# Constructing League Tables of Service Providers When the Performance of the Provider Is Correlated to the Characteristics of the Clients

*La costruzione di graduatorie di servizi pubblici quando la performance del servizio è correlata con le caratteristiche degli utenti*

Enrico Rettore
Dipartimento di Scienze Statistiche, Università di Padova
via S.Francesco 33, 35121 Padova. E-mail: rettore@stat.unipd.it

Alberto Martini
Università del Piemonte Orientale e Progetto Valutazione
Via Nizza 18, 10125 Torino. E-mail: amartini@prova.org

**Riassunto**: Uno strumento per valutare la performance dei servizi pubblici è la creazione di classifiche o graduatorie di performance. E' importante che le graduatorie riflettano solo i fattori sotto il controllo dell'organizzazione, il che può essere ottenuto utilizzando un modello di regressione. Un caso particolare riguarda servizi come la formazione professionale, dove la performance dei centri di formazione può essere correlata positivamente con le caratteristiche degli allievi. In questo caso il tradizionale approccio dell'aggiustamento tramite un modello di regressione con effetti casuali rischia di appiattire eccessivamente le graduatorie. In alternativa proponiamo un modello in cui la correlazione tra performance del centro e caratteristiche degli allievi è esplicitamente riconosciuta. Una simulazione Monte Carlo mostra come lo stimatore ad effetti casuali produce stime distorte dei parametri di interesse, mentre lo stimatore proposto non soffre di tale distorsione.

**Keywords**: random effects, within-group estimator, Monte Carlo simulation

## 1. Why league tables of organizational performance?

There is a growing body of literature dealing with the measurement of organizational performance (Gormley and Weimer, 1999). This literature is responding to a fundamental concern, that of assessing how well public and non-profit organizations are delivering their services to the clients they serve. *League tables* are one tool that can be used in making this assessment. A league table consists of a ranking of organizations providing the same type of service: organizations are ranked from best to worst performers. The goal is to identify the best performers to serve as benchmarks as well as to stimulate the worst toward a better performance.

One important area of use of league tables is the education and training field, which we will use as a leading example. In this field there is increasing demand for comparing the performance of different schools in terms of outcomes such as scholastic achievement attained by their students, or for comparing training centers in terms of the occupational outcomes experienced by their trainees. In the US performance standards have been used since 1983 to evaluate the performance of local organizations delivering training services under JTPA. In Italy much attention has been given recently by the popular press to the ranking of public Universities according to their performance along a number of dimensions. Similar demands arise in other areas of public intervention, such as health care, social services and a wide array of municipal services.

The performance obtained by an organization broadly depends on two sets of factors: one which is outside of the control of the organization (the environment, the characteristics of the clients), and one which is at least in principle under its control (organizational structure; quality and quantity of resources; talent, effort and management skills). Some of the statistical issues involved in constructing league tables concern the need for *adjusting* the ranking to eliminate the effect of factors the organization cannot control. Ideally the ranking should reflect only differences in the second type of factors. The literature speaks of *adjusted performance measures* (APMs) (Brooks, 2000).

The basic statistical procedure to construct APMs is quite simple: (i) identify the measurable dimension(s) for which a ranking has to be produced (*e.g.,* test scores, placement rates, student satisfaction, cost per client); (ii) perform separate least squares regressions using as dependent variable each relevant measure of outcome and as explanatory variables those factors which are believed to affect the outcome but are outside of the control of the organization; (iii) for each organization compute regression *residuals*, which should represent the component of the outcome that depend on the performance of the organization; (iv) use the residuals to construct the "adjusted" league table, which should identify "good" and "bad" organizations on the basis of their own merit.

There are many methodological difficulties to be resolved in implementing this basic idea. In this paper we deal with the problem that arises often in the training field, when the characteristics of the trainees are correlated to the characteristics of the training center. The effect of the outside factors, if properly controlled for, is eliminated by the regression. However, in so doing, we are also eliminating from the residuals part of the variability in the outcome which is due to the performance of the training center. This produces league tables that are too "compressed", i.e. underestimate the differences in performance between providers. Moreover, the ranking itself can be seriously affected, with organizations in the wrong position: we do not deal with this issue here due to space constraints.

## 2. League tables with correlated client and provider characteristics

The formal set-up of the problem is as follows. We have $J$ service providers (job training centers) providing broadly similar services to a given population of clients. The objective of the service is to improve some outcome(s) observable for each client after the service has been provided or while is being provided. For example, job training centers might want to improve the employability of their trainees, measured in terms of length of time needed to find employment after the training has terminated or in terms of likelihood to be employed at a certain date in the future.

