum_{k=2}^{U} n_k \alpha_k^b + (\sum_{k=1}^{U} n_k)\alpha_q^b}{\sum_{k=1}^{U} n_k}\right] \\
&= \left(\frac{1}{\sum_{k=1}^{U} n_k}\right)^2 \mathrm{Var}\left[-\sum_{k=2}^{U} n_k \alpha_k^b + (\sum_{k=1}^{U} n_k)\alpha_q^b\right] \\
&= \left(\frac{1}{\sum_{k=1}^{U} n_k}\right)^2 \sum_{i=2}^{U}\sum_{j=2}^{U} \omega_{ijq}
\end{aligned} \tag{3.22}$$

where

$$\omega_{ijq} \quad = \quad \left\{ \begin{array}{cc}
n_i n_j \mathrm{Cov}(\alpha_i, \alpha_j) & \text{if } i \neq q \text{ and } j \neq q \\
-(-n_i + \sum_{k=1}^{U} n_k) n_j \mathrm{Cov}(\alpha_i, \alpha_j) & \text{if } i = q \text{ and } j \neq q \\
-n_i(-n_j + \sum_{k=1}^{U} n_k) \mathrm{Cov}(\alpha_i, \alpha_j) & \text{if } j = q \text{ and } i \neq q \\
(-n_q + \sum_{k=1}^{U} n_k)^2 \ \mathrm{Cov}(\alpha_q, \alpha_q) & \text{if } i = j = q
\end{array} \right\}$$

3.9 Fixed-effects model vs. HEFCE method II

Tables 3.2 and 3.3 complete the analysis began in Section 3.6. They compare the standard error and z-score results of a non-model-based quality assessment approach ($\hat{D}_i$ using local variance estimation) against a fixed-effects assessment method (α_q^w) in the Medium World.

There is relatively good agreement between the two methods with regard to estimated SEs of the quality effects. However, some differences in the quality estimates are noted at the smaller universities: around 0.5% difference at Institution 5 and 0.8% at Institution 10. The quality

assessment z-scores can be readily calculated for each method, (Estimate / SE{ Estimate }). These are given in Table 3.3.

As with the parameter estimates and their SEs, the z-scores of each method do not vary excessively. In this example, the largest deviation occurs at Institution 2, where the already extremal z-score moves from 6.06 using a D_i method, to 4.66 using a model-based approach. It seems that the two methods tally well when assessing most universities, but we need to examine these methods under a different set of conditions. Section 3.10 studies the z-scores produced by both approaches in the real data, i.e., our Big World.

University (i)	n_i	$\widehat{SE}\left(\hat{D}_i\right)$	$\widehat{SE}\left(\hat{\alpha}_i^w\right)$
1	6831	0.00310	0.00344
2	3314	0.00432	0.00543
3	1113	0.00987	0.00960
4	2205	0.00681	0.00663
5	289	0.01383	0.01904
6	3238	0.00546	0.00531
7	2889	0.00602	0.00567
8	2292	0.00695	0.00649
9	1031	0.01089	0.00990
10	116	0.02135	0.03026

Table 3.2: A comparison of SEs for $\hat{D}_i$ and $\hat{\alpha}_i^w$.

University (i)	n_i	$\widehat{z}_i^D$	$\widehat{z}_i^{\alpha^w}$
1	6831	9.79	9.44
2	3314	6.06	4.66
3	1113	−3.10	−3.27
4	2205	−4.05	−4.41
5	289	4.93	3.74
6	3238	−2.31	−2.47
7	2889	−4.73	−5.11
8	2292	−3.38	−3.85
9	1031	−4.20	−4.76
10	116	2.06	1.57

Table 3.3: A comparison of z-scores for $\hat{D}_i$ and $\hat{\alpha}_i^w$

3.10 Model vs non-model-based : Big World z-scores

Consider using the Big World (Section 2.2) where there are 17,799 PCF levels and 284,399 students in 165 universities. We can estimate institutional quality assessments using a model-based approach (FE multilevel model: Section 3.5), or by using a non-model-based approach with a local variance technique (Section 2.5 and 2.3). Figure 3-1 shows the institutional assessment differences for the two methods.

There is a reasonable correlation between the two sets of z-scores: 0.95. There is however a large shrinkage effect in the model-based scores for the exceptionally good universities. The local estimation method believes these universities are more exceptional than the FE approach does. The graph seems to indicate that the local estimation's failure to correctly measure each cell's progression rate variance is rewarding the top performing institutions. These results raise concerns over the local variance estimation approach and some calibration of the method is required. Chapter 4 examines the performance of the local estimation method, alongside a number of other techniques.

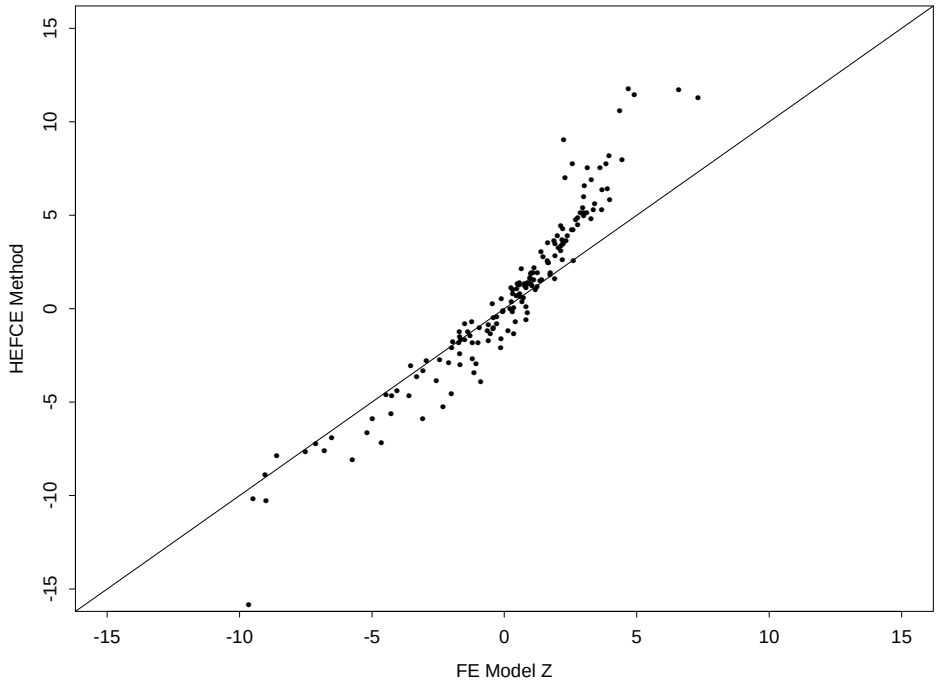

Figure 3-1: The variation between FE model and the local non-model-based z-scores.

Chapter 4

Calibration

4.1 The simulation idea

Now that I have developed a model where the SCF appears directly, I can adjust and tailor the SCF in a simulation world to act in any way we would like. The idea behind null simulations is as follows. Imagine a world where there are no university differences. In such a world, no university is good and no university is bad. However if university quality was measured we would expect, due to random fluctuation, that some universities z-scores would be larger than expected in absolute value. In this null world, the university z-scores would come from a Normal distribution with a mean of 0 and a SD of 1, by design. Normal distributional theory states that 5% of the time a z-score will be bigger than 1.96 in absolute value, i.e., if we choose 1.96 as a cut-off for deciding whether a university is unusual or not, 5% of the time a university will be unlucky and tagged as unusual when it's not. Therefore, using the 1.96 cut-off in my null world, perfect normal behaviour dictates that 5% of the time we discover an unusual establishment, i.e., an unusually bad place 2.5% of the time and an unusually good place 2.5% of the time.

So the algorithm for null simulations is as follows:

1. Fit the FE model (Section 3.5) to some selected data, e.g the Small World. Note that all the α terms are set to zero as there are no university differences;

2. Generate a "potential to drop out" for each student ($\hat{y}_{ij}$) depending on the student's characteristics. Use the original outcomes to produce these potentials;

3. For each student in the dataset, simulate whether the student dropped out or progressed, dependent on her drop out potential;

4. Calculate the z-scores for each university from the generated outcomes using the methods described in Section 2.3 and the appropriate non-model-based variance estimation technique (e.g., a local variance approach - Section 2.5). Compare each z-score to the chosen cut-off (1.96 - for the forthcoming examples);

5. Record how many good and bad universities have been noted based on the z-score cut-off;

6. Repeat steps 3 through 5 as many times as necessary for desired simulation accuracy; and

7. Combine each generated set of results to produce a calibration for the method studied. If the method is correctly calibrated, the method's z-scores should have a mean of zero and SD of 1. There will be 2.5% good universities and 2.5% bad universities on average.

4.2 A binary outcome model for calibration

The fixed-effects model suggested in Section 3.5 assumes a linear link function for regression and this can be a problem when using a binary outcome. The linear model can produce $\hat{y}_{ij}$ values of below 0 or above 1, i.e., the probability of a success is negative or more than 100%, which does not make statistical or practical sense. Therefore a more appropriate approach for calibrating a binary outcome model is a logistic multilevel structure:

$$\begin{gathered}(y_{ij} \mid p_{ij}) \overset{\text{indep}}{\sim} \text{Bernoulli}\,(p_{ij})\,, \\ \log\left(\tfrac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \textstyle\sum_{k=1}^{p} \beta_k x_{ijk} + \alpha_i, \\ \textstyle\sum_{i=1}^{N} n_{i+}\,\alpha_i = 0,\end{gathered}$$

with the same parameter definitions as in Equation 3.6.

This model is more difficult to fit than the basic linear fixed effects model but this does not cause a problem in null simulations because $\alpha_i = 0\ \forall i \in (1, \ldots, U)$. This means the adjustment model is fully saturated and therefore, both approaches (linear or logistic link functions) produce the same $\hat{y}_{ij}$'s for each student because the predicted $\hat{y}_{ij}$'s are just equal to $\hat{p}_{\cdot j}$, the observed national progression rate in PCF category j, for students falling into that PCF cell. So when all the university effects ($\hat{\alpha}_i$) are equal, as long as students are in the same PCF category they will have the same predicted $\hat{y}_{ij}$.

4.3 Local estimation results

Introduction

In all of the following sections, I am using the null simulation technique described in Section 4.1. The local variance method from Section 2.5 is used to estimate $\text{Var}(\hat{p}_{kj})$ from Equation 2.13. All results on calibration are based on the original non-model-based approach given in Section 2.3.

Small World

The local variance estimation method was first studied in the Small World described in Section 2.2. This world has N = five universities and M = four PCF categories. There are only four possible levels of "potential to drop out" for a student because there are only four PCF categories. The algorithm in Section 4.1 was used and 2,000 simulation runs were performed.

Results:

- z-scores had a mean of 0.064 (0.006) with a SD of 1.08 (0.011) (Monte Carlo standard errors in parentheses);
- z-scores were of an approximate Gaussian shape;
- 1.8% (0.1%) of universities were identified as bad, i.e., $\hat{z}_i \leq -1.96$;
- 4.1% (0.2%) of universities were identified as good, i.e., $\hat{z}_i \geq 1.96$; and so
- 5.9% (0.3%) of the time a university was identified as unusual, i.e., $|\hat{z}_i| \geq 1.96$.

The mean of the z-scores is a little on the positive side and the overall SD is a little large, 1.08 compared to a target of 1.00. In this situation, the local variance method seems to highlight too many universities as good, 4.1% compared to a target of 2.5%, and doesn't tag enough universities as bad, 1.8% compared to 2.5%. This all means that the percentage of unusual universities is 0.9% over its target value of 5.0%.

Therefore in the Small World, the local variance method shows some asymmetry towards identifying too many good universities. One of the assumptions of a superpopulation sampling argument (Fleiss (1981)) was that the proportion of successful outcomes was close to 0.5. Student progression (a successful outcome) is over 92% in the Small World and this seems to be a prime candidate for why the asymmetry in the results is seen.

Small World ($p = 0.5$)

To test whether the asymmetry seen in the Small World was due to the very high 92% progression rate in the data, I created a fake Small World where the overall progression rate was approximately 50%. This meant generating faked progression outcomes for each student in the Small World, where the overall target progression rate was 50%. So now we have $\hat{p}_{..} \approx 0.50$ rather than $\hat{p}_{..} \approx 0.92$. The 2000 simulations were rerun in this 50% Small World.

Results: Results:

- z-scores had a mean of 0.000 (0.005) with a SD of 1.04 (0.009);
- z-scores were of an approximate Gaussian shape;
- 2.6% (0.2%) of universities were identified as bad, i.e., $\hat{z}_i \leq -1.96$;
- 2.5% (0.2%) of universities were identified as good, i.e., $\hat{z}_i \geq 1.96$; and therefore
- 5.1% (0.2%) of the time a university was identified as unusual, i.e., $|\hat{z}_i| \geq 1.96$.

These results are a better tally with the target results compared to the original Small World analysis. The target z-score mean equated exactly with the simulation mean and there is significant improvement in the SD towards 1.00. The 2.5% mark was achieved for identification of good universities and nearly achieved for the badly performing universities. This mean that the method in this generated world was only just over estimating the proportion of unusual universities (5.1% against 5.0%).

Big World

The results for the local variance estimation technique look promising in the Small World set-up. The results are excellent dealing with a success rate near 50%, with a drop in effectiveness as the progression rate moves away from 50%. This drop in efficiency isn't dramatic and the asymmetric results seen in the actual Small World are still acceptable.

What would happen to the method under a different set of conditions? Given that we are really interested in the quality assessment of universities for the whole of UK HE rather than a world with only five universities, the method should be tested on the whole UK student population for entry in 1996/1997. This means using the Big World defined in Section 2.2, with 284,399 individuals in 165 HE institutions. All eight possible PCFs are taken into account when analysing the 90.1% population progression rate. By crossing all of these PCFs, $2 \cdot 2 \cdot 21 \cdot 13 \cdot 3 \cdot 3 \cdot 3 \cdot 2 = 58{,}968$ potential student types can be identified. There were 17,799 different student profiles found in the 1996/1997 dataset, i.e., around 30% of the potential student profiles. Although 284,399 students sounds a very large number, there is only an average of $\frac{284399}{165 \cdot 17799} \approx 0.1$ students per cell in the $N \times M$ grid (Table 2.2). Therefore, using the notation in Section 2.1, $N = 165$, $M = 17,799$, $p_{..} = 0.901$ and $n_{++} = 284,399$.

With this set-up using the local variance estimator, the results are as follows:

- z-scores had a mean of 0.101 (0.003) with a SD of 1.60 (0.007);
- z-scores were of an approximate Gaussian shape;
- 8.2% (0.1%) of universities were identified as bad, i.e., $\hat{z}_i \leq -1.96$;
- 10.9% (0.1%) of universities were identified as good, i.e., $\hat{z}_i \geq 1.96$; and
- 19.1% (0.1%) of the time a university was identified as unusual, i.e., $|\hat{z}_i| \geq 1.96$.

These results are dramatically worse than any of the Small World results. There is excessive identification of unusual universities, with nearly 19% being marked as unusual rather than the expected 5%. The SD of the z-scores is much larger than expected and does not compare favourably with the Small World results. These results imply that the local $\hat{\text{SE}}(\hat{D}_i)$ values must be much too small. It seems that the $N \times M$ grid has become too sparse for the local variance method and it appears that sensible estimates for the variance of cell progression rates cannot be gained using only local progression rate information. This is mainly because the cell progression rates are based on very few people, which contradicts one of the assumptions in a superpopulation sampling approach. With an increasing number of PCF adjustors in the analysis, the local estimation technique will become less and less efficient.

4.4 Other estimation techniques

Motivation for alternative techniques

It appears that the local variance technique fails as the data matrix gets sparser and sparser. This is mainly because the estimate of the cell proportion ($\hat{p}_{kj}$), used to calculate the variance

of the probability of cell success ($\hat{V}(\hat{p}_{kj})$), becomes very inaccurate as it is based on a small number of individuals. A more robust method would be to "borrow strength" or information from other related cells within the data matrix (i.e., include more individuals in the estimation of the cell proportion). There are a variety of different methods for borrowing strength to estimate $\hat{V}(\hat{p}_{kj})$ for use in Equation 2.13, and these are described in below. Each of these methods has the same flavour for estimating the variance of the cell proportion: rather than using the local cell proportion (p_{kj}) in the standard variance of a proportion equation, i.e., $V(\hat{p}_{kj}) = \frac{p_{kj}(1-p_{kj})}{n_{kj}}$, substitute something else in for the p_{kj} and keep the same denominator, n_{kj}, the number of individuals within cell $_{kj}$.

Global estimate

The first substitution involves replacing the local success rate ($\hat{p}_{kj}$) with the global (overall) success rate ($\hat{p}_{..}$). This means that each and every cell variance estimated is identical if there were the same number of individuals in each data matrix cell. The method borrows strength from the whole table and this may cause some biases as some local cell true proportions might be dramatically different to the global mean. The produces a global estimator ($\hat{V}^g(\hat{p}_{kj})$) of the cell proportion variance:

$$\hat{V}^g(\hat{p}_{kj}) = \frac{\hat{p}_{..}(1-\hat{p}_{..})}{n_{kj}}. \tag{4.1}$$

γ shrinkage estimate

Rather than just using either the local cell proportion ($\hat{p}_{kj}$) or the overall proportion ($\hat{p}_{..}$), I study the effects of using a combination of both to provide an estimate for the cell proportion variance. The estimate is made up of γ percent of the global estimate and $1-\gamma$ percent of the local cell proportion. This creates a estimate that is the global proportion shrunk back towards the local cell proportion by means of the γ value. For our initial studies, I set γ as 0.5, i.e., a variance estimate ($\hat{V}^\gamma(\hat{p}_{kj})$) which was based on a proportion halfway between the global and local rates:

$$\hat{V}^\gamma(\hat{p}_{kj}) = \frac{\hat{p}^*_{kj}\left(1-\hat{p}^*_{kj}\right)}{n_{kj}}, \tag{4.2}$$

where $\hat{p}^*_{kj} = \gamma\hat{p}_{..} + (1-\gamma)\hat{p}_{kj}$.

University estimate

The university estimate substitutes the university success rate ($\hat{p}_{k.}$) in place of $\hat{p}_{kj}$ for cells in university k. This means regardless of which PCF category cells fall in, as long as the cells are associated with the same university, the same success rate is used to estimate ($V(\hat{p}_{kj})$). So in this case, the original estimated cell proportion variance is replaced with the university variance estimate, $\hat{V}^u(\hat{p}_{kj})$:

$$\hat{V}^u(\hat{p}_{kj}) = \frac{\hat{p}_{k.}\,(1-\hat{p}_{k.})}{n_{kj}}. \tag{4.3}$$

Connected cell estimate

This estimator calculates an estimated cell rate ($\hat{p}^{cc}_{kj}$) by combining information on all the students in university k and all the students in PCF category j. This new rate, $\hat{p}^{cc}_{kj}$, is used in the substitution to create a connected cell variance estimate, $\hat{V}^{cc}(\hat{p}_{kj})$. This is called a connected cell estimate because it combines information from cells in the same university or same PCF category as the target cell, and creates a success rate estimate based on many more people than n_{kj}.

$$\hat{V}^{cc}(\hat{p}_{kj}) = \frac{\hat{p}^{cc}_{kj}\left(1-\hat{p}^{cc}_{kj}\right)}{n_{kj}}, \tag{4.4}$$

where $\hat{p}^{cc}_{kj} = (n_{k+}\hat{p}_{k.} + n_{+j}\hat{p}_{.j} - n_{kj}\hat{p}_{kj})/(n_{k+} + n_{+j} + n_{kj})$.

ANOVA estimate

The ANOVA variance estimator is based on the theory used in ANOVA.

Substitute the following estimate in for the $\hat{p}_{kj}$ in the variance equation:

- Start with the global rate, $\hat{p}_{..}$;
- Take off the difference between the global rate and the cell university rate, $(\hat{p}_{..} - \hat{p}_{k.})$. This leaves you with the university success rate, $\hat{p}_{k.}$;
- Now take off the difference between the global rate and the cell PCF rate, $(\hat{p}_{..} - \hat{p}_{.j})$. $\hat{p}_{.j} + \hat{p}_{k.} - \hat{p}_{..}$ becomes your estimate to be substituted for $\hat{p}_{kj}$ in the original $\frac{p(1-p)}{n}$ equation:

$$\hat{V}^a(\hat{p}_{kj}) = \frac{(\hat{p}_{.j} + \hat{p}_{k.} - \hat{p}_{..})\,(1-(\hat{p}_{.j} + \hat{p}_{k.} - \hat{p}_{..}))}{n_{kj}}. \tag{4.5}$$

Limit estimators

Introduction to the Limit Estimators

The university variance estimate could potentially work because each estimated university progression rate is based on a relatively large number of students, i.e., not less than ten or so. A similar estimate, based on PCF categories, would not be as effective because as the matrix gets sparser and sparser, the numbers of individuals in each PCF category falls and falls to produce some very small categories. In some cases only one or two students of that specific PCF type exist. For these Limit variance estimators, I use the shrinkage estimator to estimate the local cell proportion if the student numbers for the cell's PCF category is greater than 20, i.e.,

we think there is some useful information from the PCF category. In the Limit 1 estimator, if the numbers are less than 20 in the cell's PCF category, i.e., little useful PCF category information, I use the overall progression rate, $\hat{p}_{..}$, as the cell estimate. In the Limit 2 estimate, with less than 20 individuals present in the PCF category, I use the university progression rate. The cell progression rate variance estimators for Limit 1 and Limit 2 are $\hat{V}^{l1}(\hat{p}_{kj})$ and $\hat{V}^{l2}(\hat{p}_{kj})$ respectively:

Limit 1 Estimate

$$\hat{V}^{l_1}(\hat{p}_{kj}) = \frac{\hat{p}_{kj}^{l_1}\left(1-\hat{p}_{kj}^{l_1}\right)}{n_{kj}}. \tag{4.6}$$

where $\hat{p}_{kj}^{l_1} = \hat{p}_{..}$ if $n_{+j} < 20$ and
$\hat{p}_{kj}^{l_1} = \hat{p}_{kj}^{\gamma}$ otherwise.

Limit2 Estimate

$$\hat{V}^{l_2}(\hat{p}_{kj}) = \frac{\hat{p}_{kj}^{l_2}\left(1-\hat{p}_{kj}^{l1}\right)}{n_{kj}}. \tag{4.7}$$

where $\hat{p}_{kj}^{l_2} = \hat{p}_{k.}$ if $n_{+j} < 20$ and
$\hat{p}_{kj}^{l_2} = \hat{p}_{kj}^{\gamma}$ otherwise.

4.5 Performance of alternatives in various situations

The eight potential variance estimation methods were studied in the four different worlds defined in Section 2.2. These different scenarios cover a broad range of the potential data matrix structures, allowing me to obtain an idea of how the approaches would perform in other sets of data. In the two smaller worlds, 2000 simulations were completed but in the larger two situations, due to the speed of the simulations, only 500 runs were simulated. The results are given in Table 4.1.

In the Small, Medium and Published worlds, the γ method and the Limit methods produce exactly the same z-scores for each simulation and university and thus, exactly the same tail behaviour. These three methods perform relatively well in the smaller worlds giving an overall misclassification rate of 5.4% (compared to the 5.0% target). The inequality between the two tails favours identifying average universities as excellent. In the Big world, the performance of the three methods begins to vary but not dramatically. The Limit methods have a slightly low misclassification rate (4.8%), with the Limit 1 method having excellent equality in the tails (2.4% vs 2.4%). The γ method tends to misclassify a little too much (5.2%) with a little inequality between the tails (2.5% vs 2.7%).

In all the worlds, the local method always overestimated the number of extremal institutions, rising from 5.9% misclassified in the Small world to a huge 19.1% in the Big world. Nearly a

fifth of universities are classed as extremal in this Big world when only 5.0% are expected to be identified.

The ANOVA method also consistently overestimates and has a similar performance to the local method in the Small and Medium worlds. It performs slightly better than the local method in the Published and Big worlds but still has poor misclassification rates (10.5%, 9.0%). The university cell method has some large inequalities between the tails for the four worlds but has reasonable misclassification rates (4.4% to 5.6%). The performance of the method appears to rely quite heavily on the sparseness of the set-up, i.e., there is a monotone drop in misclassification rates as the size of the data-grid increases. The connected-cell approach has good equality in the tails and misclassification rates for the smaller worlds, but struggles (along with many of the other methods) in the Big world especially on tail inequality (1.7% against 2.5%).

Considering only this dataset, each method has its advantages and disadvantages depends on the data set-up (i.e., which world is being used). All the method are affected by the high progression rates causing inequalities in the tail behaviour. The local and ANOVA approaches can be rejected as they perform considerably worse than all the other methods in every world. In the Big world, the γ and Limit approaches seem to cope best with the data structure and high progression rates.

	World(%)											
	Small			Medium			Published			Big		
Runs	2000			2000			500			500		
Est.	Low	High	Both	L	H	B	L	H	B	L	H	B
Global	2.8	2.4	5.2	2.6	2.3	4.8	2.6	2.1	4.7	2.2	1.7	3.9
	(0.2)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(0.1)
Local	1.8	4.1	5.9	2.2	3.6	5.8	2.5	4.4	6.9	8.2	10.9	19.1
	(0.1)	(0.2)	(0.3)	(0.1)	(0.1)	(0.2)	(0.0)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Uni.	1.7	4.0	5.6	1.9	3.2	5.2	1.7	3.5	5.2	1.5	2.9	4.4
	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.2)	(0.0)	(0.1)	(0.1)	(0.0)	(0.1)	(0.1)
C. Cell	2.5	2.3	4.8	2.5	2.7	5.2	2.4	2.4	4.8	1.7	2.5	4.2
	(0.2)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.0)	(0.1)	(0.1)	(0.0)	(0.1)	(0.1)
ANOVA	1.8	3.8	5.6	2.2	3.6	5.8	4.1	6.4	10.5	3.7	5.3	9.0
	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
$\gamma^{0.5}$	2.0	3.3	5.4	2.3	2.8	5.0	2.2	2.6	4.8	2.5	2.7	5.2
	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.2)	(0.0)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Limit 1	2.0	3.3	5.4	2.3	2.8	5.0	2.2	2.6	4.8	2.4	2.4	4.8
	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.2)	(0.0)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Limit 2	2.0	3.3	5.4	2.3	2.8	5.0	2.2	2.6	4.8	2.2	2.6	4.8
	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.2)	(0.0)	(0.1)	(0.1)	(0.0)	(0.0)	(0.1)

Table 4.1: The results of the variance estimates in the four worlds.

4.6 Performance of alternatives: $p = 0.5$

To provide an idea of how the overall success rate of the data affects the various approaches, I re-create the four worlds with an overall artificial success rate of around 50%. This will allow

us to examine the effects, on the misclassification rates, of losing the asymmetric behaviour of the tails. Table 4.2 shows the results of these simulations.

When the progression rate is reduced to 0.5, nearly all the tail inequalities seen in the original data results disappear. The difference between the two tails is now never more than 0.2% and in general is either 0.0% or 0.1%. Therefore the performance of the approaches can now be assessed solely against the target overall misclassification rate of 5.0%. All the methods react well in the Small world, with all the misclassification rates being between 4.8-5.1% compared with a target of 5.0%. In the Medium world, all the methods produce the same misclassification rate (5.1%) except for the local variance method which performs badly by identifying unusual universities 5.5% of the time rather than the "hoped" 5.0%. In the Published world the story is very similar to the Medium world, with only the local variance method doing very badly. In the Big world, there is more variation in the misclassification rates with the local approach identifying 16.4% extremal institutions. The remaining rates range between 4.6% and 6.2%, which given the sparseness of the data, are close to the target misclassification rate.

	World(%)											
	Small			Medium			Published			Big		
Runs	2000			2000			500			500		
Est.	Low	High	Both	L	H	B	L	H	B	L	H	B
Global	2.5	2.3	4.8	2.5	2.5	5.0	2.5	2.4	4.9	2.3	2.2	4.6
	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.1)
Local	2.6	2.5	5.1	2.8	2.7	5.5	3.2	3.1	6.2	8.2	8.2	16.4
	(0.2)	(0.2)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Uni.	2.5	2.5	5.0	2.6	2.5	5.1	2.5	2.5	5.0	2.4	2.3	4.6
	(0.2)	(0.2)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.1)
C. Cell	2.5	2.3	4.8	2.5	2.5	5.1	2.5	2.4	4.9	2.3	2.3	4.6
	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.1)
ANOVA	2.5	2.5	5.1	2.6	2.5	5.1	2.5	2.5	5.0	2.8	2.7	5.5
	(0.2)	(0.2)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
$\gamma^{0.5}$	2.5	2.4	4.9	2.6	2.6	5.1	2.6	2.6	5.2	3.1	3.1	6.2
	(0.2)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Limit 1	2.5	2.4	4.9	2.6	2.6	5.1	2.6	2.6	5.2	2.8	2.8	5.6
	(0.2)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Limit 2	2.5	2.4	4.9	2.6	2.6	5.1	2.6	2.6	5.2	2.8	2.8	5.6
	(0.2)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)

Table 4.2: The results of the variance estimates with a 50% rate.

4.7 Implication of results on a general dataset

It is all well and good to proclaim results for our case-study dataset but what are the implications when a general dataset needs to be analysed? What is the best approach for somebody who wishes to generate her own set of performance indicators using the non-model-based approach?

Both the local and ANOVA variance techniques can be rejected as sensible methods because both perform badly in all situations. The university approach seems to be unable to deal with rates away from 50% so this leaves us with a choice of four methods to select from: the connect cell approach; the shrinkage or γ approach; and the two Limit methods. There are some concerns over the complexity of the Limit estimates as they require a careful choice of the what constitutes a "valid" establishment, i.e., how do you pick where to set your limit? These Limit approaches are based on a combination of the global, γ and university estimates which means an extra level of calculation and complexity.

The performance of the two remaining methods are similar. The $\gamma^{0.5}$ estimation seems to provide slightly better results with overall rates away from 50% and the connect cell technique giving slightly improved results with a 50% success rate. The key is that the γ approach is more flexible as the γ parameter can be modified depending on the data set-up and success rates. The $\gamma^{0.5}$ results are very good in nearly all the set-ups and, with the possibility of tweaking the shrinkage parameter, the results can be improved upon depending on the accuracy required.

I recommend using the γ variance estimation approach when using non-model-based methods. Setting the parameter to 0.5 provides good results in nearly all situations. When more accurate results are required, modification of this parameter will help. Section 4.8 examines the effect of adjusting the γ parameter in a variety of situations.

4.8 Shrinkage estimate

Introduction to the γ shrinkage estimate

The choice of γ in the shrinkage estimate is not arbitrary. The effects of a single value of γ (say 0.5) vary depending on the conditions of the data set-up. When $\gamma = 0$ in the shrinkage estimate, the shrinkage method is identical to the local variance method. When $\gamma = 1$, the shrinkage method is identical to the global variance method. As we have seen previously in Sections 4.5 and 4.6, the local method works well when there is little or no sparseness in the data matrix and the global method works better when the data matrix has a degree of sparseness. This implies that the correct value of γ is dependent on the sparseness of the data matrix, along with the overall success rate.

Even when the specific data set-up is known, there are problems in deciding the value of γ. In the perfect world, where all universities have no quality differences, the γ calibration has multiple aims (as in Section 4.1):

- To make the proportions of universities flagged as "poor" (i.e., a z-score of less than -1.96) be as close to 2.5% as possible;
- To place 2.5% of the universities in the higher tail as well (i.e., a z-score of more than 1.96);
- To make the two tails of the flagging system as symmetric as possible. This means making the proportion of universities flagged as "poor" approximately equal to those universities

flagged as "good";

- To make the total proportion of "unusual" universities close to 5.0%.

γ effects in the Small World

γ Value	Poor(%) ≤ -1.96	Good(%) ≥ 1.96	Unusual(%)
0.0 - Local	1.81	3.81	5.62
0.1	1.84	3.67	5.51
0.2	1.87	3.59	5.46
0.3	1.94	3.39	5.33
0.4	1.96	3.19	5.15
0.5	1.99	3.07	5.06
0.6	2.12	2.83	4.95
0.7	2.19	2.59	4.78
0.8	2.37	2.43	4.80
0.9	2.56	2.26	4.82
1.0 - Global	2.71	2.11	4.82

Table 4.3: Gamma effects in the Small World.

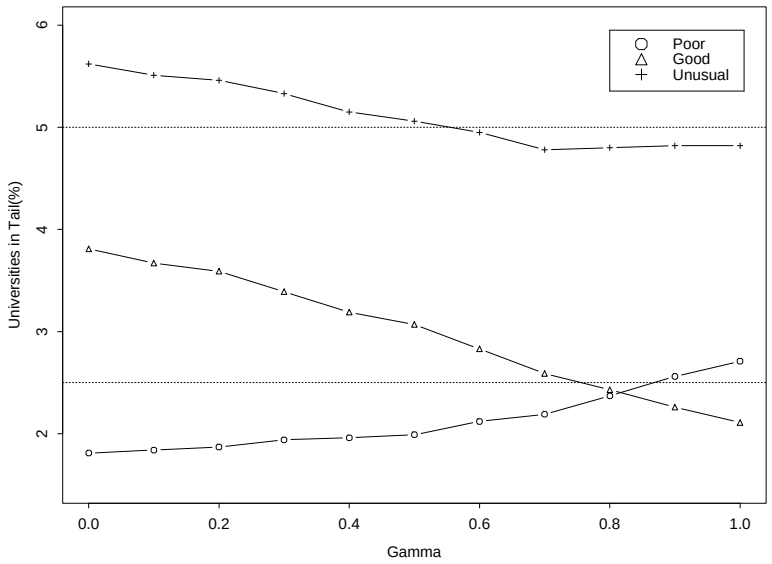

Figure 4-1: Relationship between γ and tail behaviour: Small World.

I examined the effects of changing the value of γ in a number of situations. The first of these was with the five universities and four PCF categories of the Small World set-up. There are

two empty cells out of 20 (= five x four) in this set-up, which means that the level of sparseness is 10%. The overall success rate is around 92% in this data.

The tail behaviour of the university quality assessment are given in Table 4.3 and Figure 4-1. These results are based on 2000 simulations in an artificial world where there are no quality differences between the universities. Note that, due to simulation variation, the local, $\gamma^{0.5}$ and global results from Table 4.1 do not exactly match the $\gamma^{0.0}$, $\gamma^{0.5}$ and $\gamma^{1.0}$ results given in Table 4.3. This applies to Tables 4.3 - 4.5.

$\gamma^{0.0}$ (or the local variance approach) produces the worst results in the Small World. It has relatively large misclassification rate (5.6%) and there is asymmetry in the tails. Symmetry is achieved in the tails when $\gamma = 0.8$ but the target misclassification rate occurs around $\gamma = 0.55$. Therefore we need a γ that is a tradeoff between these two targets. A γ between 0.6 and 0.8 seems to be the optimal choice.

γ effects in the Published World

In this second set-up I examine the Published World (Section 2.2). We now have all 165 universities but we are adjusting on only two PCFs, subject of study and entry qualifications. These PCFs produce 272 different student categories, so there are $272 * 165 = 44,880$ potential different university vs PCF crosses. This crossing actually produces 19,419 non-empty cells meaning that the level of sparseness is $\frac{44,880-19,419}{44,880} = 57\%$. Table 4.4 and Figure 4-2 show the effects of the γ parameter in the Published World.

A similar pattern to the Small world is seen in these results. As before, there is no perfect parameter selection as the target misclassification rate occurs when γ is set to around 0.35 and equality in the tails is achieved when γ is around 0.75. Therefore the best choice of γ in this situation is dependent on which of these two targets is more important. Setting the parameter to around 0.5 provides a good balance between the two conditions.

γ Value	Poor(%) ≤ -1.96	Good(%) ≥ 1.96	Unusual(%)
0.0 - Local	2.24	4.49	6.85
0.1	2.23	3.74	5.98
0.2	2.17	3.31	5.47
0.3	2.11	3.02	5.13
0.4	2.09	2.76	4.85
0.5	2.11	2.56	4.67
0.6	2.13	2.43	4.56
0.7	2.21	2.27	4.48
0.8	2.27	2.19	4.46
0.9	2.36	2.13	4.49
1.0 - Global	2.49	2.06	4.56

Table 4.4: Gamma effects in the Published World.

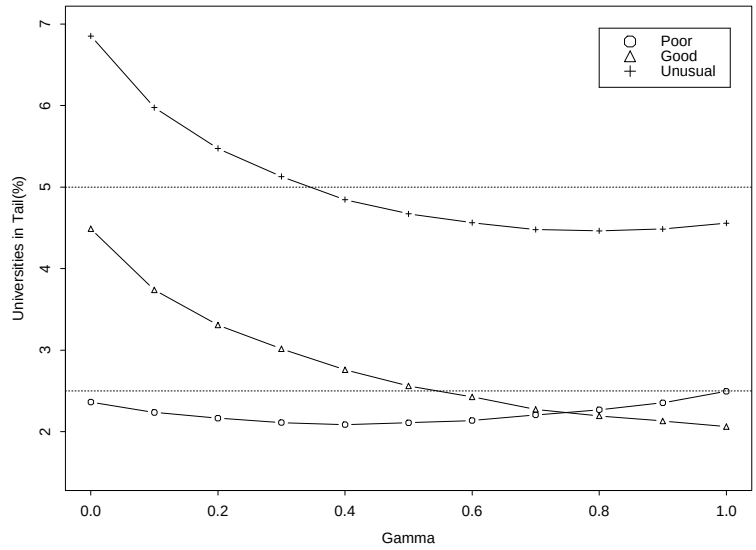

Figure 4-2: Relationship between γ and tail behaviour: Published World.

γ effects in the Big World

The third analysis studies the Big World where the students from the 165 universities are included. All eight PCFs are adjusted for, giving rise to 17,799 different PCF categories. There are $165 * 17,799 = 2,936,835$ potential different PCF by university cell options. In the data, 129,890 of these cells contain at least one student. The level of sparseness is $\frac{2,936,835-129,890}{2,936,835} = 96\%$, which means that the data is spread very finely over the whole grid.

γ Value	Poor(%) ≤ -1.96	Good(%) ≥ 1.96	Unusual(%)
0.0 - Local	8.11	10.86	18.97
0.1	5.64	7.57	13.21
0.2	4.26	5.49	9.74
0.3	3.41	4.23	7.64
0.4	2.87	3.38	6.25
0.5	2.49	2.84	5.34
0.6	2.24	2.43	4.68
0.7	2.13	2.19	4.32
0.8	2.09	2.02	4.11
0.9	2.09	1.89	3.97
1.0 - Global	2.18	1.83	4.01

Table 4.5: Gamma effects in the Big World.

Due to the very sparse nature of the data, the local ($\gamma^{0.0}$) variance estimation approach works very ineffectively, producing an overall misclassification rate of 19%. This means that low γ values do not perform very well in this data. The overall misclassification rates drop to a reasonable level only when the γ parameter is larger than around 0.5. When $\gamma \geq 0.5$, the tail

symmetry also improves dramatically as well. In data this sparse, an optimal γ appears to be around 0.5 – 0.6 as the rate is close to 5.0% and the tail asymmetry is at a minimum. Table 4.5 and Figure 4-3 show the γ effects in the Big World (based on 500 simulations).

