Testing Procedures for Multilevel Models with Administrative Data

Verifiche di Ipotesi per Multilevel con Dati Amministrativi

Giorgio Vittadini[1]
Department of Statistics
University of Milano-Bicocca, Italy
giorgio.vittadini@unimib.it
Maurizio Sanarico
Noustat s.r.l.
m.sanarico@noustat.it
Paolo Berta
C.R.I.S.P.
crisp.statistica@unimib.it

Abstract: Recent Relative Effectiveness studies of the Health Sector have strongly criticized hierarchical ranking in hospitals. As an alternative, they propose a multi-faceted approach which evaluates the quality and characteristics of Hospital services. In this direction, the use of administrative data has proven highly useful. This data is less precise than clinical data but performs more effectively in describing general situations.
The numerosity of the population renders all the parameters significant in linear model tests. We must therefore utilize resampling schemes in order to verify the hypotheses concerning the significance of the parameters in opportunely drawn subsamples.

Keywords: Hospitals Effectiveness, Administrative Data, Testing Procedure for Multilevel Models, Heterogeneous Samples, Bootstrap.

1 Hospital Effectiveness with Administrative Data

Several recent statistical papers deal with risk-adjusted comparisons on the basis of mortality or morbidity outcomes corrected by means of Multilevel models in order to take into account different case-mix of patients (Goldstein and Spiegelhalter (1996);Vittadini *et al.* (2003); Vittadini *et al.* (2004)). These papers, extremely accurate from the methodological point of view, are all based on small samples of patients with particular pathologies. Other medical papers propose risk-adjusted comparisons as a method for evaluating quality and effectiveness of health structures (Iezzoni (1997)).
Moreover, in some countries private or public External Health Agencies gather, ad-hoc, larger data sets and use linear and logistic models in order to validate quality indicators (AHRQ (2003); JCAHO (2004)). In other cases, they benchmark health structures by means of risk-adjusted comparisons (CIHI (2003); National Health Service (2004)).
Recently, this use of risk adjusted comparisons for benchmarking health structures has been strongly criticized (Lilford *et al.* (1994)). In particular, it has been stated that: “The sensitivity of an institution’s position in league tables to the method of risk adjustment used suggests that comparisons of outcomes are unlikely to tell us about the quality of

[1] via Bicocca degli Arcimboldi, 8 - 20126 Milano

care". Therefore it has been suggested that "The agencies should facilitate the development and dissemination of a database for best practice and improvement based on the results for primary and secondary research."(Lilford *et al.* (1994)). In this direction, the use of administrative data, used by payors to pay bills and manage operations, can be very useful. In fact: "These data are typically computerized, making it easy to collect and use large quantities of information Administrative data have been used to examine geographic variation in utilization of surgical and medical procedures, monitor the use of health services, assess the effects of a policy change on health expenditures, evaluate the relationships between hospital death rates and hospital characteristics" (Damberg *et al.* (1998)).

Which problems do linear models and, in particular, multilevel models involve when they are used with administrative data? Besides problems connected with the accuracy of data (i.e. coding accuracy, timing of diagnoses uncertainty etc.) (Damberg *et al.* (1998)), there is a relevant methodological problem. It is known that when there are large data sets the significance tests associated with linear models refuse the null hypothesis in all cases. In fact the sample size influences the results, and beyond a certain threshold, it is the only determining factor of the test. Every explicative variable seems to be significant for explaining the outcomes, and this result is particularly misleading for the topics mentioned above (Vroman Battle and Rakow (1993)).

Therefore we need appropriate testing procedures able to verify hypotheses regarding the significance of explicative variables in samples drawn from the population associated with administrate data. In general terms, we must devise inference methods in heterogeneous samples collected from large data sets (Duncan and Moon (1998)). The conclusion obtained for tests connected with the Multilevel Model used for the evaluation of healthcare institutions, can be generalized for Linear models in a more general context.

2 The Model

Let us consider a number of outcomes obtained from hospital discharge forms. These outcomes are binary variables and due to the hierarchical structure of **n** observations in a Logistic Multilevel Model.

