A possible drawback for the method could appear the choice of the smoother and of the bandwidth parameter, a critical point in nonparametric estimation. Notwithstanding, this choice is somewhat less important for *submodel plots*. In the comparison of two nonparametric estimates, if the same estimator with the same bandwidth parameter is used for both curves, then the pointwise bias in the estimates will be canceled. In a different setting, Bowman and Young (1996) showed that the comparison of nonparametric curves yields a remarkable stability of results over a wide (reasonable) range of the bandwidth parameters.

In addition, we note that the procedure could be time consuming in the presence of many covariates. However, the method can be strongly enhanced by a dynamic graphics device that allows for an easy and quick change of the projection directions.

Further work will involve a general use of the proposed method to compare complex multivariate regression models, not necessarily restricted to parametric models (e.g. generalized additive models, projection pursuit regression models).

References

Bedrick E.J., Tsai C.L. (1994). Model Selection for Multivariate Regression in Small Samples. *Biometrics*, 50, pp. 226-231.

Bowman A., Young S. (1996). Graphical comparison of nonparametric curves. *Applied Statistics*, 45, pp. 83-98.

Brown, B.W. (1980). Prediction analysis for binary data, *Biostatistics Casebook*, Miller, R.J., Efron, B., Brown, B.W. & Moses, L.E. (Eds.), Wiley, New York.

Collett D. (1991). *Modelling binary data*. Chapman and Hall, 1991.

Cook R.D. (1998). *Regression Graphics*. Wiley, New York.

Cook R.D., Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley, New York.

Jovanovic, B.D. and Hosmer, D.W. (1997) A simulation of the performance of Cp in model selection for logistic and Poisson regression, *Computational Statistics and Data Analysis*, 23, pp. 373-379.

On the Use of Multivariate Regression Models in the Context of Multilevel Analysis

Giorgio Vittadini

Department of Statistical Sciences - University of Milan-Bicocca. E-mail: giorgio.vittadini@unimib.it

Abstract: The use of Multivariate Regression Models with mixed data to evaluate and decompose relative effectiveness of different social agencies presents numerous problems. The solution proposed is to use the Seemingly Unrelated Equations Models (SURE) in the framework of Multilevel Analysis, following quantification of the response variables by means of simultaneous Multidimensional Scaling methods. An example is provided.

Keywords: Multivariate Regression Analysis, Mixed Data, Multilevel Analysis, Interaction Parameters, SURE Analysis, Relative Effectiveness.

1. Study with mixed indicators

In the studies about the effectiveness evaluation of a social service, the outcome is defined as a long-term result of the output onto a particular aspect of the user. Such an outcome can be measured by an appropriate set of indicators.
Of particular relevance is the relative effectiveness which evaluates different social agencies (hospitals, schools, and etc...). In such cases, the outcome depends on the explicative variables connected with the users and on other variables concerning the agencies. Outcomes can be described by a set of mixed indicators or by latent variables, obtained with measurement errors from those indicators[1].
To this aim the following models have been utilized:

- Multivariate Models with mixed variables that take into account associations and correlations among outcomes.
- Multilevel Models which clearly reveal the relative effectiveness of every social agency.

This paper proposes an integration of the two families of instruments.

2. Multivariate Models and relative effectiveness

From the interpretative point of view, it is reasonable to calculate simultaneously the value of outcomes which describe different but correlated aspects of the "state of well being" of the user. The Multivariate Models are suitable for this purpose.

[1] For a systematic explanation of these concepts see Gori and Vittadini (1999), pp. 135-146.

There have been numerous attempts in the literature to utilize Multivariate Models with mixed variables for outcome studies[2]. The most sophisticated model is Fitzmaurice-Laird (1997), a Multivariate Regression Model that analyzes the linear dependence of a set of mixed response variables (binary, politomic, and continuous) from explicative variables[3]. In the Fitzmaurice-Laird model (1997) we can consider the conditional expected values of every dependent variable with respect to a set of explicative variables, by means of Marginal Regression Models (linear functions for the quantitative dependent variables and logit functions for categorical and binary responses are used). Using the likelihood methods applied to the joint density of the quantitative and qualitative dependent variables, one gets a solution, which considers the associations between dependent and explicative variables[4].