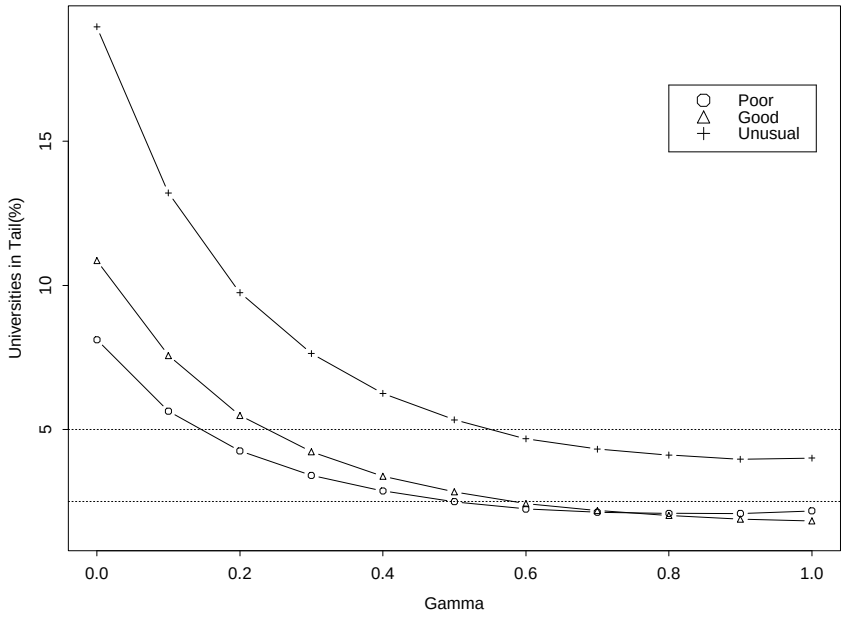

Figure 4-3: Relationship between γ and tail behaviour: Big World.

Selecting a γ value in a general dataset

The first clear message from the different worlds is that as the data grid gets sparser and sparser, a low γ value choice becomes less and less favourable. With only 5% of the grid filled, a low γ choice could produce an overall misclassification rate of nearly 20% when only 5% was expected. By combining this information with the results from Sections 4.5 and 4.6, in most cases γ values around 0.5 – 0.6 work well.

When the overall progression rate is near 50%, a slightly higher range of γ values should be favoured as γ values of around 0.5 tend to produce marginally too high overall misclassification rates. This is highlighted by comparing the $\gamma^{0.5}$ results against the global results ($\gamma^{1.0}$) given in Table 4.2. So, in general, these rules give a reasonable guide for selecting the γ value that will produce precise quality assessment results. If even more precise results are required then similar simulation studies should be carried out on the dataset in question by using the calibration methods described in Section 4.1.

Chapter 5

University quality assessments for 1996/1997

5.1 The real results: non-model-based approach

In Chapter 4 we saw that the shrinkage variance estimation approach produced well calibrated results for non-model-based assessment. Now, rather than having a world with no university differences, I examine the actual data for 1996/1997 to identify which universities have exceptionally bad or good progression rates given their student intake. Tables A.1 - A.4 are ordered by quality assessment z-score, with those universities doing extremely well in terms of progression having a positive $\hat{D}_i$ and thus a large positive z-score. (A positive $\hat{D}_i$ implies that university i's observed progression rate is larger than its expected progression rate, based on its student intake.)

A number of different styles of league table have been published over the years and the effects of these approaches are discussed on this HE dataset:

- Very crude performance tables would only concentrate on the $\hat{O}$ column in the data, i.e., the university with the highest progression rate has performed the best. This approach rewards universities that have received students who are expected to do well in terms of progression. It does not attempt to measure how the universities have affected these students, i.e., it does not identify universities that took in students who had a high probability of drop-out and managed to keep these students in HE next year. Using this basic approach, Institution 14, Institution 79 and some of the medical colleges would look very good. Institution 33 and Institution 110 have very low observed progression rates and these are perceived to be poorly performing using this crude approach.

- A second potential set of tables would include the expected, or benchmark, value and the institutions would be ordered on the difference between the observed and expected, i.e., the $\hat{D}_i$ values. The problem with this approach is that no account is taken of how variable

the $\hat{D}_i$'s are. Some of the smaller institutions are especially prone to having large variation in their differences. Under this ranking system, establishment such as Institution 165 or Institution 55 look particularly bad, being placed in the same "category" as Institution 110. All three institutions have a negative difference of around 4% but Institution 110 has 3,650 students compared to Institution 165's or Institution 55's 150 individuals. On the other side of the coin, Institution 33's positive 6% difference looks exceptional but this is only based on 92 students.

- The complete analysis is to provide tables that show: the observed rates; the expected rates; the difference between these rates; the SE for this difference; and the significance of the difference, when compared with its SE. These tables attempt to take into account variation in the $\hat{D}_i$'s, so that a large positive $\hat{D}_i$ doesn't necessarily mean a university is identified as excellent. They try to take $\hat{D}_i$ variation into account and provide an estimate of how likely this result is pure chance. Under these conditions, Institution 9 and Institution 110 are identified as areas of concern. This approach does not mean that smaller universities cannot achieve an extremal status; for example Institution 101 with only 639 students is found to be unusually bad, and Institution 71 college with only 913 students is identified as exceptional. I recommend this approach as the most effective league-table methodology.

5.2 A potential university summary

Introduction

The overall tables cannot give a complete picture of what is happening within a university. Some student cohorts might be performing exceptionally well and others might be performing exceptionally badly, making the university look average. For greater clarity, we need to "open" the box and examine the finer detail of the university's performance. Goldstein (2001) describes this effect as "differential effectiveness" and warns that this type of information can be masked by standard league tables. This example of a potential summary shows our quality assessment approach can be used to help with the differential effectiveness problem, and thus can only improve institutional standards.

The following section provides some examples of what could be offered to universities to help understand where there are potentially interesting areas of their performance. The report is based on the actual results for Poppleton (note, data is real but the institutional name has been changed) in 1996/1997. This type of analysis can also be used as a data quality check as misclassified students and invalid data can be easily spotted. It was discovered that Poppleton had made a mistake with their data entry and a whole selection of students had been misclassified. The Poppleton analysis helped to highlight this data problem. The section also describes the calculations that were used to produce such a university summary.

Poppleton University

The material between the △ symbols is an example of text that could be sent to a new university.

△ This summary gives the findings on Poppleton's 1st year progression rate for students who commenced a programme of study during academic year 1996/1997. A student is classed as having successfully progressing into her 2nd year if she is still present at a higher-education establishment at the start of academic year 1997/1998.

- Poppleton successfully progressed **80.2%** of its starting students for 1996/1997.
- The progression rate for all the universities was **90.1%**.
- After taking into account the **qualifications of entry** and **subject of study** for Poppleton's students, the university's expected progression rate is **85.9%**.

Your progression benchmark is 85.9% and your actual progression performance is 80.2%. This means you are **under performing** against your benchmark percentage by **5.7%**. This 5.7% difference has been identified as **statistically and practically significant** based on your university profile.

This significant difference could be due to one of two effects:

- Poppleton is not performing as well as it should do with its student population, compared to how the rest of the country's universities perform with similar students to Poppleton's.
- Poppleton's student cohort and/or the university itself is unusual in ways not taken into account by the analysis, i.e., issues not relating to student qualifications or subject of study.

If you can provide us with any information on factors that Poppleton think affect its students progression from their first year, we will be happy to consider and analyse them.

The following information identifies student types where Poppleton's progression rate is practically significantly different to the UK population (Table 5.1). These student types should be examined by the university to discover why Poppleton is performing differently to the rest of the UK higher education establishments.

The largest area of concern relates to the 267 students, with some type of higher education qualification on entry, studying Social Sciences and Law. These students are not progressing as well as expected. This low progression rate could be due to a number of factors:

- There may be an issue on the level and pitch of the material within their course that is not helpful to students with a higher education qualification.
- The level of lectures within the whole university may not be tailored towards students with a higher education qualification. This does not seem to be the case in Poppleton as other subject areas are progressing higher education students significantly better than Social Sciences and Law, e.g., Mathematical Sciences, Engineering and Combined Subjects.

Student Profile		Poppleton		Population		
Qual.	Sub.	n	Prog. Rate	Prog. Rate	Difference	SE
Higher Education	Social Sciences & Law	267	11.2%	77.3%	-66.1%	1.9%
Unknown	Business Admin. & Librarianship	31	35.5%	81.5%	-46.0%	8.7%
Unknown	Engineering & Technology	34	41.2%	84.1%	-42.9%	8.6%
Unknown	Combined & Subjects	39	46.2%	73.7%	-27.6%	8.1%
Unknown	Social Sciences & Law	22	45.5%	81.8%	-36.3%	10.9%
Unknown	Math. Sciences & Computing	25	48.0%	80.0%	-32.0%	10.2%

Table 5.1: Areas of concern.

- There may be a misclassification of student, e.g., these students might mainly be on a single year course and not tagged as being on a single year course.

The other major concern over Poppleton's low progression performance are those students whose qualifications are unknown. These students, in general and especially within Business Admin., Engineering, Social Sciences, Mathematical Sciences and Combined Subjects, have a significantly lower progression rate than expected. This may be down to how Poppleton selects its students at entry, i.e., Poppleton are selecting students in the unknown qualifications category that are completely different to the rest of the population's unknown students. Another possible explanation is that Poppleton are categorising their unknown qualification students differently to the rest of the universities' unknown students. For example, a student with a foreign degree might be classed as unknown within Poppleton but be classed as having a higher education qualification within another university.

Student Profile		Poppleton		Population		
Qual.	Sub.	n	Prog. Rate	Prog. Rate	Difference	SE
A-Level Points 8-9	Math. Sciences & Computing	48	95.8%	86.4%	+9.5%	2.9%
Other	Allied to & Medicine	48	93.7%	84.1%	+9.6%	3.5%
GNVQ	Arts &Design	28	96.4%	89.2%	+7.2%	3.6%

Table 5.2: Areas of excellence.

Poppleton has some excellent progression performances for this year (Table 5.2). The Mathematical and Computer Science subject areas are progressing certain types of students with an excellent record. Students with non-standard qualifications studying subject allied to medicine are also noted as progressing with an exceptional record. These areas should be examined and analysed in order to learn more about why excellent progression rates are being obtained.

Subject Area	n	Progression Rates				
			Qualification, Subject Adjusted for		8 Adjustors	
		Observed	Expected	Difference	Expected	Difference
Subjects allied to medicine	437	89.7%	86.7%	3.0%	86.6%	3.1%
Biological sciences and Physical sciences	141	85.1%	86.8%	-1.7%	84.2%	0.9%
Agriculture and related subjects	72	80.6%	87.7%	-7.2%	82.2%	-1.7%
Mathematical sciences and Computer science	401	85.0%	83.9%	1.2%	83.6%	1.4%
Engineering and Technology	353	77.1%	81.9%	-4.8%	80.5%	-3.5%
Architecture, Building and Planning	44	79.5%	85.3%	-5.8%	82.9%	-3.3%
Social studies and Law	488	41.2%	81.7%	-40.5% *	71.3%	-30.1% *
Business studies and Librarianship	294	81.6%	87.2%	-5.6%	85.1%	-3.4%
Creative arts and Design	449	89.5%	89.5%	0.0%	89.5%	0.0%
Education	256	96.1%	92.0%	4.1% *	92.5%	3.6% *
Combined Subjects	723	86.7%	86.1%	0.6%	86.2%	0.5%

Table 5.3: Subject area analysis.

Note: * indicates that the difference is statistically significant.

When Poppleton's subject areas are separately examined (Table 5.3), there are two principal departments of interest. The observed progression rate in the Social Studies and Law department is significantly (practically and statistically) lower than expected (41.2% compared against a benchmark of 81.7%). This is mainly due to the 267 unusual students noted in Table 5.1. The Education department should be praised as its observed progression rate is significantly higher than expected (96.1% against 92.0%). The progression of these Education students is very encouraging and their staff should be congratulated for its efforts. $\triangle$

The mechanics behind the university summary

The process for breaking the university results down into finer detail is:

1. Calculate each PCF category progression rate using the selected PCFs in the model. These are the probabilities of progression for a student with certain PCF characteristics;

2. Calculate the difference for each individual between her observed progression status and her predicted progression probability;

3. Split the students into different sections depending on what variable you are breaking the results on;

4. Calculate the mean difference and the standard deviation of this mean difference separately for each section;

5. Find the standard error of the difference by dividing the standard deviation of the mean difference by the square root of the number of people within that section; and

6. Create a z-score using the normal rules and compare with the appropriate limit.

Another way to picture this approach is to consider the whole dataset, with one row for each individual and a number of key columns. The first is the observed progression rate and this will be one or zero, depending on whether the student progressed or not. The second column is her expected progression probability, given all the PCFs adjusted for. For each individual a third column, $\hat{\Delta}$, can be calculated and this is the difference between the individual's observed and expected columns. The students can now be separated into the appropriate sections (e.g., subject areas). The dataset for a certain section has the following form:

$$\begin{pmatrix} \hat{O} & \hat{E} & \hat{\Delta} & \text{Sect} \\ 0 & 0.70 & -0.70 & 7 \\ 1 & 0.90 & 0.10 & 7 \\ 1 & 0.86 & 0.14 & 7 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0.78 & -0.78 & 7 \end{pmatrix} \tag{5.1}$$

The mean of the $\hat{\Delta}_i$ form the mean difference between the $\hat{O}_i$ and $\hat{E}_i$ for the section in question (seven in this example) and the SD of these $\hat{\Delta}_i$ create the associated SD for the difference. The significance of this difference from zero can then be calculated.

This approach allows all the actors in the quality assessment to examine why (and how) a university achieved its progression status. This finer detail can be used as a tool for improving student progression across the whole sector, with good and bad progression practice identified at a very low level. So rather than saying Institution 46 has excellent student progression, we can say what influences Institution 46's excellent progression rate.

5.3 Model- vs. non-model-based comparison

Introduction

Section 3.5 showed how model-based $\hat{\alpha}_i$'s could be used to mimic the behaviour of the non-model-based $\hat{D}_i$'s. Furthermore, in Section 3.9 I showed how the inferred standard errors (and the associated z-scores) for a non-model-based local variance estimation approach reproduced the model-based results in the Medium World. It has been shown that this local estimation approach is flawed under many circumstances, usually producing results that identify too many unusual universities. The following parts of this section look at how the improved shrinkage variance estimation approach tallies with a model-based approach.

Fixed-effects model vs HEFCE method III: the Medium World

To complete the analysis started in Section 3.9, I repeat the process replacing the local variance estimation approach with the shrinkage (γ set to 0.5) approach for the non-model-based SEs. Table 5.4 compares the two sets of standard errors produced for the methods and Table 5.5 shows how the z-scores vary depending on the method of choice.

In this Medium World, there is very good agreement between the non-model-based and model-based methods for both the standard errors of the university effects and their associate z-scores. In all ten institutions, the SEs from the shrinkage approach are a closer match to the model-based approach than with the non-model-based local variance technique. The same pattern is noted in the z-scores.

University (i)	n_i	$\widehat{SE}\left(\hat{D}_i\right)$	$\widehat{SE}\left(\hat{\alpha}_i^w\right)$
1	6831	0.00317	0.00344
2	3314	0.00474	0.00543
3	1113	0.00978	0.00960
4	2205	0.00675	0.00663
5	289	0.01683	0.01904
6	3238	0.00542	0.00531
7	2889	0.00589	0.00567
8	2292	0.00674	0.00649
9	1031	0.01055	0.00990
10	116	0.02639	0.03026

Table 5.4: A comparison of SEs for $\hat{D}_i$ and $\hat{\alpha}_i^w$

University (i)	n_i	$\widehat{z}_i^D$	$\widehat{z}_i^{\alpha^w}$
1	6831	9.58	9.44
2	3314	5.52	4.66
3	1113	−3.13	−3.27
4	2205	−4.09	−4.41
5	289	4.04	3.74
6	3238	−2.33	−2.47
7	2889	−4.83	−5.11
8	2292	−3.48	−3.85
9	1031	−4.33	−4.76
10	116	1.67	1.57

Table 5.5: A comparison of z-scores for $\hat{D}_i$ and $\hat{\alpha}_i^w$.

Model vs non-model-based : the Published World

A more important question to ask is how would this model-based approach affect the Big World results. Let us set the results from running our non-model-based approach using a $\gamma_{0.5}$

variance estimator as "truth". We have already seen that this γ approach has well calibrated tail behaviour for estimating good and bad universities. We can now examine how far from the "truth" the FE model-based results are.

In the non-model-based world, if I use the Bonferroni (e.g., Johnston and Wichern (1982)) cut-off, based on 165 comparisons, for identifying good ($z \geq 3.61$) and bad universities ($z \leq$ -3.61), there are 16 good, 129 average and 20 bad universities. The Bonferroni approach is developed from a probability inequality of the same name and allows the overall error rate of a series of comparisons to be controlled. To gauge how much effect using a model-based approach has on defining universities, I compared how the status of each individual university changes from the true, non-model-based, status to the new status using FE linear multi-level modelling.

		Non-Model Based		
		Bad	OK	Good
FE	Bad	20	3	0
	OK	0	123	1
	Good	0	3	15

Table 5.6: Status changes in the Published World.

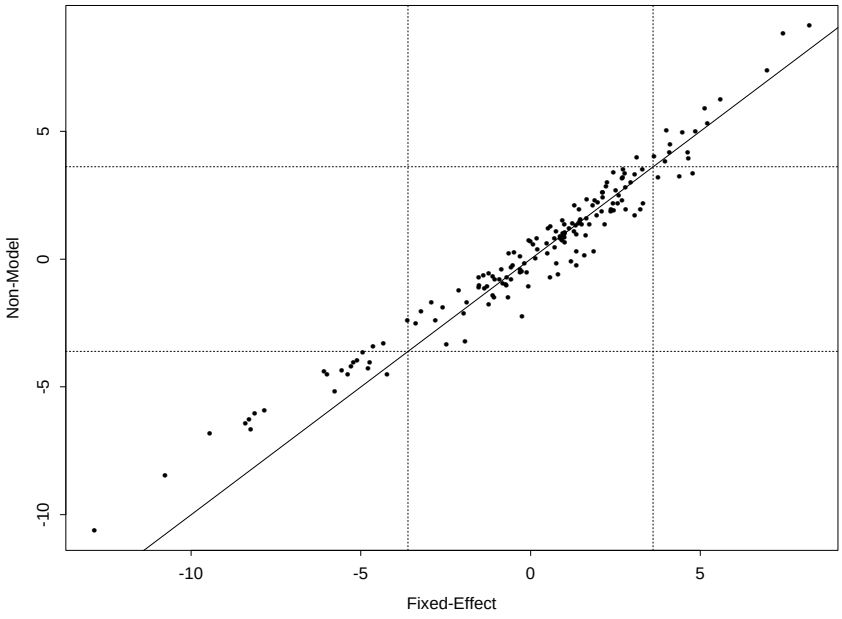

Figure 5-1: The link between the model-based and non-model-based z-scores.

With the linear multi-level approach, 4.2% (7 out of 165) of the universities are misclassified compared with the non-model-based technique. All of the bad universities are discovered and only one of the good universities is pushed into the average category. There is never an occasion

when a university is identified as good when it is in fact bad, or vice-versa. Figure 5-1 shows that the two sets of z-scores are highly correlated (0.975). The dotted lines on the figure represent the Bonferroni cut-off points (± 3.61) for the z-scores. If a point appears in the bottom left rectangle created by these lines, the university is defined as bad using both methods. Similarly if the university appears in the centre or top right rectangle, its status is average or good respectively using the two methods. If the points appear in any of the other rectangles on the figure, this means that there is a difference in opinion between the two methods. The numbers of points in these rectangles tally with the "misclassified" universities given in Table 5.6. As you can see from the plot, if an institution is misclassified this usually occurs when the university z-score is very close to the cut-off point, e.g., these are establishments which had a z-score of 3.58 using one method and a z-score of 3.65 using the other.

5.4 Breaking the analysis up: a regional perspective

There is some debate over whether certain groups of universities should be included in the same analysis as the whole population. This concern arises when some universities are perceived to be so different in character to the rest of the institutions, that however much adjustment is made (i.e., more and more PCFs included) the true underlying differences in the style of institution is never taken into account. This concern has been raised over some regional institutions. The tradition of the these universities means that a different teaching and social ethos has been created and there are many issues about including them in a UK wide analysis. Therefore it is important to note and record how sensitive the results are when these regional institutions are considered separately and in a whole UK study.

In a world with only the specific regional universities included, there are 7069 different student types created from 29231 students in 21 universities. Potentially a separate analysis on these regional universities could be done, which would mean that they would be omitted from a complete UK analysis. Table 5.7 shows the potential results for these universities, based on the non-model-based approach described in Section 2.3.

With a regional only set-up, a single university is identified as good (Institution 144) and two universities are classed as bad (Institution 9 and Institution 78), based on the Bonferroni z-score cut-off for 21 universities (± 3.03). When all 165 UK universities are assessed, Institution 144 remains as the only excellent regional university (based on a new Bonferroni cut-off for 165 universities of 3.61). However, on the other extreme, Institution 78 and Institution 9 are still tagged as bad in the UK analysis but now Institution 64 and Institution 145 are also identified as unusually poor for student progression. These two universities cause some concern when considering whether to use a regional only analysis.

Consider a worst case scenario for a separate or whole analysis problem. Imagine if all the regional institutions were exceptionally good in terms of progression statistics relative to the rest of the university population. In a complete UK analysis, all the regional institutions would be identified as particularly good (and rightly so). In a regional analysis only, maybe a couple of universities would be rewarded with an excellent tag. Potentially some universities might be

classed as unusually bad.

So is there a problem with the analysis and, if so, what issues need to be addressed? The real problem is deciding the question that needs to be answered. Are we interested in the regional universities relative to the performance in that region or relative to the whole UK? The comparative analysis (Regional vs UK) attempts to answer the two questions posed. A baseline needs to be set-down. If it so happens that the regional universities are doing best due to hard work and a bit of luck, then they should be rewarded as being identified as excellent in a complete UK analysis. But if the regional universities use completely different methods to the rest of the UK that can't be measured, then a separate analysis needs to be undertaken. If the alternative methods could be measured in some formal and fair way, then they could be adjusted for in a complete UK analysis by including additional predictors/adjustors.

	Non-Model Based z-scores	
Inst	Regional Only	Complete UK
2	−0.21	−1.91
9	−5.94	−8.11
15	−0.64	−0.36
21	1.50	−2.60
22	−0.18	−2.57
23	0.46	1.05
28	−1.36	−0.37
31	1.44	−1.05
37	1.02	2.38
55	0.24	−1.90
60	2.01	1.02
64	−2.57	−5.80
76	1.80	−0.60
78	−4.12	−7.97
88	0.57	0.88
142	1.93	1.77
143	1.06	1.23
144	6.36	5.37
145	−1.94	−6.79
149	2.58	3.48
155	2.08	0.01

Table 5.7: Regional only vs complete UK analysis.

Chapter 6

Variation in the quality assessments

6.1 Introduction

How certain are we that we are right when we identify a university as "bad" or "good"? What we would like to have is a standard error for each university's z-score. If a university has a z-score of, say, -1.00, and thus an status of OK, how likely is it that the university is in fact a bad university with regard to progression? I am essentially interested in the effect on a university's z-score if a few of that university's students move from 0 to 1 (in terms of progression status) or vice-versa. This is like asking what would have happened to the z-score if Miss U.N. Sure had decided to drop out after getting five numbers and the bonus ball on the lottery, winning 100,000 pounds, rather than staying on after only getting four numbers after winning 25 pounds. This university got lucky in terms of progression performance as the student's lottery luck kept her at the university rather than something the university did.

The university z-score gives us an impression of how likely it is for a specific university's $\hat{D}$ to be significantly difference from zero. A large absolute z indicates a high probability that the university is different in performance from an "average university". Another approach is to use a graphical method to examine the variation in the $\hat{D}_i$'s by plotting the associated confidence intervals for each $\hat{D}_i$. Figure 6-1 shows a simple graphical solution which plots $\hat{D}_i \pm 1.96\,\widehat{SE}\left(\hat{D}_i\right)$ for each university after sorting the $\hat{D}_i$ from smallest to largest. 30 universities have their 95% "quality interval" entirely below zero and 44 entirely above zero, but only three and zero universities have their entire interval below -0.03 and above 0.03, respectively (the HEFCE practical significance cut-off).

Marshall and Spiegelhalter (1998) investigated a similar approach which examined how much variation occurred in institutional ranks based upon parameters similar to our $\hat{D}_i$'s. I could do an equivalent analysis for ranks in our approach but the primary focus of quality assessment is to discover institutions performing significantly better or worse than the overall average.

As with the $\hat{D}_i$'s, one problem with looking at z-scores is that they are based on wildly

different sample sizes. There is a feel of a hierarchical structure to quality assessments: the variation in the $\hat{D}_i$'s produces the z-scores, which in term have variation of their own. For example, in my Small World with only five universities and four PCF categories, Institution 33 had $\hat{z}_i = -1.89$ and Institution 118 had $\hat{z}_i = +2.39$, and there is no way to tell just by looking at the z-scores that I am much less certain about underlying quality at Institution 33 ($n_{i+} = 55$ students) than at Institution 118 ($n_{i+} = 1,344$). The aim of this chapter is to try and get a handle on how much the z-scores can vary by and what effects these variations have on how the universities are assessed.

These z-score variations can be examined: by considering adjusting for a different combination of PCFs; using simulation runs to produce the z-score over and over again; by bootstrapping the data directly (rather than bootstrapping the PCFs); changing the model structure (i.e., removing interaction terms where possible); or by dealing with missing values in a different manner. All of these issues are discussed in this chapter.

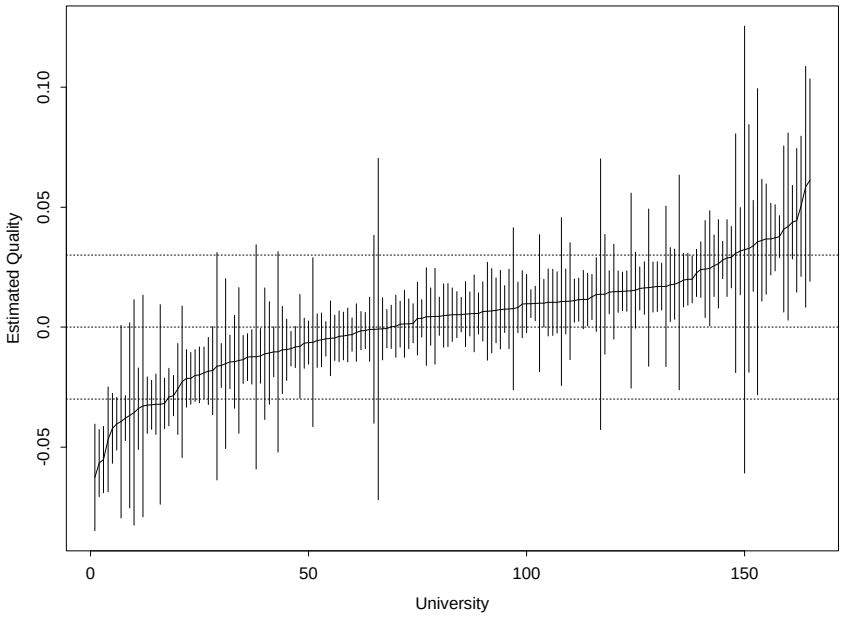

Figure 6-1: Estimated quality $\hat{D}_i \pm 1.96\,\widehat{SE}\left(\hat{D}_i\right)$.

6.2 PCF effect

Motivation for analysis

Under perfect conditions, I would like to take every valid PCF into account and this would give us an accurate picture of what is happening within universities. How much would the university z-scores vary if a new PCF was discovered? So I would like to perform a z-score

sensitivity analysis based on the number of PCFs in the model. The z-scores are calculated using a non-model-based approach with a shrinkage variance estimator. A FE model-based approach produced similar results to those found in the forthcoming sections.

There are two approaches open to me:

- Adjust for all eight PCFs available and assume that those quality assessment results are "truth". Then systematically remove PCFs to see how different the results are based on this reduced number of adjustors.
- Add PCFs to the original eight PCFs. After some further research, HEFCE were able to offer us two new PCFs for consideration: the living arrangements of the students during their first year; and the student's ethnic origin. Now rather than considering the eight PCF results as "truth", we consider models that include the additional PCFs as truth and see how much these results vary from our chosen eight PCF results. Rosenbaum and Rubin (1983) attempted a similar analysis by trying to establish the effect of an unobserved binary PCF on a binary outcome observational study. Rosenbaum restricted himself to the scenario where there were only two treatments (equivalent to only two institutions in our case) and only a single additional binary covariate.

Table 6.1 shows how much influence additional PCFs have on the z-scores. It is based on shrinkage variance estimation, and row k averages over all $\binom{8}{k}$ possible subsets of PCFs. As more and more PCFs are adjusted for, the university quality assessment z-scores become less extremal in nature, i.e., more of the variation at a university is being explained by the adjustors. For example, if I only took student gender into account, then the university would be "blamed" for the remaining variation and the institution would be accountable for the variation caused by subjects of study, student qualification or class status. Therefore if not all the PCFs are included in the modelling process, her progression rate influence is blamed on the university and it's true progression performance is clouded (i.e., the potential for institutional misclassification is increased.) The table highlights how much effect additional PCFs have on the quality assessment results.

Additional PCFs

The Living Variable

The first additional variable available to us recorded the living arrangements of the student in her first year. This variable has five levels: the student lived in university-owned property (including off-campus accommodation); student lived in the parental home; student owned her own accommodation; the student's living arrangements were not known; and the student was not in attendance at the institution (e.g., a student studying a full-time distance learning course).

Table 6.2 shows the marginal distribution for this living variable and the associated progression rates for each living type.

Number of PCFs	Models	Mean M	z-scores Mean	SD	% "Bad"	% "Good"	% Unusual
0	1	1.0	0.860	6.20	0.212	0.280	0.509
1	8	6.1	0.762	5.29	0.185	0.258	0.442
2	28	31.1	0.678	4.64	0.165	0.227	0.392
3	56	132.4	0.602	4.17	0.150	0.200	0.351
4	70	483.9	0.531	3.81	0.139	0.176	0.316
5	56	1,498.9	0.461	3.53	0.127	0.154	0.282
6	28	3,933.3	0.389	3.30	0.116	0.135	0.252
7	8	8,906.4	0.314	3.10	0.111	0.114	0.225
8	1	17,799.0	0.236	2.90	0.097	0.073	0.170

Table 6.1: How adding PCFs affects the z-scores.

Living Status	n	Progression Rate
Institution maintained property	128580	0.943
Parental/Guardian home	36353	0.870
Own Home	57024	0.867
Other	23466	0.877
Not Known	37840	0.860
Not in attendance at the institution	1136	0.870
Total	284399	0.901

Table 6.2: The living status for students.

Splitting the students by the original eight PCFs and the living variable (in a fully saturated fashion) produced 45,570 different student types. Using a Bonferroni cut-off ($|z| \geq 3.61$), we can examine how the status of the universities change after the addition of the living PCF. Table 6.3 show the universities status for the two methods.

		8 + Living Good	OK	Bad	Total
Original 8	Good	7	5	0	12
	OK	6	130	1	137
	Bad	0	6	10	16
Total		13	141	11	165

Table 6.3: Status change on addition of the living variable.

This leads to: an overall misclassification rate of 10.9%; a "Bad but not called Bad" rate of 9.1%; and a "Good but not called Good" rate of 46.2%. The table shows that, if we consider the eight + living PCF results as truth, the majority of "truly" poor universities are identified with only eight PCFs (10 out of 11). On the other hand, using only the eight PCFs means that only seven of the 13 good universities are discovered. In total 147 of the universities are

correctly categorised.

The Ethnic Variable

The ethnic variable describes the ethnic origin of the individual and has nine different levels that are listed in Table 6.4. The highest progression rates occur with those individuals from a Chinese background. The largest group are from a White background and this group of 223,336 individuals have an above average success rate. Students from a Black and Bangladesh backgrounds have the most disappointing rates, with only around 84% progressing. These differences could be down to a number of factors ranging from cultural differences, subject of study, entry qualifications, or attitude of individuals at university to specific ethnic origins.

Splitting the students by the original eight PCFs and the ethnic variable produces 40,420 different student types. Table 6.5 shows how the universities change in status when the ethnic variable is taken into account. This leads to: an overall misclassification rate of 5.5%; a "Bad but not called Bad" rate of 11.8%; and a "Good but not called Good" rate of 31.3%. The change in results isn't as dramatic as with the living variable. Misclassification occurs in nine universities (half the number produced by not including the living variable), with 5 out of 16 "good" universities being classed as OK when only the original eight PCFs are used. The number of bad universities not identified increases to two from a potential population of 17.

Ethnic Type	n	Progression Rate
Chinese	2607	0.925
White	223336	0.908
Indian	10258	0.905
Unknown	9150	0.896
Other(NI Black)	7346	0.878
Information Refused	15036	0.872
Pakistani	5599	0.868
Bangladesh	1593	0.842
Black	9474	0.839
Total	284399	0.901

Table 6.4: Breakdown of ethnic background.

		8 + Ethnic			
		Good	OK	Bad	Total
Original 8	Good	11	1	0	12
	OK	5	130	2	137
	Bad	0	1	15	16
Total		16	132	17	165

Table 6.5: Change in status on addition of ethnic class.

Both Variables

I can also adjust for the eight PCFs and the two additional PCFs as well. This produces 76,308 unique student types. This means that the data is spread extremely finely over the whole "university by PCF" grid but the non-model-based method still works and is valid. I am now assuming that the true results are produced by adjusting for all ten PCFs. This analysis produces Table 6.6.

This leads to:

- An overall misclassification rate of 10.9%;
- A "Bad but not called Bad" rate of 18.1%;
- A "Good but not called Good" rate of 36.4%.

The overall misclassification rate is still relatively small (given I am assuming I have missed two PCFs completely) with only 18 universities being incorrectly identified. There are some concerns with over a third of the good universities being marked as OK.

		8 + 2 extra			
		Good	OK	Bad	Total
Original 8	Good	7	5	0	12
	OK	4	131	2	137
	Bad	0	7	9	16
Total		11	143	11	165

Table 6.6: Adjusting for ten PCFs.

Implications

In most cases, it is those universities that are very close to the z-score cut-off point (defined by Bonferroni) that change status. In this chapter I examine methods that allow me to determine how often these universities, "close" to the cut-off, will fall either side of it. The results given in this section are certainly encouraging for identification of universities, with a maximum misclassification rate of 10.9%. In this data the methods do, however, struggle to discover all the good universities with up to a 36% rate of good universities being marked as average/OK. There are not as many concerns with regard to finding bad universities, with misclassification rates from 9.1 to 18.1%.

Sensitivity to omitting PCFs

Note: Results in Tables 6.7-6.8 are averaged across all possible removals in each row. Bonferroni cut-off based upon $|\hat{z}| \geq 3.61$. The HEFCE cut-off is based on $|\hat{z}| \geq 3.00$ and $|D| \geq 3\%$.

Number of PCFs Removed	Overall Misclassification (%) Using Cutoff	
	Bonferroni	HEFCE
0	0.00	0.00
1	7.65	4.92
2	11.52	7.64
3	15.24	11.21
4	19.07	14.71
5	19.66	18.97
6	21.90	24.39
7	23.79	31.21
8	38.79	39.39

Table 6.7: Overall misclassification: omitting PCFs.

Rather than study the effects of adding two additional PCFs to our original set, I can also ask how sensitive our finding are to omitted PCFs. One empirical answer takes the world based on eight PCFs as truth and asks how close working with only seven, six, ... PCFs comes to reproducing that truth.

When the Bonferroni or HEFCE cutoffs are used, omitting one of the eight PCFs leads to an average overall misclassification rate of only 5–8% (8–12 universities out of 165), and even with two missing PCFs the HEFCE cutoff has an average error rate of only 7.6% (Table 6.7).

Number of PCFs Removed	Good But Not Called Good (%) Using Cutoff	
	Bonferroni	HEFCE
0	0.00	0.00
1	8.33	9.38
2	14.29	14.73
3	17.86	20.54
4	19.88	25.00
5	20.39	29.02
6	20.54	33.93
7	21.88	40.63
8	25.00	50.00

Table 6.8: Good institutional misclassification: omitting PCFs.

When classification errors occur with the Bonferroni or HEFCE cutoffs, they almost all involve incorrectly labelling a university as "good" when actually it's "OK" (Table 6.8); the average rate of failing to identify "bad" universities is only 0–2% with the HEFCE cutoff up to and including six omitted PCFs (Table 6.9). These results need to be studied in finer detail as obviously some PCFs are more important in identifying truth than others. I concentrate on the effect of removing a single PCF from our original eight. Tables B.1-B.8 show how the results vary depending on which PCF is removed. The pseudo-R^2 values in the table are a measure of the predictive power of each PCF in a logistic regression of progression status on the variable

in question (e.g. STATA (2001)).

Number of PCFs Removed	Bad But Not Called Bad (%) Using Cutoff	
	Bonferroni	HEFCE
0	0.00	0.00
1	4.69	1.04
2	6.47	0.60
3	7.37	0.74
4	8.21	0.71
5	8.04	0.89
6	7.37	1.79
7	7.81	4.17
8	12.50	8.33

Table 6.9: Bad institutional misclassification: omitting PCFs.

The tables show quite clearly that student entry qualifications and subject of study have the greatest influence on generating the "truth". When the qualification PCF is omitted (Table B.3) using a Bonferroni cut-off, 17.6% of the universities are misclassified. This compares to an average, one PCF removed, rate of 7.7%. When the subject PCF is removed (Table B.4), a high misclassification rate is still noted: 14.5%. The overall misclassification rates vary between 2.4% and 7.9% if other PCFs are taken out. When only the state school PCF is removed, only four out of 165 universities are misdefined and all 16 bad institutions are identified. The next subsection examines models that always include the qualification and subject PCFs as it is likely that, in any general study, PCFs with a large influence will not be "forgotten". Therefore I need to analyse scenarios where less important PCFs are omitted from models but the principal PCFs are always included.

Omitting PCFs from models including qualifications and subject

Number of PCFs Removed	Overall Misclassification(%)	Bad But Not Called Bad(%)	Good But Not Called Good(%)
0	0.00	0.00	0.00
1	4.85	4.17	5.56
2	6.02	5.83	10.56
3	7.21	7.19	12.92
4	8.12	9.17	13.33
5	9.09	11.46	12.50
6	9.69	12.50	16.67

Table 6.10: Results with models involving qualifications and subject, using 3.61 cutoff.

Given how influential qualifications and subject of study are on student progression, we need to review the models that contain at least these two PCFs. Essentially I am trying to repeat the analysis of the previous subsection but with a little prior information (or common sense) on which PCFs should be present.

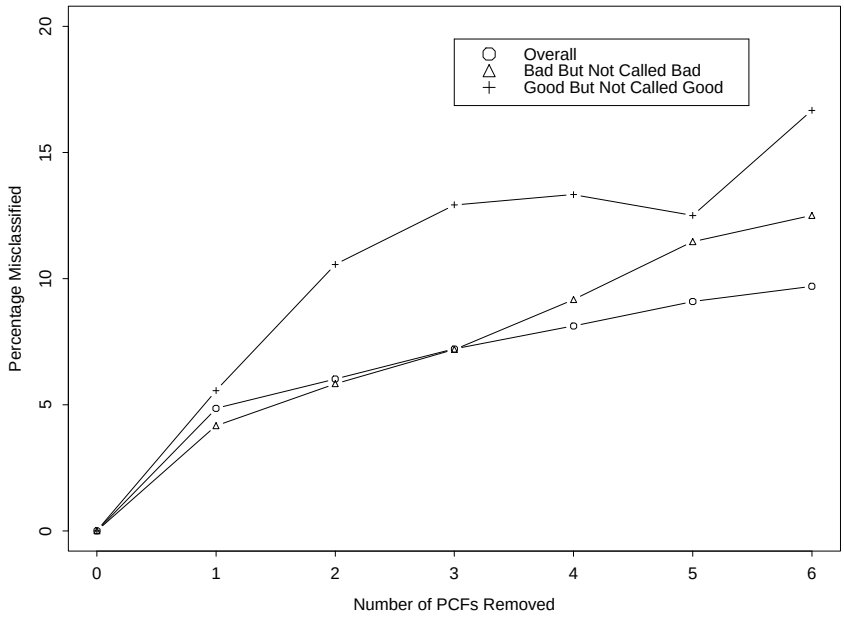

Figure 6-2: Effect of PCF omission on qualification and subject models.