We call $y_{ij}$ the outcome obtained by the *i-th* trainee who attended the *j-th* training center, $j=1,J$ and $i=1,n_j$. The centers are assumed to differ along two sets of dimensions: the quality/effectiveness of the service they provide and the characteristics of their clients. The latter represent the severity of the "problem" the service is trying to ameliorate (unemployment, social exclusion, disability). Thus, the ranking must be performed in a way that holds each providers "harmless" for the fact of serving clients with worse characteristics. We want the ranking to reflect entirely (and only) differences in the observed outcomes which are due to differences in the quality of the service provided.

The most simplistic way to rank the providers would be to use the average outcome experienced by their clients after the service. By computing the average outcome per center $\bar{y}_j$, $j=1,J,$ one can straightforwardly rank the centers according to the average outcome experienced by their trainees. The obvious objection is that this is an unadjusted comparison and we cannot tell whether the position in the league table is due to the quality of the trainees or the performance of the center. The challenge is to disentangle the "effect" of the providers from the "effect" of the differences among their clients. A standard regression approach can be used to deal with this problem:

$$y_{ij} = \beta x_{ij} + u_j + \varepsilon_{ij} \qquad (1)$$

where $x_{ij}$ are clients observable characteristics relevant for the outcome. The error term is composed of a term $u_j$ which represents the quality of the performance the training center $j$ and it is common to all trainees in center $j$, and a component specific to each trainee, $\varepsilon_{ij}$. The estimated coefficients $\beta$ allow us to construct a ranking *adjusted* for trainees characteristics. The ranking is made using for each center the average residual, i.e. the average outcome minus the average expected outcome based on clients characteristics: $\bar{y}_j - \beta \bar{x}_j$. It should be noted that with this procedure we only eliminate the effect due to *observable* trainee characteristics. We do not deal with the issue of unobservables here.

Rather, we deal with another problem, the possibility that the correlation between $x_{ij}$ and $u_j$ is non zero. This possibility arises in many concrete situations. For example, trainees with better than average characteristics might be better informed and thus more able to choose the best training center ("trainees self-selection"). Centers that attract better students perhaps because of their geographical location also tend to attract better teachers ("teachers self-selection"). Finally, centers with better teachers and managers are in a position to recruit better students and "weed out" less promising cases ("creaming"). The mechanisms can be different, *but the common result is that providers performance and clients quality can be positively correlated.*

If we did not adjust through regression, the league table would unjustly rank the better center in a higher position by "adding" the effect of the better students. On the other hand, by using traditional regression methods, we obtain the opposite bias: the ranking penalizes the better providers by subtracting part of their good performance and rewards bad performers by being too lenient for the fact of having bad students.

Formally, this is seen by observing that (1) implies:

$$E\{y_{ij} \mid x_{ij}\} = \beta x_{ij} + E\{u_j \mid x_{ij}\} \qquad (2)$$

where the conditional mean on the RHS is a non-trivial function of $x_{ij}$ due to the correlation between $x_{ij}$ and $u_j$. By regressing $u_j$ on $x_{ij}$, its variance is partly accounted for by $x_{ij}$. Then, the residual variability across centers with respect to $u_j$, left after accounting for the observable heterogeneity $x_{ij}$, understates the true variability of $u_j$. This is exactly what we want to avoid.

Blundell and Windmeijer (1997) solve the problem in the linear case by exploiting a simple *within-group* estimator drawn from the panel data models literature. This solution exploits the linearity of equation (1) to get rid of the center specific effect $u_j$ by means of the so called *within-group* transformation. This allows to obtain a consistent estimate of $\beta$. Then, resting on such consistent estimate it is straightforward to obtain an estimate of $\sigma_u^2$. This is not a feasible solution in a non linear model. To see why, let the outcome in (1) be the employability of the subjects which is not directly observable. Instead, we observe the binary outcome $y_{ij}^*$ equal to 1 if $y_{ij} > 0$ and zero otherwise, with $y_{ij}^* = 1$ meaning that the *i-th* trainee from the *j-th* center is employed at a specific date after the completion of the program. The non-linearity resulting from this set up implies that we cannot get rid of the center specific effects $u_j$ by resorting to the *within-group* data transformation as in the linear case. To mimic the *within-group* estimator we should include in the regression as many dummy variables as the centers included in the analysis which is infeasible if the number of centers $J$ is large; and it also provides an inconsistent estimate as $J$ grows to infinity but the number of student per center remains fixed--a classic result in the non-linear panel data literature (Hsiao, 1986).