There are still some unresolved problems in this approach:

a) First, the Gaussian multivariate distribution hypothesis for the dependent quantitative variables or the use of binomial distribution hypothesis for the qualitative variables can give unsatisfactory results with respect to the distributions of the indicators.

b) Secondly, direct use of mixed indicators may be inappropriate to describe the outcomes. Most importantly in many cases we have ordinal indicators obtained from surveys or experiments. Data obtained must be transformed in quantitative objective measures independent upon type of sample, difficulty of the trials and ability of the individuals (Rasch 1960, Vittadini 1999). Secondly, from the statistical point of view latent variables are more indicative to express the outcome (Gori and Vittadini 1999, Vittadini 1999).

c) In order to measure the relative effectiveness of every single social agency (effectiveness A) and decompose it into effectiveness due to the observable resources (effectiveness B) and to non-observable capacity of the management (effectiveness C), it is necessary to utilize a Multilevel approach. In fact, in a case in which the effectiveness C was identified with dummy variables instead of random variables (used in Multilevel analysis), there was collinearity with the variables inherent to the effectiveness B (Gori and Vittadini 1999).

d) In covariance models, where the dependent variable is measured after a treatment at the time t=1, this variable is introduced as an explicative variable covariate at the time t=0. When more treatments are implemented in order to take into account the effects of interaction between these covariates and the treatments, interaction parameters are also introduced (Cox and McCullagh, 1982). In relative effectiveness studies, interactions between agencies and the initial state of the user also exist. Therefore the variation of the outcome with respect of the initial state of the user is not identical for every initial state and every social agency (Fleiss 1986, Rice and Leyland 1996). Thus interaction parameters should be introduced also in Multivariate Regression Models with mixed variables. However, there are difficulties in creating adequate maximum likelihood functions, in finding computational solutions and in proposing the interpretation of the results.

[2] For example: using Marginal Multivariate Regression Models with discrete variables, taking into account the associations between outcomes, the relationship between vision loss and demographic characteristics, educational level, and access to medical care was studied (the sample of 5000 people aged 40 years and above was collected by the *Baltimore Eye Survey)*; the relationship between diarretic and respiratory infections and the lack of *vitamin A* was also studied (the sample of 3000 children was collected by the *Indonesian Children Study*); the genetic association between asthmatic parents and children was studied in a sample of 100 people [Liang Zeger Qaqish (1992)]. Using an Alternate Logistic Regression Model the positive effects of a particular therapy on epileptic patients were analyzed in a sample of 52 individuals, collected by *John Hopkins Hospital* [Carey-Zeger-Diggle (1993)]. Using a Marginal Regression Model with mixed variables for Clustered Ordinal Measurement, the post-operative neurological improvement of 144 patients, followed for ten years, was analyzed [Heagerty and Zeger (1996)].

[3] Using this model the health consequences of air pollution were studied in a sample of 7243 children between 9 and 10 years old. Data were compiled by *the Harvard Six Cities Study.* The effects of parents' psychological disorders on the development of their children was studied in a sample of 56 children. Data were compiled by *the St. Louis Research Project.*

[4] The solutions are characterized by the relative asymptotic efficiency with respect to the case in which the equations were considered one by one. They are robust with respect to misspecification of association pairs between binary variables. They solve the problem of the missing values.

3. The SURE Model in the framework of multilevel analysis

In order to resolve the problems illustrated above, we propose the utilization of Multivariate Models with quantitative dependent variables characterized by stochastic links between different equations, in a Multilevel contest (Hox 1995). In order to quantify all the observed indicators and simultaneously obtain the outcomes as their linear transformations, we utilize the appropriate Multidimensional Scaling instruments and the Vectorial Space Decomposition Method named Restricted Regression Component Decomposition (RRCD) (Haagen and Vittadini 1998, Vittadini 1999). In this way we resolve the problem reported in section 2b. Among the quantitative dependent variable models with stochastic links between different equations, we chose the Seemingly Unrelated Equation Models (SURE). Furthermore, the introduction of a Multilevel framework hypothesizes that "there is a hierarchical data set, with one single dependent variable that is measured at the lowest level and explanatory variables at all existing levels..." and that in every equation "the intercept and slope coefficients are assumed to vary across the schools; for that reason they are often referred to as random coefficients" (Hox 1995),(Aitkin and Longford 1986). This fact allows studying completely the problem. Therefore, relatively to j-th outcome ($j=1,\ldots,h$) at the time t_1 (after the service allocation) drawn on the i-th ($i=1,\ldots\ldots,n_\ell$) user to whom the service is allocated by the ℓ-th social agency ($\ell = 1,\ldots,q$; $n = n_1+\ldots+n_\ell+\ldots+n_q$), the SURE model in a Multilevel context equation has the following structure:

$$y_{ji\ell} = \alpha_{j\ell} + \sum_{k=1}^{p} \beta_{kj} x_{kji\ell} + e_{ji\ell} \tag{1}$$

$$\alpha_{j\ell} = \delta_{j\ell} + \sum_{w=1}^{\ell} \delta_{wj\ell} \cdot f_{wj\ell} + m_{j\ell}; \quad \beta_{kj} = v_{kj\ell} + \sum_{w=1}^{\ell} v_{wkj\ell} \cdot f_{wkj\ell} + \chi_{kj\ell}; \tag{1a}$$

with: $y_{ji\lambda}$ outcome; $x_{kji\lambda}$ explicative k-th ($k=1,...,p$) variable belonging to the set of outcome explicative variables $X_{kj\lambda}$; $\alpha_{j\lambda}$ and β_{kj} random coefficients which identify the effectiveness A; $e_{ji\lambda}$ random residual; $\delta_{j\lambda}$, $\delta_{wj\lambda}$, $v_{kj\lambda}$, $v_{wkj\lambda}$ parameters; $f_{wj\lambda}$ ($w=1,...,t$) specific observable w-th characteristic of the social agency; $\sum_{w=1}^{t}\delta_{wj\lambda}\cdot f_{wj\lambda}$ and $\sum_{w=1}^{t}v_{wkj\lambda}\cdot f_{wkj\lambda}$ technological relationships which identify the effectiveness B; $m_{j\lambda}$ and $\chi_{kj\lambda}$ random residuals which identify effectiveness C. From (1a), (1) we obtain:

$$y_{ji\lambda}=\delta_{j\lambda}+\sum_{w=1}^{t}\delta_{wj\lambda}\cdot f_{wj\lambda}+\sum_{k=1}^{p}v_{kj\lambda}\cdot x_{kji\lambda}+\sum_{k=1}^{p}\left(\sum_{w=1}^{t}v_{wkj\lambda}\cdot f_{wkj\lambda}\right)x_{kji\lambda}+$$
$$+m_{j\lambda}+\sum_{k=1}^{p}\chi_{kj\lambda}\cdot x_{kji\lambda}+e_{ji\lambda} \qquad (2)$$

Defined the $g_{j\lambda}=\sum_{k=1}^{p}\chi_{kj\lambda}\cdot x_{kji\lambda}$ random part $\pi_{ji\lambda}=\left[m_{j\lambda}+g_{j\lambda}+e_{ji\lambda}\right.$ contains the random error structure of the equation (Hox 1994). The matrix model of the explicative variables is of full rank and non-stochastic; the expected values of the residuals are null and their variances are finite; the contemporaneous covariances between the diverse residuals are constant; the intertemporal covariances are null for every couple of residuals. The contemporaneous covariances among residuals postulate stochastic relationships between the equations of the model (Srivastava and Giles, 1987).

The variances and covariances of the random error structure $\pi_{j\lambda}$ have the following structure:

$$\sigma^2_{\pi_{j\lambda}}=\sigma^2_{m_{j\lambda}}+\sigma^2_{g_{j\lambda}}+\sigma^2_{e_{j\lambda}}+2\sigma_{m_{j\lambda}g_{j\lambda}}+2\sigma_{m_{j\lambda}e_{j\lambda}}+2\sigma_{g_{j\lambda}e_{j\lambda}} \quad (j=1,..., h; \quad =1,...,q) \ (3)$$

Then there are: the $\sigma_{\pi_{j\lambda}\pi_{jm}}$ covariances inherent to the same outcome, for different social agencies λ and m ($j=1,..., h;$ λ,$m=1,...,q$, λ $\neq m$); the $\sigma_{\pi_{j\lambda}\pi_{z\lambda}}$ covariances inherent to different outcome j and z for equal social agency λ ($j, z=1,...,h, j\neq z;$ $=1,...,q$); the $\sigma_{\pi_{j\lambda}\pi_{zm}}$ covariances between different outcomes j and z and different social services λ and m. Introducing in every equation of the model (1) the initial state of the user we obtain:

$$y_{ji\lambda}=\alpha_{j\lambda}+\sum_{k=1}^{p}\beta_{kj}x_{kji\lambda}+\gamma_{j\lambda}\cdot y_{ji\lambda(t_0)}+e_{ji\lambda};\quad \gamma_{j\lambda}=\tau_{j\lambda}+\lambda_{j\lambda}\cdot s_{j\lambda}+u_{j\lambda} \qquad (4)$$

with $\gamma_{j\lambda}$ random coefficient inherent to initial state of the user, $\tau_{j\lambda},\lambda_{j\lambda}$ parameters; $s_{j\lambda}$ explicative variable, $u_{j\lambda}$ random residual. Expressing (4) in compact terms we obtain:

$$y_{ji\lambda}=\delta_{j\lambda}+\sum_{w=1}^{t}\delta_{wj\lambda}\cdot f_{wj\lambda}+\sum_{k=1}^{p}v_{kj\lambda}\cdot x_{kji\lambda}+\sum_{k=1}^{p}\left(\sum_{w=1}^{t}v_{wkj\lambda}\cdot f_{wkj\lambda}\right)x_{kji\lambda}+$$
$$+\tau_{j\lambda}\cdot y_{ji\lambda(t_0)}+\lambda_{j\lambda}\cdot s_{j\lambda}\cdot y_{ji\lambda(t_0)}+m_{j\lambda}+g_{j\lambda}+u_{j\lambda}\cdot y_{ji\lambda(t_0)}+e_{ji\lambda} \quad (5)$$

with $\tau_{j\lambda}\cdot y_{ji\lambda(t_0)}$ the effects of interaction between the observable variables inherent to the social service λ-th and the initial state j-th of the user i-th. From the random error structure of the equation $\pi^*_{ji\lambda}=\left[m_{j\lambda}+g_{j\lambda}+u_{j\lambda}\cdot y_{ji\lambda(t_0)}+e_{ji\lambda}\right]$ contains the random error structure of the equation; from that we obtain the variances and covariances of the random error structure which take into account the interactions.

4. Possible further developments

We observed that the integration between the two models resolves the problems reported in paragraph 3. Some problems still remain open:
1) The models postulate linear relationships, whereas non-linear relationships can exist between explicative and dependent variables.
2) In the case of numerous parameters there is difficulty in constructing robust and consistent estimations of the parameters.
3) The normality hypothesis of the errors is often inadequate. In fact, "specific distributional assumption regarding the disturbances of the SURE model it has been one of normality" and "its violation may have serious consequences for our inferences" (Judge, Griffiths, Hill and Lee, 1980). The problem is partially resolvable if the observed distributions are symmetrical (conservative leptocurtic or liberal platicocurtic). In this case, using the Generalized Linear Models, such distributions may be studied as an extension of Gaussian distribution. The problem becomes more serious when the distribution is not symmetrical, especially in the presence of numerous parameters. In this case, the methods based on Multivariate Ranks and the bootstrap methods (which have been used for the SURE models) appear inadequate (Rilstone and Veall, 1995), because if the starting distribution is not normal the new distribution will also be not normal. The method for constructing inferences based on iterative simulation and Bayesian inference for models with not normal distribution of errors appears more proficient (for example using Monte Carlo Markov Chain methods by means of algorithms such as Metropolis Hastings and Gibbs sampling) (Gelman and Rubin, 1992, Geyer, 1992). Such methods are commonly employed, but they have never been used in the case of SURE models.

4) The problem of the selection bias is unresolved in simpler models[5] as in multivariate models.

5. An example of relative effectiveness study

In this example the relative effectiveness of 5 retirement homes (RSA) of different sizes in Lombardy[6], is studied. The sample was made up of 225 residents. By means of the Functional Independent Measure (FIM) scale we measured motor, cognitive, sfinteric skills in two instances; simultaneously we collected other explicative variables.