Table 6.10 and Figure 6-2 show the effects of PCF removal from these more restricted models. As expected, the average misclassification rates are a lot lower compared against the models from the preceding subsection. Even with all six possible PCF removed, the overall misclassification rate is less than 10%. This means that, given a large number of PCFs have been forgotten or removed from the adjustment process, the identification process remains fairly stable. The same asymmetry seen in the "bad" and "good" identification results is also seen in these results, with a relatively high proportion of good universities being marked as average when all six PCFs are omitted.

Moving from a model with all PCFs present to a model with a single PCF removed produces the largest jump in misclassification (the overall rate moves from 0.00 (obviously) to 4.9%). Removal of further PCFs has a lesser effect on the identification rates, with the models involving two to five PCFs removed all having similar results. The results from this section and the previous one are encouraging as the results seem relatively stable when considering PCF omission.

A maximum number of PCFs?

I have already discussed that it is important to adjust for as many PCFs as possible, but do the methods restrict us to how many PCFs we can use? The non-model-based local variance method fails when students in a PCF category all fail or all pass at a certain university causing the expected progression rate for all those students to be zero or one. This does not occur in the

smaller worlds because it would mean that each and every student at that university, who fell into specific PCF category, would have to be present in a nation-wide category of student who all passed or failed. In our dataset, the local method fails in the Big world when around 40,000 student types are identified (or ten PCFs are adjusted for). The local method also breaks at a specific university if all the student's observed and expected value are matched identically.

In the Big World, the shrinkage method has yet to break down but would fail under these conditions:

- A university's students became completely isolated from the remaining population. The method is then unable to calculate a valid SE for the university's $\hat{D}_i$.
- The whole population all pass or all fail.

This does not mean that the method is producing completely valid results under all other conditions but it is capable of producing results of some description in all other cases. Section 4.8 discusses how valid the method results are under a variety of conditions, relating to how sparse the data grid is or/and the proportion of successes/failures.

6.3 Bootstrapping

Bootstrapping is a commonly used simulation method for generating standard errors for parameter estimates of interest. Both Efron and Tibsharani (1993) and Manley (1997) discuss the idea behind the bootstrapping approach:

1. Rather than using the original data, sample observations from the original data;
2. Calculate your parameters of interest from this generated data; and
3. Repeat 1 and 2 until a reasonable level of accuracy is achieved, given the time available.

I am interested in examining the variation in the universities' z-scores. The bootstrap approach for my data involves:

1. Use the Published World as the original dataset.
2. At university i, sample with replacement n_{i+} (the number of students at that university). A new simulated dataset is now created with 284,399 students with the correct number of students at each university. This dataset is a version of the original dataset with some students omitted and some students appearing more than once.
3. Use the non-model-based approach (with shrinkage variance estimation) to generate each university's z-score.
4. Repeat steps 2-3 until sufficient simulations have been completed.
5. Standard errors for the z-scores can now be computed from the standard deviations of the z-scores across simulated datasets.

Tables C.1 - C.4 show the results of a bootstrap analysis adjusting for only two PCFs, student qualifications and subject of study based on a target of 1,000 bootstrap replications. The tables show, for each university, the percentage of simulations the university is found to be good, bad or average, alongside the simulation z-score mean and SD. The "true" status and z-score is also given: the true is defined to be the university's quality assessment results based on the original Published World.

At some institutions, the number of valid simulations was less than 1,000. An invalid simulation was produced when all the university's students progressed and these occurred at small universities which had very high original progression rates (e.g., Institution 11 has only 222 students and 98% of them progressed). The bootstrap mean z-scores were highly correlated with the true university z-scores: 0.990. The z-score standard deviations varied between 0.96 (Institution 140) and a massive 8.7 (Instiution 11). Intuitively a standard deviation of 8.7 seems very large and we shall see in Section 6.4 (by another method) that the SD should be a little smaller. This unusual result is caused because the institutional progression rate is very close to 1.0 and the binomial assumptions start to break-down.

Some institutions are always defined as unusual and we can be reasonably sure that there is something out of the ordinary going on at those universities (e.g., Institution 1 is always classed as excellent, Institution 110 is always marked as poor). Some institutions are always defined as average (e.g., Institution 94 or Institution 152) and we can be fairly certain, given the data provided, that there is no cause for concern at these establishments. If I define success as a university where the percentage of time its classification in the simulations matches its "true" status is at least (say) 75%, then the success rate across the 165 universities was 82%. For example, at Institution 3 the true status is bad (z-score of -4.02) and 79% of the simulations class Institution 3 as bad, so this is termed as a success because I have correctly identified the university more than 75% of the time.

6.4 Non-null simulations

A more robust approach than the bootstrap method is based on non-null simulations, in which data sets are generated with the FE model (Section 3.5):

$$\begin{array}{l} y_{ij} = \beta_0 + \sum_{k=1}^{M} \beta_k x_{ijk} + \alpha_i + e_{ij}, \\ e_{ij} \overset{\text{IID}}{\sim} N\left(0, \sigma_e^2\right), \quad \sum_{i=1}^{N} n_{i+}\, \alpha_i = 0, \end{array} \tag{6.1}$$

for various choices of $\alpha_i \neq 0$.

We have already seen that the α_i's are the model equivalent to the non-model-based D_is. Therefore this method creates an idea of z-score variance developed initially from a model-based structure, but then using the standard non-model-based methods to create estimates for D_i.

To create a variance for the α-scores, the following steps are taken:

1. Fit the model to the dataset of interest (i.e., using the Published World data), via maximum likelihood. Software packages such as `GLIM`, `SAS` or `STATA` are capable of fitting such

models. These software packages cannot place the $\sum_{i=1}^{N} n_{i+}\,\alpha_i = 0$ restriction directly so one of the α_i must be set as a baseline group, i.e., $\alpha_b = 0$ where b is the baseline PCF category.

2. Use the output from this baseline model fitting (the α parameters and their covariance matrix) and the information on restriction conversion given in Section 3.8 to produce valid $\hat{\alpha_i}$ and associated standard errors. These results can be used to find the actual $\hat{\alpha}_i$ for each university.

3. Use the α and β parameters from the model to create an estimated probability of progression for each student in the dataset.

4. Simulate a progression result for each student in the dataset using her own individual probability of progression based on her PCF characteristics and university of study.

5. Calculate each D_i and, in turn, the associated university z-score using the standard non-model-based shrinkage method given in Section 4.8.

6. Repeat steps 4-5 for an appropriate number of simulations (say q), depending on time and accuracy considerations.

Using this technique leads to a series of q z-scores for each university, which can be used in a similar manner to the bootstrap simulations, i.e., calculate a SD for each z-score and other university z-score summaries. For illustration I fit this model to the entire data set, via maximum likelihood, but using only the $M = 272$ PCF categories based on entry qualifications and subject employed in the HEFCE December 1999 publication of PIs, obtaining $\hat{\alpha}_i$ values (i.e., the Published World). I then set $\alpha_i = \hat{\alpha}_i$ and generated 1,000 random replications of all 165 universities (with the same n_{i+} values as in the actual data), keeping track of the mean and SD of the $\hat{z}_i$ scores and the percentage of time each university was classified as "bad", "OK", and "good" (with the original z-scores on the Published data as "truth"). Tables D.1 - D.4 show the results of this non-null analysis.

The means of the z-scores tracked the "true" HEFCE values almost perfectly (and the correlation was 0.9998). The SDs of the z-scores ranged from 0.49 (Institution 14) to 1.14 (Institution 107), with a mean of 0.96. As before, if I define a success as a university where the percentage of time its classification in the simulations matches its "true" status is at least 75%, then the success rate across the 165 universities was 85%.

How do the bootstrap and non-null results compare? Figure 6-3 shows the link between the non-null and bootstrap z-score means. There seems to be good agreement between the two methods in terms of z-score means. The correlation between the two sets of means is 0.99. Figure 6-3 clearly shows some shrinkage of the non-null z-score means back towards zero when compared with the bootstrap. This is mainly because the bootstrap method struggles to deal with estimating z-scores when a university's progression rate approaches one. The bootstrap is more dependent on the binomial assumptions being made about individual proportions.

Figure 6-4 confirms a difference in the two methods. The figure shows the relationship between the two sets of z-score SDs produced by the bootstrap and non-null approaches. The bootstrap SDs are always larger than the non-null SD estimates. This is once again related to the inability to deal with institutional proportions as they approach one. If a simulation run produces a university that has nearly a 100% progression rate, the binomial variance approximation fails and will produce overestimates of the SDs. This near 100% progression rate occurs very frequently with the bootstrap as each simulation run is forced to chose from a restrictive population.

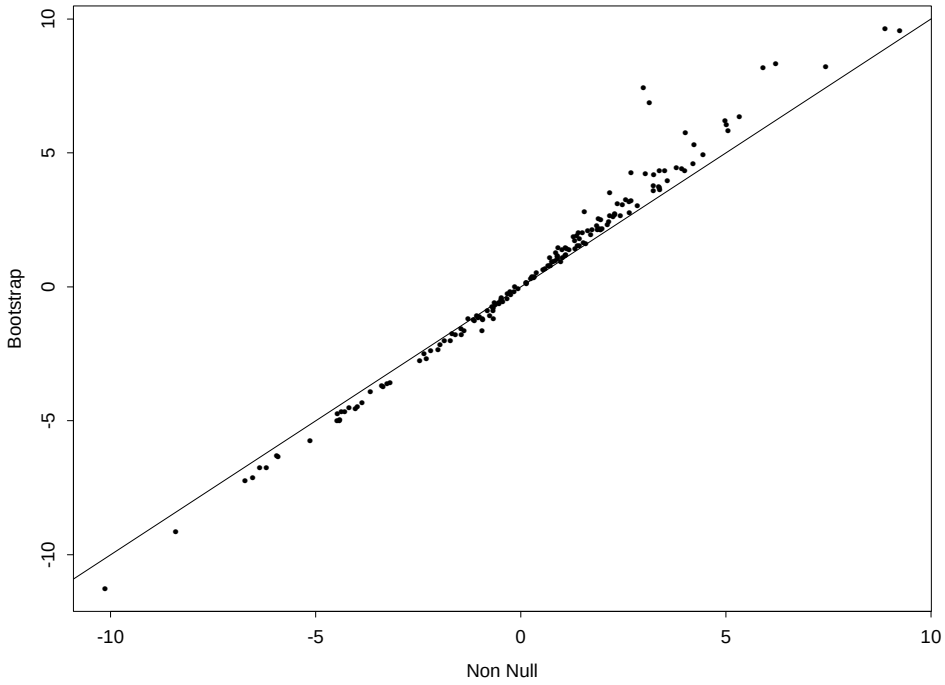

Figure 6-3: Link between non-null and bootstrap z-score means.

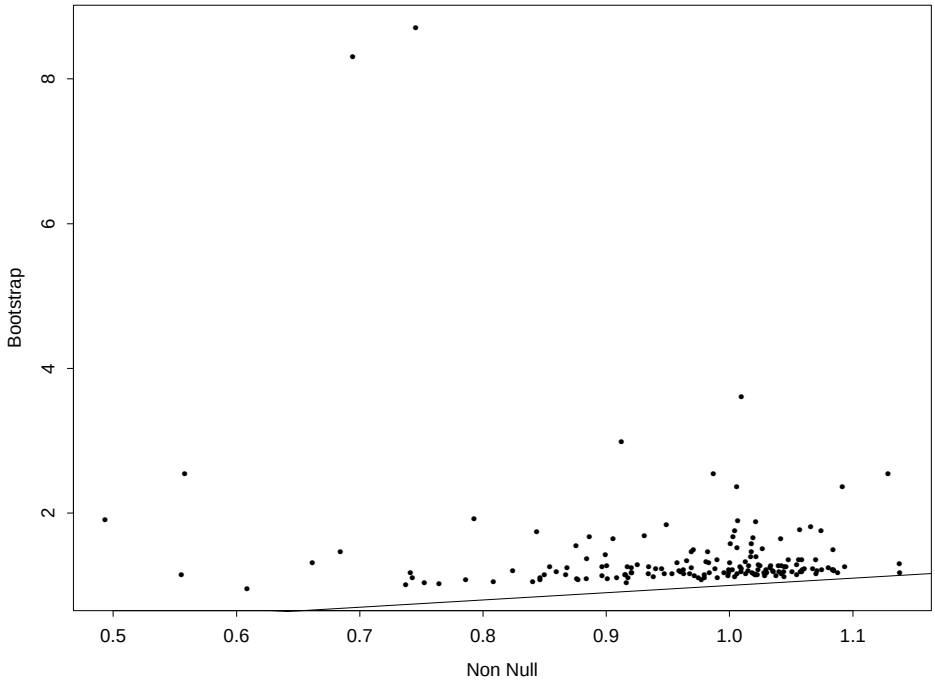

Figure 6-4: Link between non-null and bootstrap z-score SDs.

6.5 Reducing the number of interactions for models

Introduction

When trying to establish a university's quality, we have already seen in Section 3.2 that fitting a fully-saturated linear fixed-effects model produces similar results to the non-model-based approach. In normal circumstances, if a statistician was told that an outcome was being modelled using a number of predictors and all their possible interactions, she would probably suggest that there is serious inefficiency in the process, i.e., not all the terms in the model equation are required.

There are some advantages of using a fully-saturated approach for explaining progression rates. Firstly, with all interactions included, the universities are unable to "claim" that their results would have been different if additional interaction effects had been present as there are no extra interactions. Secondly, `GLIM`'s eliminate command allows fully-saturated models to be fitted with much more ease compared to models, with a large number of terms, that are not fully-saturated.

In the Medium World, the fully-saturated model essentially contains 36 predictors: one constant, six main effects, thirteen two-way interactions, twelve three-way interactions and four four-way interactions. The simplest question we can ask is how many of these interaction

terms are required to reproduce the fully-saturated results to a reasonable level? In the Big World, there are eight PCFs making eight-way interaction terms possible. In this case, there are eight potential models that could be considered: main effects, up to two-way interactions, ..., up to seven-way interactions and fully-saturated. A statistician might also decide rather than include all x-way interactions in the model, to include some x-way interactions but remove the non-statistically significant ones. She would then consider using a forward (or backward) step-wise approach to predictor selection.

Medium World

University	Main Effects Only			2-Way			3-Way		
	α	SE	z-Score	α	SE	z-Score	α	SE	z-Score
1	0.032	0.003	9.36	0.032	0.003	9.33	0.033	0.003	9.38
2	0.026	0.005	4.88	0.026	0.005	4.69	0.025	0.005	4.64
3	-0.032	0.010	-3.36	-0.032	0.010	-3.34	-0.031	0.010	-3.27
4	-0.029	0.007	-4.34	-0.029	0.007	-4.41	-0.029	0.007	-4.43
5	0.071	0.019	3.71	0.071	0.019	3.75	0.071	0.019	3.74
6	-0.013	0.005	-2.49	-0.013	0.005	-2.45	-0.013	0.005	-2.48
7	-0.029	0.006	-5.08	-0.029	0.006	-5.10	-0.029	0.006	-5.11
8	-0.025	0.007	-3.83	-0.025	0.006	-3.86	-0.025	0.006	-3.86
9	-0.046	0.010	-4.68	-0.047	0.010	-4.72	-0.047	0.010	-4.74
10	0.048	0.030	1.59	0.049	0.030	1.63	0.048	0.030	1.57

Table 6.11: Model effects in the Medium World I.

University	Fully Saturated			Non-Model Based		
	α	SE	z-score	D	SE	z-score
1	0.033	0.003	9.44	0.030	0.003	9.50
2	0.025	0.005	4.67	0.026	0.005	5.17
3	-0.031	0.010	-3.28	-0.031	0.009	-3.23
4	-0.029	0.007	-4.41	-0.028	0.007	-4.19
5	0.071	0.019	3.73	0.068	0.019	3.57
6	-0.013	0.005	-2.47	-0.013	0.005	-2.37
7	-0.029	0.006	-5.09	-0.028	0.006	-4.99
8	-0.025	0.006	-3.87	-0.024	0.006	-3.64
9	-0.047	0.010	-4.75	-0.046	0.010	-4.60
10	0.047	0.030	1.57	0.044	0.030	1.46

Table 6.12: Model effects in the Medium World II.

The non-model-based results were compared against FE models with varying levels of interactions in the Medium World. If I assume temporarily that the non-model-based results are the "truth", we can compare the other approaches to see how much they vary from these "true" assessments (Table 6.11 - 6.12). A Bonferroni cut-off for 10 comparisons is 2.81 so universities with an absolute z-score larger than this point are classed as unusual. In all cases, a university is never misclassified i.e., all methods produce the same status for each university (Good: Institution 1; Institution 2; and Institution 5. Average: Institution 6; and Institution

10. Poor: Institution 3; Institution 4; Institution 7; Institution 8; and Institution 9). There is very little change in the z-score behaviour implying that moving from a non-model-based approach to a FE model-based approach with a reduced number of interactions has little effect on quality assessment. The main effects seem to explain the majority of the variation in progression performances.

Published World

I can repeat this analysis in the Published World. This world only adjusts for two PCFs so there are only two potential model-based alternatives to the non-model-based approach: fully-saturated; or main-effects only. As before, I assume that the non-model-based results are "truth". Table 6.13 shows the misclassification rates of the 165 universities for the model-based approaches.

Model Description	Incorrect University Status(Out of 165)	Misclassified
Linear Full	6	4%
Linear Main	14	8%

Table 6.13: Model effects in the Published World.

The misclassification ranges from 4% to 8% which are relatively small and the universities misclassified will be those whose z-score falls very close to the Bonferroni cut-off in each approach. There is some information in the interaction term (qualifications * subject), with an additional eight institutions misclassified when it is omitted. In this set-up, the cost of placing the interaction term into the modelling is outweighed by its predictive properties.

Big World

Model Description	Incorrect University Status(Out of 165)	Misclassified
Linear Full	4	2%
Linear Two-Way	13	8%
Linear Main	21	13%

Table 6.14: Model effects in the Big World.

I cannot complete a fully investigation of interaction effects in the Big World as the majority of models have too many parameters to adjust for and the calculations would taken excessive time. The fully-saturated model can be calculated as the eliminate command in `GLIM` allows a single predictor to be "removed" from the analysis. In a fully-saturated set-up, the student's PCF category can be written in a single parameter and this can be removed from the analysis. These restrictions mean we can calculate misclassification results (compared against

the non-model-based approach) for main effects only, including two-way interactions and a fully-saturated model.

The misclassification rates range from 2 to 13 % for the possible model-based approaches (Table 6.14). Moving from a non-model-based to a model-based technique based on a fully saturated multilevel linear model means only 4 from 165 universities are misclassified. The removal of interaction terms has a greater effect with an additional 9 universities being misclassified when a two-way interaction model is used rather than a fully-saturated one. The effect is as dramatic when the two-way interaction terms are also lost, producing a main effects only structure.

Implications of the interaction terms

Interactions can and will have an effect for quality assessments at a given set of institutions. The importance of including interactions varies depending on the data structure. If it is important that people believe you have adjusted for everything, then the fully-saturated approach should be used (model- or non-model-based). Both approaches can be relatively easily implemented using `GLIM` (in the model-based approach) or using the process described in Section 2.3 (in the non-model-based approach). If a reduction in interactions is needed, perhaps because a random-effects (RE) methodology is required (see Section 8.1), then the number of x-way interactions required for reasonable prediction increases as the number of PCFs and data size increases. With larger datasets (i.e., more than four PCFs adjusted for and more than 50,000 observations) it is recommended that at least two-way interactions are used, and further interactions if time allows.

6.6 Sensitivity to missing values

Introduction to missing values

Some of the eight PCFs examined have a large degree of missingness present. 37% of student have a school type that is unknown and over 30% have an undefined social class. From discussions with HEFCE and studying the data in general, it seems that a large factor of missing data for students is related to whether they are mature or not. This implies that if a mature student has some missing categories in the other PCFs, it's highly probably that the missingness is due to the fact that the student is mature. In these cases, the missing PCF category acts as another proxy for identifying mature students.

Yang et al. (ming) examine the effects of non-random missing data in a multilevel dataset relating to student A-level attainment. They use the dataset used by Yang and Woodhouse (2001) and examine how adjusting for certain PCFs affects the assumptions of missingness. They conclude that student subject choice has a high correlation with student attainment and fitting models that include terms relating to the combination of subjects a student has chosen can help with dealing with non-random missing data.

To examine the effects of missing categories in my HE data, I look at three cases. This will allow us to discover how sensitive the data are to missing values:

- Run the analysis with all the information present (i.e., as normal), treating "missing" as another category in the PCF;
- Remove all students who have any missing information in any of the categories. This means that the methods are only performed on students who have "perfect" data;
- We can assume that if a mature student has missingness then that is due to her being mature and not another unmeasured factor. Keeping these students helps to increase the information in the dataset. Missing categories in young students cannot be attributed to them being mature (because they're not) and must be down to another, potential unmeasured, factor and these student are omitted.

  The effects of these different treatments are examined in the Medium and Big Worlds.

Medium World

Table 6.15 shows how the number of students vary at each institution: when all students are included; when "young" students with a missing category are removed; and when all students with a missing category are removed.

University	Original n	Without Non-Mature Missing n	Without Missing n
1	6831	4334	2051
2	3314	2918	2535
3	1113	994	773
4	2205	1815	1345
5	289	140	53
6	3238	2466	1532
7	2889	2132	1260
8	2292	1936	1055
9	1031	789	545
10	116	111	86
Overall	23318	17635	11235

Table 6.15: Variation in numbers depending on missingness: Medium World.

The table gives a good guide to the data quality and proportion of mature students at each university. Institution 1 loses over 4,000 students (around 70% of the original total) when individuals with any missing categories are taken out, whereas Institution 10 and Institution 3 lose only around 25-30% of their original intake. Institution 5 has very few total "perfect" students with its levels dropping from 289 to a small 53. HEFCE hope to improve the data by including a missing rate in their future tables (i.e., for each institution to give a value that is an indication of how many of their students have missing categories).

Tables 6.16 - 6.17 show how the non-model-based results vary depending on how missingness is dealt with. Table 6.18 shows how the universities' status changes (assuming that status from

the full original data are truth.) The principal message from these tables is that with less data, more and more universities become average (i.e., there isn't enough evidence to identify unusual institutions). When the non-mature students are removed, two of the bad universities are now found to be OK (Institution 8 and Institution 9). If even more students are removed (i.e., all students with a missing category) then all five bad universities are found to be OK. In this case, only one university can be identified as unusual (Institution 2, where the z-score moves from 5.17 to 5.75). It is rare for a university z-score to be pushed further from zero if a number of its students are removed. In most cases, removal of students produces a shrinkage effect in the quality assessment z-scores.

	Original					Without Non-Mature				
University	$\hat{O}$	$\hat{E}$	$\hat{D}$	$\hat{SE}(\hat{D})$	$\hat{z}$	$\hat{O}$	$\hat{E}$	$\hat{D}$	SE($\hat{D}$)	$\hat{z}$
1	0.900	0.870	0.0304	0.00320	9.50	0.892	0.863	0.0289	0.00419	6.91
2	0.922	0.896	0.0262	0.00507	5.17	0.923	0.898	0.0253	0.00502	5.04
3	0.850	0.881	−0.0306	0.00947	−3.23	0.851	0.880	-0.0293	0.01031	-2.84
4	0.859	0.887	−0.0276	0.00659	−4.19	0.861	0.887	-0.0265	0.00740	-3.58
5	0.938	0.870	0.0681	0.01909	3.57	0.936	0.857	0.0789	0.0244	3.24
6	0.860	0.873	−0.0126	0.00532	−2.37	0.870	0.871	-0.0043	0.00616	-0.69
7	0.848	0.876	−0.0284	0.00569	−4.99	0.841	0.874	-0.0339	0.00694	-4.88
8	0.844	0.868	−0.0235	0.00645	−3.64	0.847	0.865	-0.0184	0.00729	-2.53
9	0.834	0.880	−0.0458	0.00996	−4.60	0.856	0.880	-0.0242	0.01175	-2.06
10	0.940	0.896	0.0440	0.03014	1.46	0.937	0.896	0.0407	0.02725	1.49
Overall	0.878					0.902				

Table 6.16: Dealing with missingness: Medium World I.

	Without Missing				
University	$\hat{O}$	$\hat{E}$	$\hat{D}$	$\hat{SE}(\hat{D})$	$\hat{z}$
1	0.894	0.901	−0.0071	0.00596	−1.18
2	0.936	0.908	0.0278	0.00484	5.75
3	0.875	0.896	−0.0210	0.01066	−1.97
4	0.901	0.905	−0.0042	0.00754	−0.56
5	0.981	0.902	0.0790	0.03187	2.48
6	0.903	0.898	0.0055	0.00696	0.78
7	0.890	0.903	−0.0128	0.00800	−1.60
8	0.873	0.892	−0.0195	0.00911	−2.15
9	0.879	0.902	−0.0229	0.01292	−1.77
10	0.942	0.912	0.0297	0.02851	1.04
Overall	0.876				

Table 6.17: Dealing with missingness: Medium World II.

Big World

I can now repeat a similar analysis but replacing the Medium World with the Big World (i.e., all eight PCFS adjusted for). As before, I take the z-scores (and their associated university status) produced from the original data with no students removed to be "truth" and compare

		Without Non-Mature			Without			
		Good	OK	Bad	Good	OK	Bad	Total
	Good	3	0	0	1	2	0	3
Original	OK	0	2	0	0	2	0	2
	Bad	0	2	3	0	5	0	5
Total		3	4	3	1	9	0	10

Table 6.18: Status changes derived from missingness: Medium World.

how much the institutional status changes depending on the rules of removal (Table 6.19).

		Without Non-Mature			Without Missing				
		Good	OK	Bad	Good	OK	Bad	NA	Total
	Good	7	5	0	6	6	0	0	12
Original	OK	3	134	0	3	119	2	13	137
	Bad	0	3	13	0	10	6	0	16
Total		10	142	13	9	135	8	13	165

Table 6.19: Status changes derived from missingness: Big World.

In this larger analysis, some universities have very few students to start with. Removal of students from these small institutions produces establishments with 100% (or 0%) success rates and thus an invalid quality assessment z-score, and this is how the 13 not applicable institutions are produced in the "without missing" analysis. The same features are noted in this larger world that were seen in the Medium World, i.e., a shrinkage effect back towards zero for the z-scores when students are removed. There are 28 (17% of the 165 universities) unusual institutions using the original data. This number drops to 23 (14%) when non-mature students with a missing category are removed. When all the missing category students are excluded, only 11% of the 152 valid institutions are found to be unusual.

There is some sensitivity to student removal in the data. With the smallest possible data (i.e., all missing students removed) the method manages to find six of the twelve good universities. It struggles a little more with the bad universities, finding only six from sixteen. These are encouraging results but there are two cases where average universities are identified as bad and three cases where average universities are found to be good.

Chapter 7

Non-model-based alternatives

7.1 Introduction to the alternative approaches

The original approach involving indirect standardisation to the university cohort (Section 2.3) is a non-model-based approach. There are other non-model-based approaches that could be used and this chapter attempts to describe and examine a few of the alternative approaches. Each of these methods has its advantages and disadvantages over the original approach. Section 7.2 examines the effects of using a ratio to express a university's quality assessment rather than using the original difference method. Section 7.3 looks at the effect of removing university i's information when creating its benchmark, i.e., a studentized ideal. Direct standardisation has been mentioned earlier (Section 2.3) and it was shown that direct standardisation cannot be performed on data of this nature due to missing cells. Section 7.4 looks at a modified direct standardisation approach that allows for missing cells.

7.2 A ratio approach

Motivation for a ratio approach

Epidemiologists regularly use adjustment methods in a similar fashion to the indirect methods described in Section 2.3 to produce observed and expected rates in their studies. However, it is not normal practice to calculate the difference between the $\hat{O}_i$ and $\hat{E}_i$ for institution i. They prefer to use a ratio approach where they calculate $\frac{\hat{O}_i}{\hat{E}_i}$ for each institution i. A common example of this is the standard mortality ratio (SMR), used to identify unusual death rates in parts of a population (Clayton and Hills (1993)). The main distinction between the ratio and the difference approaches is that the ratio is a relative measure.

The forthcoming parts of this section examine the work that has been completed by others on the ratio technique. The ratio standard errors used are compared against the difference SE for validity and equivalence. The effects of using a ratio approach in a number of different scenarios is also contrasted against the difference methodology.

Research into the ratio $\frac{\hat{O}_i}{\hat{E}_i}$

Hosmer and Lemeshow (1995) suggested three methods for calculating the standard error of the ratio, $\frac{\hat{O}_i}{\hat{E}_i}$, in quality assessment. DeLong et al. (1997) suggest that the following method is the easiest of the three suggestions to implement and use. They indicate that all three methods produce similar results in most situations. A more precise standard error is also suggested, which involves a great deal more computation and can only be used when an additional "out of sample" dataset is available.

DeLong et al. (1997) also use fixed and random-effects (discussed in Section 8.1) modelling to examine institutional performance with a logistic link, but no mention is made of key issues, e.g., interaction term effects (Section 6.5) and problems using a non-linear link (Section 8.2). They find that a FE set-up is anti-conservative and suggest that a logistic RE set-up provides a realistic assessment of institutional performance as it can deal with adjusting for predictors, constrains the individual risk probabilities between 0 and 1 and it is robust to small institution sample sizes.

The less complex 95% CI suggested for the ratio by Hosmer and Lemeshow is:

$$\begin{aligned} 95\%\ \text{CI}(\frac{\hat{O}_i^*}{\hat{E}_i^*}) &= \frac{\hat{O}_i^*}{\hat{E}_i^*} \pm 1.96\ \hat{\text{SE}}(\frac{O_i^*}{E_i^*}) \\ &= \frac{\hat{O}_i^*}{\hat{E}_i^*} \pm 1.96 \frac{\sqrt{\sum_{j=1}^{n_{i+}} \hat{p}_{ij}(1-\hat{p}_{ij})}}{\hat{E}_i^*} \end{aligned} \tag{7.1}$$

which implies:

$$\begin{aligned} \hat{z}_i^r &= \frac{1 - (\frac{\hat{O}_i^*}{E_i^*})}{\hat{\text{SE}}(\frac{O_i^*}{E_i^*})} \\ &= \frac{1 - \frac{\hat{O}_i^*}{\hat{E}_i^*}}{\frac{\sqrt{\sum_{j=1}^{n_{i+}} \hat{p}_{ij}(1-\hat{p}_{ij})}}{\hat{E}_i^*}} \\ &= \frac{\hat{E}_i^* - \hat{O}_i^*}{\sqrt{\sum_{j=1}^{n_{i+}} \hat{p}_{ij}(1-\hat{p}_{ij})}} \end{aligned} \tag{7.2}$$

where, for university i: $\hat{O}_i^*$ is the observed institutional number of successes; $\hat{E}_i^*$ is the expected institutional number of successes; n_{i+} is the number of students at the university; $\hat{p}_{ij}$ is the predicted success rate for the j^{th} student at the university; and $\hat{z}_i^r$ is the ratio quality assessment z-score for the university.

This does not complete the formula for $\hat{Z}_i^r$ as both $\hat{O}_i^*$ and $\hat{E}_i^*$ can be redefined using the standard $\hat{O}_i$, the observed success rate at university i, and $\hat{E}_i$, the expected success rate at university i. In our notation, given that y_{ij} is the 0/1 outcome for student j at university i:

$$\hat{O}_i = \frac{\sum_{i=1}^{n_{i+}} y_{ij}}{n_{i+}} \quad \text{and} \quad \hat{E}_i = \frac{\sum_{i=1}^{n_{i+}} \hat{p}_{ij}}{n_{i+}}$$

whereas

$$O_i^* = \sum_{i=1}^{n_{i+}} y_{ij} \quad \text{and} \quad E_i^* = \sum_{i=1}^{n_{i+}} \hat{p}_{ij}$$

in the DeLong et al. (1997) notation. This implies that:

$$O_i^* = n_{i+} O_i \quad \text{and} \quad E_i^* = n_{i+} E_i \tag{7.3}$$

So replacing O_i^* and E_i^* in Equation 7.2 with the appropriate n_i, O_i and E_i forms gives:

$$\begin{aligned} \hat{z}_i^r &= \frac{\hat{E}_i^* - \hat{O}_i^*}{\sqrt{\sum_{j=1}^{n_{i+}} \hat{p}_{ij}(1-\hat{p}_{ij})}} \\ &= \frac{n_{i+}\hat{E}_i - n_{i+}\hat{O}_i}{\sqrt{\sum_{j=1}^{n_{i+}} \hat{p}_{ij}(1-\hat{p}_{ij})}} \\ &= \frac{-\hat{D}_i}{\frac{\sqrt{\sum_{j=1}^{n_{i+}} \hat{p}_{ij}(1-\hat{p}_{ij})}}{n_{i+}}}, \end{aligned} \tag{7.4}$$

where $\hat{D}_i$ is the original difference estimate as in Equation 2.9.

This shows that the approach by DeLong et al. (1997) was based on only accounting for variance in the observed progression rates and does not consider variance in the expected rates. This is because Equation 7.4 implies that the standard error of the $\hat{D}_i$ term is the denominator of this equation, $\frac{\sqrt{\sum_{j=1}^{n_{i+}} \hat{p}_{ij}(1-\hat{p}_{ij})}}{n_{i+}}$. The denominator term can be re-written as follows:

$$\begin{aligned} \hat{\text{SE}}(\hat{D}_i) &= \frac{\sqrt{\sum_{j=1}^{n_{i+}} \hat{p}_{ij}(1-\hat{p}_{ij})}}{n_{i+}} \\ &= \sqrt{\left(\frac{1}{n_{i+}}\right)^2 \sum_{j=1}^{n_{i+}} \hat{p}_{ij}(1-\hat{p}_{ij})} \\ &= \sqrt{\left(\frac{1}{n_{i+}}\right)^2 \text{Var}(O_i^*)} \\ &= \sqrt{\text{Var}\left(\frac{O_i^*}{n_{i+}}\right)} \end{aligned} \tag{7.5}$$

but $O_i^* = n_{i+}O_i$, so

$$\hat{\text{SE}}(\hat{D}_i) \quad = \sqrt{\text{Var}\,(O_i)} \quad = \text{SE}(O_i) \tag{7.6}$$

This formulation is restrictive as it assumes no variance in the expected rates, i.e., it assumes that the model and its parameters are correct, and it does not allow for covariance between $\hat{O}$ and $\hat{E}$.

Small World results

University	n	$\hat{R}_i = \frac{\hat{O}_i}{\hat{E}_i}$	$\hat{\text{SE}}(\hat{R}_i)$	$\hat{z}_i^r$
1	6831	0.98	0.0036	−5.56
14	2917	1.06	0.0049	12.22
33	55	0.89	0.0450	−2.46
42	901	0.94	0.0101	−6.08
118	1344	1.01	0.0073	2.03

Table 7.1: Ratio method results for the Small World.

University	$\hat{D}_i$	$\hat{\text{SE}}$(Original)	$\hat{z}$(Original)	$\hat{\text{SE}}$(Eq 7.4)	$\hat{z}_i^r$
1	−0.018	0.0020	−9.05	0.0033	−5.56
14	0.056	0.0035	15.84	0.0046	12.22
33	−0.100	0.0456	−2.18	0.0405	−2.46
42	−0.056	0.0098	−5.75	0.0093	−6.08
118	0.014	0.0063	2.19	0.0068	2.03

Table 7.2: Method comparison using ratio and difference approaches: Small World.

NB In all tables in this section, the original method is non-model-based with $\gamma^{0.5}$ variance estimation.

I can examine how the ratio approach results vary against the difference results in a number of situations. I begin with the Small World, which has four PCF categories and five universities. Testing the ratio method in a practical way will help to contrast how the ratio and difference measures describe the institutional quality assessments. Table 7.1 displays how the results would look if we used the ratio approach described in the previous subsection. Table 7.2 then compares the derived SE from our difference approach and from Equation 7.4.

The results show that the ratio method does produce different z-scores to the original method and provides smaller values with the more extremal z-scores. As we have already seem, the principal difference is down to the fact that the ratio SEs are only based on the observed progression rates and take no account of variability in the expected rates. The order of the quality assessments is not preserved over the two methods, with Institution 14 doing the best in both approaches but Institution 42 or Institution 1 are ranked lowest depending on which method is chosen. If Bonferroni cut-offs are used, both methods agree on all the universities' status.

In terms of appearance, it is a question of taste and what type of message needs to be put across. For example, Institution 1 is under performing by around 1.8 percentage points using difference methodology. When the ratio approach is used, Institution 1 is only achieving 98% of its perceived benchmark. The overall message is the same, Institution 1 is under performing, but the two approaches give a difference flavour to the performance indicator.

Medium World results

The analysis on the Small World is repeated in the Medium World, with ten universities and 36 PCF categories. Tables 7.3 and 7.4 compare the two quality assessment methods. The order of the z-score is preserved over both methods ranging from Institution 1 marked as the best down to Institution 7 as the worst. As in the Small World, the biggest differences in scores are at the extremities with Institution 1's score changing from 9.50 (original) to 7.51 (ratio related). The status of the universities remains constant across both approaches (Bonferroni based).

University	n	$\hat{R}_i = \frac{\hat{O}_i}{\hat{E}_i}$	$\hat{\text{SE}}(\hat{R}_i)$	$\hat{z}_i^r$
1	6831	1.03	0.0046	7.51
2	3314	1.03	0.0059	4.98
3	1113	0.97	0.0109	−3.18
4	2205	0.97	0.0076	−4.12
5	289	1.08	0.0226	3.46
6	3238	0.99	0.0067	−2.17
7	2889	0.97	0.0069	−4.68
8	2292	0.97	0.0081	−3.34
9	1031	0.95	0.0114	−4.56
10	116	1.05	0.0314	1.57

Table 7.3: Ratio method results for the Medium World.

University	$\hat{D}_i$	$\hat{\text{SE}}$(Original)	$\hat{z}$(Original)	$\hat{\text{SE}}$(Eq 7.4)	$\hat{z}_i^r$
1	0.030	0.0032	9.50	0.0040	7.51
2	0.026	0.0051	5.17	0.0053	4.98
3	-0.031	0.0095	-3.23	0.0096	-3.18
4	-0.028	0.0066	-4.19	0.0067	-4.12
5	0.068	0.0191	3.57	0.0197	3.46
6	-0.013	0.0053	-2.37	0.0058	-2.17
7	-0.028	0.0057	-4.99	0.0061	-4.68
8	-0.023	0.0065	-3.64	0.0070	-3.34
9	-0.046	0.0097	-4.60	0.0100	-4.56
10	0.044	0.0301	1.46	0.0282	1.57

Table 7.4: Method comparison using ratio and difference approaches: Medium World.

Big World results

The most interest regarding the variation in the methods occurs when all 165 universities are compared. I examine the Big World with all eight PCFs adjusted for. Given there are 165

institutions, it is not practical to provide tables similar to those given in Small and Medium worlds. Instead I can compare the association between the two sets of z-scores graphically. This correlation is given in Figure 7-1. Table 7.5 shows how the status of the universities vary depending on which method is used.

		Difference Status			
		Good	OK	Bad	Total
Ratio Status	Good	9	2	0	11
	OK	3	134	1	138
	Bad	0	1	15	16
Total		12	137	16	165

Table 7.5: Status changes between ratio and difference methods: Big World.

NB Status based on a Bonferroni cut-off of 3.61.

