Performance assessment and league tables. Comparing like with like.

N. T. Longford, SNTL, Leicester* and
D. B. Rubin, Harvard University, Cambridge, MA

Abstract

We formulate performance assessment as a problem of causal analysis and outline an approach based on the missing data principle for its solution. It is particularly relevant in the context of so-called league tables for educational, health-care and other public-service institutions. The proposed solution avoids comparisons of institutions that have substantially different clientele (intake).

Key words: Caliper matching, causal analysis, multiple imputation, non-ignorable assignment, performance indicators, potential outcomes.

JEL Classification: C14 — Econometric and statistical methods: Semiparametric and nonparametric methods

*N. T. Longford, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain. Email: NTL@SNTL.co.uk

Performance assessment and formation of league tables of public-service institutions have, over the last few years, become important statistical activities, as documented by the Journal of the Royal Statistical Society, Series A (Smith, 1990; Goldstein and Spiegelhalter, 1996; Deeley and Smith, 1998; Spiegelhalter, 1999; Spiegelhalter *et al.*, 2002; Stone, 2002; Bailey and Hewson, 2004; Bird, 2004; Bratti *et al.*, 2004; Draper and Gittoes, 2004; Bird *et al.*, 2005; and Smith and Street, 2005). Target setting is a focus of the UK Government Departments responsible for the performance of these institutions, and satisfying the targets, improving the ranking and securing a better assessment appear to be imperatives for the institutions' senior management teams. These management goals often compete for resources with the aspects of clinical priorities that are not represented among the targets. Discussion of the target setting 'culture' featured in the 2005 electoral campaign in the UK, and the jury is still out on whether and how the culture contributed to the outbreak of methicillin-resistant Staphylococcus aureus (MRSA) infections and other public service calamities and controversies in the UK.

Publication of the league tables of the UK universities, National Health Service Trusts, local authorities, railway companies and the like, is by now a well established regularly occurring event. On the one hand, the desire for transparency motivates publication of league tables, perceived as a comprehensive way of comparing institutions of a similar kind in an easy-to-understand format. On the other hand, the statistics community generally recognises, but often does not argue persuasively, that such an ordered list, being subject to uncertainty, is misleading. A sign of ultimate admission of this is the acknowledgement that the assessment (e.g., the assigned position in a league table) is more important than the (latent, imperfectly observed) performance. A natural progression of this state of affairs is that institutions will concentrate more and more on improving the (manifest) assessment, while ignoring improvements in

their (latent) performance. Measured assessment is generally easier to improve than actual performance, especially in the short term. As a result, some institutions may be rated higher and higher, but the quality of the service they provide could well stagnate.

Secondary schools in the UK are probably a case in point. They have become very successful (on average) in generating General Certificates of Secondary Education (GCSE), and it is widely accepted that that is what matters. Careful matching of students with subjects they study and alteration of the curricula to match the contents of examinations are bound to be important factors, competing with the ideal of equal opportunity, in which all students learn as much as their interests and capacities permit. Learning and skills acquired no longer count unless they come with a GCSE!

We want to discuss another deficiency of the league tables, namely that they neither compare nor even attempt to compare 'like with like'. For illustration, we consider a performance assessment of schools or universities based on their students' outcomes. In our interpretation of 'like with like', comparing two schools makes sense only when many students who attended one could conceivably have attended the other. In most of the statistical literature on performance assessment, this issue is addressed by adjustment, typically using linear regression. Such modelling is burdened by numerous caveats related to the distributional assumptions and the functional form assumed for the (multilevel) regression. More thorough search among the models may reward the analyst with a better fitting model but, when model uncertainty is ignored, the relevant selected-model-based estimators are neither unbiased, nor efficient, and their precision is often grossly overestimated (Longford, 2005a).