Therefore given the variable $\mathbf{Y}_{ij}$ = Bin (n, π_{ij}), we fit the following model

$$Logit(\pi_{ij}) = f(X,Z) = \gamma_{00}+\gamma_{10}\mathbf{X}_{ij}+\gamma_{01}\mathbf{Z}_j+\gamma_{11}\mathbf{X}_{ij}\mathbf{Z}_j+u_{1j}\mathbf{X}_{ij}+u_{0j}+\varepsilon_{ij} \quad (1)$$

$$\pi_{\mathrm{ij}} = \frac{1}{1+\exp(-\mathrm{f}(\mathbf{X},\mathbf{Z}))} \quad (2)$$

Before the estimation of the model, we applied a process aimed at optimising the univariate relationship between the outcomes and the predictors.

Such a process consisted of a discretization based on the automatic identification of the linear intervals in each relationship, assigning an indicator variable to each interval.

In this way, non-linearities were captured and modelled, thereby enhancing the expressive power of the independent variables.

The model building initiates with a model including only the dummies; in this phase we facilitated a chunk elimination; the second step consisted in adding all the remaining predictors. A backward elimination completed the model building process.

A positive side effect of the transformation process was the weakening of the evidence calling for the inclusion of interactions. A possible explanation of this fact is that most of the variability is retained by the main effects after transformation, and that non linearity and second-order interaction compete for the same information.

3 Problems Involved in Tests with Large Data Sets

Because of the high number of degrees of freedom, all first level effects turned out to be significant, whereas this is not true in the case of the hospital level effects.

This result is inherent in the classical estimation procedure, designed for small to medium samples (at most some thousands), and it is due to the indefinite narrowing of the standard errors as sample increases.

The imbalance between the first and second level analysis has been overcome by utilizing an empirical testing procedure looking first at the size of effects, then at their relative variability (gauged by confidence intervals) and finally the statistical significance of the effects.

This procedure adheres to the following logic: first evaluate the practical importance, then the extent of statistical variability associated with the effect (uncertainty), then the significance is used as a standard measure of the deviation to a null effect.

The next step in the present work tries to go beyond what is described above.

Statistical inference based on very large sample (> 100.000), containing many heterogeneous groups, leads to irrelevance of statistical testing because of the exceeding power. We think that the real interest is to disentangle the complex data structure.

In summary “failing to reject the null hypothesis” is not the same as “accepting the null hypothesis” or as “rejecting the alternative hypothesis” (Vroman Battle and Rakow (1993)) because of the large size of the sample, the null hypothesis is rejected but this does not mean that the alternative hypothesis of significance is accepted.

4 The New Proposal

We propose a scheme of analysis in which we first attempt to discover and represent the heterogeneity, then we model the detailed data incorporating the structure which has emerged, performing the inference using a number of competing approaches: conducting the analysis within each sub sample, patching together the results using and comparing a standard approach, a Bayesian approach, resampling-based approaches.

The standard approach has many drawbacks: in this paper it is used to provide a reference point. The Bayesian approach, by modelling the probabilities directly, seems to be immune to the problems discussed above Albert and Chib (1951). However, apart from being computationally very expensive, it is not clear how this approach performs in the presence of large samples.

The structured resampling-based approach is an attempt to overcome the problems described by combining resampling-based techniques (for example, various versions of Bootstrap or Boosting) and a representation of the heterogeneity in the sample which can be obtained either by a data-driven approach (useful for achieving a statistically representative analysis of the heterogeneity) or by a knowledge-driven approach (useful to test hypotheses or to explore specific well-identified sub-samples) (Di Ciccio and Efron (1996); Efron (1996)).

Multilevel models represent the statistical relationships existing between a given dependent variable or response function and a set of predictors, taking into account the objects of different size to which such predictors are associated and the relationships (most of the time hierarchical) between these objects. This allows taking into account the different sources of variability in the data correctly. However the multilevel paradigm is not able to capture all of the variability and heterogeneity in the data. For example, it is not able to explain the heterogeneous behaviour of a given agent (hospital) considered from the response function conditional to the set of predictor, relative to other similar agents. Being alike after modelling out the variability associated to the multi-level model amount, they are probably alike from a managerial point of view.

Given this, we considered a more complementary approach in which no predefined structure was super-imposed on the observations. This methodology has been called cluster-weighted modelling or soft-clustering. The idea is to combine the results in order to highlight cases that behave in a well-characterized way (belonging to a single domain of influence or cluster) and cases whose response has characteristics partially shared by more than one cluster.

