A Simulation Study on Hong Kong Diploma of Secondary Education
School-based Assessment Moderation

August 2010

Introduction

1. When expert judgment is employed in the SBA moderation process for a HKDSE subject, a representative sample of student works is obtained from each school, which will then be re-assessed. The average score of the re-assessed student works could be used to reflect the school performance level on SBA. Such a measure would be adopted in the sample review for HKDSE Liberal Studies, which is used as a check against statistical moderation results based on written examination.

2. In general, only some few samples could be obtained from a school for re-assessment; while each of them may have quite a lot of candidates. On the other hand, a large number of schools could be involved. To improve the assessment accuracy on the school performance level on SBA, one of the common statistical techniques is to share information across different schools.

3. In the following, we introduce some technical background about historical development on sharing information across different groups to improve accuracy and show the steps of conducting a simulation study in order to gauge the actual benefits of using Hierarchical Bayesian modeling in expert judgment. The findings of the study are encouraging that the total squared error, on average, is reduced by 30%, as compared with the use of simple average scores.

Technical Background

4. Let us consider the following setting that is already a classical example to show the benefits of sharing information across $k$ different groups or units. Suppose that

$$Y_i | \theta_i \sim N(\theta_i, \sigma^2)$$

where $\sigma^2$ is assumed to be known and $i=1,2,\ldots,k$. 
The model could be interpreted under the context of SBA moderation as follows:

\( Y_i \) is the simple average scores on the sample from a school after re-assessment of the student works,
\( \theta_i \) is the ‘true’ SBA school performance level which could be, in principle, obtained by accurately and exhaustively re-assessing all student works from the school concerned, and
\( \sigma^2 \) is the variance which is supposed to be known and the same across different groups (such an assumption will be relaxed later in the simulation study).

When considering each group alone, the maximum likelihood estimate (MLE) for \( \theta_i \) is simply \( Y_i \).

Sharing Information across Groups – James Stein Estimator

5. Charles Stein shocked the statistical world in 1955 with his proofs that MLE (i.e., simple average score) was no good under this setting. An estimator, known as James–Stein estimator can be shown to dominate, or outperform, the MLE by borrowing information across different groups. The James-Stein estimator has a number of forms. One of them is shown below:

\[
\bar{Y} + [1 - \frac{(k-3)\sigma}{\sum (Y_i - \bar{Y})^2}](Y_i - \bar{Y})
\]

or

\[
\frac{(k-3)\sigma}{\sum (Y_i - \bar{Y})^2}\bar{Y} + [1 - \frac{(k-3)\sigma}{\sum (Y_i - \bar{Y})^2}]Y_i
\]

where \( \bar{Y} \) is the overall average of \( Y_i \).

6. Therefore the James-Stein estimator could be regarded as the weighted average of \( \bar{Y} \) and \( Y_i \) where the weightings depend on the variation of \( Y_i \) across schools. Compilation of the James-Stein estimator for a group requires information from other groups as well. When the variation of \( Y_i \) is large, the effect of sharing information across groups will be not much and the estimator will be close to \( Y_i \); and vice versa. It has been proved that the risk of the James-Stein estimator (i.e., the expected value of total squared error) is always less than the MLE provided that the number of groups, \( k \) is larger than 3.
7. It has found that the James-Stein estimator could be regarded as a particular form of empirical Bayes estimator. Moreover, it is well accepted that full Bayesian modeling is a better alternative, as compared with the empirical Bayes approach. It is because full Bayesian modeling replaces maximization required in estimation by the use of integration, and deals with effectively the nonidentically distributed cases (i.e., sampling variances are different), and automatically injects uncertainty induced by estimating extra (hyper-) parameters. In this regard, full Bayesian modeling is suggested in expert judgment for SBA moderation in HKDSE.

**Full Bayesian Modeling**

8. To improve assessment accuracy, hierarchical Bayesian modeling is suggested to be employed so as to share information across different schools.

9. Let $Y_i$ be the simple average of scores from external assessors for a school $i$. The hierarchical model is set up as follows:

$$Y_i \sim N(\theta_i, \sigma_i^2/n_i) \text{ for } i = 1,\ldots,m \text{ (i.e., there are } m \text{ schools and } n_i \text{ is the sample size)}$$

$$\theta_i \sim N(\mu, \tau^2) \quad \text{ (i.e., all } \theta_i \text{ are sampled from a super-population)}$$

$$1/\sigma_i^2 \sim \text{gamma}(v_0/2, v_0\sigma_0^2/2) \quad \text{ (i.e., all } \sigma_i^2 \text{ are sampled from a super-population)}$$

10. Some non-informative priors $p(\mu)$, $p(\tau^2)$, $p(v_0)$, $p(\sigma_0^2)$ could be set up for these unknown parameters $\mu$, $\tau^2$, $v_0$, and $\sigma_0^2$. The hierarchical structure is represented in the diagram below.

