Teaching Styles Revisited - Teaching Statistical Modelling in R

Hinde, John
National University of Ireland Galway,
Statistics Group, School of Mathematics, Statistics and Applied Mathematics
University Road, Galway, Ireland
E-mail: `john.hinde@nuigalway.ie`

Introduction

The paper on the reanalysis of data on teaching styles (Aitkin et al, 1981) provided a case-study on the power and usefulness of statistical modelling techniques. There was a dual aim of providing further insight into a classic dataset, that had led to much controversy in the educational research community, and of developing and disseminating new modelling techniques. Here we consider the use of this paper and associated dataset to illustrate how it is now easy to use such extensive case studies in teaching statistical modelling. This enables students to follow a substantial, multi-faceted paper in an active way, providing motivation for learning about a range of modelling methodologies and considering extended and alternative analyses. The R system with its wide range of packages, provides an ideal vehicle for this and equally the case-study is a useful focus for developing proficieny in R.

Teaching Styles Study

The teaching styles data came from a study (Bennet, 1976) of the relationship between *Teaching Style* and *Pupil Performance* in primary schools in the UK. The original findings, and major conclusion, was that formal teaching methods were associated with greater pupil progress in the basic skills of reading, writing and mathematics.

There are two main components to the data. In the first phase a questionnaire on 28 items (in 6 areas) of classroom behaviour was sent to all 871 primary schools in Lancashire & Cumbria for completion by 3rd and 4th year teachers. This had an 88% response rate and subsequent analyses were based on 468 4th (final) year teachers. The items, recoded to give 38 basic binary items of classroom behaviour, were intended to capture different styles of teaching, from very formal and structured through to a more progressive, informal approach. The original analysis used principal components to reduce dimensionality prior to cluster analysis, Various cluster solutions were studied and a 12 cluster solution was chosen, with clusters that were roughly ordered from formal to informal, although 78 of the teachers were still not well classified.

From the 12 cluster solution, 37 teachers were selected to represent seven of the clusters, where these could be classified as 2 informal, 3 mixed, and 2 formal. Subsequently, 950 pupils in the classes of the selected teachers were followed for one year, with the following data being collected on the pupils

- pre-tests on entry to class in reading, mathematics, and English;

- post-tests on the same three subjects prior to exit from the school (at end of 4th year);

- personality tests — 15 personality variables measured.

The main analysis then involved testing the differences between the three teaching styles as represented by the cluster allocations (formal/mixed/informal) using an analysis of covariance of the post-test scores adjusted for the pre-test scores. They identified highly significant differences between outcomes for all three of the styles, with formal being significantly better than informal. This original analysis

led to much media and political debate and a call for a return to more formal classroom methods. These conclusions were strongly criticised and there was considerable controversy amongst educational researchers about the findings. This led to the subsequent reanalysis of the data as reported in Aitkin et al (1981).

Teaching Styles Study Reanalysis

The broad thrust of the reanalysis was to use a statistical modelling approach to each stage of the analysis, instead of the more ad hoc methods of the original study. In place of the principal components analysis and cluster procedure a latent class model was used to give model based clustering of the teacher questionnaires. These teaching style estimates were then used in a variance components model of the pupil progress data, where it was now found that teaching style was not a significant factor. Some formal analysis of the pupil personality data was also undertaken. In the following sections we highlight each of these steps and illustrate how the analyses are now easily reproduced using the R system and suggest how there is scope for yet further investigation. At the time, the model fitting in the reanalysis required considerable software development and the model fitting, although extensive, was constrained by the computing power available at the time; this limited, to some extent, the scope for intensive model investigation and validation.

Clustering Teachers — Latent Class Model

A latent class model was used to provide a direct clustering of the teachers using the full set of 38 binary items. The model assumes *conditional independence* of items within each class with the response probability model given by

$$P(\mathbf{X} = \mathbf{x}) = \sum_{j=1}^{K} \pi_j \prod_{l=1}^{38} P(X_l = x_l | j, \theta_{jl}).$$

This approach results in a probabilistic assignment of teachers to clusters and item importance is assessed simultaneously in the model, rather than using principal components as as pre-processing data reduction step. Model estimation is easily implemented using an EM algorithm, although care has to be taken with potential local maxima and one approach is to use a number of different starting points to attempt to find the best solution. At the time of the reanalysis this was time-consuming, but it now standard and fast practice in the latent class procedure in R. For example, for a 3 class solution, motivated by the three final groups in the original study, we have

```
class3x<-poLCA(f,styles,nclass=3,nrep=10)   ### try a number of starts
Model 1: llik = -9789.05 ... best llik = -9789.05
Model 2: llik = -9790.939 ... best llik = -9789.05
Model 3: llik = -9789.05 ... best llik = -9789.05
Model 4: llik = -9789.05 ... best llik = -9789.05
Model 5: llik = -9797.303 ... best llik = -9789.05
Model 6: llik = -9789.05 ... best llik = -9789.05
Model 7: llik = -9789.05 ... best llik = -9789.05
Model 8: llik = -9799.223 ... best llik = -9789.05
Model 9: llik = -9790.939 ... best llik = -9789.05
Model 10: llik = -9798.57 ... best llik = -9789.05
.......

Estimated class population shares
 0.3975 0.3012 0.301
```

showing clearly the problem of local maxima in the log-likelihood, but giving a best solution very close to that in the reanalysis paper where the estimated class population shares (π_j) were $(0.366, 0.312, 0.322)$. The solution found here is slightly superior to that in the 1981 paper, but not substantially different. However, it should be noted that both solutions are rather different from the cluster labelling used in the original study.