In the final part of this section, we define soft-clustering methodology in further detail. First, the basic assumptions in the approach:

- The clusters do not interact or describe the data locally with respect to the maximum of the joint probability.
- There is no prior information. An arbitrary cost function E$_{ij}$, is used to express the energy associated to z$_i$ in the cluster C$_j$with centre μ_j^z.
- An iterative process with many clusters utilized to achieve a satisfactory partitioning of the data space through a sequential fusion of the clusters.

The probability that point $x_i \in C_j$ belongs to cluster C$_j$ is expressed by p$_{ij.}$The total average cost is therefore:

$$< E >= \sum_{j=1}^{M} \sum_{i=1}^{N} p_{i\,j}\, E_{i\,j} \tag{3}$$

Equation (3) acts as a boundary condition to the data distribution. To find a stable distribution we follow the maximum entropy principle during each step of the iterative process. The p$_{ij}$ which maximizes entropy is:

$$H = -\sum_{j=1}^{M} \sum_{i=1}^{N} p_{i\,j}\, \log\, (p_{i\,j}) \qquad \begin{cases} \sum\limits_{i=1}^{N} p_{ij} = 1 \\ \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{M} p_{ij} E_{ij} =< E >= cost \end{cases} \tag{4}$$

the Boltzmann distributions are:

$$p_{i\,j} \;=\; \frac{\exp\,[-\beta\; E_{ij}]}{Z_j} \qquad Z_j \;=\; \sum_{i=1}^{N} \exp\,(-\beta\, E_{i\,j}) \tag{5}$$

where the distribution function β is a Lagrange multiplier. Using a thermodynamic analogy, if $\beta \quad \propto$ 1/T, where T is the a “temperature” of the system, with increasing β, the system tends to be frozen and only the closer points influence each other. With decreasing β we have a more disordered system (observations have a higher degree of interaction). The assumption of independence between the clusters and the p$_{ij}$ of different clusters allows us to define the free energy F$_j$ for the cluster C$_j$:

$$F_j \; = \; -\frac{1}{\beta} \; \log \, Z_j \quad \frac{\partial F_j}{\partial \mu_j^2} \; = \; 0 \qquad \forall j \tag{6}$$

Considering the squared euclidean distance:

$$E_{i\,j} \; = \; |\, z_i \, - \, \overrightarrow{\mu}_i \,|^2 \; = \; |\, y_i \, - \, \mu_i^{\,y} \,|^2 \, + \, |\, x_i \, - \, \mu_i^{\,x} \,|^2 \tag{7}$$

we obtain:

$$\mu_j^{\,z} \; = \; \sum_{i=1}^{N} \frac{z_i \; \exp \left[-\beta \; \left(z_i \, - \, \mu_j^{\,z} \right)^2 \right]}{\sum\limits_{i=1}^{N} \exp \left[-\beta \; \left(z_i \, - \, \mu_j^{\,z} \right)^2 \right]} \tag{8}$$

Equation (8) cannot be solved analytically. A solution can be obtained through fixed-point iteration of the following formula:

$$\mu_j^{\,z} \; (n+1) \; = \; \sum_{i=1}^{N} \frac{z_i \; \exp \left[-\beta \; \left(z_i \, - \, \mu_j^{\,z} \; (n) \right)^2 \right]}{\sum\limits_{i=1}^{N} \exp \left[-\beta \; \left(z_i \, - \, \mu_j^{\,z} \, (n) \right)^2 \right]} \tag{9}$$

which is typically iterated until a stable $\mu_j^{\,z}$ is obtained. The process converges to a local minimum with respect to specified initial conditions and β, which reflects the number of clusters used to represent the data.

This topic, currently under investigation, is the application of the cluster-weighted modelling of all data, taking into account the patient as well as the hospital level, constructing a multi-level cluster weighted framework of analysis, able to answer some very interesting questions without the necessity of using ad-hoc procedures. Below, we give a brief account of the theory.