The Rubin's causal model (Holland, 1986) bypasses several of these deficiencies and recognises the principal source of difficulty — non-ignorable allocation of units (students) to treatments (schools); see Rubin (1978, 1991 and 2005)

and Rosenbaum and Rubin (1983). Each student is associated with a *potential outcome* for every school that he or she could conceivably have attended. Of these, only one outcome is observed for every student, for the school attended. Thus, the data have the form $(W, \mathbf{Z}, \mathbf{Y})$, where W indicates the school attended, $\mathbf{Z}$ is a set of background variables and $\mathbf{Y} = \left(Y^{(1)}, \ldots, Y^{(S)}\right)$ are the potential outcomes for the schools $s = 1, \ldots, S$. For classical statistical inference, the values of $\mathbf{Z}$ and $\mathbf{Y}$ are *fixed*, and randomness rests solely with W, assuming that in a replication of the reality students would attend different schools, but their outcomes in these schools are 'hard-coded', waiting to be revealed if the student attends that school. The background variables $\mathbf{Z}$ should be selected so as to ensure, or make palatable, the assumption that the assignment of students to schools, W, is *ignorable*; that it does not depend on the potential outcomes after conditioning on $\mathbf{Z}$:

$$(W \,|\, \mathbf{Z}, \mathbf{Y}) \sim (W \,|\, \mathbf{Z}), \tag{1}$$

for the corresponding conditional distributions.

Students with some backgrounds would never contemplate attending certain schools — students with such backgrounds do not occur in these schools. We can reflect this is our set-up by declaring $Y_j^{(s)}$ as undefined for all implausible combinations of students j and schools s.

In the standard missing-data formulation, we regard the enumeration of the variables $(W, \mathbf{Z}, \mathbf{Y})$ as the complete data, and the incomplete data comprise completely recorded W and $\mathbf{Z}$, but only Y_W instead of $\mathbf{Y}$. Relying on the assignment process being ignorable, (1), the values of $\mathbf{Y}$ can be completed by one of several (multiple) imputation methods (Rubin, 2004; Longford 2005b), such as hot-deck (matching on the background $\mathbf{Z}$) and propensity matching. A key principle in this process is that the method should not be informed in any way by the (observed or unobserved) values of the outcomes $\mathbf{Y}$.

Any two schools s_1 and s_2 could be compared straightforwardly, by the values of their summaries T (such as the means) of their potential outcomes, restricted to the students attending one of the schools, $T(Y_{s_1} \,|\, W = s_1)$ and $T(Y_{s_2} \,|\, W = s_1)$, *if* these outcomes were observed completely and were defined. If the complete-data comparison cannot be evaluated because some values $Y_j^{(s)}$ are not defined, the comparison cannot and *should not* be made. When a few values of $Y^{(s_2)}$ are not defined, we may resort to some compromise, such as reducing the summary T to students whose potential outcomes are defined and comparisons of pairs of schools to those students who could attend either school in the pair.

Because hardly any student would seriously contemplate enrolment at Oxford University and a 'new' UK university with a mainly vocational orientation as equal alternatives, this device would preclude any comparison of all the UK universities. This may lead to 'indirect' comparisons, when universities A and C are not comparable, but both pairs (A,B) and (B,C) are, or even to more complicated linking, and to finding contradictions with transitivity.

We propose to address this problem by *caliper matching*, in which each school is associated with $2K$ of its closest 'rivals', and its assessment is restricted to a comparison of these $2K + 1$ institutions. For example, an ordering of the schools may be defined according to the academic quality of their students' backgrounds, and for the school in position h, the schools in positions $h - K$, $h - K + 1, \ldots, h - 1, h + 1, \ldots, h + K$ are declared as its rivals. As the focal school is in the middle of its rivals, the comparisons can (but do not have to) be made using elementary complete-data methods, supplemented by multiple imputation for the unobserved potential outcomes. The uncertainty about the comparisons can be estimated by the between-imputation variance. (There is no within-imputation variance because the complete-data analysis is without any variation.) The size of the caliper, $2K + 1$, should be set so that there would

be very few undefined values $Y_j^{(s)}$ for schools s and students j in any set of $2K$ rival schools; each of these sets of schools would be nearly exchangeable. Very small K, $K = 1$ or 2, should be avoided because too little comparing would take place.