Table 1: *The variables of the model*

DEPENDENT VARIABLES (t=1 2° draft)	EXPLICATIVE VARIABLES (t=0 1° draft)
Qualitative: Latent Variables 11 indicators of motor skills → FM1; 5 indicators of cognitive skills → FC1; 2 indicators of sfinteric skills → FS1	Qualitative: 11 indicators of motor skills → FM0; 5 indicators of cognitive skills → FC0; 2 indicators of sfinteric skills → FS0; Quantitative: Nosological Macroclassification = MCN. Qualitative: Provenience = PRO; Dummy: Vision (2 modalities / 1 dummy) = V_1, Hearing (3/2) = U_1,U_2; Social Services (5/4)=RSAI (I=1,2,3,4), Interaction FQ0 (Q=M,C,S)-RSAI (I=1,2,3,4)

By means of the ALS Multidimensional Scaling methodology (Young, de Leeuw, Takane, 1976) the ordinal modalities of the motor, cognitive, sfinteric skills indicators are transformed into quantitative values and simultaneously 6 latent variables are obtained by means of RCDR. Therefore we have a model in which the dependent variables are the latent variables FM1, FC1, FS1. In table 3 we observe that the results obtained with the SURE method compared with those obtained with the Ordinary Least Squares (OLS) method present significant differences. Moreover, using the fifth RSA as a reference point we observe that in the first equation no RSA has more effectiveness, in the second we find significant differences between the RSA2 and the RSA4, in the third the effectiveness of the RSA1 and RSA4, RSA2 and RSA3 are equivalent. Finally we observe (Fig.1) that distributions of FM1, FC1, FS1 are almost symmetrical and transformable into normal distributions.

Table 2: *Explained variance of the latent variables*

	FM	FC	FS
t=1	78.14	79.71	98.96
t=0	78.89	83.09	99.36
ρ^2	68.70	60.10	60.20

[5] Copas and Li (1997) examine a case in which two independent subsamples of equal numerousity of patients are not subdivided among the medical treatments in the random mode; also in this simple case the robustness and validity of estimates is heavily influenced by the non randomness of the subsamples. Gori (1992) proposes a study of this problem in two cases of two social agencies, but an extension to the case of more than two social agencies is still lacking at this point.

[6] Having considered the universe of the social agencies, the parameters inherent to the social agencies are fixed effects. Indeed we do not decompose the effectiveness as postulated in section 2c. Therefore there was no need to utilize the Multilevel Analysis.

Figure 1: *Cumulative Distributions of the latent variables*

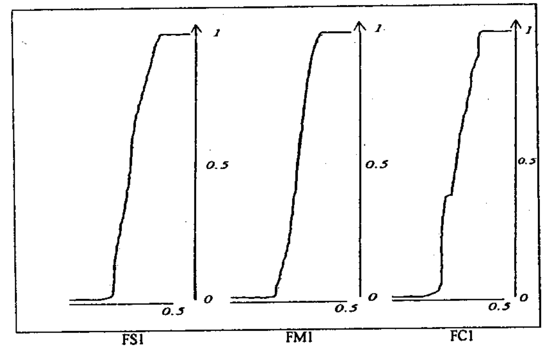

Table 3: *Estimations and standard errors of the parameters*

Variables		Estimation	Standard error		Estimation	Standard error		Estimation	Standard error	
INT		0.179013	0.42328		-0.293252	0.32140		-0.035074	0.22631	
F0		0.702301	0.05986		0.567080	0.07067		0.708070	0.07376	
V1		-0.316453	0.37974		0.111068	0.29559		-0.027746	0.20211	
U1		0.00904327	0.34456		0.153666	0.26650		-0.140215	0.18541	
U2		-0.171813	0.44592		-0.670021	0.34592		-0.069402	0.23843	
MCN		-0.427725	0.27051		-0.284346	0.21095		0.096989	0.14511	
PR0		-0.214866	0.27587		-0.187318	0.21251		-0.210229	0.14605	
RSA1		0.368478	0.27334		0.371480	0.13613		-0.037974	0.17453	OLS
RSA2		0.195361	0.18386		0.319548	0.14581		0.036307	0.17500	
RSA3		0.090738	0.08220		0.134334	0.09586		-0.011763	0.10565	
RSA4		0.183009	0.07963		0.185559	0.09942		0.092408	0.09613	
I1		0.895490	0.86308		0.665183	0.39349		-0.254697	0.31801	
I2		0.800830	0.48131		0.752619	0.41328		0.219688	0.08680	
I3		0.244861	0.33049		0.867803	0.25647		0.219688	0.08680	
I4	FM1	0.509220	0.29733	FC1	0.657681	0.22866	FS1	0.196530	0.15638	
INT		0.281961	0.42152		-0.301730	0.32210		-0.031446	0.22452	
F0		0.601513	0.07562		0.548824	0.08614		0.598249	0.08639	
V1		-0.286105	0.37513		0.182117	0.29680		-0.00573215	0.19915	
U1		-0.064152	0.34047		0.126007	0.26819		-0.157545	0.18363	
U2		0.067896	0.44814		-0.600076	0.35046		-0.00624369	0.23609	
MCN		-0.381401	0.26656		-0.229918	0.21223		0.104402	0.14313	
PR0		-0.141028	0.27578		-0.121240	0.21459		-0.184113	0.14494	
RSA1		0.365700	0.34861		0.365283	0.16760		0.024663	0.20496	SURE
RSA2		0.131445	0.24659		0.287541	0.18086		0.145475	0.19391	
RSA3		0.285164	0.10458		0.219554	0.11782		0.222087	0.12381	
RSA4		0.117353	0.10384		0.101879	0.12355		0.081229	0.11511	
I1		0.468137	1.02470		0.602479	0.39918		-0.313396	0.30692	
I2		-0.053030	0.75421		0.414974	0.48207		0.195712	0.08617	
I3		0.046121	0.33419		0.778766	0.25960		0.195713	0.08617	
I4		0.501007	0.29661		0.653265	0.22941		0.221006	0.15801	