A convenient solution, mimicking that proposed by Chamberlain (1980), is to model the dependence of the center specific "effect" on the trainees characteristics. Let $u_j$ depends on $x_{ij}$ *only through their center specific average* $\bar{x}_j$ according to the model: $u_j = \vartheta \bar{x}_j + \nu_j$, with $\nu_j$ representing the center heterogeneity uncorrelated to the trainees observable heterogeneity. The assumption is that part of the center characteristics cannot be observed by the students in choosing the center and/or cannot be exploited in selecting teachers or students. By substituting the above expression in (1) we get:

$$y_{ij} = \beta x_{ij} + \vartheta \bar{x}_j + \nu_j + \varepsilon_{ij}\,. \qquad (3)$$

Few remarks are in order. First, the number of additional regressors we need to include in the equation to deal with the correlation between $u_j$ and $x_{ij}$ does not depend on the number of centers since it is just equal to the number of regressors originally included in model (1). Second, a straightforward way to test for the presence of correlation between $u_j$ and $x_{ij}$ amounts to test the null hypothesis $\vartheta = 0$. Finally, once an estimate has been obtained for β, θ and $\nu_j$, $j=1,J$ the true heterogeneity across centers is immediately recovered by evaluating the sum $\hat{\vartheta} \bar{x}_j + \hat{\nu}_j$. To illustrate how the correlation between providers quality and clients characteristics affects estimated ranking and how the proposed approach deals with the problem, we have resorted to a Monte Carlo simulation.

## 3. Results of the Monte Carlo simulation

To exemplify the performance of the Random Effects estimator adjusted as in (3) to account for the correlation between $u_j$ and $x_{ij}$ we performed a simulation exercise. We simulated 500 pseudo samples setting the number of service providers at $J$=100 and the number of clients per center at n=10 and 20. We included just one explanatory variable in the model, drawing its values from a Normal r.v. with unit variance and center specific mean drawn from a standard Normal ($x_{ij} \sim N(\mu_j, 1)$, $\mu_j \sim N(0, 1)$. The coefficient of $x_{ij}$ is set at $\beta=1$. As for $u_j$, it is drawn from a Normal r.v. correlated to $\mu_j$, ρ =.2,.5,.8, with (marginal) unit variance.

Both the standard Random Effects (RE) and the Adjusted Random Effects (ARE) estimators are evaluated on each pseudo sample. The ARE is implemented replacing each $\mu_j$ -- which is unknown in practice -- by its sample counterpart, $\bar{x}_j$.

The results of the simulation exercise are presented in Table 1. Apparently, ARE does not exhibit any bias in estimating both $\beta$ and $\sigma_u^2$ no matter for the degree of correlation between $u_j$ and $x_{ij}$. On the other hand, as expected RE is biased with respect to both $\beta$ and $\sigma_u^2$ with the amount of bias increasing with $\rho$. In particular, $\beta$ is overestimated while $\sigma_u^2$ is underestimated. As for the role of the sample size, while the unbiasedness of ARE does not rest on the number of clients, the bias of RE does increase as n gets smaller. It is also worth noticing that ARE is just a bit underestimating the amount of correlation between $u_j$ and $x_{ij}$ the bias resulting from having replaced $\mu_j$ by $\bar{x}_j$.

**Table 1. Simulated means (standard errors) of the Adjusted Random Effects (ARE) and Random Effects (RE) estimators (based on 500 replicates of the simulation)**

| Parameter (= true value) | Estimator | ρ = 0.2 | | ρ = 0.5 | | ρ = 0.8 | |
|---|---|---|---|---|---|---|---|
| | | n=10 | N=20 | N=10 | n=20 | n=10 | n=20 |
| θ (≈ρ) | ARE | .19 (.13) | .19 (.11) | .45 (.13) | .49 (.11) | .74 (.12) | .77 (.10) |
| β (=1) | ARE | 1.01 (.07) | 1.00 (.05) | 1.01 (.08) | 1.01 (.05) | 1.00 (.08) | 1.00 (.05) |
| | RE | 1.04 (.07) | 1.03 (.05) | 1.10 (.08) | 1.07 (.05) | 1.17 (.08) | 1.11 (.05) |
| $\sigma_u^2$(=1) | ARE | 1.01 (.11) | 1.00 (.09) | 1.01 (.13) | 1.01 (.10) | 1.00 (.13) | 1.01 (.10) |
| | RE | .98 (.10) | .98 (.09) | .92 (.11) | .94 (.09) | .80 (.11) | .85 (.08) |

## References


Blundell R. and F. Windmeijer (1997), "Cluster Effects and Simultaneity in Multilevel Models", *Health Economics,* Vol. 6.

Brooks A. (2000), "The Use and Misuse of Adjusted Performance Measures", *Journal of Policy Analysis and Management*, Vol. 19, No. 2

Chamberlain G. (1984), 'Panel data', in Griliches Z. and M.D. Intriligator (eds.) *Handbook of Econometrics*, Vol. 2, Elsevier Science, Amsterdam.

Gormley W. and D. Weimer (1999), *Organizational Report Cards,* Harvard University Press, Cambridge.

Hsiao C. (1986), *Analysis of Panel Data*, Cambridge University Press, Cambridge.