11. Based on the hierarchical modeling, information could be shared across schools for estimating $\theta_i$ and $\sigma_i^2$. For schools with extreme values of sample mean and
Sample variance, the corresponding estimates of $\theta_i$ and $\sigma_i^2$, which could be obtained using standard techniques known as Markov Chain Monte Carlo method (MCMC), will be pulled towards the corresponding overall estimates ($\mu$ and $\sigma_0^2$).

**Simulation Study based on 2010 ASL Liberal Studies**

12. The analytic proof showing advantages obtained from sharing information across groups is only readily available in some simple settings, such as the James-Stein or empirical Bayes estimator, where sampling variances are assumed to be known and equal. For full Bayesian modeling, analytic results on the advantages obtained from sharing information across groups are not easily obtained. In the following, we use summary statistics for 2010 ASL Liberal Studies exam to conduct a simulation study whose objective is verify whether there is any gain in assessment accuracy by the use of full Bayesian modeling and the corresponding magnitudes.

13. The expert judgment method was adopted for SBA moderation in 2010 ASL Liberal Studies (LS) examination. From each school, five pieces of student works were selected for re-assessment in order to determine the school performance level. Totally there were 196 schools attending the examination. For these schools, we calculate their averages and variances of scores assigned by external assessors. Based on the distributions of these summary statistics, we could use parametric models to fit the distributions, which are shown below.
14. In the above diagrams, the densities of the observations are shown as histograms and the curves are the corresponding fitted parametric models. For school averages, the distribution is fitted using a normal distribution. For school variances, the distribution of their inverses is fitted using a Gamma distribution. Based on these parametric models, we conduct a simulation study as follows:

15. Step (1): In accordance with the distributions of school averages and school variances based on 2010 ASL LS exam, we generate 500 pairs of \((\theta_i, \sigma_i^2)\), representing 500 virtual schools. Note that after generating \((\theta, \sigma^2)\), the values of these parameters are fixed throughout the simulation process.

Step (2): For each pair of \((\theta_i, \sigma_i^2)\), we generate 6 \(Y_{ij}\) values representing a sample from each virtual school; i.e.,

\[ Y_{ij} \sim N(\theta_i, \sigma_i^2). \]

Step (3): From the generated values in Step (2) for a virtual school, we calculate the sample mean and sample variance. Given the sample mean, \(Y_i\), we can derive the squared error for each virtual school and sum them up over different schools as follows:

\[ \text{Total sq. error (sample means)} = \sum (Y_i - \theta_i)^2 \]

Step (4) We apply hierarchical Bayesian modeling to the \(Y_i\)'s and obtain a set of means for these 500 schools \(\{Y_i^{(B)}\}\). Accordingly, we can compute the total squared error for sample means with Bayesian adjustments as follows:

\[ \text{Total sq. error (sample means with Bayesian adjustments)} = \sum (Y_i^{(B)} - \theta_i)^2 \]

Step (5) Repeat Step (2) – (4) for 100 times. The averages of total squared errors could be calculated respectively for sample means and sample means with Bayesian adjustments.

**Simulation Results**

16. From the simulation, the average of total squared errors (over 100 trials) for sample mean with Bayesian adjustments \(Y_i^{(B)}\) is found to be 3467.1 which
corresponds to a some 30% reduction when comparing with that for simple sample mean $Y_i$ (4859.2). Thus, a prominent improvement is resulted by using hierarchical Bayesian modeling. The total squared errors over these 100 trials are graphically displayed below.

**Boxplots of Total Squared Errors**

17. From the boxplots above, the superiority of sample mean with Bayesian adjustments, in terms of total squared errors, is obvious, as compared with that of simple sample mean.

**Summary and Conclusions**

18. We would like to highlight the following points for consideration:

(a) Sharing information across groups to improve accuracy was proposed a long time ago (in 1955) by Charles Stein. Analytic results of the improvement could be available in some simple settings.

(b) In later development, full Bayesian modeling is proposed to cater more general situations. However, the analytical results on the improvement could not be easily obtained.
(c) In this study, we use simulation and base on summary statistics from ASL LS exam in order to assess any advantages obtained from sharing information across groups by means of hierarchical Bayesian modeling. It is found that a 30% reduction of the average of total squared errors is resulted using hierarchical Bayesian modeling.

(d) Such an approach could be used in the moderation method based on expert judgment and also in re-marking of student works for outlier detection in statistical moderation based on written exam results.