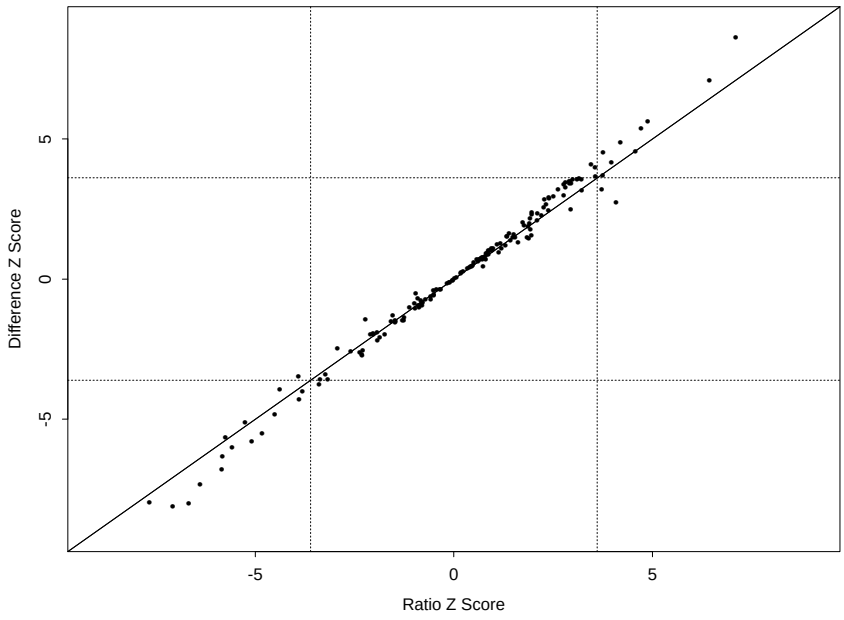

Figure 7-1: Correlation between ratio and difference z-scores: Big World.

Figure 7-1 shows that the scores are very highly associated with a correlation of 0.994. There appears to some shrinkage of the ratio z-scores back towards zero, compared against the difference scores. On the whole, the two sets of scores match very closely and by looking at Table 7.5 we can see the methods only disagree on seven universities. The implications of these results are that: if statistically valid results are necessary, then the difference approach should

be used as the SE it is based upon has a sounder statistical background. If, however, a ratio style presentation is preferred then results will not be dramatically different from the "valid" difference results, with a small misclassification rate.

7.3 Studentized quality assessment

Motivation for a studentized approach

The studentized method takes into account that some universities can look better (or worse) than they actually are because the universities take very unusual students. In the basic derivation, a university that has 90% of a certain type of student in the whole country (i.e., in a specific PCF category) is going to have an expected value close to their observed value because the university contributes so much to their own expected value. In the studentized derivation, this expected value is only based on students of a similar type from other universities. The problem with the studentized method is that the expected value is now based on fewer observations.

To illustrate the difference between the basic and studentized derivations, the following simple example is provided: Consider three universities, two PCF categories and a grid set-up as follows (this data is taken from the original HEFCE data 1996/1997) (Tables 7.6 and 7.7).

	PCF Categories		Weighted
University	Male	Female	Row Mean
27	0.914	0.943	0.932
33	0.800	0.800	0.800
145	0.802	0.870	0.843
Weighted Column Mean	0.804	0.871	0.844

Table 7.6: Studentized example: proportions.

	PCF Categories		University
University	Male	Female	Sum
27	35	53	88
33	5	50	55
145	1678	2437	4115
PCF Category Sum	1718	2540	4258

Table 7.7: Studentized example: numbers.

Consider this as the whole population of universities. If we use the basic derivation to calculate the implied quality difference (D_3) for Institution 145, the following formula's apply:

$$\hat{D}_3 = \hat{O}_3 - \hat{E}_3 \quad = \quad 0.843 - E_3$$

$$
\begin{aligned}
&= 0.843 - [1678(0.804) + 2437(0.871)] \\
&= 0.843 - 0.844 \\
&= -0.01
\end{aligned}
$$

In the studentized case, the overall PCF category rates (0.804 and 0.871) are replaced by PCF category rates ignoring those students in Institution 145, so the $\hat{D}_i^s$ value (the studentized $\hat{D}_i$ value) is now calculated as follows:

$$
\begin{aligned}
D_3^s = O_3 - E_3^s &= 0.843 - E_3^s \\
&= 0.843 - [1678(0.900) + 2437(0.871)] \\
&= 0.843 - 0.884 \\
&= -0.41
\end{aligned}
$$

The difference has changed from -0.01 in the basic case to -0.41 in the studentized case. The estimate for $\hat{D}_3^s$ is based on less observations than $\hat{D}_3$ and this means that the SE of the difference will change and in turn, the z-score.

The differences can be seen with relative ease in this simplified case study, but with many more universities and PCF categories, comparisons become much more difficult. The main interest is to see how this studentized approach performs in correctly identifying good and poor universities. As in the basic derivation case, this means creating an artificial world with no quality differences between universities and seeing how well the method is calibrated against the 2.5% benchmarks.

Derivation of the studentized approach

As I did for the basic indirect approach (Section 2.3), a formulaic derivation is found for the studentized method: For university i, define $\hat{D}_i^s$ as the studentized $\hat{D}$ value and $\hat{E}_i^s$ as the studentized expected value. Also define $\hat{p}_{ij}^s$ as the overall progression rate for PCF category j, ignoring university i. We know that:

$$
\hat{D}_i^s = \hat{O}_i - \hat{E}_i^s = \frac{1}{n_{i+}} \sum_{j=1}^{M} n_{ij} \left(\hat{p}_{ij} - \hat{p}_{ij}^s \right) \tag{7.7}
$$

but now

$$
\hat{p}_{ij^s} = \frac{1}{n_{+j} - n_{ij}} \left\{ \left(\sum_{k=1}^{N} n_{kj} \hat{p}_{kj} \right) - n_{ij} \hat{p}_{ij} \right\} \tag{7.8}
$$

which implies that

$$\begin{aligned}
\hat{D}_i^s &= \frac{1}{n_{i+}} \sum_{j=1}^{M} n_{ij} \left\{ \hat{p}_{ij} - \frac{1}{n_{+j} - n_{ij}} \left[\left(\sum_{k=1}^{N} n_{kj} \hat{p}_{kj} \right) - n_{ij} \hat{p}_{ij} \right] \right\} \\
&= \frac{1}{n_{i+}} \sum_{j=1}^{M} \left\{ n_{ij} \hat{p}_{ij} - \frac{n_{ij}}{n_{+j} - n_{ij}} \sum_{k=1}^{N} n_{kj} \hat{p}_{kj} + \frac{n_{ij}^2}{n_{+j} - n_{ij}} \hat{p}_{ij} \right\} \\
&= \frac{1}{n_{i+}} \sum_{j=1}^{M} \left\{ \left(n_{ij} + \frac{n_{ij}^2}{n_{+j} - n_{ij}} \right) \hat{p}_{ij} - \sum_{k=1}^{N} \frac{n_{ij} n_{kj}}{n_{+j} - n_{ij}} \hat{p}_{kj} \right\} \qquad (7.9)
\end{aligned}$$

Now consider the double summation over k and j as we did in the basic derivation.
If $i = k$:

$$\begin{aligned}
k \text{ Summation Term} &= \left\{ n_{ij} + \frac{n_{ij}^2}{n_{+j} - n_{ij}} - \frac{n_{ij} n_{ij}}{n_{+j} - n_{ij}} \right\} \hat{p}_{kj} \\
&= \frac{n_{ij} \hat{p}_{kj}}{n_{i+}} \qquad (7.10)
\end{aligned}$$

If $i \neq k$:

$$k \text{ Summation Term} = -\frac{n_{ij} n_{kj}}{n_{i+}(n_{+j} - n_{ij})} \hat{p}_{kj} \qquad (7.11)$$

As before, the $\hat{D}_i^s$s can be written as a weighted sum:

$$\hat{D}_i^s = \sum_{j=1}^{M} \sum_{k=1}^{N} \lambda_{ikj}^s \hat{p}_{kj}$$

where

$$\lambda_{ikj}^s = \left\{ \begin{array}{cc} \frac{n_{ij}}{n_{i+}} & \text{if } i = k \\ -\frac{n_{ij} n_{kj}}{n_{i+}(n_{+j} - n_{ij})} & i \neq k \end{array} \right\} \qquad (7.12)$$

The problem with this methodology is that one university might have all of the students that fall into a specific PCF category. In this case, the formulas would fail as the overall progress rate for that PCF category, ignoring the university in question, would be undefined as there are no other students except for those in the university in question.

One way to partially solve this problem is to assume that the difference between the university rate for that category is the same as the overall rate for the category ignoring the specific university students, i.e for the categories where $n_{ij} = n_{+j}$ assume $\hat{p}_{ij} = \hat{p}_{\cdot j}^s$. This means that

λ^s_{ikj} has a modified form:

$$\lambda^s_{ikj} = \left\{ \begin{array}{cc} 0 & \text{if } n_{ij} = n_{+j} \\ \left\{ \begin{array}{cc} \frac{n_{ij}}{n_{i+}} & \text{if } i = k \\ -\frac{n_{ij}\, n_{kj}}{n_{i+}(n_{+j} - n_{ij})} & i \neq k \end{array} \right\} & \text{if } n_{ij} \neq n_{+j} \end{array} \right\} \tag{7.13}$$

Performance in the Small World

Variance Estimate	Tail	Basic %	Student %
Global	Low	2.8	2.8
	High	2.4	2.4
	Overall	5.2	5.2
Local	Low	1.8	1.8
	High	4.1	3.9
	Overall	5.9	5.7
γ 0.5	Low	2.0	2.0
	High	3.3	3.2
	Overall	5.4	5.2

Table 7.8: Studentized tail behaviour: Small World.

The tail behaviour in a null simulations is examined for the studentized approach. These results are compared to the tail behaviour for the "original" approach, i.e., the results given in Section 4.5. Three variance estimation techniques were compared: local (described in Section 2.5); global and shrinkage (Section 4.4). As before, the target percentages are 2.5%, 2.5% and 5.0% (Low, High and Overall). The first world examined is the Small World, based on 2,000 simulations and the results are given in Table 7.8.

The studentized method reacts in a very similar way to the original method, with little or no difference in the null simulation tail behaviour. The two sets of global results are identical, there is only a small 0.2% difference in the local technique results and equivalent differences are noted in the favoured γ approach. With this set-up, there appears to be no real difference between the two methods.

Performance in other selected worlds

Further comparisons between the two methods were made in the larger Published and Big Worlds. Due to the increased size of the datasets involved, Tables 7.9 and 7.10 are based on 500 simulations. As in the Small World, the studentized method has very similar tail behaviour to the original basic method in these two larger worlds. In only one set of comparisons do the results vary by more than 0.1%. This occurs in the Big World when local variance estimation techniques are used.

In conclusion, there appears to be little difference between studentized and unstudentized

methods in terms of tail properties, both techniques produce reasonable results when the shrinkage approach is used.

Variance Estimate	Tail	Basic %	Student %
Global	Low	2.6	2.5
	High	2.1	2.1
	Overall	4.7	4.6
Local	Low	2.5	2.4
	High	4.4	4.4
	Overall	6.9	6.8
γ 0.5	Low	2.2	2.1
	High	2.6	2.6
	Overall	4.8	4.7

Table 7.9: Studentized tail behaviour: Published World.

Variance Estimate	Tail	Basic %	Student %
Global	Low	2.2	2.1
	High	1.7	1.8
	Overall	3.9	3.9
Local	Low	8.2	8.8
	High	10.9	11.3
	Overall	19.1	20.1
γ 0.5	Low	2.5	2.4
	High	2.7	2.8
	Overall	5.2	5.2

Table 7.10: Studentized tail behaviour: Big World.

Table 7.11 and Figure 7-2 compare the basic and studentized z-scores produced from a non-model-based approach with $\gamma^{0.5}$ variance estimation. Out of the 165 universities only 7 change status based on a Bonferroni cut-off and all 7 move from being classed as ok in the basic approach to good in the studentized approach. This sounds like quite a large shift of numbers but Figure 7-2 shows that these 7 institutions are very close to the ± 3.61 cut-off in both methods and the movement in their z-scores is only very minor.

		Studentized			
		Good	OK	Bad	Total
Basic	Good	12	0	0	12
	OK	7	130	0	137
	Bad	0	0	16	16
Total		19	130	16	

Table 7.11: Status changes between the basic and studentized approaches.

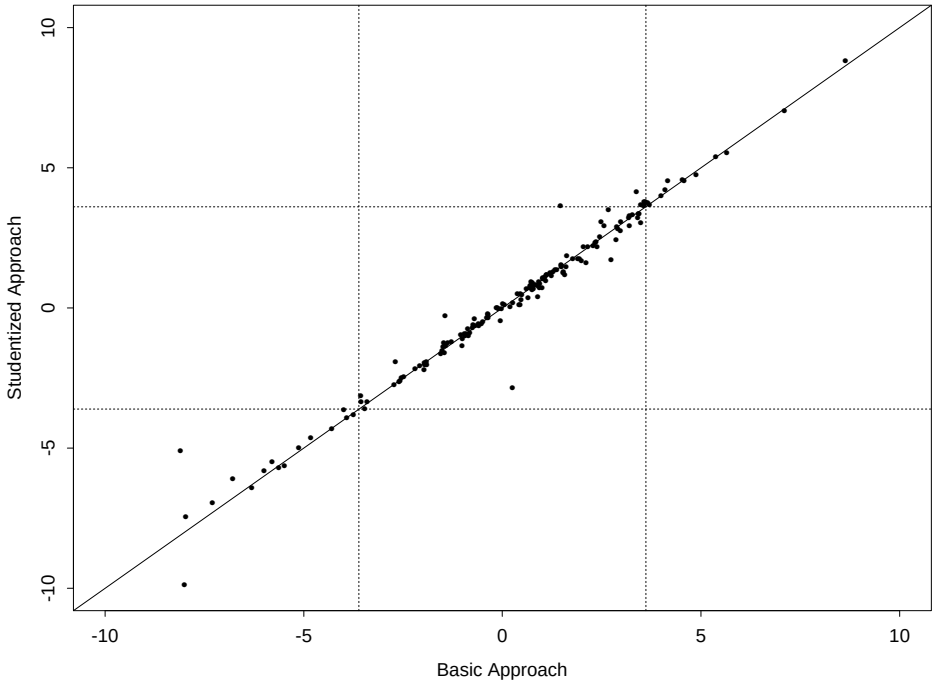

Figure 7-2: Comparison of z-scores from the basic and studentized approaches: Big World.

7.4 Fuzzy direct standardisation

The fuzzy direct idea

In Section 2.3, direct standardisation was introduced. It asks an alternative question to indirect: "What would the observed overall progression rate have been at this university if its progression rates in the PCF categories had been what they were, but its distribution of students across PCF categories had instead matched the national distribution"?

This means replacing a university's distribution of students across PCF categories with the national distribution of students, when calculating the expected progression rate for a university. A university progression rate is now required for each student type (defined by their PCF category) present in the national distribution. At a large number of universities, university specific information on each PCF type is not available as many institutions don't have certain student types that are present in the national cohort. For example, in the Small World Institution 33 does not have any young female or male students, meaning that no Institution 33 specific progression rates can be provided for those type of students. Direct standardisation fails if this scenario occurs.

I can, however, create a modified direct standardisation approach which can be examined. When no PCF category information is available at a university, i.e., there are no students of that

type there, estimate the progression rate of those students with the national PCF progression rate for students of that type. This provides us with a fuzzy direct standardisation approach. The following sections describe the derivation and the "null simulation" tail behaviour of the method.

Derivation

The method for creating fuzzy direct quality assessment results is as follows:

Define $\hat{D}_i^f$ as the fuzzy direct standardisation D value for university i and E_i^f as the direct standardised expected value.

$$\hat{D}_i^f = O_i - E_i^f = \frac{1}{n_{i+}} \sum_{j=1}^{M} \{n_{ij}\hat{p}_{ij}\} - E_i^f \tag{7.14}$$

In fuzzy standardisation,

$$E_i^f = \frac{1}{n_{i+}^f} \sum_{j=1}^{M} n_{+j}\hat{p}_{ij} I_{n_{ij}\neq 0}$$

where

$$n_{i+}^f = n_{++} - \sum_{j=1}^{M} n_{+j} I_{n_{ij}\neq 0} \tag{7.15}$$

This implies that:

$$\begin{aligned} D_i^f &= \frac{1}{n_{i+}} \sum_{j=1}^{M} \{n_{ij}\hat{p}_{ij}\} - \frac{1}{n_{i+}^f} \sum_{j=1}^{M} n_{+j}\hat{p}_{ij} I_{n_{ij}\neq 0} \\ &= \sum_{j=1}^{M} \frac{1}{n_{i+}} \{n_{ij}\hat{p}_{ij}\} - \sum_{j=1}^{M} \frac{1}{n_{i+}^f} n_{+j}\hat{p}_{ij} I_{n_{ij}\neq 0} \\ &= \sum_{j=1}^{M} \left\{ \frac{n_{ij}\hat{p}_{ij}}{n_{i+}} - \frac{1}{n_{i+}^f} n_{+j}\hat{p}_{ij} I_{n_{ij}\neq 0} \right\} \\ &= \sum_{j=1}^{M} \left\{ \frac{n_{ij}}{n_{i+}} - \frac{n_{+j}}{n_{i+}^f} I_{n_{ij}\neq 0} \right\} \hat{p}_{ij} \end{aligned} \tag{7.16}$$

So now D_i^f can be written as:

$$\hat{D}_i^f = \sum_{j=1}^{M} \lambda_{ij}^f \hat{p}_{kj}^f$$

where

$$\lambda_{ij}^{f} = \left\{ \begin{array}{cc} \frac{n_{ij}}{n_{i+}} & \text{if } n_{ij} = 0 \\ \frac{n_{ij}}{n_{i+}} - \frac{n_{+j}}{n_{i+}^{f}} & n_{ij} \neq 0 \end{array} \right\} \tag{7.17}$$

but this simplifies to:

$$\lambda_{ij}^{f} = \left\{ \begin{array}{cc} \frac{n_{ij}}{n_{i+}} - \frac{n_{+j}}{n_{i+}^{f}} & \text{if } n_{ij} \neq 0 \\ 0 & \text{otherwise} \end{array} \right\}$$

where

$$n_{i+}^{f} = n_{++} - \sum_{j=1}^{M} n_{+j} I_{n_{ij} \neq 0} \tag{7.18}$$

Performance of the fuzzy direct method

The fuzzy direct method was examined in the Small World (based on 2,000 simulations) and in the Published World (500 simulations). Given the very poor nature of the results, there was no need to run any further analysis in other worlds.

For both worlds, the fuzzy direct results are very poor. The method massively overestimates the number of unusual universities, identifying far too many good and poor institutions. The overall rate of unusualness exceeds 60% in some cases, compared against a target of 5% (Table 7.12). Even when using the favoured γ variance estimation approach, the results are still particularly poor. For example in the Published World, 17.4% of universities are found to be poor when only 2.5% should be (Table 7.13). The same pattern is seen with the identification of good universities, where the method finds a rate of 16.3% compared against the required 2.5%. The picture is clear, the fuzzy direct method does not work because of its inability to sensibly predict the behaviour of the "missing PCF student types".

Variance Estimate	Tail	Basic %	F. Direct %
Global	Low	2.7	33.9
	High	2.1	29.0
	Overall	4.8	62.9
Local	Low	1.8	40.2
	High	3.8	23.4
	Overall	5.6	63.6
γ 0.5	Low	2.0	33.8
	High	3.1	25.2
	Overall	5.1	59.0

Table 7.12: Fuzzy direct tail behaviour: Small World.

Variance Estimate	Tail	Basic %	F. Direct %
Global	Low	2.5	14.1
	High	2.1	15.6
	Overall	4.6	29.8
Local	Low	2.3	30.1
	High	4.5	27.3
	Overall	6.9	58.0
γ 0.5	Low	2.1	17.4
	High	2.6	16.3
	Overall	4.7	33.7

Table 7.13: Fuzzy direct tail behaviour: Published World.

Chapter 8

Model-based alternatives

8.1 Random-effects models

Motivation

In 1999 HEFCE were required to develop performance indicators for employment rates of graduates from each HE institution (HEFCE (2001)). The data for the indicators came from the First Destination Return (FDR) survey for students graduating in the 1999-2000 academic year. The FDR collects information about the activities of all students graduating from full-time HE courses. Two indicators were created as there is a difficulty in how to deal with students going into further study. The first indicator deemed students as having a successful outcome if they were found to be in work or further study six months after graduation. The second indicator removed students who went on to further studies from the analysis. The modelling used eight student level PCFs: age on entry; entry qualifications; subject of study; gender; ethnic group; social background; degree classification; and whether or not the student was on a sandwich course. The key difference between the student progression work and this employment study is that the employment model also included three institutional level adjustors: two based on employment in the locality, and the average A-level (or Scottish Highers) score at student entry. These institutional level predictors mean that a non-model based approach was impossible and a FE technique also fails. The results of this employment study are given later in this chapter. Smith et al. (2000) use an econometric view to develop ideas on employment performance indicators using data from the first destinations of graduates from pre-1992 UK HE institutions. They use a pictorial approach for identifying unusual institutions (in terms of employment outcomes) with a particular focus on unadjusted and adjusted ranks of each university. Random-effects (RE) models are common-place in quality assessment approaches including Normand et al. (1997), who use a three-level RE logistic model to examine the variation in US medical care providers treating a cohort of elderly heart attack patients.

The fixed-effects (FE) models described in Section 3.5 are constrained because of the need to formally include institutional identifiers in the modelling process. This feature means that no

institutional level predictors can be adjusted for in the quality assessment process when using FE methods. If higher level predictors were included in the FE regression models, then the design matrix (X) for the model would not have linearly independent columns, making X impossible to invert and thus, the associated regression methods invalid (i.e., $\beta = (X^T X)^{-1} X^T Y$ cannot be formed).

In a large number of cases, institutional predictors are required (as in the employment analysis) and this makes a FE method impossible, so another quality assessment approach is required. The non-model-based method explored in Section 2.3 cannot deal with institutional level predictors, whereas a RE set-up can. In a RE model, institutional dummies are not required to formally appear in the modelling process but they are included as the institutional level residuals. A key point to note is that any institutional level predictors not formally included in the RE modelling form part of the institutional level residuals. So it is very important that all appropriate institutional level predictors are included in the fixed part of the RE model.

RE models can produce statistically valid SEs for each institutional level residual and thus a valid RE assessment z-score can be produced when institutional level predictors are adjusted for. The forthcoming sections describe the model structure of a REs model and then go on to examine how the results produced by RE models differ from a non-model-based method and a FE approach in a number of different set-ups.

The random-effects model

A random-effects model version of the linear fixed-effects model is:

$$\begin{aligned} y_{ij} &= \beta_0 + \sum_{k=2}^{M} \beta_k x_{ijk} + a_i + e_{ij}, \\ e_{ij} &\overset{\text{IID}}{\sim} N\left(0, \sigma_e^2\right), \\ a_i &\sim N\left(0, \sigma_a^2\right) \forall i \in (1, \ldots, N), \end{aligned} \tag{8.1}$$

where: y_{ij} is the outcome; x_{ijk} is the PCF "carrier" k for student j in university i; n_{i+} are the number of students in university i; σ_a^2 and σ_e^2 are the variance components for the university and overall effects; and the a_is are the random-effects model equivalent to the fixed-effects α_is, i.e., the "perceived" quality of university i.

This random-effects model can be examined using `MLwiN` (Rasbash et al. (2000)). `MLwiN` can produce statistically valid SEs for the university level residuals(a_i) and so an assessment z-score can be formed from the residual (a_i) and its associated standard error(SE(a_i)). This z-score is the random-effects equivalent of the HEFCE $\hat{D}_i$ z-score and FE $\hat{\alpha}_i$ z-score. More information on Bayesian multilevel modelling can be found from many sources including Draper (ming) and Goldstein (1995).

Small World results

Initially I concentrate on comparing fully saturated FE and RE models. The same interaction effects can be seen in a RE approach that were shown in the FE methodology (described in Section 6.5). I began by comparing the two sets of results in the Small World where there are few universities in a relatively large dataset (over 10,000 observations). Table 8.1 shows how the parameter (α or a) estimates, along with the associated z-scores, are affected by the approach used. The second table (8.2) examines how the status of each university changes depending on the approach used and a Bonferroni z-score cut-off ($|Z| \geq 2.56$).

		Fixed-Effect			Random-Effects (RIGLS)		
University (i)	n_i	$\hat{\alpha}_i$	$\widehat{\text{SE}}(\hat{\alpha}_i)$	$\hat{z}$-score	$\hat{a}_i$	$\widehat{\text{SE}}(\hat{a}_i)$	$\hat{z}$-score
1	6831	-0.0210	0.00215	-9.76	-0.0023	0.0261	-0.09
14	2917	0.0616	0.00435	14.17	0.0795	0.0263	3.02
33	55	-0.121	0.0356	-3.38	-0.0723	0.0354	-2.05
42	901	-0.0619	0.00844	-7.33	-0.0421	0.0268	-1.57
118	1344	0.0194	0.00683	2.83	0.0372	0.0266	1.40

Table 8.1: Random-effects results: Small World.

		Random			
		Bad	OK	Good	Total
Fixed	Bad	0	3	0	3
	OK	0	0	0	0
	Good	0	1	1	2
Total		0	4	1	5

Table 8.2: Random-effects status changes: Small World.

The first issue with REs is that, because of the structure of the approach, shrinkage is expected in the parameter estimates compared with the FE estimates. Does this mean that the same shrinkage is seen in the z-scores?

We would expect that some strength (or information) is borrowed from the other institutional scores so a degree of shrinkage in the z-scores should be noted. This will affect how the z-score cut-off is chosen. In some circles the Bonferroni adjustment is seen as an "alternative" to using REs, i.e., you can either increase the absolute value of the z-score cut-off by using Bonferroni, or decrease the size of the z-scores by using a RE approach. The two cut-off tactics are not directly comparable, i.e., if you use a RE approach this doesn't mean that 1.96 should be selected as the z-score cut-off point. Little or no work has been completed in the literature on the effects of using FE vs RE, but as we will see in the larger worlds, a slightly reduced z-score cut-off produces equivalent RE results to the FE approach.

In this Small World, every RE z-score has been shrunk back towards zero compared against the FE results. It is difficult to judge the exact effects of a RE model with so few universities but the key message is that the RE z-scores are a shrunken version of the FE scores. This result is reiterated in the status changes as only one university is found to be unusual in a RE set-up, using the slightly too large Bonferroni cut-off.

Medium World results

In the slightly larger Medium World, the shrinkage message is a little clearer. Table 8.3 shows the parameter estimates for the FE and RE model-based approaches. Table 8.4 repeats the status change analysis completed in the Small World, but now involves ten universities.

		Fixed-Effects			Random-Effects (RIGLS)		
University (i)	n_i	$\hat{\alpha}_i$	$\hat{\text{SE}}(\hat{\alpha}_i)$	z-Score	$\hat{a}_i$	$\hat{\text{SE}}(\hat{a}_i)$	z-Score
1	6831	0.0325	0.0034	9.44	0.0349	0.0124	2.82
2	3314	0.0253	0.0054	4.67	0.0277	0.0129	2.14
3	1113	-0.0314	0.0096	-3.28	-0.0265	0.0145	-1.82
4	2205	-0.0293	0.0066	-4.41	-0.0254	0.0133	-1.90
5	289	0.0712	0.0191	3.73	0.0575	0.0192	2.99
6	3238	-0.0131	0.0053	-2.47	-0.0099	0.0129	-0.77
7	2889	-0.0290	0.0057	-5.09	-0.0254	0.0130	-1.95
8	2292	-0.0250	0.0065	-3.87	-0.0213	0.0133	-1.60
9	1031	-0.0471	0.0099	-4.75	-0.0408	0.0147	-2.78
10	116	0.0475	0.0303	1.57	0.0292	0.0242	1.21

Table 8.3: Random-effects results: Medium World.

		Random			
		Bad	OK	Good	Total
Fixed	Bad	1	4	0	5
	OK	0	2	0	2
	Good	0	1	2	3
Total		1	7	2	10

Table 8.4: Random-effects status changes: Medium World.

In all cases the z-scores are drawn closer to zero using a RE approach, whilst eight of ten of the RE parameter estimates are smaller in absolute value compared with their FE counterparts.

Using the standard Bonferroni cut-off for ten institutions (2.81), the number of unusual institutions moves from eight in FEs to three in the REs set-up. Moving the RE cut-off to around 1.60 would produce equivalent results to the FE approach. The choice of z-score cut-off is critical and is dependent on which model-based approach is used and/or the number of institutions involved.

Big World results

Initially, I focus on main effects models only. In Section 6.5, we have seen the relationship between fitting fully-saturated fixed-effects models and models containing a variety of interaction terms. In this section, I study the differences between fitting a main-effect only model in the FE and RE worlds. The main effects only model in the RE world can be fitted with relative ease within `MLwiN`. Fitting RE models with further interaction terms is limited in the same way as FE models: models with a large number of terms run very slowly and can be time-consuming to set up. The main difference between the FE and RE parameter constraints is that there is

no "easy" way to fit a fully-saturated RE model, compared against the FE trick of using GLIM's eliminate command.

Figure 8-1 shows the parameter estimates for each university in both worlds. This produces a plot with 165 points, one for each university. The 45 degree line can be used to compare the two sets of results. Shrinkage can be clearly seen in the plot with the random-effects parameters significantly smaller at the two extremes of the data. The positive side of the graph shows a large deviation from the 45 degree line. A similar pattern can be seen in the comparison of the associated university z-scores (Figure 8-2).

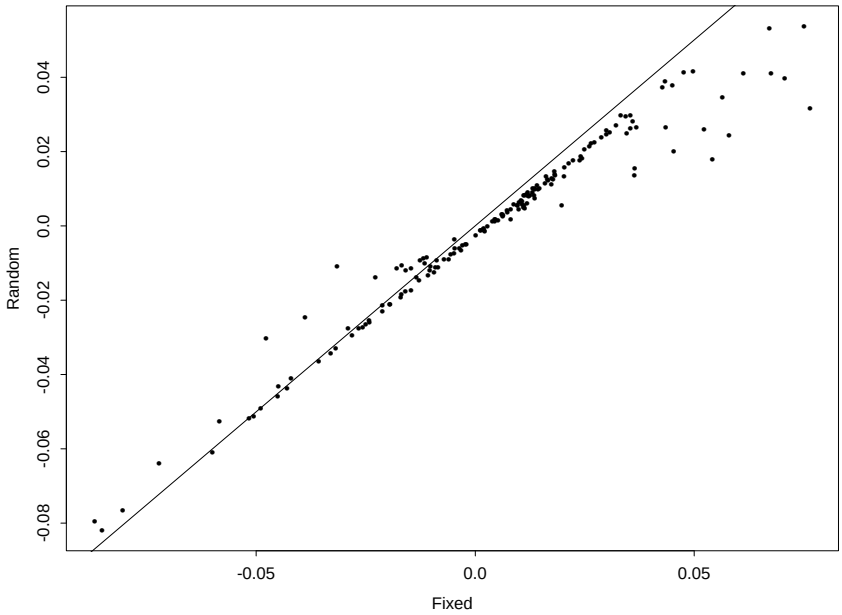

Figure 8-1: Parameter comparison, fixed vs random: Big World.

If the RE results were used rather than the FE as a method for establishing university quality, the results would differ as follows (Table 8.5):

		Random			
		Bad	OK	Good	Total
Fixed	Bad	23	0	0	23
	OK	1	121	0	122
	Good	0	8	12	20
Total		24	129	12	165

Table 8.5: Random-effects status changes: Big World.

The main difference in the results is that eight good universities in the FE world are now classified as OK. This is caused by the shrinkage effect noted in the smaller worlds and therefore a more sensible cut-off point (rather than Bonferroni: 3.61) may be called for in a RE formulation. In the Medium World, we saw that moving the cut-off from 2.8 to 1.6 produced equivalent

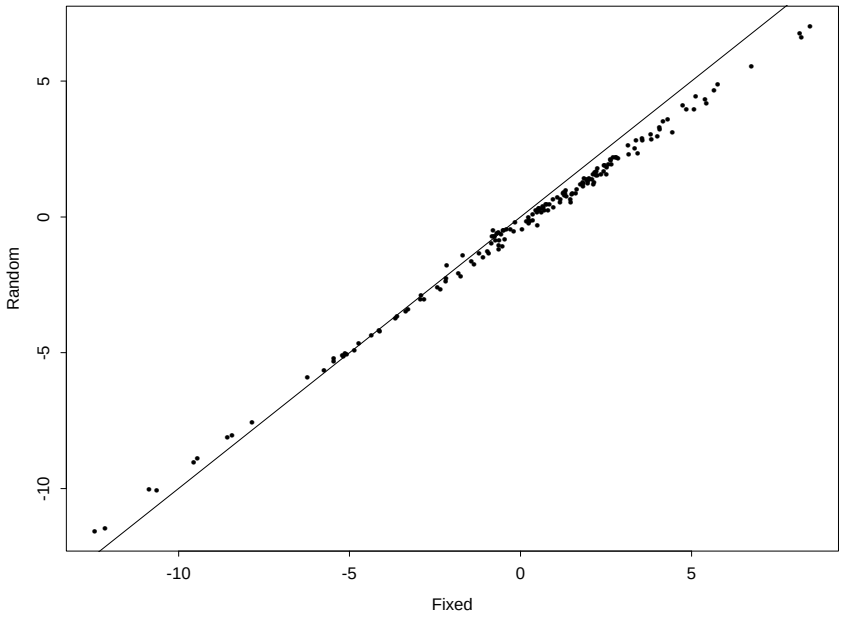

Figure 8-2: z-score comparison, fixed vs random: Big World.

institutional status' in a RE approach. With more universities (i.e., 165 rather than 10), the z-score cut-off does not need to be reduced as dramatically to produce equivalent RE results to the FE version. A problem occurs with the asymmetry of the results, reducing the RE cut-off from 3.61 to say 3.50 improves the status agreement between the two methods (Table 8.6):

		Random			
		Bad	OK	Good	Total
Fixed	Bad	23	0	0	23
	OK	1	121	0	122
	Good	0	6	14	20
Total		24	127	14	165

Table 8.6: RE status changes (3.50 RE cut-off): Big World.

Lowering the RE cut-off even further (to say 3.00) starts to produce misclassification in the "poor" tail (Table 8.7):

So the choice of z-score cut-off for RE isn't clear and a balance is required to ensure that appropriate misclassification rates are achieved in the tails. Chapter 10 examines how the choice of model affects quality assessment results compared against a gold standard, i.e., directly measured process information. This type of gold standard analysis helps us understand how z-score cut-off affects institutional quality assessment.

		Random			
		Bad	OK	Good	Total
Fixed	Bad	23	0	0	23
	OK	3	119	0	122
	Good	0	4	16	20
Total		26	123	16	165

Table 8.7: RE status changes (3.00 RE cut-off): Big World.

The employment results

I have developed a modelling approach that allows institutional level predictors and so the employment case-study mentioned earlier can now be completed. The data is based on students who obtained their first degrees in 1999-2000 and the base population for the first indicator (i.e., including those who went on into further study) was 156,821 individuals. This number is reduced to 123,499 when students in further study were excluded for the second indicator.

The published tables have the following form (Table 8.8):

	Employment Indicator 1			Employment Indicator 2		
Institution	Population	Indicator	Benchmark	Population	Indicator	Benchmark
Anglia	1,112	94	93	950	92	92
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Ulster	2,034	94	94	1,511	91	93

Table 8.8: Employment indicators: an example

The RE modelling provides us with an a_i value for each institution, which acts as the difference between the observed (indicator in the tables) and expected (benchmark) rates. The indicator rates can be found by simply studying the original dataset, and the expected rates can be inferred from the a_i value and its associated observed rate. The z-score value for each institution is recorded but not given in the tables. An institution is given a $+$ / $-$ symbol if it is found to be significantly better / worse than its benchmark (i.e., its quality assessment $|z|$-score is greater than 3.00 and the $|a_i|$ is more than 3%).

Using indicator 1 with 156 institutions in the dataset, 13 were found to be under performing and 8 were doing significantly better than expected. When those individuals who went on into further study are dropped (indicator 2), 15 institutions are now found to be performing badly, with 7 doing better than expectations.

A full version of the employment study results can be found in HEFCE (2001).

The random-effects approach: conclusions

The main advantage RE holds over FE modelling approaches is its ability to deal with higher level PCFs, i.e., institutional-level predictors such as university funding or applications per place. This advantage seems to infer that a FE or non-model-based approach isn't required as they hold no apparent benefits over REs, but this isn't strictly true. Currently there is no

way to deal with very large RE models, i.e., models that involve adjusting for a large number of parameters. The ability to adjust for many PCFs (including interactions) is critical in any quality assessment analysis and if RE holds back on the number of PCFs that can be used, FE or a non-model-based approach should be considered. The non-model-based approach is also relatively simple to follow and this should always be taken into account when deciding between the three modelling options.

The second main problem with REs is the choice of z-score cut-off. This is a problem in the other approaches but Bonferroni and other similar work seems to provide a reasonable solution to the cut-off choice. In REs, the z-score cut-off choice is more difficult and more careful thought needs to be carried out by quality assessors to understand the mechanics of REs and thus, the choice of z-score.

Completing the random-effects story: calibration

As the RE results look very similar to the FE and non-model based results, there has been no cause for concern regarding the RE effects as the alternative method results have been shown to be well behaved (e.g., Section 3.1). This theory does not necessarily show that the RE method is well tuned and I need to develop a way of calibrating the RE approach as well.

As in Section 3.1, I use the FE model where the SCF appears directly as a model term and then set all these terms ($\hat{a}_i$s) to zero, i.e. assume all universities have the same quality performance. Simulated datasets can be generated and the RE approach (i.e., with `MLwiN`) can be used to analyse these generated datasets and the results can be examined to calibrate the technique. The RE model is assumed to have the same form as Equation 8.1 and each a_i has a z-score associated with it which represents university quality. Given the assumptions of Equation 8.1, if the method is well-calibrated I would expect that 5.0% a university is found to be unusual using the standard 1.96 z-score cut-off point.

It would be normal practice to test the calibration of RE in the Small World, but there is difficulty. The Small World has relatively few universities, five, and I am generating datasets whose core assumption is that universities have no differences (i.e., there is no university-level variation). This means when this data is fitted within `MLwin` (or any other appropriate RE package) there is not enough information in the data for the method to identify any institution-level variation, i.e, this term is set to zero. The conclusion of this problem is that the RE approach cannot estimate any a_i terms from the modelling as it believes that no institution-level terms are necessary. As more and more data is included in the study, i.e., more students and institutions are included, the more justification the modelling has for including institution-level terms (even if there is no institution-level variation by design).

In the Big World, calibration is possible as there is enough information to justify inclusion on university-level terms. In the non-model-based calibration (Section 3.1), a fully-saturated model was initially examined and calibrated. In a RE approach for the Big World, a fully saturated approach would mean building 18,000 dummy variables into `MLwiN` which is too time consuming on both the model building and analysing sides of the coin. It is sensible to consider

a RE main-effects only model for calibration as this model has 41 explanatory variables, 165 institutions and 284,399 students. The 41 explanatory variables are as follows: 1 constant; 1 term for student age; 1 term for student gender; 2 terms for school type; 2 terms for parental occupation; 1 term for year of study; 2 terms for low HE participation; 12 terms for subject of study; and 20 terms for student qualifications. How these variables are broken-down is described in Section 1.6.

Using a 1.96 cut-off, the following results were obtained for the Big World (based on 500 simulation runs):

- 4.2% (1.0%) of universities are identified as good;
- 6.3% (1.4%) are identified as bad; and
- 89.5% (2.0%) are identified as ok.

This means that 11.5% of institutions are found to be unusual, more than twice the expected 5.0% rate. Given that I have used a RE approach for quality-assessment, some shrinkage in the z-scores would be expected and normally less institutions than expected from a FE approach would be tagged as unusual, i.e., less $|z|$-scores achieving greater than 1.96 and thus rates of less than 2.5% seen in the tails of the institutional distribution. These results imply an opposite effect, i.e., it seems the z-scores are inflated in the RE world. Why?