The real advantage of modern computing power lies in attempting to answer the question of how many classes should be used. Approaches adopted in 1981 included fitting sequences of models with different numbers of classes (noting possible problems with multiple maxima in the likelihood surface) and then attempting to test for different values of K; this is a non-standard testing problem and the approach adopted in the reanalysis was to use a bootstrap simulation of $-2\log\ell$ to assess significance, now implemented relatively easily and quickly. The final classes were also validated by considering the *total formality score* across the binary items and relating this to a normal mixture model across the classes, again something that is now very easily implemented.

Teaching Style and Pupil Progress

A core aspect of the reanalysis was the treatment of the pupils as nested within classes (teachers) where the teachers were treated as random normally distributed effects. The 1981 paper used the EM algorithm, but a basic variance components model is now easily fitted with the `lmer` package; for example for the performance progress in English we have

```
Linear mixed model fit by maximum likelihood
Formula: eng.s.p ~ eng.scen + (1 | teacher)
   Data: pupilf
  AIC  BIC logLik deviance REMLdev
 6224 6243  -3108     6216    6222
Random effects:
 Groups   Name        Variance Std.Dev.
 teacher  (Intercept) 10.746   3.2781
 Residual             46.730   6.8360
Number of obs: 920, groups: teacher, 36

Fixed effects:
            Estimate Std. Error t value
(Intercept) 106.45824    0.59410  179.19
eng.scen      0.74348    0.01777   41.84
```

The sensitivity to the normality assumption can be checked by using the non-parametric maximum likelihood approach described in Aitkin et al (2009) and implemented in the package `npmlreg`. A finite mass point distribution can be used in place of the normal distribution for the teacher random effects. Varying the number of mass points we have

No. points	$-2\log\ell$
1	6325.731
2	6212.712
3	6206.942
4	6206.649
5	6206.656

suggesting a 3 mass point model. The detailed output for this model gives

```
Call:  allvc(formula = eng.s.p ~ eng.scen, random = ~1 | teacher,
          data = pupilf, k = 3, random.distribution = "np")
Coefficients:
eng.scen      MASS1      MASS2      MASS3
  0.7448   102.1820   105.2298   110.3676


Component distribution - MLE of sigma:      6.828
Random effect distribution - standard deviation:        3.21772


Mixture proportions:
     MASS1        MASS2        MASS3
0.2230806   0.4065043   0.3704151
-2 log L:            6206.9
```

showing a relatively symmetric random effect distribution and very similar variance estimates to those for the normal model.

To study the relationship between teaching style and pupil progress the reanalysis used probabilistic assignments of teachers to the 3 validated latent classes (methods) in the mixed model. This is specified by

$$Y_{pqr} = \mu + \gamma x_{pqr} + \alpha_p + T_q + \epsilon_{pqr}$$

where

- x_{pqr} is pre-test score and γ denotes a common pre-test effect (this can be generalised to γ_p to allow for interaction with teaching method);

- α_p are method (teaching style) effects;

- T_q is the teacher ability assumed to be random, $T_q \sim N(0, \sigma_T^2)$;

- $\epsilon_{pqr} \sim N(0, \sigma_E^2)$ is the pupil level error term, independent of T_q.