Let us consider the data vectors as **{ y$_n$,x$_n$,z$_n$ },** where **y** is the response function, **x** the level-1 variables and **z** the level-2 variables. We then infer p(**y**,**x**,**z**) as the joint probability density. This density is expanded over a sum of cluster C$_k$, each cluster containing an input distribution, a local model, and an output distribution. The input distribution is:

$$\begin{aligned} p(\mathbf{y}, \mathbf{x}, \mathbf{z}) \; &= \sum_{k=1}^{K} p(\mathbf{y}, \mathbf{x}, \mathbf{z}, c_k) = \sum_{k=1}^{K} p(\mathbf{y}, \mathbf{x}, \mathbf{z}|c_k) p(c_k) = \\ &= \sum_{k=1}^{K} p(\mathbf{y}|\mathbf{z}, \mathbf{x}, c_k) p(\mathbf{x}|\mathbf{z}, c_k) p(\mathbf{z}|c_k) p(c_k) \end{aligned} \tag{10}$$

with the normalization condition Σ_np(c$_k$)=1. In the presence of both discrete and continuous predictors we must further partition them accordingly.

The next point is to associate a specific density to each of the terms in the formula. Normally, the conditional distributions p**(x**|**z,**c$_k$**)** and p**(z**|c$_k$**)** are taken to be Gaussian distributions with appropriate covariance matrices (for example, diagonal or structured). The output distribution p**(y**|**x,z,**c$_k$**)** depends on whether **y** is continuous-valued or discrete, and on the type of local model connecting **x**,**z** and y: f(**x**,**z**,b$_k$).

In most cases a linear function is enough, given the composition of many local functions (as many as required by the data distribution), to represent complex nonlinear functions.

5 An Application

The study is based on the administrative data provided by the Lombardy Regional Health Care Directorate regarding 1.152.266 admissions to 160 hospitals. The data consists of: regional population anagraphical records, Administrative Hospital Discharge records and hospitals' structural characteristics. Response variables are: in-hospital and post-discharge mortality, patient's discharges against medical advice, transfers to other hospitals, unscheduled hospital re-admissions, unscheduled returns to operating room. Patients' case mix and hospitals' characteristics are also collected from the same sources. We use a logistic Multilevel model to investigate best and worse practices of hospitals connected with their characteristics (i.e.: size, private vs public status, general vs specialized, etc.). The test procedures mentioned above are used in order to evaluate the significance of parameters related to explicative variables in the context of large populations. Multilevel models produce a variety of useful results, and in the Health Care Effectiveness Evaluation context level-2 residuals are particularly important. In the present case, we have 160 residuals, one for each hospital, with a considerable level of heterogeneity, indicating either a possible difference in managerial effectiveness, or some other source of variability.

Is there further information, perhaps at a higher level, not-well defined at the sampling design stage, that can reasonably account for a significant portion of the level-2 residual variation? Is it possible to individualize combinations of conditions associated with specific portions of the level-2 residual distribution?

In order to answer to these questions, a simple but effective strategy identifies a set of variables able to determine the aforementioned patterns. Then, we apply a multiple comparison Bonferroni-adjusted procedure to identify their statistical significance. Three criteria are used: the effect size, the adjusted p-value and the standard errors (confidence intervals) of the effect. In the analysis we also include second-order interaction effects. The procedure described has been well-accepted in literature, and able to obtain interpretable results consistent with experience-backed beliefs.

At this point we pursue an integrated approach in combining the Multilevel and Cluster weighted models. Therefore we present an outline of a practical application of the association between Multilevel modelling and cluster-weighted clustering, namely, to apply the soft-clustering to level-2 residuals to obtain automatically the main patterns of variation in the joint distribution of residuals.

As a further interesting result, we obtain the probability of each hospital to belong to each cluster. This allows us to discern well-characterized hospitals from other less-clearly defined ones.

The case study concerns the soft-clustering analysis for mortality rate within 30 days after discharge outcome. The procedure detected 4 clusters. All the observations were well-characterized by a single cluster with one exception, as already mentioned. The tables below show an analysis of variance on these clusters and their composition:

Parameter		**Estimate**	**Standard Error**	**t Value**	$\mathbf{Pr} > \vert\mathbf{t}\vert$
Intercept		-.3277484344	0.20144526	-1.63	0.7347222
Cluster	cluster0	0.0340194876	0.37985182	0.09	6.45
Cluster	cluster1	0.5136519229	0.22580980	2.27	0.16875
Cluster	cluster2	-.1232972155	0.25478708	-0.48	4.36875
Cluster	cluster3	0.0000000000	.	.	.
DEAS		0.6099373653	0.17460280	3.49	0.0006
PRS		0.4559875042	0.18319776	2.49	0.0965278
IRCCS		-.4917529255	0.16183465	-3.04	0.0028
UNI		-.1586361703	0.14348553	-1.11	1.8798611
PRIV		-.0270973074	0.34221433	-0.08	6.5069444