The principal reason relates to what the original FE model used for setting the calibration is based on and what the set-up of the RE model tested is. In the calibration approach, a FE model is used to create a predicted probability of success for each individual in the dataset and then the simulated datasets are generated based on these probabilities. The FE model is assumed to be fully saturated as the predicted probabilities are simply the PCF characteristic cell mean success rate for the cell into which that individual falls. The RE model contains main effects only and fails to take into account some important higher-level interaction terms. The effect on the institutional quality-assessments is similar to when key PCFs are omitted from the analysis (Table 6.1), i.e., the institutional z-scores are larger than expected because unexplained variation is blamed on the institutions rather than being taken into account by the omitted PCFs. In the Big World, the original probabilities from FE are based on around 18,000 adjustors whereas the RE model being studied only contains 41. Inflation of the z-score is to be expected.

When the 1.96 z-score is increased to 2.73 the institutional results now read as follows:

- 2.1% (0.6%) of universities are identified as good;
- 2.9% (1.0%) are identified as bad; and
- 95.0% (1.3%) are identified as ok.

The calibration can be repeated using a different baseline FE model, i.e., one that is equivalent to the RE model analysed. This involves fitting a main-effects FE model with no institutional terms and around 40 PCF effects. This provides us with a predicted probability of

success for each individual based on main-effects only that can be used to calibrate the RE main-effects only model. 500 simulated datasets were generated for the analysis but only 239 of these sets contained enough information for a RE approach to justify inclusion of institutional terms. The reduced number of PCF terms means that there is reduced variation in the individual probabilities of success and thus less information in the modelling and its outcomes. This produced the 261 non-valid simulations.

The results for the 239 valid runs are as follows, based on a 1.96 z-score cut-off:

- 0.005% (0.006%) of universities are identified as good;
- 0.007% (0.006%) are identified as bad; and
- 0.013% (0.009%) are identified as unusual.

If the cut-off is dramatically reduced to 0.57 then more constructive results are obtained:

- 2.5% (3.0%) of universities are identified as good;
- 2.6% (2.9%) are bad; and
- 5.1% (5.8%) are identified as unusual.

The 261 non-valid runs are produced when the maximum likelihood estimate for the institution-level variance is estimated as exactly zero. With the other 239 runs produce estimates of the institution-level variance that range between 1.6e-9 and 1.7e-6. The mean variance estimate is 3.6e-6 with a standard deviation of 3.1e-6.

The RE z-score shrinkage effect is apparent now with the cut-off required to drop from 1.96 to 0.57 to achieve a 5% misclassification rate. This z-score reduction is larger than expected and there are a number of potential reasons for this unusually large reduction. The principal reason seems to be that the RE method is very close to not accepting that institution-level terms are required, i.e., more than half the runs are rejected for this reason, and so both the a_i and its associated SE terms are minimal causing a larger margin of error. Further work is required to improve this RE calibration approach and thus found what a good z-score cut-off point should be. More discussion of z-score cut-off choice is given in Section 10.4.

The main message from these calibration results is that when equivalent FE and RE quality-assessment approaches are used, the z-score cut-off for the RE approach should be significantly smaller than when the FE approach is used. How much smaller is an area of debate but an improved RE calibration method will cast more light on the subject. The lost of interaction terms when using RE can be adjusted for by increasing the cut-off point and a decision theory approach (Section 10.4) should be implemented to decide on an exact z-score choice.

It is obviously that this approach to random-effects calibration is not valid and an alternative must be developed to correctly calibrate the results. The 261 non-valid simulation runs give an indication of how small the institution-level variances are and with such minimal values for the 239 other runs, any results produced must be questioned.

8.2 Non-linear regression

The next step?

So far in a model-based approach, I have examined methods that involve a linear link function, i.e., it is assumed that the outcome is continuous rather than binary. It is natural to develop these models so that a binary outcome assumption can be made and this involves assuming a non-linear link function, e.g., the logit or probit. This gives rise to a new model structure:

$$\begin{aligned}(y_{ij} \,|\, p_{ij}) &\overset{\text{indep}}{\sim} \text{Bernoulli}\,(p_{ij})\,, \\ F\,(p_{ij}) &= \beta_0 + \textstyle\sum_{k=1}^{q} \beta_k x_{ijk} + \zeta_i\end{aligned}$$

where y_{ij} is whether student j in university i progressed into her second year; p_{ij} is the probability of progression for that student based upon PCF characteristics; $F()$ is the link function; x_{ijk} identifies whether student j in university i has PCF characteristic k; and ζ_i is either the FE standard α_i or the RE standard a_i. There are N universities.

Starr et al. (1986) compared non-linear regression (similar model form to above) and indirect standardisation to examine the effect of occupation on the fertility of workers. Starr considered logistic regression as an alternative to indirect standardisation because it allows for the consideration of a wider range of PCFs (e.g., continuous rather than just discrete variables). Starr found, in general, little or no difference between the two sets of results but recommended that both methods should continue to be explored and thus, a larger body of similar comparative studies is built up.

In a FE set-up, the α_i's have the standard restriction:

$$\textstyle\sum_{i=1}^{N} n_{i+}\, \alpha_i = 0.$$

When REs are used, the a_i's have the following condition placed upon them:

$$a_i \sim N\big(0, \sigma_a^2\big).$$

We can now fit both the RE and FE models in the Small World using four different link functions: one linear link (i.e., as we have done before) and three non-linear functions (i.e., logit, probit and complementary log-log). Tables 8.9 and 8.10 show the effect on institutional z-scores in the Small World using the four different link functions.

University	Linear	Logit	Probit	C. Log Log
1	−9.76	−11.59	−12.45	−12.75
14	14.17	11.86	13.55	14.95
33	−3.38	−4.01	−3.50	−3.05
42	−7.33	−9.47	−8.74	−7.94
118	2.83	0.13	0.71	1.18

Table 8.9: Link function effects using FE: Small World.

University	Linear	Logit	Probit	C. Log Log
1	−0.09	−0.20	−0.13	−0.12
14	3.02	3.07	2.74	2.72
33	−2.05	−1.86	−1.82	−1.82
42	−1.57	−1.64	−1.42	−1.42
118	1.40	1.35	1.25	1.26

Table 8.10: Link function effects using RE: Small World.

There are some variations in the z-scores but the order of the institutions is preserved when looking at each table individually. In Table 8.9 Institution 14 always has the most positive score, with Institution 1 always achieving the most negative. The same is true in Table 8.10 except that the positions of Institution 42 and Institution 33 are interchanged. The main message is that we need to look at the non-linear effects in other worlds to understand whether there are deterministic effects related to each link function.

Breaking all the rules

Unfortunately I cannot analyse a large number of set-ups because a binary assumption can cause problems with this type of modelling. To explain the mechanics behind the problem, I simplify our data set-up.

Imagine the following standard regression set-up:

$$\begin{matrix} Y \\ \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \end{matrix} \qquad \begin{matrix} X \\ \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \end{matrix}$$

where Y is the outcome matrix and X is the regression design matrix. To relate this set-up to quality assessment, consider Y as the outcome vector recording student progression. X contains the information on the students' characteristics and her place of study. In a more complex structure, X would have many more columns representing PCF characteristics and university indicator functions. For the purpose of this analysis, imagine that the second column of our design matrix represents whether a student is attending a certain university (the column could quite easily show whether a PCF characteristic was present). So we have a quality assessment set-up of five students in two universities with no PCF adjustment.

The standard FE linear regression set-up would be (Section 3.5):

$$Y = X\beta + \epsilon \tag{8.2}$$

$$\text{where} \quad \beta = \begin{pmatrix} \beta_0 \\ \alpha_1 \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

$$\text{and } \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad \forall i \in (0, 4).$$

The parameter estimates for β are found using $\hat{\beta} = (X^T X)^{-1} X^T Y$. The co-variance matrix, associated with the estimates, can then be found using $\text{Cov}(\beta) = \sigma^2 (X^T X)^{-1}$. These formulas (or any valid alternative method) produce estimates in a linear regression approach, given by Table 8.11.

	Parameter Estimate	SE of Estimate
β_0	0.33	0.27
α_1	0.67	0.43

Table 8.11: Non-linear problem: parameter estimates.

Therefore if, for example, the α_1 was a university dummy, i.e., there were only two universities in the set-up, then quality assessment methods based on fixed (Section 3.5) and random (Section 8.1) approaches could be completed.

The main concern of this approach for quality assessment is that the outcome is binary, i.e., can only take the values 0 or 1. Linear regression assumes that the outcome is continuous (i.e., not restricted to two values) and so one of the primary regression assumptions is broken. Data with a binary outcome suggests a non-linear regression method. As we have seen in the Small World analysis a number of non-linear link models are available but for ease, we examine the logit link function. The other non-linear functions produce similar results.

The logit set-up is as follows:

$$\begin{aligned} y &\sim \text{Bernoulli}(P) \\ \text{logit}(P) &= X\beta \end{aligned} \tag{8.3}$$

where P is the probability vector for individual success ($y_i = 1$).

Under normal conditions, the logit will reject this data for the following reason. The α_1 parameter cannot be estimated because the dummy X_1 predicts success perfectly, i.e., $p(y_i = 1 | x_{1i} = 1) = 1$ in all appropriate cases in the data. This means that for some individuals, the predicted probability of success is 1, which cannot be accepted by the logit function:

$$\text{Logit function} \quad = \text{Log}\left(\frac{p_i}{1-p_i}\right) \quad \to \infty \text{ as } p_i \to 1 \tag{8.4}$$

Using other non-linear link functions produce a similar results when an individuals predicted probability is zero. So when linear regression is used, a reasonable parameter estimate and

associated standard error can be produced for the α_1 parameter but the assumption of a continuous outcome is violated. In a non-linear set-up, it assumes a binary outcome, which is correct, but it cannot produce estimates for the α_1 parameter and its associated standard error.

In practical terms for quality assessment methods, this means that a non-linear methodology will fail when either of these two situations occurs:

- A university has either a 100% or 0% success rate. This means that that university's α parameter has the completely deterministic characteristic which causes non-linear failure; or
- A PCF characteristic determines a 100% or 0% success rate. The related β parameter will cause the non-linear failure due to the indicator's link with the binary outcome variable.

Potential non-linear solutions

On the whole, individuals who perform an analysis with a binary outcome do not recognise the completely deterministic issue as a problem and thus, does not suggest or provide any solutions to the problem. I have identified four potential routes to solve this non-linear problem but none seem to provide a satisfactory resolution:

- Use a Bayesian approach and attach an informative prior to the problematic β (or α) parameter;

  The major difficulty with this approach is that the model is very dependent on the prior choice for the parameter. It is nearly impossible to design a sensible and appropriate prior for this parameter. MCMC simulation runs using `BUGS` (Spiegelhalter et al. (2000)) confirm a large sensitivity towards the parameter prior.

- Modify the dataset so this does not occur;

  In general literature and statistical software, the accepted approach is to modify the original dataset so that it can deal with a non-linear link function. This means either removing individuals for which the design matrix(X) columns can determine success or failure exactly, or changing an individual's outcome from success to failure (or vice-versa) to ensure that the outcome cannot be determined exactly from a certain predictor. These modification options are not appropriate for my analysis as removing sets of individuals from a university's population (especially those with 100% or 0% success rates) would bias the analysis. Modifying information on individual success rates would also be completely unacceptable and, in most cases, does not solve the problem. For example if we have a single predictor dummy related to only one individual, changing a successful outcome to a failure would still ensure that the predictor can determine success or failure completely.

  In short, adjusting the dataset is not a option.

- Rather than using a single link function, use a model with a variable link function;

Using a non-linear link function requires the following set-up:

$$(y_{ij} \mid p_{ij}) \overset{\text{indep}}{\sim} \text{Bernoulli}\,(p_{ij})$$
$$p_{ij} = F^{-1}\left\{\beta_0 + \textstyle\sum_{k=1}^{p} \beta_k x_{ijk} + \zeta_i\right\}$$

where the ζ_is are either the standard α_is in a FEs approach or the standard a_is in a REs set-up.

In a normal non-linear link set-up, $F^{-1}(x)$ would take a logit or probit structure:

$$F^{-1}(x) = \frac{e^x}{1+e^x} \tag{8.5}$$

but as we have seen, this tactic fails when $\hat{p}_{ij} = 0$ or 1, i.e., $F^{-1}(x) = 0$ or 1.
I can modify the link so it becomes a mixture of link functions depending on the value of $\eta = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ijk} + \zeta_i$. One such option is:

$$F^{-1}(\eta) = \left\{ \begin{array}{ccc} 0.01 & \text{if} & \eta \leq 0.01 \\ \eta & \text{if} & 0 < \eta < 1 \\ 0.99 & \text{if} & \eta \geq 0.99 \end{array} \right\} \tag{8.6}$$

This ensures that the predicted individual probability of success never drops below 0.01 or above 0.99. Maximum likelihood methods will struggle with such a set-up as the β estimates will never converge to one single value and so the associated ML equations will iterate forever and never decide on a single value. MCMC methods can however fit such models, with very flat priors for the variables of interest.

Using my example given in the previous section, I can use `BUGS` (or other related MCMC software) to quickly examine the parameter behaviour. Using flat priors for the β parameters, i.e., $\beta_0, \beta_1 \sim$ Normal(0, 10000), a burnin of 5,000 and a monitoring run of 50,000 produced the following FE results (Table 8.12):

	Parameter Estimate	SE of Estimate
β_0	-136.6	144.1
β_1	359.1	196.4

Table 8.12: Parameter estimates for the non-linear problem.

with the predicted individual success probabilities having the following characteristics (Table 8.13):

	Estimate	SE	2.5%	Median	97.5%
p_0	0.99	0.01	0.99	0.99	0.99
p_1	0.043	0.16	0.01	0.01	0.702
p_2	0.99	0.01	0.99	0.99	0.99
p_3	0.043	0.16	0.01	0.01	0.702
p_4	0.043	0.16	0.01	0.01	0.702

Table 8.13: Progression probabilities for the non-linear problem.

It is apparent that the β priors are too flat with such a small amount of information. Further testing shows that the prior information has to be extremely strong to produce sensible estimates for the β parameters. Similar alternative link functions also produce disappointing results and this methodology does not seem to offer a sensible solution at the moment.

- Use a linear link but be aware of the continuous versus binary outcome problems;

  The quality assessment properties of using a linear link are good but, as we know, this link breaks the assumption that the outcome is binary: success or failure. Using a standard non-linear link method seems to fraught with difficulties and providing a valid non-linear link model will require more complex model assumptions and methods. For individual analysis, these complications need to be balanced against the statistical importance of breaking the binary outcome assumption. When the numbers involved are reasonably large, the binary outcome assumption is not a severe restriction as the binomial distribution tends towards to normal distribution with larger sample sizes. Currently, accepting a linear link but noting it should (if we were being strict) be non-linear appears to be the only sensible solution.

Non-linear approach: conclusion

Given the methods and work carried out so far on non-linear quality assessment approaches, there seems to be no apparent way of including the "binary outcome assumption" in the modelling. The sensible option is to use a linear link function, note that the outcome is binary and assume that the institutional assessments would not dramatically change if the non-linear option was available.

There is plenty of potential to research this further and hopefully provide a more satisfactory assessment option.

Chapter 9

Longitudinal data

9.1 Extra information: additional years

So far I have only considered analysis involving one year's worth of data. Since the start of the project, an additional year of data has been made available i.e., 1997/1998. The 1998/1999 data, the third year of the series, is due to be published in October 2001. It would be appropriate to start to consider how institutions can be compared across the years and so I would like to develop models that include any additional yearly data.

In this time series, universities appear in each year but the students within those institutions will change. Therefore the focus is on examining changes relating to the quality assessment at the university level rather than any analysis at the student level. With very few data points for each university (i.e., a quality assessment for each year), normal time series approaches cannot be applied. The standard method for very small time series is to use repeated measures (Everitt (1995)). This method is designed to deal with situations when a sequence of observations are made on each of a number of experimental units (in our case, HE institutions). Gray et al. (ming) describes a model that examines the stability of average A/AS level results for English institutions in four different years. Gray uses a combination of fixed and random-effects to establish how the performance of the institutions changes over the years. They conclude that "value-added" results for institutions are difficult to predict in a time-series framework (i.e., you can't assume that past performance is a guide to future institutional quality assessments.

Section 9.2 of this chapter describes the differences between the two years in terms of raw data and institutional quality assessments based upon our non-model-based approach (Section 2.3). I go on to repeat, on the 97/98 data, some investigative analysis carried out on 96/97 data. Section 9.3 describes a new model-based quality assessment approach that deals with more than one year's worth of data.

9.2 Comparing the two years

Data analysis 1997/1998

Tables E.1 - E.8 show the marginal progression distributions for students, broken down by each of the eight PCFs. There were slightly more students entering HE in 1997/1998 compared with 1996/1997, around 290,000 compared with 285,000. The overall progression rate has not changed very much, with a progression rate of 90.1% in 96/97 compared with a 90.4% in 97/98. There are around the same proportion of males and females in both datasets, with similar progression rates for each group across years. In general, the same progression patterns and distributions of students across PCF categories are seen in the 97/98 data that were recorded in the 96/97.

The principal differences between the two years are:

- There are less "unknowns" in both the state school (37% down to 30%) and social class (32% down to 25%) PCF variables. This is because data collection methods have become more reliable in the later year's data;
- There is an increase in the proportion of young students, 74.5% compared against 71.2% in 96/97; and
- Slightly more students are entering HE for the first time in 97/98. 88.5% of the 96/97 population were in her first year of HE study but in 97/98, this proportion increases to 91.4%.

Quality assessment comparison

I can now repeat our quality assessment analysis on the 1997/1998 data, using our non-model-based approach. Table 9.1 shows the numbers of bad, ok, and good universities in each year.

Year	Bonferroni			HEFCE		
	Bad	Normal	Good	Bad	Normal	Good
1996	20	124	16	13	133	14
1997	15	129	16	10	139	11

Table 9.1: University status' for 1996 and 1997.

The choice of cut-off does make a difference, so the HEFCE and Bonferroni cut-off results are both given. With a Bonferroni approach, there are 16 good universities in each year but 1997/1998 seems to have fewer poor universities (20 in 96/97 drops to 15 in 97/98). If a HEFCE cut-off approach is used ($|z_i| \geq 3.00$ and $|\hat{D}_i| \geq 0.03$) there is an overall reduction in the number of unusual universities compared with the Bonferroni approach. 97/98 has fewer significant institutions than 96/97, i.e., 25 in the first year and only 23 in the second. In general, the number of unusual universities drops from year one to year two.

Table 9.2 shows how each university's assessment changes between years based upon a Bonferroni cut-off. Table 9.3 repeats the analysis using the HEFCE cut-off approach.

Bonferroni		Status 1997			
		Bad	Normal	Good	Total
Status 1996	Bad	8	12	0	20
	Normal	7	110	7	124
	Good	0	7	9	16
	Total	15	129	16	160

Table 9.2: Status changes - Bonferroni cut-off.

HEFCE		Status 1997			
		Bad	Normal	Good	Total
Status 1996	Bad	4	9	0	13
	Normal	6	121	6	133
	Good	0	9	5	14
	Total	10	139	11	160

Table 9.3: Status changes - HEFCE cut-off.

Using either cut-off, there seems to be some movement in the performance of institutions over the two years. Under a Bonferroni cut-off, 33 universities have changed status from one year to the next. A similar story is found using a HEFCE cut-off approach, where 30 universities have changed in assessment

In both cut-off cases, no university changes from one extreme to the other, i.e., from bad to good or vice-versa. There seems to be a trend for universities to be "normal" in 1997/1998 compared to the 1996/1997 data. In 1996/1997 there were 124 normal universities using the Bonferroni cut-off and this number increased to 129 in 1997/1998. A similar outcome is observed when the HEFCE cut-off is used. In all cases the number of extremal universities (good or bad) does not increase in the later year. Some suggested reasons for these trends are:

- The badly performing universities have improved whilst the whole university population is developing faster than the 1996/1997 excelling institutions;
- Data quality from the universities has improved dramatically. This makes the playing field a little more level by removing the extremal universities who struggled or excelled because their data were wrong. This seems to be the most likely explanation for the trend seen in the data.
- The trend just happens to be down to random fluctuation in university quality.

In 1997/1998, a high proportion of the universities remain in their original 1996/1997 state, 79% (127 / 160) using Bonferroni and 81% using the HEFCE cut-off. The universities of most interest are those that remain as either good or bad. Being tagged as bad in two years has a lot more weight than being marked poor for one out of two years. Universities that remain good for both years may be doing something right in terms of their progression data and should be studied further. The Tables 9.4 and 9.5 given a summary of which universities appeared in an extreme state for the two years in question.

NB Table 9.4: the five universities marked with a hash(#) are also classed as good in both years using the HEFCE cut-off. There were no universities that were good in both years using the HEFCE cut-off but not good in both years using the Bonferroni cut-off.
Table 9.5: the three universities marked with an asterisk(*) are classed as bad in both years using both the HEFCE cut-off and the Bonferroni cut-off. One university was bad in both years under the HEFCE classification but not under the Bonferroni cut-off. That was Institution 101, -3.37 (145) in 1997/1998 and -5.15 (159) in 1996/1997. It escapes being bad for both years under the Bonferroni classification because its 1997/1998 z-score is above -3.61.

University	1997		1996	
	Z Score	Rank	Z Score	Rank
17	7.06	1	5.03	6
164 #	6.06	2	5.02	8
71 #	5.88	3	5.90	5
121 #	5.76	4	6.25	4
149#	5.74	6	5.04	7
144 #	5.39	7	4.95	9
30	5.08	8	4.49	10
129	4.72	9	4.19	11
1	4.70	10	9.13	1

Table 9.4: Bonferroni cut-off: universities classed as good in both years.

University	1997		1996	
	Z Score	Rank	Z Score	Rank
154 *	-13.52	159	-6.41	156
75	-7.81	158	-4.37	148
145	-7.39	157	-6.02	154
81	-5.26	154	-4.51	151
122 *	-4.67	152	-6.26	155
151	-4.22	150	-6.81	158
64 *	-4.10	149	-5.93	153
78	-3.86	148	-8.45	159

Table 9.5: Bonferroni cut-off: universities classed as bad in both years.

Repeated studies in 97/98

I can study whether similar results hold for the 1997/1998 data as were seen in the 1996/1997 analysis. Firstly I look at the tail behaviour of each of the variance estimation methods in the 97/98 data. This analysis is a repeat of the 96/97 given in Sections 4.5 and 4.6. Table E.9 shows the behaviour of the tails when the original 97/98 dataset is used (equivalent to Section 4.5) and Table E.10 gives the results when the overall target progression rate is changed to 0.5 (equivalent to Section 4.6).

The results in Table E.9 are very similar to those seen in the 1996/1997 data, with the same patterns noted. Our favoured γ approach performs well in all cases with an overall misclassification rate varying between 4.8 and 5.5. There are minor fluctuations in the overall tail rates compared with 96/97 (especially in the smaller worlds) but nothing that would indicate a need for different overall conclusions in 97/98. When the overall success rate is set at around 50%, the 97/98 results (given in Table E.10) are nearly identical to the 96/97 results. The γ estimation technique still provides a good approach for quality assessment.

Table E.11 shows how PCF removal affects institutional status misclassification rates based on a Bonferroni cut-off for the 97/98 dataset. The equivalent 96/97 results are given in Tables 6.7-6.8. The overall misclassification pattern is very similar in both years with 97/98 having a slightly higher "final" rate when no PCFs are adjusted for. On the whole, the 97/98 "bad but not called bad" rates are a little higher than in 96/97 but there is no indication of a different pattern as the differences are very minor. The opposite is true with the "good but not called good" rates where the 97/98 are slightly lower than the 96/97 rates but, as before, the differences are not dramatic. There is no evidence to suggest the results and assumptions made about the 96/97 data do not hold for the 97/98 data in terms of PCF omission.

Table E.12 continues the work from Tables E.11 by identifying the best PCF choices given the number of PCFs you can adjust for is restricted. The idea is that you are told how many PCFs you can have and you have to decide which ones produce the lowest misclassification rates (truth being the results from an eight PCF model). The analysis is completed in both 96/97 and 97/98.

Considering models with at least two PCFs, HEFCE's model of Qualification and Subject is the basis of the best model, except in 1997/1998 when only two PCFs are allowed. The misclassification rate for the qualification and subject only in 1997/1998 is 8.69%, which is only beaten by the qualifications and lowclass model. It is therefore key that, at a bare minimum, both qualifications and subject are adjusted for in any institutional quality assessment analysis.

9.3 Repeated measures

The approach

I would like to develop a modelling system where both years worth of data are included in one model structure, i.e., not to fit two separate models to create the two years worth of results.

Consider the following repeated-measures (this technique is fully described by Diggle et al. (1994)) set-up:

$$y_{ijk} = \mu + \alpha_i^t + \alpha_j^u + \alpha_{ijk}^s + \alpha_{ij}^{tu} + \text{covariates}_{ijk} + e_{ijk} \tag{9.1}$$

where α_i^t is the time effect, α_j^u is the university effect(baseline), α_{ijk}^s is the student variation, α_{ij}^{tu} is the interaction between time and university and covariates_{ijk} are the adjustment variables.

This model can be easily fitted using standard statistical methods. These methods use a single university (say $j = 1$ for ease) in a single year (say the first year, $i = 0$) as the baseline

effect, i.e., all other university and time effects are relative to university 1 in the first year's data. As we have seen in Section 3.8, the institutional effects should be measured relative to the overall population rather than a specified baseline university. With only a single year, those results can be used.

Yang et al. (2000) used a repeated measures set-up to examine data from the 1983 - 86 - 87 British Election Panel Study for three different time points. Initially they fit three separate (one for each year) two-level models and then go on to pool the three years worth of data into a single three-level repeated measures model. They conclude that the three-level repeated measures model struggles to deal with their data adequately as some voters offer a constant response over the three years and they recommend an alternative multilevel multivariate model. This constant response over time is not a significant issue in the HE data.

With a repeated measures set-up, there are multiple years and we have to decide what institution performance is measured against. The potential questions are:

1. What was the institution's performance like compared with the other performances in that year?

2. What was the institution's performance like compared against institutional performances from either year?

3. What was the two year institutional performance like compared with the other institutional performances?

The third question does not involve using repeated measures as the yearly effect can be ignored and students are considered to be from the same population regardless of which year they fell into.

To decide between questions 1 and 2, it seems more sensible to consider the institutional performance relative to other institutions in the same year (this is certainly what most league tables attempt to achieve), i.e., try to answer question 1. There is some interest in seeing whether an institution has changed in character year-on-year but there is little interest in comparing Cambridge in 1996/1997 against Bristol in 1997/1998.

A further issue is whether we are interested in the absolute difference or relative difference for each university. An absolute approach asks whether university j's performance in 97/98 changed from its performance in 96/97. A relative approach asks a slightly different question: after accounting for the overall differences between the two years, has university j's performance changed year on year? Imagine in 96/97, the University of Brecon has a z-score of 0.00 and in the following year that z-score changed to -4.00. If the overall population performance dropped in the 2nd year, are we more interested in Brecon's performance change relative to all the other universities or relative to itself? Initially I focus on the absolute change, i.e., not adjusting for the overall trend for all universities.

To answer question 1, I need to impose two restrictions on the implied quality differences for university j in year i (ϱ_{ij}^{*}) relative to the overall population. With two year's worth of data,

I have to place two side conditions on the α_j^ws:

$$\text{Restriction 1} = \sum_{j=1}^{U} n_{0j}\alpha_{0j}^w = 0 \tag{9.2}$$

$$\text{Restriction 2} = \sum_{j=1}^{U} n_{1j}\alpha_{1j}^w = 0 \tag{9.3}$$

or with x years worth of data, we require x α^w restrictions:

$$\sum_{j=1}^{U} n_{ij}\alpha_{ij}^w = 0, \quad i = 0, \ldots, x-1 \tag{9.4}$$

where n_{ij} is the number of entry students at university j in year i.

These restriction can be fitted using a FE approach and this is described below. To implement a RE methodology is much more difficult as each set of α^ws has to be normally distributed around zero. This restriction can easily be fitted to one set of α^ws (say α_0^ws) but there is no easy way to place a similar restriction on the other set of α^w's as each one of these is a combination of other model parameters. I therefore use a FE approach to examine these repeated measures models.

1. Set university 1 in year 0 as the baseline university. This means that $\alpha_{01}^u = 0$ when using the baseline α restriction to fit the repeated measures model.

2. Using the student numbers for year 0, i.e., $n_{01}, n_{02}, \ldots, n_{0U}$, calculate the weighted αs using the method described in Section 3.8. These are now defined to be α_{0j}^w.

3. Calculate for unweighted αs for year 1 (α'_{1j}) using the appropriate university, interaction and year terms using the weighted alpha for year 0 and university 1 (α_{01}^w), i.e:

$$\alpha'_{1j} = \alpha_{01}^w + \alpha_1^t + \alpha_j^u + \alpha_{1j}^{tu} \quad \forall j \in (1, \ldots U) \tag{9.5}$$

4. Convert these α' into year 1 baseline αs by taking off α'_{11} from each α':

$$\alpha_{1j}^{b\prime} = \alpha'_{1j} - \alpha'_{11} \tag{9.6}$$

After inspection, it can be shown that $\alpha_{1j}^{b\prime}$ is just university j's model effect (α_j^u) plus the interaction between the year, 1, and the university, j (α_{1j}^{tu}):

$$\begin{aligned} \alpha_{1j}^{b\prime} &= \alpha'_{1j} - \alpha'_{11} \\ &= (\alpha_{01}^w + \alpha_1^t + \alpha_j^u + \alpha_{1j}^{tu}) - (\alpha_{01}^w + \alpha_1^t + \alpha_1^u + \alpha_{11}^{tu}) \end{aligned}$$

but α_1^u and α_{11}^{tu} are zero by definition, so:

$$\begin{aligned}\alpha_{1j}^{b\prime} &= \alpha_{01}^w + \alpha_1^t + \alpha_j^u + \alpha_{1j}^{tu} - \alpha_{01}^w - \alpha_1^t - 0 - 0 \\ &= \alpha_j^u + \alpha_{1j}^{tu} \qquad \forall j \in (2, \ldots, U) \\ &= 0 \qquad \text{for } j = 1 \end{aligned} \tag{9.7}$$

5. Calculate the year1 weighted αs (α_{1j}^w) by using the student numbers for year 1, i.e., $n_{11}, n_{12}, \ldots, n_{1U}$, and the method from Section 3.8.
   So now:

$$\alpha_{11}^w = -\frac{\sum_{j=2}^U n_{1j}\left(\alpha_j^u + \alpha_{1j}^{tu}\right)}{\sum_{j=1}^U n_{1j}} \tag{9.8}$$

And $\forall j \in (2, \ldots, U)$:

$$\begin{aligned}\alpha_{1q}^w &= -\frac{\sum_{j=2}^U n_{1j}\left(\alpha_j^u + \alpha_{1j}^{tu}\right)}{\sum_{j=1}^U n_{1j}} + \alpha_{1q}^{b\prime} \\ &= -\frac{\sum_{j=2}^U n_{1j}\left(\alpha_j^u + \alpha_{1j}^{tu}\right)}{\sum_{j=1}^U n_{1j}} + \alpha_q^u + \alpha_{1q}^{tu} \end{aligned} \tag{9.9}$$

Steps 3-5 can be replaced with:

- After calculating the weighted αs for year 0, rerun the model with a new baseline university/time. Now set university 1 in year 1 as the baseline university and complete steps 1-2 again, substituting year 0 with year 1. This means calculating a weighted mean for year 1 based on the numbers from year 1. Note that the baseline αs $\alpha_{11}^u = 0$ for year 1 will not match the $\alpha_{1j}^{b\prime}$'s given in Equation 9.7 but the same year 1 weighted αs will still be produced.

Variance of the weighted estimates

We can also derive standard errors for the parameters of interest.

Let n_{i+} be the total number of students in year i, i.e., $\sum_{j=1}^U n_{ij}$.

$$\begin{aligned}\text{Var}\left(\alpha_{11}^w\right) &= \text{Var}\left(-\frac{1}{n_{1+}}\sum_{i=2}^U n_{ij}\left(\alpha_j^u + \alpha_{ij}^{tu}\right)\right) \\ &= \left(\frac{1}{n_{1+}}\right)^2 \text{Var}\left(\sum_{i=2}^U n_{ij}\left(\alpha_j^u + \alpha_{ij}^{tu}\right)\right) \\ &= \left(\frac{1}{n_{1+}}\right)^2 \text{Var}\left(n_{12}\alpha_2^u + n_{12}\alpha_{12}^{tu} + \ldots + n_{1U}\alpha_U^u + n_{1U}\alpha_{1U}^{tu}\right)\end{aligned}$$

$$= \frac{\sum_{i=2}^{U}\sum_{j=2}^{U} n_{1i}n_{1j}\left(\mathrm{Cov}(\alpha_i^u,\alpha_j^u)+2\mathrm{Cov}(\alpha_i^u,\alpha_{1j}^{tu})+\mathrm{Cov}(\alpha_{1i}^{tu},\alpha_{1j}^{tu})\right)}{n_{1+}^2} \quad (9.10)$$

$$\begin{aligned}
\mathrm{Var}\left(\alpha_{1q}^w\right) &= \mathrm{Var}\left(-\frac{\sum_{j=2}^{U} n_{1j}\left(\alpha_j^u+\alpha_{1j}^{tu}\right)}{n_{1+}}+\alpha_q^u+\alpha_{1q}^{tu}\right)\\
&= \mathrm{Var}\left(\left(-\frac{1}{n_{1+}}\right)\left(\left[\sum_{j=2}^{U} n_{1j}(\alpha_j^u+\alpha_{1j}^{tu})\right]-n_{1+}\alpha_q^u-n_{1+}\alpha_{1q}^{tu}\right)\right)\\
&= \frac{\mathrm{Var}\left(\left[\sum_{j=2,j\neq q}^{U} n_{1j}(\alpha_j^u+\alpha_{1j}^{tu})\right]+(n_{1q}-n_{1+})(\alpha_q^u+\alpha_{1q}^{tu})\right)}{n_{1+}^2}\\
&= \frac{\sum_{i=2}^{U}\sum_{j=2}^{U}\omega_{ijq}}{n_{1+}^2}
\end{aligned}$$

$$\omega_{ijq} = \left\{\begin{array}{cc}
n_{1i}n_{1j}\left(\mathrm{Cov}(\alpha_i^u,\alpha_j^u)+2\mathrm{Cov}(\alpha_i^u,\alpha_{1j}^{tu})+\mathrm{Cov}(\alpha_{1i}^{tu},\alpha_{1j}^{tu})\right) & \text{if } i\neq q \text{ and } j\neq q\\
n_{1j}(n_{1q}-n_{1+})\left(\mathrm{Cov}(\alpha_q^u,\alpha_j^u)+2\mathrm{Cov}(\alpha_q^u,\alpha_{1j}^{tu})+\mathrm{Cov}(\alpha_{1q}^{tu},\alpha_{1j}^{tu})\right) & \text{if } i=q \text{ and } j\neq q\\
n_{1i}(n_{1q}-n_{1+})\left(\mathrm{Cov}(\alpha_i^u,\alpha_q^u)+2\mathrm{Cov}(\alpha_i^u,\alpha_{1q}^{tu})+\mathrm{Cov}(\alpha_{1i}^{tu},\alpha_{1q}^{tu})\right) & \text{if } j=q \text{ and } i\neq q\\
(n_{1q}-n_{1+})^2\left(\mathrm{Cov}(\alpha_q^u,\alpha_q^u)+2\mathrm{Cov}(\alpha_q^u,\alpha_{1q}^{tu})+\mathrm{Cov}(\alpha_{1q}^{tu},\alpha_{1q}^{tu})\right) & \text{if } i=j=q
\end{array}\right\} \quad (9.11)$$

These expressions can be extended when using more than two years of data.

We now have valid expressions for university quality assessments and their associated standard errors. These can be used to calculate appropriate z-scores in a modelling set-up that includes the yearly structure of the data.

Missing values

On some occasions universities do not appear in one year's data but appear in another year. This can be due to a number of issues ranging from data problems at the university to new universities being introduced. In these cases, we have some missing information in the repeated measures data.

For my data with only two years' information, a small number of universities do not appear in both years. The methods described previously are still valid as the n_{ij}s provide enough information to adjust for missing universities (i.e., when university j is missing in year i, $n_{ij}=0$), and so these missing effects will not affect the calculations.

When repeated measures approaches are used with missing values, some interaction terms are rejected by the modelling. With regard to calculating the parameter estimates and their associated standard errors, these rejected interactions terms should be classified as being zero.

For illustration:

Consider the 1996/1997 (year 0) and 1997/1998 (year 1) data sets studying only three

medical institutions: Institution 10; Institution 160; and Institution 109. Tables 9.6 and 9.7 show how the student numbers and progression rates vary from year 0 to year 1.

University	Year 0	Year 1	Overall
10	0.94	-	0.94
160	0.96	0.95	0.96
109	0.98	1.0	0.99
Overall	0.96	0.97	0.97

Table 9.6: Progression rates: repeated measures example.

University	Year 0	Year 1	Overall
10	116	0	116
160	365	174	539
109	165	160	325
Overall	646	334	980

Table 9.7: Student numbers: repeated measures example.

Let terms where $j = 1$ represent Institution 10, $j = 2$ represent Institution 160, and $j = 3$ represent Institution 109. From the tables above, we can see that the Royal Free only has data for year 0. For covariates, we only adjust on student age and gender. This produces four covariate levels. We can fit a repeated measures model using standard statistical methods. I will use the Institution 10's students in year 0 as the baseline group. The parameter estimates for each effect are given in Table 9.8 (the covariate estimates are not given as they are not required in the quality calculations). The associated covariance matrix for these estimates is also available (Table 9.9).

Model Parameter	Estimate
α_1^t	-0.0087
α_2^u	0.0261
α_3^u	0.0416
α_{12}^{tu}	Dropped
α_{13}^{tu}	0.0274

Table 9.8: Parameter estimates: repeated measures example.

	Year	
University	0	1
Royal Free(1)	0.0000	-0.0087
St George's(2)	0.0261	0.0173
UCWM(3)	0.0416	0.0602

Table 9.9: Relative effects: repeated measures example.

We can now calculate the university and time effects relative to the baseline university and year (α'). For year 0, with Institution 10 is the overall baseline, we can use the method from Section 3.8 using year 0 student numbers.

$$\alpha_{01}^w = -\frac{365(0.0261)+165(0.0416)}{646} = -0.0253 \tag{9.12}$$

$$\alpha_{02}^{w} \quad = -0.0253 + .0261 \quad = 0.0008 \tag{9.13}$$

$$\alpha_{02}^{w} \quad = -0.0253 + 0.0416 \quad = 0.0669 \tag{9.14}$$

For year 1, we need to convert our α'_{1j}'s into $\alpha^{b\prime}_{1j}$'s. This is simply done by taking off α'_{11} from the other α'_{1j}'s. This gives $(\alpha^{b\prime}_{11}, \alpha^{b\prime}_{12}, \alpha^{b\prime}_{13}) = (0.0000, 0.0261, 0.0690)$, $\alpha^{b\prime}_{11} = 0$ by definition.

Using student numbers from year 1 and the appropriate formula's again, these can be converted in the required $(\alpha^{w}_{11}, \alpha^{w}_{12}, \alpha^{w}_{13}) = (-0.0467, -0.0206, 0.0223)$. ($\alpha^{w}_{11}$ is rejected as a parameter estimate as there are no students at that university in that year).

Yearly changes in university performance

The model structure now allows us to study whether a specific university has changed in quality performance from one year to the next. This essentially means looking at whether its year 0 performance is significantly different to its year 1 performance. This can be tested by looking to see whether the additional terms added to the α^{w}_{0j} to create the α'_{1j} are significantly different to zero.