In this model, method effects are found not to be statistically significant; this change from the original analysis can be attributed to allowing for the between class variation and hence the intra-class correlation between children in the same class. For example, for the mathematics scores:

```
Linear mixed model fit by maximum likelihood
Formula: math.s.p ~ math.scen + z2 + z3 + (1 | teacher)
   Data: pupilf
  AIC  BIC logLik deviance REMLdev
 6437 6466  -3212     6425     6421
Random effects:
 Groups    Name          Variance Std.Dev.
 teacher   (Intercept)   20.122    4.4858
 Residual                57.364    7.5739
Number of obs: 921, groups: teacher, 36


Fixed effects:
             Estimate Std. Error t value
(Intercept) 102.63308    1.21366   84.56
math.scen     0.81115    0.02192   37.01
z2           -3.60902    4.89355   -0.74
```
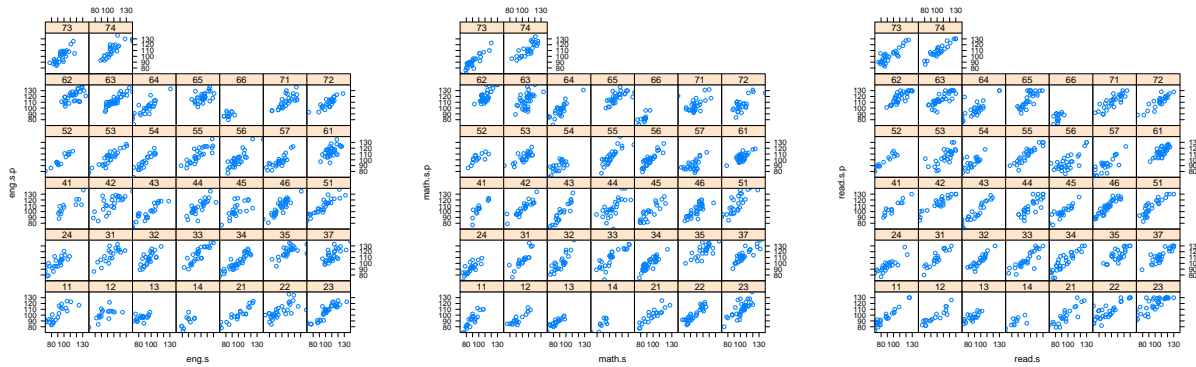
Figure 1: Pupil performance scatterplots for English, Mathematics and Reading

```
z3               0.69271     1.71107      0.40


> anova(pupil.mathz,pupil.math)
Models:
pupil.math: math.s.p ~ math.scen + (1 | teacher)
pupil.mathz: math.s.p ~ math.scen + z2 + z3 + (1 | teacher)
            Df    AIC     BIC   logLik Chisq Chi Df Pr(>Chisq)
pupil.math   4  6433.7  6453.0 -3212.9
pupil.mathz  6  6437.0  6465.9 -3212.5 0.739      2     0.691
```

In this way it is possible to reproduce the core analyses of the paper and to see that once the correct error structure is used the apparently significant methods effect of the original analysis disappear. Of course, there is also potential for other analyses and variants of the model are easily considered. For example, one could consider combining teacher and method effects with a mixed random effect structure

$$T_q \sim \text{Mixture-}N(\alpha_p, \sigma_T^2) \approx \sum_{p=1}^{3} \pi_{qp} N(\alpha_p, \sigma_T^2)$$

where π_{qp} are given by teacher clustering probabilistic assignments. This is similar to the heterogeneity model of Verbeke and Lesaffre (1996) but with fixed π_{qp} and common variance, although this could also be relaxed to separate variances, $\sigma_{T,p}^2$.

One commonly cited advantage of modern computing systems is the power and ease of use of graphics. A simple multiple scatterplot (see Figure 1) of pre- and post-test scores for each class within each of the three subjects shows some very strong patterns. This suggests that rather than a separate analysis for each subject, a common model might be more powerful.

Pupil Personality Clustering

The 15 personality variables (only 8 used in original analysis) were used to attempt to classify pupils into distinct personality types. Normal mixture model were used as a form of model-based clustering, although with only limited success due to computational problems. These analyses are now easily reproduced and extended using for example the `mclust` package of Fraley and Raftery (2002). Factor models were also used in an attempt to identify personality factors, although again with no real success, however model and computational developments in the intervening 30 years offer many possibilities.

From simple regression models for achievement, including both personality variables and pre-test scores, it was clear that there were quite marked personality effects, especially for reading. An obvious approach, although not available at the time, would be to include the personality variables (or factors) directly in the pupil achievement analysis together with teacher styles and random effects.

Discussion

Case studies such as this provide a good basis for practical work grounded in substantial and complex problems. They allow students to understand the interplay between data analysis and significant research studies, while providing training in methodology and application through hands-on work. In recent years, it has become much more common for data and code associated with published articles to be made available on journal, or author, websites and this is certainly to be welcomed. However, these examples are often much narrower in focus and methodology is often merely illustrated rather than integrated as part of a larger multi-stage study. A collection of more wide-ranging analyses of this type would form a very useful teaching resource and bring both models and software implementations into practical focus. This is likely to be particularly helpful for the weaker, or average student, while the more able student is easily led to explore extensions and alternative approaches.

The raw data from the teaching styles study is currently available on the ESRC UK Data Archive, although I hope to be able to make a more polished and usable version available in due course, ideally together with R code vignettes for the core analyses.

ACKNOWLEDGEMENT

REFERENCES

Aitkin, M., Anderson, D., and Hinde, J. (1981) Statistical modelling of data on teaching styles (with Discussion) *Journal of the Royal Statistical Society, Series A*, **144**, 419-461.

Aitkin, M., Francis, B., Hinde, J., and Darnell, R. (2009), *Statistical Modelling in R*, Oxford.

Fraley, C. and Raftery, A.E. (2002) Model-based clustering, discriminant analysis and density estimation, *Journal of the American Statistical Association*, **97**, 611-631.

Verbeke, G. and Lesaffre, E. (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217-221.