Cluster	**N Obs**	**Variable**	**Mean**	**Minimum**	**Maximum**	**Std Dev**	**Median**
cluster0	55	PRIV	1.00	1.00	1.00	0.00	1.00
		IRCCS	0.00	0.00	0.00	0.00	0.00
		DEAS	0.16	0.00	1.00	0.37	0.00
		PRS	0.15	0.00	1.00	0.36	0.00
		RR	1.00	0.24	3.05	0.59	0.06
cluster1	24	PRIV	0.00	0.00	0.00	0.00	0.00
		IRCCS	0.25	0.00	1.00	0.44	0.00
		DEAS	0.00	0.00	0.00	0.00	0.00
		PRS	0.00	0.00	0.00	0.00	0.00
		RR	1.21	0.24	0.175	0.05	1.10
cluster2	49	PRIV	0.02	0.00	1.00	0.14	0.00
		IRCCS	0.08	0.00	1.00	0.28	0.00
		DEAS	1.00	1.00	1.00	0.00	1.00
		PRS	0.00	0.00	0.00	0.00	0.00
		RR	1.16	0.27	2.17	0.41	1.11
cluster3	32	PRIV	0.03	0.00	1.00	0.18	0.00
		IRCCS	0.03	0.00	1.00	0.18	0.00
		DEAS	0.00	0.00	0.00	0.00	0.00
		PRS	1.00	1.00	1.00	0.00	1.00
		RR	1.25	0.39	2.56	0.55	1.23

The risk within clusters presents a significant heterogeneity. The presence of DEAS and emergency units remain significant in explaining the heterogeneity. Another interesting result is that most of the hospitals with a given profile have similar patterns of residual variation (hospital-specific effectiveness), whereas some are deviate from this norm. This is a significant finding, allowing the proposal of innovations for the improvement in Quality for specific health care facilities.

References

AHRQ (2003) *Guide to Inpatient Quality Indicators*, www.ahrq.gov/dat/hcup.

Albert J. and Chib S. (1951) Bayesian tests and model diagnostics in conditionally independent hierarchical models, *Journal of the American Statistical Association*, 92, 916–925.

CIHI (2003) *Hospital report 2002, Acute Care Technical Summary*, secure.cihi.ca/cihiweb/splash.html.

Damberg C., Kerr E.A. and McGlynn E. (1998) Description of data sources and related issues, in: *Health Information Systems, Design Issues and Analytical Applications*, McGlynn E.A., Damberg C.L., Kerr E.A. and Brook R., eds., RAND Health Corporation, 43–76.

Di Ciccio T.J. and Efron B. (1996) Bootstrap confidence intervals, *Statistical Science*, 11, 189–212.

Duncan C. Jones K. and Moon G. (1998) Context, composition and heterogeneity: using multilevel models in health research, *Social Science and Medicine*, 46, 97–117.

Efron B. (1996) Empirical bayes methods for combining likelihoods, *Journal of the American Statistical Association*, 91, 538–565.

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 A*, 159, 385–443.

Iezzoni L. (1997) The risk of risk adjustment, *JAMA*, 278 (19), 1600–1607.

JCAHO (2004) www.jcaho.com.

Lilford R., Mohammed M., Spiegelhalter D.J. and Thomson R. (1994) Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma, *The Lancet*, 364, 1147–1154.

National Health Service C. (2004) *A Commentary on Star Ratings 2002-2003*, www.chi.nhs.uk./ratings.

Vittadini G., Carabalona R. and Rossi C. (2004) Metodologie statistiche di valutazione dell'efficacia delle strutture sanitarie, in: *Qualità e Valutazione delle Strutture Sanitarie*, Pagano M. and Vittadini G., eds., Etas, Milano, 269–282.

Vittadini G., Rossi C. and Sanarico M. (2003) Recenti sviluppi nella metodologia statistica per la valutazione dell'efficacia degli ospedali, *Statistica*, 1, 1–24.

Vroman Battle M. and Rakow E.A. (1993) Zen and the art of reporting differences in data that are not statistical significant, *IEEE Transactions on Professional Communication*, 39,2, 75–80.