Define $C_j = \alpha'_{1j} - \alpha^{w}_{0j}$, the change in university j's performance from year 0 to year 1.

$$\begin{aligned} C_j &= \alpha'_{1j} - \alpha^{w}_{0j} \\ &= \alpha^{w}_{01} + \alpha^{t}_{1} + \alpha^{u}_{j} + \alpha^{tu}_{1j} - \alpha^{w}_{0j} \\ &= \left(\alpha^{w}_{01} + \alpha^{u}_{j}\right) - \alpha^{w}_{0j} + \alpha^{t}_{1} + \alpha^{tu}_{1j} \\ &= \alpha^{t}_{1} + \alpha^{tu}_{1j} \\ &= \left\{ \begin{array}{cc} \alpha^{t}_{1} & \text{if} \quad j = 1 \\ \alpha^{t}_{1} + \alpha^{tu}_{1j} & \text{if} \quad \forall j \in (2, \ldots, U) \end{array} \right\} \end{aligned} \tag{9.15}$$

$$\mathrm{Var}\left(C_j\right) = \left\{ \begin{array}{cc} \mathrm{Var}\left(\alpha^{t}_{1}\right) & \text{if} \quad j = 1 \\ \mathrm{Var}\left(\alpha^{t}_{1}\right) + \mathrm{Var}\left(\alpha^{tu}_{1j}\right) + 2\mathrm{Cov}\left(\alpha^{t}_{1}, \alpha^{tu}_{1j}\right) & \text{if} \quad \forall j \in (2, \ldots, U) \end{array} \right\} \tag{9.16}$$

A z-score for the change in status for each university j (Z^{c}_{j}) can be created using the standard Z equation, $\frac{C_j}{\mathrm{SE}_{(C_j)}}$. Table 9.10 shows a summary of the results examining the change in university quality from year 0 to year 1 (C_j). The top five universities on each extreme are given along with a selection of other universities. A negative C_j indicates that university j's quality has worsened over time. In this data, Institution 128 looks particularly bad as it has a very negative C_j of -0.367, which implies that the university's effect on progression has changed by around 36% from 1996/1997 to 1997/1998. This gives rise to a extremal Z^{c}_{j} of -17.5, a highly significant value in a standard normal distribution. Using the standard rules of z-score cut-off, all five "bad" universities (in terms of change) give cause for concern. HEFCE discovered that Institution 128 had a data problem with their 97/98 data - around 130 students on a one year course had been marked as not progressing when they had in fact left HE because they had graduated. On the other side of the coin, it looks as if Institution 110 has dramatically improved

their effects on progression, with around a 4% improvement in effect. This is actually down to data improvement, in 96/97 Institution 110 provided some incorrect information which made them look particularly poor in terms of progression. They did not make the same mistake when providing their data for 97/98.

University	C_j	SE(C_j)	z_j^c	Status(96 $\Rightarrow$ 97)
128	-0.367	0.021	-17.51	OK $\Rightarrow$ B
154	-0.053	0.007	-7.16	B $\Rightarrow$ B
52	-0.066	0.010	-6.44	OK $\Rightarrow$ B
4	-0.042	0.007	-6.27	G $\Rightarrow$ OK
48	-0.044	0.008	-5.79	OK $\Rightarrow$ B
$\vdots$			$\vdots$	
61	-0.014	0.008	-1.70	OK $\Rightarrow$ OK
$\vdots$			$\vdots$	
14	0.000	0.008	0.04	OK $\Rightarrow$ OK
79	0.000	0.008	0.05	OK $\Rightarrow$ OK
$\vdots$			$\vdots$	
118	0.017	0.011	1.46	OK $\Rightarrow$ OK
$\vdots$			$\vdots$	
151	0.025	0.007	3.36	B $\Rightarrow$ OK
139	0.027	0.008	3.51	OK $\Rightarrow$ OK
78	0.035	0.010	3.52	B $\Rightarrow$ B
152	0.027	0.008	3.64	OK $\Rightarrow$ G
110	0.043	0.006	6.75	B $\Rightarrow$ OK

Table 9.10: Year changes based upon a repeated measures model.

Status results produced from a non-model-based approach considering each year individually and a Bonferroni cut-off

Results for HEFCE data 1996/1997 and 1997/1998

Method comparison

The repeated measures method was performed on the 1996/1997 and 1997/1998 HEFCE data. The repeated measures z-scores for 1996/1997 had a mean of 0.16 and a SE of 3.2. The non-model-based z-scores (ignoring the yearly structure) for the same data had a mean of 0.24 and a SE of 2.90. The two sets of z-scores were highly correlated (= 0.97) and Figure 9-1 demonstrates that there is good agreement of university quality between the two methods. However, there do appear to be some shrinkage effects in the non-model-based z-scores for the extremal negative scores. This shrinkage occurs when the z-scores are less than -5 and so the university is already marked as under performing.

A similar effect is seen when the two sets of αs from the methods are compared, as shown in Figure 9-2. The correlation between the αs is 0.95.

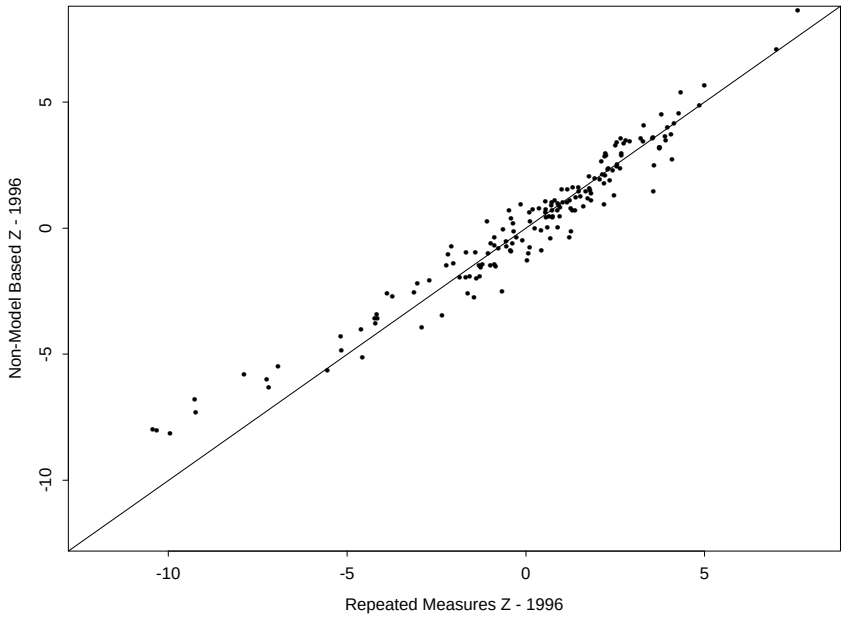

Figure 9-1: The repeated measure against the non-model-based z-scores for 96/97.

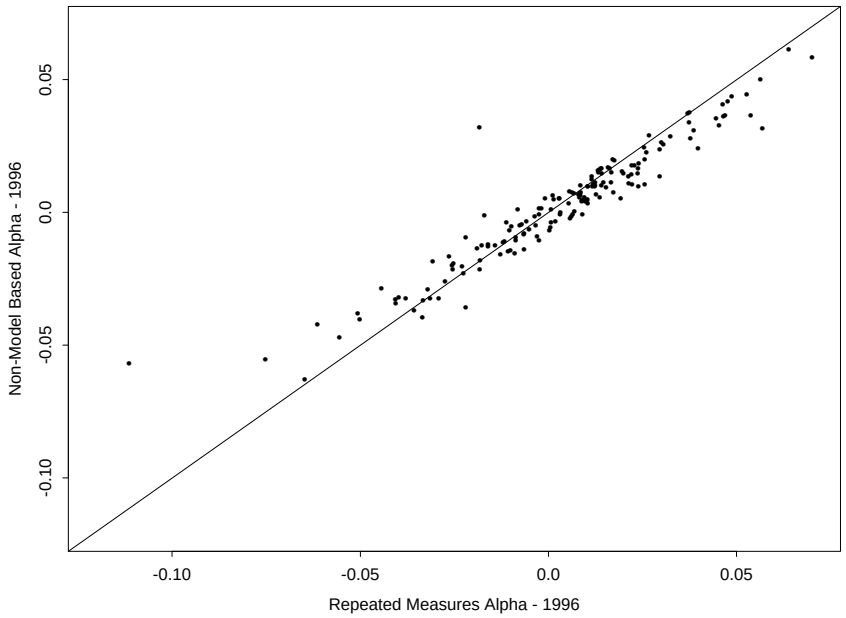

Figure 9-2: The repeated measure against the non-model-based αs for 96/97.

The 96/97 results are mirrored in the 97/98 results when comparing the two methods. Figure 9-3 shows the 0.97 correlation between the repeated measures and non-model-based approaches

using the 97/98. Note that a similar shrinkage effect is seen again in the non-model-based z-scores.

It seems that equivalent results are provided by a repeated measures approach (compared against the "non-yearly" non-model-based technique) but this repeated measures methodology allows us to examine yearly changes and effects on universities.

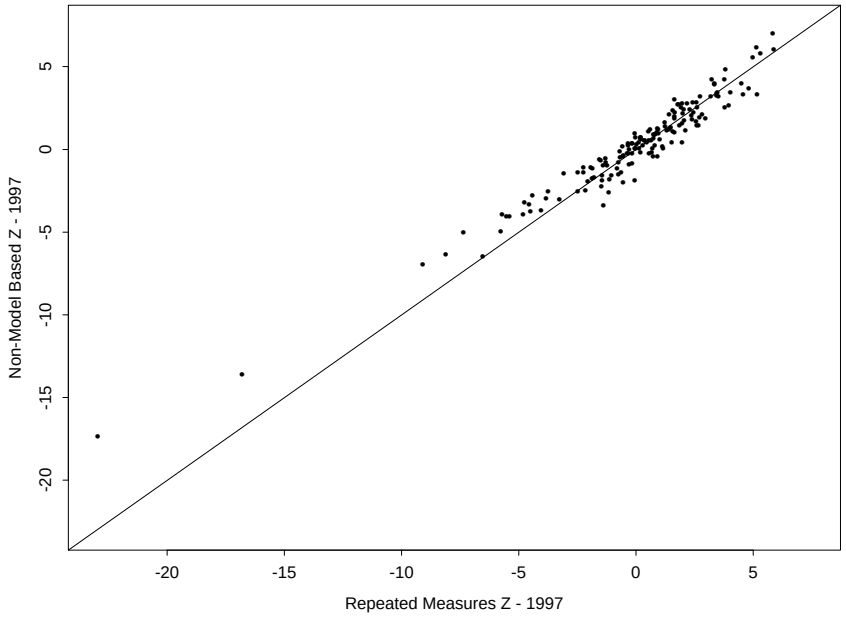

Figure 9-3: The repeated measure against the non-model-based z-scores for 97/98.

Chapter 10

Gold standards

10.1 Introduction

As we have seen, there are two principal ways to measure institutional quality: an input-output (IO) approach or by examining the institutional processes.

In the IO method, the institution is considered to be a black box, with no knowledge of what occurs within the box. The inputs are the characteristics of a person as she enters the system, e.g., sickness at admission (for hospitals); age of person; academic ability (for educational establishments); or gender. These are contrasted against the status of the individual when she leaves: e.g., mortality (for hospitals); or academic achievement(for education). Comparing the inputs with the outputs gives a view of the effectiveness of the studied institution.

With a process approach, the black box is opened and the contents are exposed. The processes within the establishment are measured directly, using expert judgement for example. This method is usually much more expensive as measurement systems have to be designed and installed at each institution. In the IO approach, many of the measurement systems for the inputs and outputs already exist and the data can be readily obtainable.

In the current literature, the difference in method expense means that there has been little work completed using process data in quality assessment. Process data is a more reliable way of measuring quality (as long as the measurement methods are effective) because the institutional quality is examined directly rather than indirectly, which occurs in the IO case.

McGraw et al. (1996) looks at how process characteristics can affect study outcomes. The article uses data from the Child and Adolescent Trial for Cardiovascular Health(CATCH) to see whether teacher characteristics and student characteristics can be used to identify changes in dietary knowledge from a number of intervention schools. Their main conclusion was that teacher characteristics had a large effect on outcome, i.e., their process data was a good predictor for outcomes.

Sound process data can be seen as a 'gold standard' for quality assessment. If an institution is conducting good practice, then good outcomes should be expected. For example, a child that is taught badly by a school should be expected to achieve lesser results compared to a child

that has received an excellent standard of tutoring.

In this chapter I examine two new datasets that have attempted to measure process data. In Section 10.2 I develop the work carried out by Kahn et al. (1990) on predicting hospital mortality rates for Medicare patients. Section 10.3 describes data with a sporting theme, examining methods for profiling soccer player performance in the English Premiership.

10.2 Medicare

Medicare description

Kahn et al. (1990) developed a system that uses characteristics of the patient at admission to predict death within 30 days of hospital admission for Medicare patients with stroke, pneumonia, myocardial infarction, congestive heart failure, and hip fracture. A two-stage cluster sample of 14,002 elderly patients (aged 65 and over) from 297 hospitals was taken from five US states, with the goal of national representiveness of the resulting patient and hospital samples. One aim of this project was to create patient-specific predictions, which could be used in identifying unexpected deaths for clinical review, and for interpreting information on unadjusted mortality rates.

Alongside input (e.g., severity of illness on entry, gender, etc.) and output (death within 30 days) data, Kahn et al. (1990) collected a number of parameters on patient process, e.g., measures of how well the patient was treated and a process score was created for each patient. This score can be used as a basis for direct measurement of institutional process.

Medicare results

Using the IO approach, we can obtain a measure of hospital quality by taking our available (and appropriate) inputs and comparing them to the output of interest. In the Medicare study, the inputs based on the patient characteristics, including age, disease type and severity, are combined into a single input score called the severity score. The output is whether the patient died within 30 days of admission to the hospital.

With this single continuous input variable, there are two IO approaches available to us. Both are confined to model-based methods, using either fixed- or random-effects modelling. The non-model-based HEFCE approach cannot be used as the individual predictors are not available and the severity scale is a continuous variable. The IO methods examined are essentially based on a main effects only model, due to the way in which the severity score was created. There is no option available to examine interaction effects as the data come directly from the later stages of study, when the severity indicator was produced.

The fixed-effects (FE) (main-effects only) quality assessment method was used to establish the performance of each of the 297 hospitals. This method created a perceived difference between observed and expected mortality rates for each hospital and an associated z-score. The z-scores varied between -2.41 and 3.74, with the Bonferroni cut-off for 297 hospitals set at 3.76. The mean of the scores was -0.001 with a standard deviation of 1.09. The α estimates

also have a very high correlation with their associated z-scores because all of the hospitals have similar small sample sizes.

The data deals with death rates rather than survival rates so a negative z-score indicates that the observed death rate was lower than the expected rate, i.e., the hospital is doing better than expected. Conversely a positive score indicates that the hospital's patients aren't surviving as well as they should do.

We would expect there to be a correlation between how well the patients were looked after (process) and the quality assessment of the hospital after adjustment (z-score). This correlation should be negative as a hospital that provides an improved treatment (process becomes more positive) should improve in their quality assessment based on adjusted death rates (z-score becomes more negative). Figure 10-1 shows the correlation between average hospital process and the hospital quality z-score.

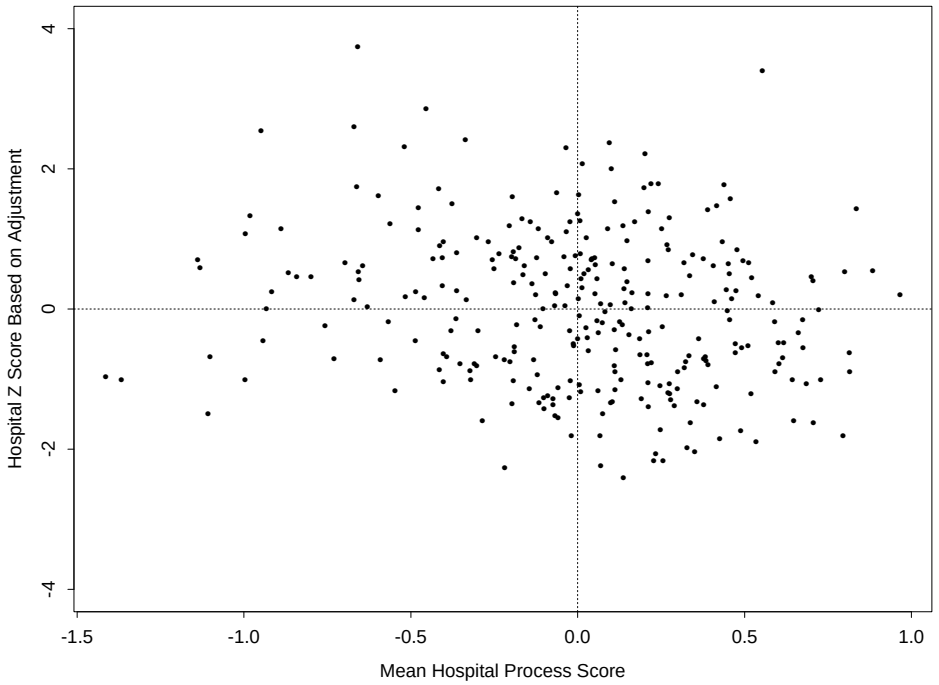

Figure 10-1: Link between process and z-score: Medicare data.

The correlation between these two variables is only -0.18, indicating that, based on results from the FE quality assessment method, an improvement in hospital caring methods leads to a small effect on patient survival. Intuitively we would expect this correlation to be much greater and the whole point of this analysis was to confirm a link between IO score and directly measured process. What possible reasons are there for this result?

The quality assessment method might be wrong, i.e., the IO approach fails to correctly identify hospitals which provide good patient care. This is highly unlikely as both the approach

and reasoning look reasonable. There is another potential answer to this question: other unmeasured factors at the hospital level could be affecting mortality. The question is now whether we can test this theory. One possible approach is as follows:

1. Calculate the quality assessment scores in the normal way, i.e., fit the FE modelling approach adjusting for the necessary input PCFs, and generate the IO z-scores.
2. Repeat the FE approach but now, rather than adjusting on the input PCFs, adjust on the individual process score for each of the institutional patients. Calculate the hospital z-scores in the normal fashion. Let us call these z-scores the process z-scores.
3. Each set of z-scores measures the unexplained variation at each hospital, after adjusting for either process or patient inputs. If both sets of scores are highly correlated, this implies that neither adjustment approaches is detecting a set of hospital PCFs that are related to mortality.

If we complete this analysis with the Medicare data, the following results are obtained. The correlation between the two sets of scores is 0.75. Figure 10-2 shows the relationship between the IO and process z-scores.

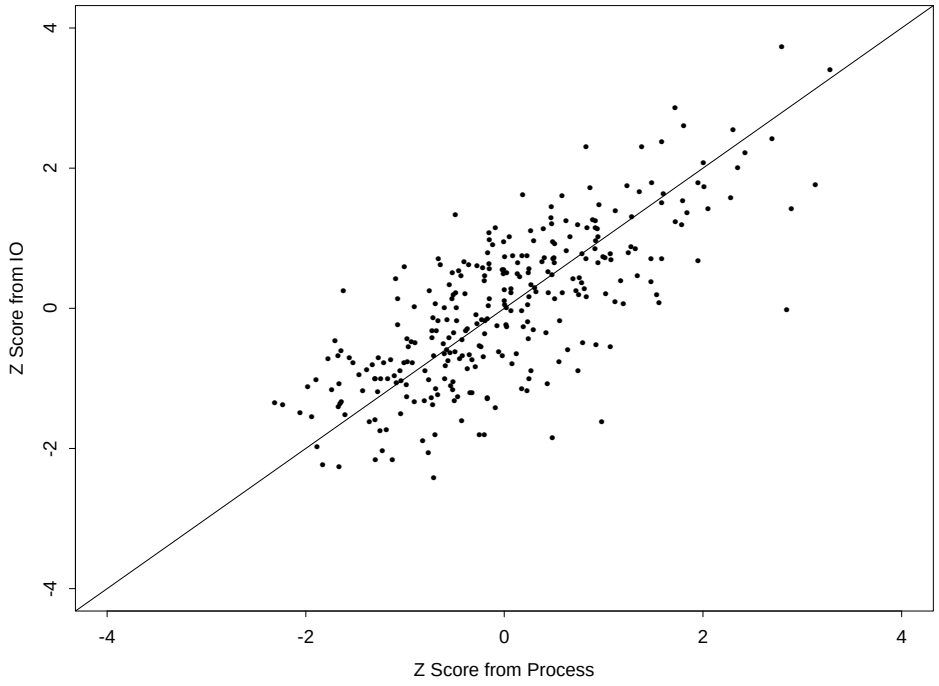

Figure 10-2: Examining unexplained variation in the Medicare data.

The high correlation indicates that there are significant unmeasured factors at the hospital level that have not been taken into account by adjusting for either process or patient inputs.

One suggestion would be a "will to live" variable because, as all of the population examined are 64 or over, some individuals have more drive to fight disease than others and a regional effect may occur. It may be that some hospitals have a high proportion of elderly patients who have a lot to live for, e.g., a hospital in a region where there is a high retirement migration rate.

The overall message from this data is that process is one of many factors that affect mortality and it is certainly not the strongest predictor. However hard hospitals try, patients still die: a depressing fact.

Similar results hold when a RE approach is used as an alternative to the FE set-up.

Recreating the Medicare data

Institutional process can be very difficult to measure using IO methods when the outcome of interest (e.g. death in the Medicare study, or progression in the HEFCE dataset) does not have a high correlation with institutional quality. In the Medicare study, the correlation between a patient's process score and her death outcome was only -0.1. This implies that the level of care given to patients had only a slight effect on their outcomes and many other factors also affected their probability of death.

In this section I try to recreate the Medicare dataset under simulation conditions, varying the correlations between a patient's outcome, her process score and her severity score based on the characteristics of the patient. Keeler et al. (1992) compared the quality of care at different types of hospitals based on explicit criteria, implicit review, and sickness-adjusted outcomes. They found that the institutional quality of care varied on a number of factors including the size of the hospital, in particular the number of beds in the hospital. With this in mind, I also take into account the correlation between hospital size and the other factors in my simulation study.

I can then examine how the relationship between the IO hospital quality assessment varies with the average process score for the hospital in question. I am creating simulation worlds where there are 10,000 patients spread across 297 hospitals, with a varying correlation matrix for patient outcome, severity, process and number of hospital beds. Essentially I am creating artificial Medicare worlds where I know how much effect hospital process has on patient outcome.

The following steps were taken to investigate the quality vs process score relationship:

1. Generate 10,000 draws (one record for each patient) from a four-dimensional multivariate normal with a mean of zero for each variable and a pre-determined covariance matrix (with the diagonal entries set to 1), based on the covariance structure seen in the original Medicare dataset. Define v_1 as the first variable in the multivariate normal, v_2, the second, ... and v_4, the final variable.

2. Create a death outcome for each patient from v_1, where a patient is recorded as dying if v_1 exceeds one and surviving if v_1 is one or less. The cut-off of one is chosen so that the overall death rate in the simulated datasets closely matches the original rate.

3. v_2 defines a patient's severity score based on her characteristics. No transform is necessary as it is essentially on a N(0,1) scale.

4. v_3 is related to the patient process, and is assumed to be the process score recorded for the patient with no transformation required. It is also used to create an average hospital process score, dependent on which institution the patient is randomly placed in.

5. v_4 defines which institution the patient is admitted to. This variable also allows us to build-in the hospital bed effect. The 297 hospitals are ordered in terms of number of beds at each institution (in the original dataset), with the smallest being first. The total number of beds in the original data is calculated and each hospital's bed quota is converted into a proportion of the whole number of beds. These proportions are then linked with quantiles in a standard normal distribution. For example, if there were a total of 100,000 beds in the dataset and the smallest institution had 100 beds available, it would make up 0.1% of the total number and the associated standard normal cut-off point would be -3.09 (the "first" 0.1% of the standard normal is below this point). The other cut-off points are determined in a similar fashion, i.e. if the second hospital had 0.2% of the total number of beds, the 2nd cut-off point would be the 0.3% (0.1% + 0.2%) quantile of the standard normal (-2.75). v_4 is then used to define which hospital each patient is at. For example, if v_4 for patient i was -3.76, then patient i would be in the first hospital (from our example). If her score was -3.00, she would be placed in the second hospital because v_{4i} fell between -3.09 and -2.75. As the hospital's are ordered by size and v_4 has a set covariance structure with the other variables, correlations between bed size and the other variables can be built into the modelling.

6. Repeat steps 1 to 5 for as many simulation runs as time allows.

For each simulation run, I record five different correlations: patient death vs patient severity score; patient death vs patient process score; patient process score vs patient severity score; hospital bed numbers vs hospital process score; and IO hospital assessment (excess mortality) vs average hospital process score. The IO hospital assessment is created by using the non-model based IO approach (with shrinkage variance estimation) and is compared against the average process score at a hospital. This is the main correlation of interest and I am interested to see what effects are seen when the other four correlations are modified. These other correlations can be changed by changing the initial covariance matrix defined in the four-dimensional multivariate normal. There is not an exact correspondence between the input (i.e. defined by the covariance matrix) and the output (i.e. defined by the correlations seen in the simulations) correlations.

So mathematically, I have:

$$\begin{pmatrix} v^*_{1ij} \\ v_{2ij} \\ v_{3ij} \\ v^*_{4ij} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.0 & c_{ds} & c_{dp} & c_{db} \\ c_{ds} & 1.0 & 0.0 & 0.0 \\ c_{dp} & 0.0 & 1.0 & c_{pb} \\ c_{db} & 0.0 & c_{pb} & 1.0 \end{bmatrix} \right) \tag{10.1}$$

for patient i in simulation run j, where v^*_{1ij} is the unmodified death outcome (i.e. not converted into a 0/1 variable), v_{2ij} and v_{3ij} are the severity and process scores for patient i respectively and v^*_{4ij} is the variable that identifies patient i's institution (unmodified). c_{ds} represents the input covariance of patient death against patient severity score, c_{dp} is the covariance of patient death against patient process score, c_{db} is the covariance between patient death and the number of beds in that patient's hospital (this is a patient level correlation), and c_{pb} is the covariance between patient process score and the number of beds (patient level again). These input covariances are set so that the output correlations in the simulation runs are close to the correlations seen in the original Medicare dataset. After some iterative testing (i.e. run a small number of simulations and change the input covariances until the output correlations are near their target) c_{ds} is set to 0.76, $c_{dp} = -0.1$, $c_{db} = -0.12$ and c_{pb} becomes 0.34. c_{ps} and c_{bs} are noted to be near zero in the data. The output correlations seen in the original Medicare dataset are given in Table 10.1.

Term	**Correlation**
Death vs Severity	0.51
Death vs Process	−0.09
Process vs Severity	−0.05
Hosp. Process vs Hosp. Beds	0.34

Table 10.1: Medicare dataset: original correlations

I can now modify the input covariances to examine the effect on the output "hospital quality assessment z-scores (excess mortality) vs the average hospital process score" correlation. Table 10.2 shows the effect on the correspondence between the IO assessment scores and the true hospital process score, of changing the original dataset correlations. Each row of Table 10.2 represents results from running 25 simulations with the associated correlations.

Potentially this analysis is only a small part of a much larger study that could examine the effects of changing the characteristics of quality-assessment datasets and is principally included to give an example of the type of work that could be carried out on this topic. These initial results imply that the correlation between hospital size and hospital process seems to have the highest importance when IO assessment is used to attempt to recreate the hospital process scores. This can be seen using numerous methods (e.g., regression with the IO vs process correlation treated as the outcome variable and the remaining correlations used as x variables, or by simply looking at the correlations between the columns of Table 10.2).

IO Assess. vs H. Proc.	Death vs Severity Score	Death vs Process	Process vs Severity	H. Beds vs H. Proc.
-0.38 (0.017)	0.50 (0.003)	-0.20 (0.005)	-0.01 (0.005)	0.72 (0.005)
-0.35 (0.024)	0.40 (0.003)	-0.07 (0.005)	-0.01 (0.004)	0.71 (0.010)
-0.35 (0.024)	0.40 (0.003)	-0.07 (0.005)	-0.01 (0.004)	0.71 (0.010)
-0.32 (0.026)	0.50 (0.003)	-0.07 (0.005)	-0.01 (0.004)	0.73 (0.009)
-0.30 (0.016)	0.46 (0.003)	-0.07 (0.005)	-0.01 (0.004)	0.72 (0.005)
-0.30 (0.008)	0.33 (0.003)	-0.07 (0.005)	0.00 (0.004)	0.72 (0.011)
-0.28 (0.019)	0.50 (0.003)	-0.07 (0.005)	-0.01 (0.004)	0.64 (0.012)
-0.28 (0.021)	0.50 (0.003)	-0.13 (0.005)	-0.01 (0.005)	0.32 (0.008)
-0.27 (0.019)	0.50 (0.003)	-0.04 (0.005)	-0.01 (0.004)	0.74 (0.009)
-0.26 (0.033)	0.53 (0.003)	-0.07 (0.005)	-0.01 (0.005)	0.72 (0.006)
-0.26 (0.013)	0.50 (0.003)	-0.20 (0.005)	-0.01 (0.005)	0.26 (0.012)
-0.25 (0.033)	0.50 (0.003)	-0.07 (0.005)	-0.01 (0.005)	0.56 (0.016)
-0.23 (0.028)	0.50 (0.003)	-0.17 (0.005)	-0.01 (0.005)	0.38 (0.012)
-0.21 (0.015)	0.33 (0.003)	-0.07 (0.005)	0.00 (0.004)	0.31 (0.013)
-0.21 (0.024)	0.33 (0.003)	-0.07 (0.005)	-0.00 (0.004)	0.49 (0.017)
-0.20 (0.021)	0.40 (0.003)	-0.07 (0.005)	-0.01 (0.004)	0.33 (0.021)
-0.19 (0.019)	0.50 (0.002)	-0.10 (0.005)	-0.01 (0.004)	0.34 (0.008)
-0.18 (0.018)	0.50 (0.002)	-0.04 (0.005)	-0.01 (0.004)	0.42 (0.013)
-0.17 (0.010)	0.46 (0.003)	-0.07 (0.005)	-0.01 (0.004)	0.33 (0.019)
-0.17 (0.031)	0.50 (0.003)	-0.07 (0.005)	-0.01 (0.004)	0.47 (0.012)
-0.14 (0.015)	0.53 (0.003)	-0.07 (0.005)	-0.01 (0.005)	0.26 (0.010)
-0.14 (0.019)	0.50 (0.002)	-0.07 (0.005)	-0.01 (0.004)	0.30 (0.009)

Table 10.2: Process effects: modifying dataset correlations

A smaller or larger sample of individuals?

In the previous section I recreated a dataset with similar properties to the original Medicare dataset. 10,000 patients were simulated for each run as this was approximately the number of individuals in the original dataset. There is no restriction on the number of individuals that can be generated in the simulations and so this presents us with the opportunity to examine the effect of having a larger or smaller dataset on process or input/output results.

Imagine generating a dataset with 100,000 patients rather than 10,000 using the simulation method described in the previous section. The proportions in each hospital are similar to the original Medicare dataset because of the quantiles trick used to generate the hospital identifier. So now we have an imaginary dataset where 100,000 people were sampled rather than 10,000, i.e., we have a lot more information. I make the assumption that this 100,000 patient dataset is truth and each hospital now has a process score that is assumed to be its real score.

We can now taken samples (without replacement) from this large dataset, sampling approximately 50%, 40%, 30%, 20% and 10% of the patients, i.e., creating datasets ranging from around 50,000 to 10,000 patients. For each sampling level, we generate 20 datasets from the large one and these 20 sets can then be analysed.

There were two key areas of these 20 datasets that were focused on: how highly correlated the

hospital process scores from the sampled datasets were with the process scores from the 100,000 patient dataset; and, correlation between the true hospital process scores (larger dataset) and the excess mortality scores (IO z-scores) from the sampled dataset. Table 10.3 shows how these correlations change with varying sample sizes. Note, at the 10% level on average around 20 hospitals have no patients in due to sampling.

Proportion Taken	Correlation Between	
	True & Sample Process	True Process & Sample Excess Mortality
100	1.000 (NA)	-0.144 (NA)
50	0.998 (0.0002)	-0.144 (0.004)
40	0.963 (0.0083)	-0.136 (0.016)
30	0.906 (0.0170)	-0.132 (0.031)
20	0.816 (0.0219)	-0.126 (0.029)
10	0.632 (0.0385)	-0.116 (0.054)

Table 10.3: The effects of sample size

The true hospital process scores are reproduced fairly accurately when 50% or 40% of the original population are used but larger losses in accuracy are seen when the sampling proportion is 30% or below. When only 10% of the original patients are sampled, the hospital process scores seen in the sample have only a 0.63 correlation with the true process scores. The drop in accuracy in reproducing the hospital process scores using an IO approach is less dramatic. Using the full data, the corelation between the excessive mortality approach (i.e., IO z-scores) and the true process scores is -0.14 but this correlation only drops to -0.12 when 10% of the original population are sampled. However, the standard deviation on this -0.12 means that these correlations range from around -0.20 to -0.03.

This analysis highlights an area where further work could be completed using other datasets and possibilities for future research are extensive. The most surprising part of this work is that the IO z-scores are not dramatically affected when 90% of the data is "lost" and the gains when all the data is available for recreating the hospital process scores using an indirect approach are not massive.

10.3 OPTA Premiership ratings

Player assessment

The aim of this study is to examine the ability and consistency of strikers in Premiership soccer.

There are two principal ways to examine the ability of a player. The first relates to a "process" measuring method. This means that the player is watched and assessed by experts and given a rating based on his performances. The OPTA ratings carry out such a system of assessment and these statistics are explained in the following section.

The second method is an IO method. This is our standard method, where a number of inputs relating to a player's goalscoring chances are measured and the outcome for the player

is also recorded. These two components are then compared to see whether the player has performed better (or worse) than expected for that match, based on the adjustors.

I consider matches in the 2000/2001 Premiership season up to and including February 8th 2001.

OPTA explanation

A specially-trained person watches a match on video and uses a computer to log each action performed by every single player on the ball, his fouls and discipline, and key decisions made by the officials. There are currently 92 distinct actions and outcomes for players that range from different kinds of shots and passes, to tackles and blocks and from different kinds of fouls and yellow cards to saves made by the goalkeeper. Every close season, the list of actions is discussed to determine the value of each element and then new categories may be added. OPTA receive the referee's copy of the video on the morning after a game and then begin the analysis. The OPTA database is the most comprehensive record of Premiership player performance in existence and offers greater insight into the game for managers, the media and fans alike.

When a match is analysed, each player's actions are recorded. For each of these actions, a player earns or loses points e.g., a goal is worth up to 500 points, whereas a short pass in a player's own half only earns five points and a foul costs a player 50 points. Over the course of a game, a player will accumulate a total number of points to give him a Game Score.

Each player's points from the appropriate games are added to give him a total, which is then divided by the number of minutes played and then multiplied by 90 (minutes), to give him an average score per game played. This is his Index Score (see Table 10.4 for an example).

Opposition	Mins Played	OPTA Game Score
Sheffield Wednesday	90	1,112
Everton	90	1,313
Coventry City	90	1,413
Southampton	0	0
Blackburn Rovers	90	1,515
Wimbledon	71	1,616
Total	431	6,969
OPTA Index Score 6,969/431 x 90(mins) = 1,455		

Table 10.4: OPTA calculation example: Dennis Bergkamp.

Input/Output approach models

To test my methods, I need to use a binary outcome. A good choice is whether the player scores at least one goal in a match. Therefore, a player is deemed to have a successful match if he scores at least one goal (Score = 1), otherwise he is deemed to have failed (Score = 0).

Another potential outcome is that the player was "involved" in at least one goal. This can be recorded as follows: Success if a player scores at least once or a player is credited with an assist for at least one goal. This outcome variable is called "Scass".

There are a variety of possible inputs, i.e., adjustors that vary for each appearance by a player in a match and could affect his goal scoring chances. It is important to remember not to include variables that directly measure a player's ability. Potential variables include:

- Weather conditions;
- Team playing at home or away;
- Opponent's defensive record;
- How many minutes did the striker play for?
- Did he come on as a substitute?
- Match importance;
- Player injury status;
- How many important games has the player been involved in recently? Player fatigue;
- Quality of striker's team;
- Other match conditions relating to goal scoring.

Some of this information is relatively easy to collect and some would take a great deal of time and effort to find. The following variables have been identified as being easy to obtain:

- Number of goals in the game;
  This variable gives an idea of how easy it was to score in the match. Its presence in any input/output model means that weather conditions and other factors relating to that specific match can be taken into account.
- Home or away;
  It is easy to identify whether a player's team were home or away for each match.
- A variable to measure the opponent's defensive record;
  A number of variables could be used here. The first is based on the opponent's last seasons position in the Premiership. For this data that would be their position at the end of the 1999/2000 season. The second is to use the number of goals they conceded in the Premiership last season. The final potential measure is to use the number of goals conceded in the Premiership during this season (i.e., for the matches in the dataset). Obviously some imputation is required for the three promoted teams in methods one and two.
- Minutes played;
  This variable is easy to find and measure.

- Substitute;
  Given the number of minutes a player has played is taken into account, it is important to measure whether he came on as a substitute or played from the beginning. A player coming on is likely to be more fresher than a player who has played the whole match. Also if a player comes on with 10 minutes to go and his team are pushing forward for a equaliser, he has an increased chance of scoring solely due to match situation. This would not be the case at the start of a match.

- Quality of striker's team;
  This is a cause for debate as the quality of a striker's team is partly down to how good a player the striker is. This is a grey-area for inclusion into the adjustors.

These variables are not easy to measure:

- Match importance;
  This is a difficult variable to measure precisely and could be highly subjective. Given that the data are for the first half of a Premiership season, it is assumed (obviously not completely correctly) that every match has equal importance to the players.

- Player injury status;
  This could also be highly subjective and there is no clear way to measure this.

- Player fatigue;
  Same difficulties as player injury status.

A number of different models have the potential to measure a strikers ability. Table 10.5 gives a summary of the models that have been considered and analysed.

Model	s1	s1 +ass	s2	s2 +ass	s3	s3 +ass	s4	s4 +ass
Outcome	Score	Scass	Score	Scass	Score	Scass	Score	Scass
Home Or Away	✓	✓	✓	✓	✓	✓	✓	✓
Minutes Played	✓	✓	✓	✓	✓	✓	✓	✓
Came On As Substitute	✓	✓	✓	✓	✓	✓	✓	✓
Goals In Match	✓	✓	✓	✓	✓	✓	✓	✓
Last Season's Goals Scored (Team)	✓	✓	×	×	×	×	×	×
Last Season's Position (Team)	×	×	✓	✓	×	×	×	×
Last Season's Goals Conceded (Opposition)	✓	✓	×	×	×	×	×	×
This Season's Goals Scored (Team)	×	×	×	×	✓	✓	×	×
This Season's Goals Conceded (Opposition)	×	×	✓	✓	✓	✓	✓	✓

Table 10.5: The potential OPTA models.

Results

The hope of this analysis is that the OPTA rating can be reproduced to a reasonable level using an input/output approach. This would mean that a less intensive method could be used to assess player performances rather than trying to directly measure the player's ability by watching him for 90 minutes in each match.

There are obvious differences of focus between the two methods that could cause variations in striker quality assessment as OPTA doesn't just concentrate on goal scoring exploits but also looks at, for example, how effective the player is at passing or tackling. Also in the IO approach, a player is rewarded equally if he scores 1 goal or 12 goals in a single match.

As I have shown before, the random-effects(RE) set-up allows for both continuous and higher-level predictors. In this set-up, match performance is nested within a player, who is nested within a team. Players are able to move teams, but this happens rarely in this data. There also exists a cross-nested structure with a number of players appearing within a specific match. In our current model, we have a two-level structure (performances within players) that has team predictors.

In a perfect world where players did not move clubs, the fixed-effects (FE) quality assessment method would be unable to analyse this data as the design matrix would not be linearly independent. This is because the team level predictors could be formed from a combination

of the player identifier dummies. A RE quality assessment method can deal with this problem and thus, as stated previously (Section 8.1), can be more flexible than the FE method. This is a cause for concern when moving from RE to FE models.

Tables F.1 - F.4 show the results for each model examining the RE equivalent to the $\hat{D}_i$s (the implied quality of a striker after taking into account the adjustors), i.e., the a_is, his associated z-scores and his correlation with the other model results and the OPTA ratings. The RE models are all based on main effects only.