The rank of a school A is estimated with reference to its rivals, and so it can be compared with the rank of another school B only when A and B have most of their respective sets of $2K$ rivals in common. Thus the rank, or another summary, reflects the school's position among its genuine competitors. The uncertainty about the summary can be represented by the empirical distribution of the summary evaluated on many replicate (multiply imputed) completed datasets. Confidence intervals can be derived from this distribution straightforwardly.

A virtue of this approach is that it implies a clear and meaningful goal for every institution: to do the best with their input (students' backgrounds). Influencing (manipulating) the input in any way does not have predictable consequences on the assessment of any given school. The comparisons do not use any regression to which schools and students with very different (background) profiles contribute, and all the details of the method are set without inspecting the outcomes, unlike in model selection. The approach requires background information, just like regression, but it uses it only locally, focussing on the ranges of values that occur in the sets of compared (comparable) institutions. We think that this is a far more principled and defensible approach than the current ones.

Acknowledgement

The research described in this manuscript was supported by the Grants No. SEC2003-04476 and SAB2004-0190 from the Spanish Ministry of Science and Technology.

References

Bailey, T. C., and Hewson, P. J. (2004). Simultaneous modelling of multiple traffic safety performance indicators by using a multivariate generalized linear mixed model. *Journal of the Royal Statistical Society* Ser. A **167**, 501–517.

Bird, S. M. (2004). Editorial: Performance monitoring in public services. *Journal of the Royal Statistical Society* Ser. A **167**, 381–383.

Bird, S. M. (Chair of the Working Party), Cox, D. R., Farewell, V. T., Goldstein, H., Holt, T., Smith, P. C. (2005). Performance indicators: good, bad, and ugly. *Journal of the Royal Statistical Society* Ser. A **168**, 1–27.

Bratti, M., McKnight, A., Naylor, R., and Smith, J. (2004). Higher education outcomes, graduate employment and university performance indicators. *Journal of the Royal Statistical Society* Ser. A **167**, 475–496.

Deely, J. J., and Smith, A. F. M. (1998). Quantitative refinements for comparisons of institutional performance. *Journal of the Royal Statistical Society* Ser. A **161**, 5–12.

Draper, D., and Gittoes, M. (2004). Statistical analysis of performance indicators in UK higher education. *Journal of the Royal Statistical Society* Ser. A **167**, 449–474.

Goldstein, H., and Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society* Ser. A **159**, 385–443.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–960.

Longford, N. T. (2005a). Editorial: Model selection and efficiency. Is 'Which model . . . ?' the right question? *Journal of the Royal Statistical Society* Ser. A **168**, 469–472.

Longford, N. T. (2005b). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician.* Springer-Verlag, New York.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomisation. *Annals of Statistics* **6**, 34–58.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* **46**, 1213–1234.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys.* 2nd ed. Wiley and Sons, New York.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modelling, decisions. 2004 Fisher Lecture. *Journal of the American Statistical Association* **100**, 322–331.

Smith, P. (1990). The use of performance indicators in the public sector. *Journal of the Royal Statistical Society* Ser. A **153**, 53–72.

Smith, P. C., and Street, A. (2005). Measuring the efficiency of public services: the limits of analysis. *Journal of the Royal Statistical Society* Ser. A **168**, 401–417.

Spiegelhalter, D. J. (1999). Surgical audit: statistical lessons from Nightingale and Codman. *Journal of the Royal Statistical Society* Ser. A **162**, 45–58.

Spiegelhalter, D. J., Aylin, P., Best, N. G., Evans, S. J. W., and Murray, G. D. (2002). Commissioned analysis of surgical performance by using routine data: lessons from the Bristol inquiry. *Journal of the Royal Statistical Society* Ser. A **165**, 1–31.

Stone, M. (2002). How not to measure the efficiency of public services (and how one might). *Journal of the Royal Statistical Society* Ser. A **165**, 405–422.