These results indicate that there is very little difference (in terms of correlation) between the a-values and their associated z-scores. This is because the number of appearances for each striker is usually less than 15 and these numbers do not vary much between strikers. This means that when the numeric information is combined with a striker's success rate, the standard error for the as does not vary much from player to player. This similarly between the as and their z-scores exists for all models. Table F.4 shows this pattern. For ease, I will focus on the z-scores rather than the a-scores.

Which model is the best match with the OPTA ratings? Table F.3 shows that model four compares the best with the OPTA ratings. Concentrating on the z-scores, in both the Score and Scass outcome models, the correlation with the OPTA ratings is 0.59. The other three models have correlations ranging between 0.38 and 0.49. Table F.3 also shows that there is only a small difference in results when deciding between using the Score outcome or the Scass outcome for modelling player performance. The correlation with OPTA does not change (or changes very little) when using either Score or Scass outcomes. In individual models, the correlation between Score and Scass z-scores is always around 0.86. This seems to infer that the z-scores do change if a score/assist outcome is used rather than just a scoring outcome, but not with regard to correlation with OPTA. This Score versus Scass difference should be monitored. The correlation between model z-scores, when the Score outcome is used, is high in all cases. The lowest correlation between models exists between models three and four, with the highest between models one and two. All correlations vary between 0.91 and 0.98. The matrix is shown in Table F.2.

For the remaining OPTA analysis, we focus on trying to recreate the OPTA rating using an input/output approach. For ease, we choose model 4s to work with as this seems to have the highest link with the OPTA system. This model does not include an adjustment for team quality and so does not "punish" players in higher-standard teams.

With only a scoring outcome as success, we can use a FE quality assessment method to examine the link between RE and FE results. A concern when moving from a RE to FE method is dealing with predictors at the higher levels. As mentioned in Section 8.1, high-level predictors can cause the design matrix for FEs to become linearly dependent, mainly due to the need to include high-level dummy variables (in this case, player identifiers). In this dataset, a number of players move between clubs which means that linearly independence is not lost when team predictors are included in the adjustment method. It does however mean that the design matrix could contain some highly correlation columns and this must be watched for. For model four, we include a team-level predictor only for the opponent's goals conceded. There is

no direct link between a player and which opponents he plays and so the design matrix linear independence is not affected here. The results below show how the RE z-scores change when a FE set-up is used and when more than just main-effects are used in the FE world.

When only main-effects are used in a FE world, the FE z-score correlate highly with the RE z-scores. The correlation is 0.99.

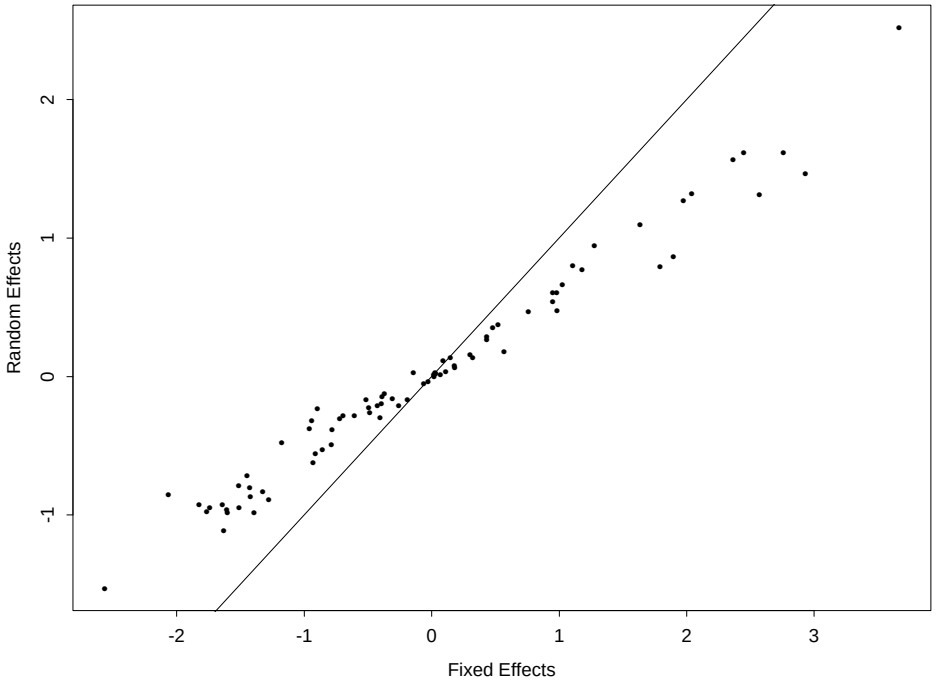

Figure 10-3: The RE and FE z-scores produced from a main effects only model.

Figure 10-3 shows how highly correlated the z-scores are. The solid line indicates the 45 degree line for the graph and clearly shows how the RE z-score are shrunk back towards the centre in comparison to the FE results.

Including additional interaction terms into the model provides greater model information but as we have seen previously, does not always provide any additional information in the z-scores. This is the case in model 4's data, where including all two-way interactions produces z-scores that are highly correlated with the main-effect only z-scores (cor = 0.993). With three-way interactions, the correlation with main effects decreases by a very small amount producing a correlation of 0.988, four-way interaction produces 0.987 and all interactions included still correlates at 0.987 with the original main effects only FE z-scores. It seems clear that only main-effects are required to produce consistent results. The comparison between FE and RE results are consistent with the findings from previous FE and RE comparisons.

The original non-model-based method would not be able to deal with model four as some of the predictors are continuous: this season's opponents goals conceded; minutes played; and

goals in the match. In order to perform the non-model-based approach, these variables will have to be converted. This conversion will provide the model with less information than in the FE and RE quality assessment approaches but the approach itself is viewed to be simpler to follow and so has other advantages highlighted previously.

The conversions are as follows:

- This season's opponents goals conceded;
  This has been converted into a three category variable.
  Category one opponents are those who have conceded less than 30 goals this season.
  Category two opponents have conceded less than 40 goals but greater than or equal to 30 goals this season. Category three opponents are those who have conceded greater than or equal to 40 goals this season.

- Minutes played;
  Players have been separated into 4 groups.
  Category one players played less than 25 minutes of a match.
  Category two players have played more than 24 minutes but less than 65 minutes.
  Category three players have played more than 64 minutes but less than 90 minutes.
  Category four players have played the whole match (90mins).

- Goals in a match; Matches have been separated into three types;
  Low scoring matches - matches where two or less goals were scored. Average scoring matches - matches where three, four or five goals were scored. High scoring matches - matches where more than five goals were scored.

Therefore, the non-model-based analysis has five variables which produce 92 different match appearance types for 80 players. The non-model-based z-scores produced correlated very highly with the FE results for both a main effects set-up (0.96) and a fully saturated set-up(0.96). This indicates that very little information was lost when the continuous variables were converted into categorical variables for the modelling process.

In all cases (RE, FE and non-model-based) the z-scores created were all similar. This tallies well with previous findings.

OPTA vs Input/Output

For ease, the RE z-scores are considered for the IO approach as all IO method produce similar z-scores. Figure 10-4 shows clearly that the two rating systems are fairly highly correlated (0.58). Those at the extremes of the two scales, in general, appear at the same extremes in the alternative system. Points to note:

- With so few appearances for each player, the z-scores do not indicate a quality difference for any single player;

- Both OPTA and IO identify some players who are placed in "unusual" rating positions, in comparison to the view of the general football watching public.

- The OPTA system takes no account of measurement error, i.e., the ratings similar to just using the equivalent $\hat{D}_i$'s in the HEFCE dataset. No effort is made to establish how variable the OPTA score is. The IO approach tries to take this into account.
- Following on from the last point, OPTA could be seen as more of a form guide rather than an ability guide.
- OPTA focuses more on player involvement in the team rather than goal scoring efforts (although this does play a large role).

The player assessments from this analysis are given in Tables F.5 - F.6.

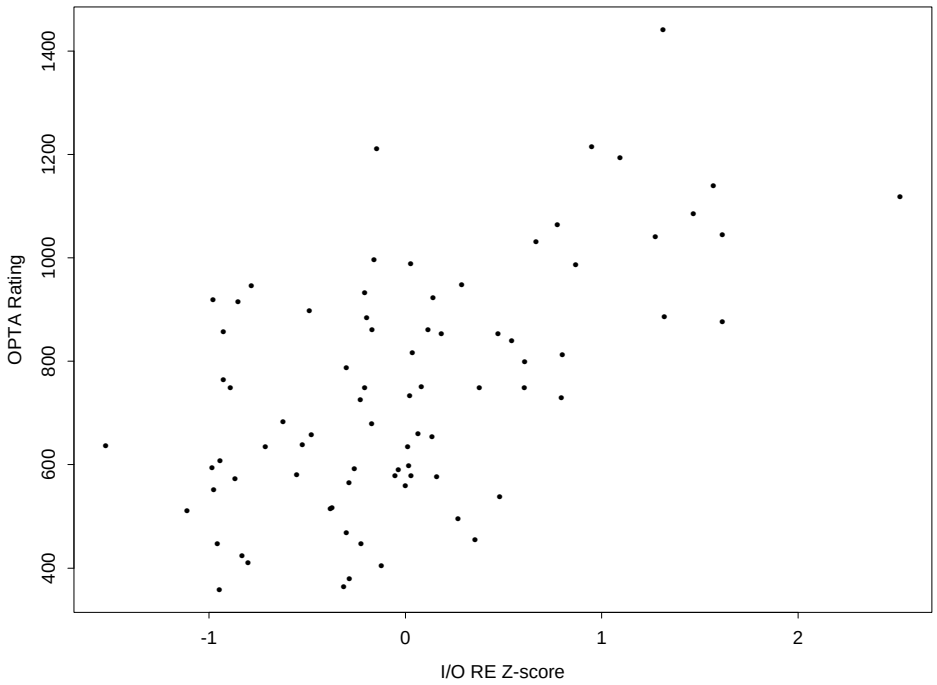

Figure 10-4: The link between the OPTA ratings and the IO approach.

10.4 Deciding on a z-score cut-off point

The choice of where to place the quality assessment z-score cut-off is not arbitrary. I have used two cut-offs for investigating quality differences in the universities' case study: the HEFCE cut-off and the Bonferroni cut-off.

The HEFCE approach is considered as it uses both a practical and statistical approach for establishing extremal institutions. Combining achieving a $|z|$-score ≥ 3.00 and an absolute difference between O and E rates of at least 3% means that the general public gain a good feel of why a university is unusual (i.e., they can see that being 4 percentage points worse

than expected is bad but might query why an institution whose rate is only 0.5% worse than expected has been identified as being poor) and there is some statistical confirmation as well (i.e., the z-scores). One of the key issues is how to decide whether 3.00 is a good choice of z-score cut-off?

In the Bonferroni approach (discussed by Miller (1966)), the following idea is presented: If a single hypothesis is true (e.g., university i is significantly different from 0 (or "an average university")), a significant difference ($P \leq 0.05$) will be seen by random chance once in 20 trials. When there are 20 independent trials (and the original hypothesis applies in each case), the chance that at least one test is significant by chance is now no longer 0.05, but increases dramatically to 0.64. Bonferroni modifies the z-score cut-off choice to attempt to adjust for this multiple comparisons problem. Perneger (1998) however does point out that the Bonferroni multiple comparisons method is not an excellent solution in many problems. This paper points out that Bonferroni is designed for situations where all hypothesis are true simultaneously, which obviously isn't the case for our case-studies (e.g., not all universities will be significantly difference from the average). Bonferroni also increases the likelihood of type II errors so that some true institutional differences from the average will be missed. Perneger recommends that the best way to deal with a multiple comparisons problems is to simply describe the tests of significance that have been performed and explain why they were carried out.

Either method produces a lot of issues on the choice of z-score cut-off. So how do we decide on a good choice for the cut-off? The type of problem can be applied in a decision analysis approach. Smith (1988) states that the goal of decision analysis is "the identification of a decision which is expected to best satisfy the stated objectives". There are four steps in an analysis of this type:

1. Establish the objectives;
2. Set-down all the decisions possible;
3. Quantify any problem uncertainty; and
4. Measure the costs associated with each possible decision.

The objectives of my HE case-study are not clear and further work with HEFCE would need to be carried out in order to establish firm parameters for both the objectives of the study and the resources available. In general terms, HEFCE's aims include identification of unusual institutions and further investigations into why these universities perform as they do. The variable (or decision) we can adjust or modify is where to set the z-score cut-off point. Consider two extremes:

- Setting the z-score cut-off point at 1.00 would identify far too many universities as unusual and so the outcomes (or cost) of this would be: a large number of universities would have their reputations unfairly questioned and the reliability of the HEFCE tables would drop; a large amount of money would be wasted examining universities whose methods were not unusual and did not need to be studied; and it would become very difficult to disseminate good (or bad) practice.

- Putting the z-score cut-off point to, say, 10.00 would mean very few establishments were identified as unusual meaning that bad (or good) universities pass through the net and are classed as average. This would mean that too much time and resources would be used on looking at the routines of a couple of institutions, rather than getting a better picture by including other appropriate establishments, and it would also mean that a number of unusually good and bad universities were not recognised as such.

A large proportion of this thesis is dedicated to point three of this process, problem uncertainty and this would also have to be taken into account. Barnetson and Cutright (2000) examines performance indicators as conceptual technologies, i.e., the writers try to study how indicators are chosen and their effects on institutions. Barnetson identifies six key areas relating to performance indicators: value; definition; goals; causality; comparability; and normalcy. These factors along with many others would have to be considered in any decision analysis involving HE institutions.

It is evident that there is a lot of give and take in this problem and a decision analysis approach would help to formalise the process. To set this problem up in a more mathematical manner, you require a loss function $L(d, \theta)$ indicating how much you will lose if decision $d \in D$ was made and the outcome was θ. The function needs to be defined for all values of $d \in D$ and θ. Then for each decision $d \in D$, the probability mass function $p(\theta|d)$ is inferred with $p(\theta|d) \geq 0$ and $\sum_\theta p(\theta|d) = 1$. The aim is then to find the decision d^* that minimises $L(d) = \sum_\theta L(d, \theta)p(\theta|d)$. I intend to pursue formalisation along these lines in future research.

Chapter 11

Summary and further work

11.1 Overall summary

In this thesis I have principally used data based on all higher education students starting in Autumn 1996. The dataset contained 284,399 individuals separated into 165 UK institutions and is used to illustrate potential quality assessment approaches for multilevel data. I have shown that the original HEFCE approach can be mathematically formalised and have developed the method to enable a large number of potential confounding factors to be adjusted for. After providing sensible forms of the observed (O_i) and expected (E_i) success rates at each institution of interest (i), a key part of my work gives a statistically valid form of the standard error of the difference between these two rates ($D_i = O_i - E_i$). A successful calibration method for the standard error of this difference has also been shown and I have discovered that a shrinkage (γ) approach provides the best technique for single cell variance estimation. Many academic papers have examined the unexplained institutional variation (D_i) but as this thesis has developed a sensible method for finding the standard error of this difference (and thus the associated z-score, z_i), I have provided methods for examining the variation of the z_i, i.e., at a level below the D_i. The more popular ratio ($R_i = \frac{O_i - E_i}{E_i}$) approach for institutional assessment has been shown to be statistically flawed using certain R_i standard error estimation techniques.

I have shown a model-based equivalent to indirect standardisation does exist, i.e., a fixed-effects fully-saturated regression model with a linear link function (Equation 3.6). I also highlight how general statistical methods can be easily used to fit a regression with the required weighted α restriction that provides important institutional assessments and shows the relationship between fixed- and random-effects approaches. Random-effects methods are a relatively new advance in quality assessment and this thesis helps to show more clearly the modelling options for quality-assessment and how these different options are linked to each other. I have also shown that non-linear regression (fixed- or random-effects) isn't always the best option when dealing with a binary outcome variable as the method can fail in some quality assessment cases. The use of a linear link with a binary outcome variable breaks some statistical assumptions but can offer a possible assessment approach when non-linear models are not valid.

The methods provided can be used in a variety of situations and they work well with a varying level of sparseness in the data-grid. The methods have been shown to be well calibrated in: the Small World (where there is a 10% level of sparseness, 2 cells missing out of 20); the Medium World (16% sparseness, 56 cells out of 360); the Published World (57% sparseness, 25461 cells out of 44880); and the Big World (where there is a massive 96% of cells with no individuals in, 2,806,945 cells out of 2,936,835). Very few alternative quality assessment methods are capable of dealing with a high degree of sparseness.

The issue of differential effectiveness is a major problem with league tables. In the thesis, I suggest a statistical and practical reporting system for universities that offers one solution to the identification of differential effectiveness within institutions. A statistical method for league-table assessment comparing different years has also been developed and examined. A repeated-measures set-up is used where institutions act as a constant factor across the years but institutional individuals change year on year.

A complaint about some adjustment processes is that the expected rate for institutions is partially based upon that institution's own performance, i.e., if the institution is very large then its expected value is pulled closer to its observed value due to the gravity of the establishment. I present an alternative studentized way of calculating an institution's expected rate which is not based on that institution's own performance.

An extensive gold-standard analysis is given which demonstrates some difficulties with both input/output and process quality assessment techniques. This gold-standard analyses show that the link between IO and process techniques varies depending on the case-study in question and other key correlations within that study. It has been shown that recreating institutional process from outcomes can be nearly impossible in a variety of cases, i.e., regardless of how well the establishment does its job, its outcomes may not be dramatically affected.

11.2 Further Work

This work could be extended in several directions and the following is a list of potential areas of future study.

- A sensible and statistical solution to dealing with a non-linear link function in a binary response quality assessment model is required. The difficulty with a non-linear link is highlighted in Section 8.2 which involves a failure of the modelling process when a single PCF category or institution has a 100 or 0 % success rate. Some potential lines of enquiry are given in the section and are summarised as follows: attach an informative prior to the problematic parameter of interest; modify the dataset; use a combination of link functions to replace the non-linear link function; or use a linear link function.

- The RE calibration needs further work to establish reasonable z-score cut-offs for RE analysis, especially when a reduced number of interactions are used. Section 8.1 highlights the areas that need more extensive work.

- The gold standard work (Section 10) can be extended in a number of ways. Further gold standard datasets are required but due to the lack of good institutional process information, analysis of a similar nature to the Medicare (Section 10.2) and OPTA (Section 10.3) datasets could be difficult. There is also a need to expand on the simulated runs which looked at how the size of the sampling and data influenced the accuracy of the z- and process scores for institutions, and how the correlations within the quality assessment datasets affected institutional quality prediction. The number of simulation runs needs to be increased quite dramatically to produce better conclusions on these gold standard effects.

- With a large dataset (e.g., the Big World) and a large number of PCFs, it becomes very difficult to fit models that are not fully saturated but have a large number of interaction terms (i.e., models with up to x-way interactions where $x \geq 3$). Suitable methods have yet to be developed for these models in both fixed- and random-effects set-ups. The issue of a large number of interactions in models is a particular problem with random-effects modelling. Further thought is required to develop techniques to fit such models. One advantage to fitting these types of model would be that an extended version of Table 6.14 could be produced (Table 11.1):

Model Description	Incorrect University Status(Out of 165)	Misclassified
Full	4	2%
Seven-Way	?	?
Six-Way	?	?
Five-Way	?	?
Four-Way	?	?
Three-Way	?	?
Two-Way	13	8%
Main	21	13%

Table 11.1: Extended interaction effects in the Big World.

- The longitudinal analysis (Section 9.3) provides some insight into the absolute year-on-year change in an institutional performance, i.e., it looks solely at the change in a institution's performance regardless of what the underlying year-on-year trend is for all universities in the dataset. It should be possible to adapt the repeated measures construction to look at relative institutional change in performance between years, i.e., after taking account of what the overall universities' trend is, how has a specific institution varied between the years in question?

- The list of PCFs for student progression has not been exhausted. This thesis focused more on the methods for quality assessment rather than the dataset itself but further work could be completed on producing a more complete list of PCFs and then identifying

which ones could sensible be include. For example, whether the student came through clearing might be considered but the data is currently not available to HEFCE.

- If the appropriate information could be found, the decision analysis in Section 10.4 should be completed. This would provide a sensible guide on where the z-score cut-off point should be chosen for institutional quality assessment and the effects of such a choice. This is specific to this higher education problem but similar analyses can be carried out with alternative datasets.

- My case-study data is highly detailed, i.e., provides all the information to a individual level. In many cases, complete individual level is not available and aggregate data at the institutional level data is the only possible information option. Ecological and epidemiological studies are principally based on higher-level aggregate data and a number of papers have explored how to analyse this type of data (e.g., Wakefield and Salway (2001)). There is therefore the potential to examine the differences and play-off between studies that have a great detail of information (i.e., my case-study and methods) and studies with less information (i.e., aggregate studies and methods). This would involved aggregating the higher education data to an institutional level and then comparing the results from an aggregate technique with my original lower-level approaches.

Appendix A

The real results

Key: n is the entry class size; $\hat{O}$ is the observed university progression rate; $\hat{E}$ is the expected university progression rate, based on adjusting for all eight PCFs; $\hat{D}$ is the difference between the observed and expected university progression rates; $\hat{SE}$ is the non-model-based standard error for the $\hat{D}$s, using a $\gamma^{0.5}$ variance estimation approach; $\hat{z}$ is the inferred z-score based on the calculated SE; and Sig is the significance of the z-score, with the following rule: H is the HEFCE cut-off, where a university is marked as unusual if its $|D|$ is greater than 3% and the associated z-score is greater than 3.00; * marks excellent universities, which have a z-score greater than the positive Bonferroni cut-off; and ** marks poor universities, whose z-score is less than the negative Bonferroni cut-off. The Bonferroni cut-off in this example is based on 165 comparisons and is ± 3.61.

Inst	n	$\hat{O}$	$\hat{E}$	$\hat{D}$	$\hat{SE}$	$\hat{z}$	Sig
9	1031	0.83	0.89	-0.06	0.007	-8.11	** H
110	3658	0.80	0.84	-0.04	0.005	-8.00	** H
78	1728	0.81	0.87	-0.06	0.007	-7.97	** H
151	2981	0.80	0.84	-0.04	0.006	-7.31	** H
145	4115	0.84	0.87	-0.03	0.004	-6.79	**
154	3126	0.84	0.88	-0.03	0.005	-6.31	** H
7	2889	0.85	0.88	-0.03	0.005	-6.00	** H
64	1501	0.82	0.86	-0.04	0.007	-5.80	** H
101	639	0.85	0.91	-0.06	0.011	-5.63	** H
122	2519	0.82	0.85	-0.03	0.006	-5.49	** H
86	1836	0.88	0.91	-0.03	0.006	-5.12	** H
4	2205	0.86	0.89	-0.03	0.006	-4.82	**
25	748	0.81	0.86	-0.05	0.011	-4.30	** H
3	1113	0.85	0.88	-0.03	0.009	-3.99	** H
137	2192	0.91	0.94	-0.02	0.005	-3.92	**
147	2881	0.85	0.87	-0.02	0.005	-3.75	**
141	2496	0.88	0.89	-0.02	0.005	-3.56	

Table A.1: The real results: the worst universities.

Inst	n	$\hat{O}$	$\hat{E}$	$\hat{D}$	$\hat{SE}$	$\hat{z}$	Sig
8	2292	0.84	0.87	-0.02	0.006	-3.56	
139	2927	0.92	0.94	-0.02	0.005	-3.47	
62	2270	0.87	0.89	-0.02	0.006	-3.41	
54	405	0.90	0.93	-0.03	0.009	-2.73	
81	3395	0.84	0.85	-0.01	0.005	-2.69	
21	1435	0.86	0.88	-0.02	0.007	-2.60	
22	2009	0.91	0.92	-0.01	0.006	-2.57	
6	3238	0.86	0.87	-0.01	0.005	-2.54	
133	4343	0.94	0.95	-0.01	0.004	-2.48	
49	2525	0.85	0.86	-0.01	0.006	-2.19	
84	2427	0.86	0.88	-0.01	0.006	-2.08	
50	2209	0.92	0.93	-0.01	0.005	-1.97	
42	901	0.86	0.88	-0.02	0.009	-1.96	
165	216	0.88	0.91	-0.04	0.020	-1.96	
2	3314	0.92	0.93	-0.01	0.004	-1.91	
55	176	0.88	0.92	-0.04	0.019	-1.90	
58	162	0.88	0.91	-0.03	0.021	-1.55	
68	127	0.89	0.93	-0.04	0.023	-1.52	
45	844	0.89	0.91	-0.01	0.010	-1.49	
106	1983	0.86	0.87	-0.01	0.006	-1.48	
74	3981	0.85	0.86	-0.01	0.004	-1.46	
135	178	0.92	0.94	-0.02	0.016	-1.45	
10	116	0.94	0.97	-0.03	0.023	-1.42	
70	5990	0.88	0.88	0.00	0.004	-1.37	
40	2134	0.95	0.95	-0.01	0.005	-1.29	
31	853	0.89	0.90	-0.01	0.009	-1.05	
47	1979	0.94	0.94	-0.01	0.006	-1.01	
113	704	0.85	0.86	-0.01	0.011	-1.00	
90	3398	0.87	0.88	0.00	0.005	-0.95	
126	2522	0.85	0.86	-0.01	0.006	-0.94	
77	355	0.89	0.90	-0.01	0.015	-0.91	
69	4773	0.94	0.94	0.00	0.004	-0.87	
103	206	0.89	0.91	-0.02	0.018	-0.86	
162	355	0.90	0.91	-0.01	0.014	-0.81	
67	2412	0.94	0.94	0.00	0.005	-0.75	
73	594	0.90	0.91	-0.01	0.011	-0.73	
75	3085	0.82	0.82	0.00	0.005	-0.71	
138	98	0.94	0.96	-0.02	0.024	-0.69	
76	1062	0.91	0.91	0.00	0.008	-0.60	
107	2527	0.86	0.86	0.00	0.006	-0.59	
134	130	0.89	0.90	-0.01	0.023	-0.53	
140	121	0.98	0.99	-0.01	0.021	-0.49	
156	1357	0.97	0.97	0.00	0.006	-0.39	
28	3435	0.95	0.95	0.00	0.004	-0.37	
131	5013	0.89	0.89	0.00	0.004	-0.36	
15	231	0.90	0.91	-0.01	0.018	-0.36	
116	3414	0.95	0.95	0.00	0.004	-0.15	
102	1626	0.89	0.89	0.00	0.007	-0.13	
91	1561	0.92	0.92	0.00	0.006	-0.10	

Table A.2: The real results: bottom half of the middle ground.

Inst	n	$\hat{O}$	$\hat{E}$	$\hat{D}$	$\hat{SE}$	$\hat{z}$	Sig
127	142	0.89	0.89	0.00	0.020	-0.05	
112	65	0.91	0.91	0.00	0.036	-0.02	
155	2992	0.93	0.93	0.00	0.005	0.01	
94	1409	0.95	0.95	0.00	0.006	0.06	
52	1662	0.86	0.86	0.00	0.007	0.20	
104	2009	0.87	0.87	0.00	0.005	0.26	
13	2720	0.89	0.88	0.00	0.005	0.26	
51	4261	0.88	0.88	0.00	0.004	0.38	
130	595	0.92	0.91	0.00	0.010	0.43	
109	365	0.96	0.96	0.00	0.010	0.45	
160	165	0.98	0.97	0.01	0.017	0.46	
96	1223	0.90	0.90	0.00	0.008	0.47	
27	88	0.93	0.92	0.01	0.028	0.49	
82	198	0.89	0.88	0.01	0.018	0.61	
111	604	0.92	0.91	0.01	0.010	0.65	
33	55	0.80	0.77	0.03	0.047	0.69	
132	308	0.92	0.91	0.01	0.014	0.70	
92	995	0.91	0.91	0.01	0.008	0.72	
118	1344	0.95	0.94	0.00	0.007	0.73	
153	1745	0.92	0.92	0.00	0.006	0.73	
19	149	0.93	0.91	0.02	0.020	0.75	
53	1539	0.90	0.90	0.00	0.007	0.76	
12	832	0.92	0.91	0.01	0.009	0.78	
80	1470	0.90	0.89	0.01	0.007	0.78	
148	136	0.93	0.92	0.02	0.022	0.83	
88	369	0.93	0.92	0.01	0.012	0.89	
26	920	0.93	0.92	0.01	0.008	0.89	
34	1003	0.90	0.90	0.01	0.008	0.92	
43	2188	0.89	0.88	0.01	0.006	0.93	
98	3684	0.96	0.96	0.00	0.004	0.94	
161	254	0.90	0.88	0.02	0.016	1.00	
60	197	0.90	0.88	0.02	0.017	1.02	
38	1525	0.90	0.89	0.01	0.007	1.02	
23	1633	0.91	0.90	0.01	0.006	1.05	
152	3241	0.88	0.88	0.01	0.005	1.08	
128	420	0.91	0.89	0.01	0.013	1.09	
32	3545	0.95	0.94	0.00	0.004	1.10	
89	65	0.94	0.90	0.04	0.032	1.12	
124	2808	0.94	0.94	0.01	0.005	1.21	
143	112	0.93	0.90	0.03	0.025	1.23	
115	98	0.95	0.92	0.03	0.026	1.27	
18	2850	0.96	0.96	0.01	0.004	1.32	
157	1199	0.95	0.94	0.01	0.007	1.37	
79	3443	0.98	0.97	0.01	0.004	1.46	
146	1232	0.94	0.93	0.01	0.007	1.48	
85	1211	0.95	0.94	0.01	0.007	1.48	
150	499	0.98	0.96	0.01	0.010	1.49	
29	2941	0.89	0.88	0.01	0.005	1.53	
24	2726	0.88	0.87	0.01	0.005	1.54	
39	679	0.99	0.97	0.01	0.007	1.57	

Table A.3: The real results: top half of the middle ground.

Inst	n	$\hat{O}$	$\hat{E}$	$\hat{D}$	$\hat{SE}$	$\hat{z}$	Sig
158	1679	0.92	0.91	0.01	0.006	1.60	
41	2047	0.87	0.86	0.01	0.006	1.63	
142	902	0.96	0.95	0.01	0.008	1.77	
16	1681	0.93	0.92	0.01	0.006	1.90	
72	963	0.92	0.90	0.02	0.008	1.94	
136	2353	0.93	0.92	0.01	0.005	1.99	
114	429	0.91	0.89	0.02	0.012	2.04	
87	1954	0.94	0.93	0.01	0.005	2.11	
105	172	0.94	0.90	0.04	0.019	2.15	
59	964	0.94	0.93	0.02	0.008	2.29	
95	92	0.97	0.91	0.06	0.025	2.33	
36	210	0.95	0.91	0.04	0.017	2.36	
37	456	0.94	0.92	0.02	0.010	2.38	
56	1125	0.94	0.92	0.02	0.007	2.45	
61	2499	0.97	0.96	0.01	0.004	2.49	
163	3611	0.90	0.89	0.01	0.004	2.56	
65	3009	0.90	0.89	0.01	0.005	2.67	
14	2917	0.99	0.98	0.01	0.004	2.74	
120	239	0.95	0.92	0.04	0.013	2.86	
66	723	0.91	0.89	0.03	0.009	2.88	
93	106	0.98	0.92	0.06	0.021	2.91	
99	2315	0.89	0.87	0.02	0.005	2.97	
11	222	0.98	0.93	0.04	0.015	2.98	
83	2092	0.95	0.93	0.02	0.005	3.18	
100	2468	0.97	0.96	0.01	0.005	3.20	
117	433	0.93	0.89	0.04	0.011	3.20	H
48	2708	0.89	0.88	0.02	0.005	3.27	
1	6831	0.90	0.89	0.01	0.003	3.37	
97	2844	0.89	0.87	0.02	0.005	3.42	
108	3633	0.89	0.88	0.01	0.004	3.43	
5	289	0.94	0.89	0.05	0.015	3.44	H
125	2210	0.89	0.87	0.02	0.006	3.48	
149	507	0.96	0.93	0.03	0.009	3.48	H
119	648	0.93	0.90	0.03	0.010	3.56	H
20	3482	0.90	0.88	0.02	0.005	3.57	
30	3819	0.91	0.90	0.01	0.004	3.57	
63	1023	0.93	0.90	0.03	0.008	3.59	
35	1975	0.95	0.93	0.02	0.005	3.66	*
129	3559	0.95	0.93	0.01	0.004	3.70	*
123	1445	0.93	0.91	0.03	0.006	3.99	*
159	3186	0.89	0.87	0.02	0.005	4.10	*
164	1685	0.95	0.93	0.02	0.006	4.16	*
57	1751	0.90	0.87	0.03	0.006	4.53	*
17	2294	0.96	0.94	0.02	0.005	4.57	*
121	949	0.95	0.91	0.04	0.008	4.88	* H
144	1173	0.95	0.91	0.04	0.007	5.37	* H
71	913	0.96	0.91	0.04	0.008	5.65	* H
44	3262	0.95	0.93	0.03	0.004	7.10	*
46	3573	0.91	0.88	0.04	0.004	8.63	* H

Table A.4: The real results: the best universities.

Appendix B

Omitting a single PCF

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	14	2	0	12	1	0
OK	2	130	0	0	142	0
Good	0	5	12	0	2	8
(False Neg. %)	(12.5)		5.5	(0.0)		1.8
Overall Error %						

Table B.1: Low HE participation: pseudo–R^2 with progression .019.

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	15	2	0	12	1	0
OK	1	131	0	0	138	1
Good	0	4	12	0	6	7
(False Neg. %)	(6.3)		4.2	(0.0)		4.8
Overall Error %						

Table B.2: Parental occupation: pseudo–R^2 .004.

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	15	9	0	12	4	0
OK	1	111	2	0	127	3
Good	0	17	10	0	14	5
(False Neg. %)	(6.3)		17.6	(0.0)		12.7
Overall Error %						

Table B.3: Entry qualification: pseudo–R^2 .049.

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	15	4	0	12	1	0
OK	1	116	2	0	129	0
Good	0	17	10	0	15	8
(False Neg. %)	(6.3)		14.5	(0.0)		9.7
Overall Error %						

Table B.4: Subject of study: pseudo–R^2 .009.

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	16	1	0	12	1	0
OK	0	133	0	0	140	0
Good	0	3	12	0	4	8
(False Neg. %)	(0.0)		2.4	(0.0)		3.0
Overall Error %						

Table B.5: State school: pseudo–R^2 .021.

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	15	2	0	11	1	0
OK	1	129	4	1	143	0
Good	0	6	8	0	1	8
(False Neg. %)	(6.3)		7.9	(8.3)		1.8
Overall Error %						

Table B.6: Year of program: pseudo–R^2 .001.

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	16	2	0	12	0	0
OK	0	128	0	0	142	1
Good	0	7	12	0	3	7
(False Neg. %)	(0.0)		5.5	(0.0)		2.4
Overall Error %						

Table B.7: Age: pseudo–R^2 .020.

	True Status Using Cutoff					
	3.61			HEFCE		
	Bad	OK	Good	Bad	OK	Good
Classified	16	137	12	12	145	8
Bad	16	2	0	12	0	0
OK	0	131	0	0	141	1
Good	0	4	12	0	4	7
(False Neg. %)	(0.0)		3.6	(0.0)		3.0
Overall Error %						

Table B.8: Gender: pseudo–R^2 .004.

	True Status Using Cutoff					
	3.61			HEFCE		
Classified	Bad	OK	Good	Bad	OK	Good
Bad	9.2	1.8	0.0	7.2	0.7	0.0
OK	0.5	76.4	0.6	0.1	83.5	0.5
Good	0.0	4.8	6.7	0.0	3.7	4.4
Overall Error		7.7			4.9	

Table B.9: Overall: averaging across all eight omitted PCFs.

Appendix C

Bootstrap results

Inst	n	True Status	True z	Bootstrap z	Bootstrap SD	Times Identified As Good	Times Identified As Average	Times Identified As Bad	Valid Sims
1	6831	G	9.13	9.58	1.5	100	0	0	1000
2	3314	OK	-1.06	-1.14	1.1	0	99	1	1000
3	1113	B	-4.02	-4.53	1.1	0	21	79	1000
4	2205	B	-4.48	-4.94	1.1	0	12	88	1000
5	289	OK	3.18	4.21	1.7	63	37	0	998
6	3238	B	-4.20	-4.52	1.1	0	22	78	1000
7	2889	B	-6.64	-7.11	1.1	0	0	100	1000
8	2292	B	-3.65	-3.90	1.1	0	42	58	1000
9	1031	B	-3.94	-4.33	1.2	0	27	73	1000
10	116	OK	-1.49	-1.54	1.1	0	97	3	1000
11	222	OK	3.15	6.88	8.7	84	16	0	894
12	832	OK	1.08	1.45	1.4	7	94	0	1000
13	2720	OK	-0.80	-0.88	1.2	0	99	1	1000
14	2917	OK	3.23	3.58	1.9	50	50	0	1000
15	231	OK	-0.39	-0.42	1.3	0	100	0	1000
16	1681	OK	2.21	2.73	1.2	22	78	0	1000
17	2294	G	5.30	6.34	1.3	99	1	0	1000
18	2850	OK	1.71	1.95	1.1	7	93	0	1000
19	149	OK	0.63	1.08	1.7	5	95	0	1000
20	3482	G	4.00	4.33	1.2	72	28	0	1000
21	1435	B	-4.01	-4.47	1.2	0	22	78	1000
22	2009	OK	-1.75	-2.01	1.1	0	92	8	1000
23	1633	OK	0.90	1.13	1.3	3	97	0	1000
24	2726	OK	0.73	0.81	1.2	1	99	0	1000
25	748	B	-4.37	-4.98	1.2	0	11	89	1000
26	920	OK	1.40	2.02	1.5	13	87	0	1000
27	88	OK	0.82	0.98	1.2	2	98	0	1000
28	3435	OK	0.32	0.39	1.1	0	100	0	1000
29	2941	OK	0.83	0.93	1.2	1	99	0	1000
30	3819	G	4.49	4.94	1.3	86	14	0	1000
31	853	OK	-0.64	-0.65	1.3	0	99	1	1000
32	3545	OK	1.95	2.13	1.1	8	92	0	1000
33	55	OK	-0.68	-0.58	1.2	0	100	0	1000

Table C.1: Bootstrap results 1.

Inst	n	True Status	True z	Bootstrap z	Bootstrap SD	Times Identified As Good	Average	Bad	Valid Sims
34	1003	OK	1.07	1.17	1.2	3	97	0	1000
35	1975	G	3.82	4.45	1.1	79	21	0	1000
36	210	OK	2.60	2.77	1.1	22	79	0	1000
37	456	OK	2.63	3.22	1.4	38	63	0	1000
38	1525	OK	0.59	0.68	1.4	2	98	0	1000
39	679	OK	1.60	2.08	2.6	26	74	1	1000
40	2134	OK	-0.59	-0.62	1.2	0	100	0	1000
41	2047	OK	1.29	1.44	1.2	4	96	0	1000
42	901	OK	-2.38	-2.68	1.2	0	78	23	1000
43	2188	OK	1.20	1.39	1.3	4	96	0	1000
44	3262	G	7.39	8.23	1.4	100	0	0	1000
45	844	OK	-1.13	-1.22	1.2	0	98	2	1000
46	3573	G	8.83	9.65	1.4	100	0	0	1000
47	1979	OK	-0.17	-0.16	1.1	0	100	0	1000
48	2708	OK	2.09	2.33	1.2	16	84	0	1000
49	2525	OK	-2.04	-2.14	1.2	0	88	12	1000
50	2209	OK	-1.49	-1.78	1.2	0	94	6	1000
51	4261	OK	-0.69	-0.73	1.2	0	99	1	1000
52	1662	OK	0.24	0.29	1.2	0	100	0	1000
53	1539	OK	1.11	1.42	1.3	5	95	0	1000
54	405	OK	0.47	0.65	1.8	5	94	1	1000
55	176	OK	-2.13	-2.33	1.2	0	87	13	1000
56	1125	OK	2.17	2.65	1.3	22	78	0	1000
57	1751	OK	3.53	3.96	1.2	59	41	0	1000
58	162	OK	-1.08	-1.05	1.2	0	98	2	1000
59	964	OK	3.32	4.34	1.5	69	32	0	1000
60	197	OK	2.13	3.51	2.5	40	60	0	988
61	2499	OK	2.17	2.43	1.0	12	88	0	1000
62	2270	OK	-3.30	-3.60	1.2	0	50	50	1000
63	1023	OK	3.52	4.34	1.4	69	31	0	1000
64	1501	B	-5.93	-6.35	1.2	0	1	99	1000
65	3009	OK	3.34	3.72	1.3	53	47	0	1000
66	723	OK	3.01	4.22	1.8	62	39	0	1000
67	2412	OK	-0.69	-0.77	1.1	0	100	0	1000
68	127	OK	-1.41	-1.63	1.5	0	96	4	996
69	4773	OK	-0.26	-0.26	1.1	0	100	0	1000
70	5990	OK	-1.90	-2.01	1.1	0	92	8	1000
71	913	G	5.90	8.20	1.7	100	0	0	1000
72	963	OK	1.36	1.80	1.3	10	90	0	1000
73	594	OK	-0.52	-0.53	1.3	0	99	1	1000
74	3981	OK	-2.40	-2.50	1.2	0	82	18	1000
75	3085	B	-4.37	-4.64	1.2	0	19	81	1000
76	1062	OK	-0.46	-0.55	1.2	0	100	1	1000
77	355	OK	-0.32	-0.29	1.2	0	100	0	1000

Table C.2: Bootstrap results 2.

		True		Bootstrap		Times Identified As			Valid
Inst	n	Status	z	z	SD	Good	Average	Bad	Sims
78	1728	B	-8.45	-9.15	1.1	0	0	100	1000
79	3443	OK	3.37	3.62	1.5	50	50	0	1000
80	1470	OK	0.69	0.80	1.3	1	99	0	1000
81	3395	B	-4.51	-4.73	1.2	0	18	82	1000
82	198	OK	-0.69	-1.20	2.4	2	91	7	952
83	2092	OK	3.22	3.79	1.1	55	45	0	1000
84	2427	OK	-2.52	-2.75	1.2	0	77	23	1000
85	1211	OK	1.72	2.15	1.3	11	89	0	1000
86	1836	B	-4.50	-4.99	1.2	0	12	88	1000
87	1954	OK	2.30	2.68	1.2	21	79	0	1000
88	369	OK	0.76	0.94	1.3	2	98	0	1000
89	65	OK	1.42	2.02	3.6	12	88	0	997
90	3398	OK	-1.23	-1.20	1.2	0	98	2	1000
91	1561	OK	0.65	0.80	1.2	1	99	0	1000
92	995	OK	1.33	1.73	1.3	8	92	0	1000
93	106	OK	3.02	7.45	8.3	83	17	0	774
94	1409	OK	0.32	0.36	1.1	0	100	0	1000
95	92	OK	2.42	3.08	1.7	30	70	0	982
96	1223	OK	1.35	1.54	1.3	6	94	0	1000
97	2844	OK	1.51	1.66	1.2	5	95	0	1000
98	3684	OK	1.95	2.18	1.0	9	91	0	1000
99	2315	OK	3.40	3.77	1.3	54	46	0	1000
100	2468	G	3.93	4.42	1.0	77	23	0	1000
101	639	B	-5.15	-5.74	1.1	0	3	97	1000
102	1626	OK	0.37	0.54	1.2	1	99	0	1000
103	206	OK	-0.77	-0.88	1.7	1	97	3	996
104	2009	OK	2.83	3.19	1.9	41	59	0	1000
105	172	OK	2.29	3.13	1.8	31	69	0	999
106	1983	OK	-0.23	-0.18	1.4	0	100	1	1000
107	2527	OK	0.28	0.33	1.3	1	99	0	1000
108	3633	OK	-0.41	-0.40	1.1	0	100	0	1000
109	365	OK	0.87	1.07	1.9	9	91	1	1000
110	3658	B	-10.60	-11.27	1.2	0	0	100	1000
111	604	OK	0.11	0.17	1.2	0	100	0	1000
112	65	OK	0.04	0.18	1.2	1	99	0	996
113	704	OK	-1.11	-1.26	1.4	0	96	4	1000
114	429	OK	1.55	2.79	2.4	30	70	0	998
115	98	OK	1.50	1.60	1.2	6	94	0	999
116	3414	OK	0.14	0.13	1.1	0	100	0	1000
117	433	OK	2.50	3.24	1.5	39	62	0	1000
118	1344	OK	0.95	1.17	1.2	3	97	0	1000
119	648	G	4.05	5.77	1.8	91	9	0	1000
120	239	OK	2.68	4.26	3.0	55	45	0	993
121	949	G	6.25	8.32	1.6	100	0	0	1000
122	2519	B	-6.26	-6.73	1.2	0	1	99	1000
123	1445	G	4.16	5.32	1.5	89	12	0	1000

Table C.3: Bootstrap results 3.

Inst	n	True Status	True z	Bootstrap z	Bootstrap SD	Times Identified As Good	Average	Bad	Valid Sims
124	2808	OK	1.39	1.55	1.1	2	98	0	1000
125	2210	OK	2.35	2.67	1.2	22	78	0	1000
126	2522	OK	-1.69	-1.75	1.2	0	94	6	1000
127	142	OK	-0.95	-1.63	2.5	2	82	16	908
128	420	OK	1.02	1.40	1.5	8	92	0	1000
129	3559	G	4.19	4.60	1.1	81	20	0	1000
130	595	OK	0.82	1.29	1.6	8	92	0	1000
131	5013	OK	-1.04	-1.11	1.1	0	99	1	1000
132	308	OK	0.92	1.48	1.7	9	91	0	1000
133	4343	OK	-2.23	-2.36	1.1	0	87	13	1000
134	130	OK	-0.78	-1.08	1.9	1	94	4	948
135	178	OK	-1.01	-1.18	1.6	1	95	4	998
136	2353	OK	1.89	2.14	1.2	11	89	0	1000
137	2192	OK	-3.32	-3.71	1.1	0	45	55	1000
138	98	OK	-0.18	0.01	1.6	2	98	0	972
139	2927	OK	-3.23	-3.57	1.0	0	50	51	1000
140	121	OK	-0.51	-0.50	1.0	0	100	0	1000
141	2496	B	-4.26	-4.66	1.1	0	16	84	1000
142	902	OK	1.92	2.56	1.2	20	80	0	1000
143	112	OK	1.00	1.20	1.2	3	97	0	1000
144	1173	G	4.95	6.20	1.3	98	2	0	1000
145	4115	B	-6.02	-6.29	1.4	0	2	98	1000
146	1232	OK	1.86	2.30	1.3	15	85	0	1000
147	2881	OK	-3.41	-3.67	1.2	0	46	54	1000
148	136	OK	0.82	0.95	1.1	2	98	0	1000
149	507	G	5.04	5.86	1.4	95	5	0	1000
150	499	OK	1.35	1.90	1.3	10	90	0	1000
151	2981	B	-6.81	-7.22	1.2	0	0	100	1000
152	3241	OK	-0.56	-0.59	1.2	0	100	0	1000
153	1745	OK	1.92	2.17	1.2	12	88	0	1000
154	3126	B	-6.41	-6.73	1.2	0	1	99	1000
155	2992	OK	0.97	1.08	1.2	1	99	0	1000
156	1357	OK	-0.09	-0.05	1.2	0	100	0	1000
157	1199	OK	1.96	2.50	1.3	20	80	0	1000
158	1679	OK	2.22	2.62	1.3	21	79	0	1000
159	3186	OK	2.84	3.04	1.3	31	69	0	1000
160	165	OK	0.23	0.38	1.2	1	99	0	997
161	254	OK	1.19	1.87	1.8	13	87	0	1000
162	355	OK	-1.00	-1.21	1.3	0	98	2	1000
163	3611	OK	1.95	2.13	1.2	11	89	0	1000
164	1685	G	5.02	6.05	1.3	98	2	0	1000
165	216	OK	-1.70	-1.80	1.1	0	94	6	1000

Table C.4: Bootstrap results 4.

Appendix D

Non-null results

		True		Simulated		Times Identified As		
Inst	n	Status	z	z	SD	Good	Average	Bad
1	6831	G	9.13	9.23	1.1	100	0	0
2	3314	OK	-1.06	-1.04	0.9	0	100	0
3	1113	B	-4.02	-4.04	1.0	0	34	67
4	2205	B	-4.48	-4.41	1.0	0	21	79
5	289	OK	3.18	3.25	1.0	35	65	0
6	3238	B	-4.20	-4.19	1.0	0	29	72
7	2889	B	-6.64	-6.54	1.0	0	0	100
8	2292	B	-3.65	-3.67	1.1	0	49	51
9	1031	B	-3.94	-3.87	1.0	0	39	61
10	116	OK	-1.49	-1.46	0.8	0	100	0
11	222	OK	3.15	3.14	0.7	28	72	0
12	832	OK	1.08	1.09	1.0	1	99	0
13	2720	OK	-0.80	-0.81	1.0	0	100	0
14	2917	OK	3.23	3.24	0.5	22	78	0
15	231	OK	-0.39	-0.34	1.0	0	100	0
16	1681	OK	2.21	2.29	0.9	8	92	0
17	2294	G	5.30	5.33	0.9	98	3	0
18	2850	OK	1.71	1.71	0.7	0	100	0
19	149	OK	0.63	0.71	0.9	0	100	0
20	3482	G	4.00	3.99	1.0	64	36	0
21	1435	B	-4.01	-3.98	1.0	0	34	66
22	2009	OK	-1.75	-1.71	1.0	0	98	2
23	1633	OK	0.90	0.89	1.0	1	100	0
24	2726	OK	0.73	0.71	1.1	1	99	0
25	748	B	-4.37	-4.42	1.0	0	21	79
26	920	OK	1.40	1.40	1.0	1	99	0
27	88	OK	0.82	0.84	1.0	1	100	0
28	3435	OK	0.32	0.33	0.8	0	100	0
29	2941	OK	0.83	0.79	1.1	1	99	0
30	3819	G	4.49	4.44	1.0	80	20	0
31	853	OK	-0.64	-0.62	1.0	0	100	0
32	3545	OK	1.95	1.94	0.8	3	98	0
33	55	OK	-0.68	-0.64	1.1	0	100	0

Table D.1: Non-null results 1.

		True		Simulated		Times Identified As		
Inst	n	Status	z	z	SD	Good	Average	Bad
34	1003	OK	1.07	1.07	1.0	1	100	0
35	1975	G	3.82	3.79	0.9	58	42	0
36	210	OK	2.60	2.64	1.0	16	84	0
37	456	OK	2.63	2.69	1.0	17	83	0
38	1525	OK	0.59	0.59	1.0	0	100	0
39	679	OK	1.60	1.63	0.6	0	100	0
40	2134	OK	-0.59	-0.53	0.9	0	100	0
41	2047	OK	1.29	1.33	1.1	2	98	0
42	901	OK	-2.38	-2.30	1.1	0	89	11
43	2188	OK	1.20	1.18	1.0	1	99	0
44	3262	G	7.39	7.43	0.9	100	0	0
45	844	OK	-1.13	-1.15	1.0	0	100	0
46	3573	G	8.83	8.88	1.0	100	0	0
47	1979	OK	-0.17	-0.17	0.9	0	100	0
48	2708	OK	2.09	2.11	1.1	7	93	0
49	2525	OK	-2.04	-1.97	1.1	0	94	6
50	2209	OK	-1.49	-1.45	0.9	0	99	1
51	4261	OK	-0.69	-0.70	1.0	0	100	0
52	1662	OK	0.24	0.24	1.1	0	100	0
53	1539	OK	1.11	1.13	1.0	1	99	0
54	405	OK	0.47	0.54	1.0	0	100	0
55	176	OK	-2.13	-2.01	0.9	0	95	5
56	1125	OK	2.17	2.17	0.9	6	94	0
57	1751	OK	3.53	3.57	1.1	49	51	0
58	162	OK	-1.08	-1.08	1.0	0	100	0
59	964	OK	3.32	3.38	1.0	40	60	0
60	197	OK	2.13	2.17	1.1	10	90	0
61	2499	OK	2.17	2.14	0.7	2	98	0
62	2270	OK	-3.30	-3.26	1.0	0	65	36
63	1023	OK	3.52	3.51	1.0	45	55	0
64	1501	B	-5.93	-5.91	1.0	0	1	99
65	3009	OK	3.34	3.37	1.0	39	61	0
66	723	OK	3.01	3.04	1.1	29	71	0
67	2412	OK	-0.69	-0.66	0.9	0	100	0
68	127	OK	-1.41	-1.38	1.0	0	99	1
69	4773	OK	-0.26	-0.33	0.9	0	100	0
70	5990	OK	-1.90	-1.86	1.0	0	95	6
71	913	G	5.90	5.90	0.9	99	1	0
72	963	OK	1.36	1.42	1.0	1	99	0
73	594	OK	-0.52	-0.48	1.0	0	100	0
74	3981	OK	-2.40	-2.35	1.1	0	89	11
75	3085	B	-4.37	-4.37	1.1	0	24	76
76	1062	OK	-0.46	-0.43	1.0	0	100	0
77	355	OK	-0.32	-0.24	1.0	0	100	0

Table D.2: Non-null results 2.

		True		Simulated		Times Identified As		
Inst	n	Status	z	z	SD	Good	Average	Bad
78	1728	B	-8.45	-8.41	1.0	0	0	100
79	3443	OK	3.37	3.38	0.7	37	63	0
80	1470	OK	0.69	0.72	1.0	0	100	0
81	3395	B	-4.51	-4.47	1.1	0	22	78
82	198	OK	-0.69	-0.67	1.0	0	100	0
83	2092	OK	3.22	3.23	0.9	34	67	0
84	2427	OK	-2.52	-2.46	1.0	0	87	13
85	1211	OK	1.72	1.74	0.9	1	99	0
86	1836	B	-4.50	-4.48	1.0	0	18	82
87	1954	OK	2.30	2.30	0.9	8	92	0
88	369	OK	0.75	0.76	0.9	0	100	0
89	65	OK	1.42	1.51	1.0	2	98	0
90	3398	OK	-1.23	-1.28	1.0	0	99	1
91	1561	OK	0.65	0.66	0.9	0	100	0
92	995	OK	1.33	1.32	1.0	1	99	0
93	106	OK	3.02	2.99	0.7	17	84	0
94	1409	OK	0.32	0.32	0.8	0	100	0
95	92	OK	2.42	2.48	0.8	10	91	0
96	1223	OK	1.35	1.38	1.0	1	99	0
97	2844	OK	1.51	1.52	1.0	2	98	0
98	3684	OK	1.95	1.98	0.8	2	99	0
99	2315	OK	3.40	3.36	1.1	39	61	0
100	2468	G	3.93	3.92	0.8	66	34	0
101	639	B	-5.15	-5.13	1.0	0	6	94
102	1626	OK	0.37	0.38	1.0	0	100	0
103	206	OK	-0.77	-0.68	1.0	0	100	0
104	2009	OK	2.83	2.64	1.0	17	83	0
105	172	OK	2.29	2.36	0.9	11	89	0
106	1983	OK	-0.23	-0.25	1.1	0	100	0
107	2527	OK	0.28	0.27	1.1	0	100	0
108	3633	OK	-0.41	-0.48	1.0	0	100	0
109	365	OK	0.87	0.90	0.8	0	100	0
110	3658	B	-10.60	-10.13	1.0	0	0	100
111	604	OK	0.11	0.14	0.9	0	100	0
112	65	OK	0.04	0.13	1.0	0	100	0
113	704	OK	-1.11	-1.13	1.1	0	99	1
114	429	OK	1.55	1.55	1.1	3	97	0
115	98	OK	1.50	1.58	0.9	2	98	0
116	3414	OK	0.14	0.14	0.8	0	100	0
117	433	OK	2.50	2.56	1.0	14	86	0
118	1344	OK	0.95	0.90	0.9	0	100	0
119	648	G	4.05	4.01	1.1	66	34	0
120	239	OK	2.68	2.69	0.9	16	84	0
121	949	G	6.25	6.22	1.0	99	1	0
122	2519	B	-6.26	-6.20	1.1	0	1	99
123	1445	G	4.16	4.22	1.0	74	26	0

Table D.3: Non-null results 3.

		True		Simulated		Times Identified As		
Inst	n	Status	z	z	SD	Good	Average	Bad
124	2808	OK	1.39	1.43	0.9	1	99	0
125	2210	OK	2.35	2.42	1.1	12	88	0
126	2522	OK	-1.69	-1.68	1.1	0	97	3
127	142	OK	-0.95	-0.94	1.0	0	100	0
128	420	OK	1.02	1.01	1.0	1	99	0
129	3559	G	4.19	4.19	0.9	74	26	0
130	595	OK	0.82	0.85	1.0	0	100	0
131	5013	OK	-1.04	-1.01	1.0	0	100	1
132	308	OK	0.92	0.91	1.0	1	100	0
133	4343	OK	-2.23	-2.19	0.8	0	96	4
134	130	OK	-0.77	-0.76	1.0	0	100	0
135	178	OK	-1.01	-0.94	0.9	0	100	0
136	2353	OK	1.89	1.87	1.0	4	96	0
137	2192	OK	-3.32	-3.35	0.9	0	60	40
138	98	OK	-0.18	-0.15	0.9	0	100	0
139	2927	OK	-3.23	-3.18	0.9	0	66	34
140	121	OK	-0.51	-0.48	0.6	0	100	0
141	2496	B	-4.26	-4.29	1.0	0	24	76
142	902	OK	1.92	1.89	0.8	2	98	0
143	112	OK	1.00	1.10	1.0	1	99	0
144	1173	G	4.95	4.98	1.0	92	8	0
145	4115	B	-6.02	-5.95	1.1	0	2	98
146	1232	OK	1.86	1.85	0.9	3	97	0
147	2881	OK	-3.41	-3.38	1.0	0	58	43
148	136	OK	0.82	0.97	0.9	0	100	0
149	507	G	5.04	5.05	0.9	95	6	0
150	499	OK	1.35	1.37	0.7	0	100	0
151	2981	B	-6.81	-6.72	1.0	0	0	100
152	3241	OK	-0.56	-0.54	1.0	0	100	0
153	1745	OK	1.92	1.92	1.0	5	95	0
154	3126	B	-6.41	-6.36	1.0	0	0	100
155	2992	OK	0.97	1.01	1.0	1	100	0
156	1357	OK	-0.09	-0.07	0.7	0	100	0
157	1199	OK	1.96	1.95	0.9	4	96	0
158	1679	OK	2.22	2.25	1.0	9	92	0
159	3186	OK	2.84	2.84	1.0	23	77	0
160	165	OK	0.23	0.27	0.6	0	100	0
161	254	OK	1.19	1.28	1.1	2	98	0
162	355	OK	-1.00	-0.93	1.1	0	100	1
163	3611	OK	1.95	1.94	1.0	6	94	0
164	1685	G	5.02	5.02	0.9	94	6	0
165	216	OK	-1.70	-1.59	1.0	0	98	2

Table D.4: Non-null results 4.

Appendix E

1997 Analysis

Gender	Freq.	Prog. Rate
Female	150676	91.8%
Male	139216	88.9%

Table E.1: Gender 97/98.

Age	Freq.	Prog. Rate
Young	215834	92.4%
Mature	74058	84.7%

Table E.2: Age 97/98.

State School	Freq.	Prog. Rate
Fee Paying	35126	94.1%
State School	166364	92.5%
Unknown	88402	85.0%

Table E.3: State school attendance 97/98.

Class Status	Freq.	Prog. Rate
Not Low	161453	92.8%
Low	57112	90.6%
Unknown	71327	84.8%

Table E.4: Parental occupation 97/98.

Geographical Participation	Freq.	Prog. Rate
Not Low Participation	243114	91.1%
Low Participation	39220	87.3%
Unknown	7558	84.0%

Table E.5: Low HE geographical participation 97/98.

Year of Program	Freq.	Prog. Rate
1st Year	264936	90.6%
Not 1st Year	24956	88.0%

Table E.6: Year of program 97/98.

Subject	Freq.	Prog. Rate
Medicine, Dentistry & Veterinary science	6779	97.9%
Languages & Humanities	30734	92.4%
Biological sciences & Physical sciences	37708	91.9%
Subjects allied to medicine	16465	91.8%
Education	14017	91.8%
Agriculture & related subjects	2493	90.9%
Social studies & Law	35117	90.8%
Creative arts & Design	25026	90.4%
Business & administrative studies & Librarianship	38246	88.9%
Combined subjects	34301	89.4%
Architecture, Building & Planning	6772	88.7%
Mathematical sciences & Computer science	21368	88.2%
Engineering & Technology	20866	86.8%

Table E.7: Subject of study 97/98.

Qualifications	Freq.	Prog. Rate
A Pts 29-30	17006	98.1%
A Pts 27-28	12698	97.0%
A Pts 25-26	14146	96.8%
A Pts 23-24	15243	96.4%
A Pts 21-22	15790	95.4%
A Pts 19-20	16684	94.4%
A Pts 17-18	16807	93.6%
A Pts 15-16	16121	92.3%
A Pts 13-14	15463	91.6%
A Pts 11-12	14172	90.2%
A Pts 9-10	12778	89.3%
A Pts 5-8	18314	87.7%
A Pts 0-4	7998	86.5%
Access/Foundation	19488	86.3%
Higher Education	24570	86.1%
GNVQ3+	11155	85.8%
A Pts Not Known	8085	85.4%
BTEC/ONC	13005	85.2%
Unknown	7014	83.4%
Others	7810	79.5%
NONE	5545	78.6%

Table E.8: Student entry qualifications 97/98.

	World(%)											
	Small			Medium			HEFCE			Big		
N Runs	2000			2000			500			500		
Estimate	Low	High	Both	L	H	B	L	H	B	L	H	B
Global	3.2	2.9	6.1	2.4	2.3	4.8	2.6	2.2	4.8	2.3	1.8	4.1
Local	1.6	3.6	5.2	2.1	3.2	5.3	2.4	4.5	6.9	8.3	10.9	19.3
Uni.	1.4	3.5	4.9	1.9	2.9	4.8	1.7	3.6	5.3	1.5	3.0	4.4
C. Cell	2.3	2.2	4.5	2.3	2.8	5.1	2.4	2.6	5.0	1.7	2.6	4.3
ANOVA	1.8	3.1	4.9	2.0	3.1	5.1	4.2	6.2	10.4	3.8	5.5	9.3
$\gamma(0.5)$	1.9	3.1	5.0	2.2	2.6	4.8	2.2	2.7	4.9	2.6	2.9	5.5
Limit 1	1.9	3.1	5.0	2.2	2.6	4.8	2.2	2.7	4.9	2.5	2.6	5.1
Limit 2	1.9	3.1	5.0	2.2	2.6	4.8	2.2	2.7	4.9	2.2	2.8	5.0

Table E.9: Performance of alternatives: 1997/1998 data.

	World(%)											
	Small			Medium			HEFCE			Big		
N Runs	2000			2000			500			500		
Estimate	Low	High	Both	L	H	B	L	H	B	L	H	B
Global	2.3	2.3	4.6	2.5	2.5	5.0	2.5	2.5	5.0	2.3	2.2	4.6
Local	2.4	2.3	4.7	2.7	2.6	5.3	3.2	3.1	6.3	8.2	8.2	16.4
Uni.	2.4	2.3	4.7	2.5	2.5	5.0	2.6	2.5	5.1	2.4	2.3	4.6
C. Cell	2.3	2.3	4.6	2.5	2.5	5.0	2.5	2.5	5.0	2.4	2.2	4.6
ANOVA	2.5	2.3	4.8	2.6	2.5	5.0	2.6	2.5	5.1	2.8	2.7	5.5
$\gamma(0.5)$	2.3	2.3	4.6	2.6	2.5	5.0	2.6	2.6	5.2	3.1	3.1	6.2
Limit 1	2.3	2.3	4.6	2.6	2.5	5.0	2.6	2.6	5.2	2.9	2.8	5.6
Limit 2	2.3	2.3	4.6	2.6	2.5	5.0	2.6	2.6	5.2	2.9	2.8	5.7

Table E.10: Performance of alternatives: 1997/1998 data, $p = 0.5$.

Number of PCFs Removed	Overall Misclassification(%)	Bad But Not Called Bad(%)	Good But Not Called Good(%)
0	0.00	0.00	0.00
1	5.05	4.81	8.33
2	8.70	4.95	11.90
3	12.57	6.59	16.67
4	17.00	7.47	22.14
5	21.66	7.69	26.79
6	26.57	7.42	30.95
7	31.60	7.69	35.42
8	36.02	7.69	41.67

Table E.11: Omitting PCFs from all models in the 97/98 data (3.61 cut-off).

Number of PCFs Removed	1996/1997 PCFs Removed	Mis. Rate(%)	1997/1998 PCFs Removed	Mis. Rate(%)
0	Full Model	0.00	Full Model	0.00
1	State	2.42	Age	0.62
2	Gender State *or* Gender G. Participation	4.24	Gender Age	1.86
3	Gender Age State	4.24	Gender Age State	3.11
4	G. Participation Age Yr Program State	6.06	Gender Age Yr Program State	4.35
5	Gender Age Yr Program State G. Participation	7.27	Gender Age Yr Program State G. Participation	4.97
6	Gender Age Yr Program State G. Participation Low Class	9.69	Gender Age Yr Program State G. Participation Subject	8.07
7	Gender Age Yr Program State G. Participation Low Class Subject	18.18	Gender Age Yr Program State G. Participation Low Class Subject	9.94
8	All PCFs	38.79	All PCFs	36.02

Table E.12: PCF models that produce the minimum misclassification rates.

Appendix F

OPTA

Model	Mean a	SE(a)	Mean z-score	SE(z-score)
s1	0.0000	0.039	0.0041	0.70
s1(+assist)	0.0000	0.057	0.0046	0.85
s2	0.0000	0.040	0.0037	0.71
s2(+assist)	0.0000	0.057	0.0041	0.86
s3	0.0000	0.026	0.0019	0.55
s3(+assist)	0.0000	0.047	0.0024	0.76
s4	0.0000	0.045	0.0068	0.77
s4 (+assist)	0.0000	0.065	0.0084	0.94

Table F.1: Descriptive statistics for the as and z-scores: OPTA.

Model	s1	s2	s3	s4
s1	1.00	0.98	0.95	0.96
s2	0.98	1.00	0.95	0.97
s3	0.95	0.95	1.00	0.91
s4	0.96	0.97	0.91	1.00

Table F.2: Correlation matrix for models s1, s2, s3 and s4.

Model	a Cor. with OPTA	z-score Cor. with OPTA	Between Score & Scass a-variable	z-score
s1	0.48	0.48	0.87	0.87
s1(+assist)	0.48	0.47		
s2	0.49	0.49	0.87	0.87
s2(+assist)	0.49	0.48		
s3	0.38	0.38	0.86	0.86
s3(+assist)	0.39	0.39		
s4	0.58	0.58	0.88	0.88
s4(+assist)	0.59	0.58		

Table F.3: Examining the links between OPTA, the Score and Scass Models.

Model	Correlation Between a and z-score
s1	0.998
s1(+assist)	0.997
s2	0.998
s2(+assist)	0.997
s3	0.999
s3(+assist)	0.998
s4	0.997
s4(+assist)	0.996

Table F.4: Correlation matrix between a and z-score.

Team	Player Name	OPTA	IO
Ipswich	Stewart	1119	2.52
Man U	Cole	876	1.62
Chelsea	Hasselbaink	1045	1.61
Man U	Sheringham	1139	1.57
Leeds	Keane	1086	1.47
Charlton	Johansson	886	1.32
Leicester	Gunnlaugsson	1442	1.31
West Ham	Di Canio	1042	1.27
Arsenal	Henry	1194	1.09
Man U	Solskjaer	1216	0.95
Everton	Jeffers	987	0.87
Southampton	Beattie	814	0.80
Derby	Strupar	730	0.79
Liverpool	Heskey	1065	0.77
Liverpool	Owen	1032	0.67
Middlesbrough	Boksic	800	0.61
Sunderland	Phillips	749	0.61
Aston Villa	Joachim	840	0.54
West Ham	Suker	538	0.48
West Ham	Kanoute	853	0.47
Everton	Cadamarteri	750	0.38
Coventry	Aloisi	455	0.36
Chelsea	Zola	948	0.29
Charlton	Svensson	495	0.27
Newcastle	Cort	854	0.18
Aston Villa	Dublin	578	0.16
Chelsea	Gudjohnsen	923	0.14
Charlton	Hunt	655	0.14
Charlton	Pringle	862	0.12
Charlton	Bartlett	751	0.08
Leeds	Viduka	660	0.06
Man C	Wanchope	817	0.04
Sunderland	Dichio	579	0.03

Table F.5: OPTA analysis: player assessments 1.

Team	Player Name	OPTA	IO
Ipswich	Armstrong	988	0.03
Tottenham	Ferdinand	734	0.02
Newcastle	Ameobi	598	0.02
Sunderland	Quinn	635	0.01
Derby	Christie	560	0.00
Everton	Campbell	590	-0.04
Leicester	Akinbiyi	579	-0.05
Leicester	Mancini	405	-0.12
Everton	Ferguson	1211	-0.15
Man U	Yorke	997	-0.16
Liverpool	Litmanen	862	-0.17
Tottenham	Rebrov	680	-0.17
Arsenal	Wiltord	885	-0.20
Newcastle	Shearer	749	-0.21
Liverpool	Fowler	932	-0.21
Coventry	Zuniga	447	-0.23
Newcastle	Gallacher	726	-0.23
Coventry	Roussel	593	-0.26
Leeds	Huckerby	380	-0.29
Leicester	Cresswell	565	-0.29
Southampton	Pahars	787	-0.30
Bradford	Saunders	470	-0.30
Aston Villa	Angel	365	-0.31
West Ham	Diawara	518	-0.37
Leicester	Benjamin	515	-0.38
West Ham	Camara	658	-0.48
Arsenal	Bergkamp	898	-0.49
Middlesbrough	Ricard	640	-0.53
Leicester	Eadie	581	-0.55
Bradford	Carbone	683	-0.62
Bradford	Blake	635	-0.71
Newcastle	Lua Lua	947	-0.78
Derby	Sturridge	410	-0.80
Middlesbrough	Deane	425	-0.83
Leeds	Bridges	915	-0.85
Southampton	Davies	573	-0.87
Leeds	Smith	749	-0.89
Everton	Moore	765	-0.93
Tottenham	Iversen	858	-0.93
Bradford	Ward	609	-0.94
Southampton	Rosler	358	-0.95
Man C	Goater	447	-0.96
Charlton	Lisbie	552	-0.98
Arsenal	Kanu	919	-0.98
Bradford	Windass	594	-0.98
Derby	Burton	511	-1.11
Man C	Wright-Phillips	636	-1.53

Table F.6: OPTA analysis: player assessments 2.

Bibliography

Aitkin, M., D. Anderson, and J. Hinde (1981). Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society, Series A 144*, 148–161.

Anderson, S., A. Auquier, W. Hauck, D. Oakes, W. Vandaele, and H. Wesiberg (1980). *Statistical methods for comparative studies: techniques for bias reduction*. New York: Wiley.

Barnetson, B. and M. Cutright (2000). Performance indicators as conceptual technologies. *Higher Education 40*, 277–292.

Becker, R., J. Chambers, and A. Wilks (1988). *The new S language*. Wadsworth.

Bishop, Y., S. Fiensburg, and P. Holland (1974). *Discrete multivariate analysis: theory and practice*. Cambridge, MA: MIT Press.

BMA (2000). *Clinical Indicators (League Tables)*. British Medical Association - board of science and education.

Brunsden, V. and M. Davies (2000). Why do HE students drop out? - a test of Tinto's model. *Journal of Further and Higher Education 24*, 301–310.

Burgess, J., C. Christiansen, S. Michalak, and C. Morris (2000). Medical profiling: improving standards and risk adjustment using hierarchical models. *Journal of Health Economics 19*, 291–309.

Clayton, D. and M. Hills (1993). *Statistical models in epidemiology*. Oxford University Press, Oxford.

Cochran, W. (1977). *Sampling Techniques*. Wiley: 3rd Edition.

Cochran, W. (1983). *Planning and analysis of observational studies*. Wiley.

Collett, D. (1991). *Modelling binary data*. Chapman and Hall.

Daley, J., S. Jencks, D. Draper, G. Lenhart, N. Thomas, and J. Walker (1988). Predicting hospital-associated mortality for Medicare patients. *JAMA 260*, 3617–3624.

DeLong, E., E. Peterson, D. DeLong, L. Muhlbaier, S. Hackett, and D. Mark (1997). Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine 16*, 2645–2664.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Diggle, P., K. Liang, and S. Zeger (1994). *The analysis of longitudinal data.* Oxford University Press, Oxford.

Draper, D. (1995). Inference and hierarchical modelling in the social sciences. *Journal of Educational and Behavioral Statistics 20*, 115–147.

Draper, D. (1996). Discussion of "league tables and their limitations: statistical issues in comparisons of institutional performance," by H Goldstein and DJ Spiegelhalter. *Journal of the American Statistical Association 159*, 416–418.

Draper, D. (Forthcoming). *Bayesian hierarchical modeling.* Springer-Verlag.

Efron, B. and R. Tibsharani (1993). *An introduction to the bootstrap.* Chapman and Hall, London.

Everitt, B. (1995). The analysis of repeated measures: a practical review with examples. *The Statistician 44*, 113–135.

FEFC (2000). *FE Performance indicators 1998-1999: April 2000.* www.fefc.ac.uk/data/performanceindicators.html.

Fisher, R. (1973). *Statistical methods and scientific inference.* New York, Macmillan: 3rd Edition.

Fleiss, J. (1981). *Statistical methods for rates and proportions.* Wiley.

Goldstein, H. (1995). *Multilevel statistical models.* Kendall's Library of Statistics: Arnold.

Goldstein, H. (1997). Value added tables: the less-than-holy-grail. *Managing Schools Today 6*, 18–19.

Goldstein, H. (2000). Using pupil performance data for judging schools and teachers. *British Education Research Journal.*

Goldstein, H. (2001). *League tables and schooling.* Institute of Education, University of London.

Goldstein, H. and D. Spiegelhalter (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society Series A - Statistics in Society 159*, 385–409.

Gray, J., H. Goldstein, and S. Thomas (Forthcoming). Predicting the future: the role of past performance in determining trends in institutional effectiveness at A-level. *British Educational Journal.*

HEDA (1998). *The characteristics and performance of higher education institutions.* Higher Education Division, Australia.

HEFCE (1999a). *Performance indicators in higher education 1996-1997, 1997-1998.* Higher Education Funding Council for England.

HEFCE (1999b). *Performance indicators in higher education. First report of the performance indicators steering group.* Higher Education Funding Council for England.

HEFCE (2000). *Performance indicators in higher education 1997-1998, 1998-1999.* Higher Education Funding Council for England.

HEFCE (2001). *Indicators of employment and other post-qualification activies.* Higher Education Funding Council for England.

Herring, R. (1936). The relationship of marital status in female to mortality from cancer of the breast, female genital organs and other sites. *The American society for the control of cancer 18*, 4–8.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*, 945 – 970.

Hosmer, D. and S. Lemeshow (1995). Confidence interval estimates of an index of quality performance based on logistic regression models. *Statistics in Medicine 14*, 2161–2172.

Johnston, R. and D. Wichern (1982). *Applied multivariate statistical analysis.* Prentice Hall, New Jersey.

Kahn, K., L. Rubenstein, D. Draper, J. Kosecoff, W. Rogers, E. Keeler, and R. Brook (1990). The effects of the DRG-based prospective payment system on quality of care for hospitalised Medicare patients - an introduction to the series. *JAMA 264*, 1953–1955.

Keeler, E., L. Rubenstein, K. Kahn, D. Draper, E. Harrison, M. McGinty, W. Rogers, and R. Brook (1992). Hospital characteristics and quality of care. *JAMA 268*, 1709–1715.

Manley, B. (1997). *Randomisation, bootstrap and Monte Carlo methods in biology.* Chapman and Hall, London.

Marshall, E. and D. Spiegelhalter (1998). Reliability of league tables of in-vitro fertilisation clinics: retrospective analysis of live birth rates. *British Medical Journal 3166*, 1701–1704.

McGraw, S., D. Sellers, E. Stone, J. Bebchuk, E. Edmundson, C. Johnson, K. Bachman, and R. Luepker (1996). Using process data to explain outcomes. *Evaluation Review 20*, 291–312.

Miller, R. (1966). *Simultaneous statistical inference.* New York.

NHS (2000). *Quality and performance in the NHS. Performance indicators: July 2001.* NHS Executive. www.doh.gov.uk/nhsperformanceindicators/index.htm.

Normand, S., M. Glickman, and C. Gatsonis (1997). Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association 92*, 803–814.

Perneger, T. (1998). What's wrong with Bonferroni adjustments? *British Medical Journal 316*, 1236–1238.

Rasbash, J., W. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, G. Woodhouse, D. Draper, I. Langford, and T. Lewis (2000). *A user's guide to MLwiN, V2.1.* Multilevel Models Project, Institute of Education, University of London.

Raudenbush, S. and J. Willms (1991). *Schools, classrooms, and pupils.* Academic Press, Inc.: Harcourt Brace Jovanovich.

Rosenbaum, P. and D. Rubin (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B 45*, 212–218.

Rubin, D. (1974). Estimating causal effects of treatments in randomised and non-randomised studies. *Journal of Educational Psychology 66*, 688–701.

SCEE (2001). *Education and employment fourth report. 30th Jan 2001.* House of Commons, www.parliament.the-stationery-office.co.uk/pa/cm200001/cmselect/cmeduemp/205/20502.htm.

Scheffé, H. (1959). *The analysis of variance.* Wiley, New York.

Smith, D. (1994). Evaluating risk adjustment by partitioning variation in hospital mortality rates. *Statistics in Medicine 13*, 1001–1013.

Smith, J. (1988). *Decision analysis: a Bayesian approach.* Chapman and Hall.

Smith, J., A. McKnight, and R. Naylor (2000). Graduate employability: policy and performance in higher education in the UK. *The Economic Journal 110*, 382–411.

Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks (2000). *BUGS: Bayesian inference using Gibbs sampling, Version 0.5, (version ii).* MRC Biostatistics Unit.

Starr, T., R. Dalcorso, and R. Levine (1986). A comparison of logistic regression and indirect standardisation. *American Journal of Epidemiology 123*, 490–498.

GLIM (1993). *GLIM 4 - the statistical system for generalised linear interactive modelling.* Oxford University Press.

STATA (2001). *STATA 7 documentation set.* Stata Press.

Thomas, N., N. Longford, and J. Rolph (1994). Empirical Bayes methods for estimating hospital-specific mortality rates. *Statistics in Medicine 13*, 889–903.

Tinto, V. (1975). Dropout from higher education: a theoretical synthesis of recent research. *Review of Educational Research 45*, 89–125.

Wakefield, J. and R. Salway (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A 164*, 119–137.

Wilkinson, I., J. Hattie, J. Parr, M. Townsend, M. Thrupp, H. Lauder, and A. Robinson (1999). *Influence of peer effects on learning outcomes: a review of the literature.* Ministry of Education, Wellington, New Zealand.

Yang, M., H. Goldstein, W. Browne, and G. Woodhouse (Forthcoming). Multivariate multilevel analyses of examination results. *Journal of the Royal Statistical Society, Series A*.

Yang, M., A. Heath, and H. Goldstein (2000). Multilevel models for repeated binary outcomes: attitudes and vote over the electoral cycle. *Journal of the Royal Statistical Society, Series A 163*, 49–62.

Yang, M. and G. Woodhouse (2001). Progress from GCSE to A- and AS- level: institutional and gender differences, and trends over time. *British Educational Journal 27*, 245–267.