References

Aitkin M., Longford N. (1986) Statistical modelling issues in school effectiveness studies, *Journal of the Royal Statistical Society*, 149, 1, 1-43.

Carey V., Zeger S. L., Diggle P. (1993) Modelling multivariate binary data with alternating logistic regressions, *Biometrika*, 80, 3, 517-526.

Copas J. B., Li H.G. (1997) Inference for non-random samples, *Journal of Royal Statistical Society, B*, 59, 1, 55-95.

Cox D. R., McCullagh P. (1982) Some Aspects of Analysis of Covariance, *Biometrics*, 38, 541-561.
Fitzmaurice G. M., Laird N. M. (1997) Regression models for mixed discrete and continuous responses with potentially missing values, *Biometrics*, 53, 110-122.
Fleiss J. L. (1986) *The design and analysis of clinical experiments*, Ed. Wiley.
Gelman A., Rubin D. B. (1992) Inference from Iterative Simulation using Multiple Sequences, *Statistical Science*, 7, 4, 457-472.
Geyer J. C. (1992) Practical Markov Chain Monte Carlo, *Statistical Science*, 7, 4, 473-483.
Gori E. (1992) La valutazione dell'efficienza ed efficacia dell'istruzione, *Comunicazione invitata alla XXXVI Riunione Scientifica della S.I.S.*, Pescara, 21-24 Aprile 1992, 219-229.
Gori E., Vittadini G. (1999) La valutazione dell'efficienza ed efficacia dei servizi alla persona. Impostazione e metodi, in *Qualità e valutazione nei servizi di pubblica utilità*, Etas, 121-241.
Haagen K. L., Vittadini G. (1998) Restricted Regression Component Decomposition, *Metron*, 56, 1-2, 53-75.
Heagerty P. J., Zeger S. L. (1996) Marginal regression models for clustered ordinal measurements, *Journal of the american statistical association*, 91, 435, 1024-1036.
Hox J.J. (1994) Hierarchical regression models for interviewer and respondent effects, *Sociological methods and research*, 22, 3, 300-318.
Hox J.J. (1995) *Applied Multilevel Analysis*, TT-Publikaties, Amsterdam.
Judge G. G., Griffiths W. E., Hill R. C., Lee T. S., Lütkepohl H. (1980) *The Theory and Practice of Econometrics*, Wiley, New York.
Liang K. Y., Zeger S. L., Qaqish B. (1992) Multivariate regression analyses for categorical data, *Journal of Royal Statistical Society*, B, 54, 1, 3-40.
Rasch G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Danmarks Paedagogiske Istitut, Copenaghen.
Rice N., Leyland A. (1996), Multilevel models: applications to health data, *Journal of health services research and policy*, 1, 3, 154-164.
Rilstone P., Veall M. (1996), Using bootstrapped confidence intervals for improved inferences with seemingly unrelated regression equations, *Econometric Theory*, 12, Dec 1996, 569-580.
Srivastava V. K., Giles D. E. A. (1987) *Seemingly Unrelated Regression Equations Models*, Marcel Dekker, New York
Vittadini G. (1997) Una metodologia statistica per la performance evaluation dei servizi alla persona di pubblica utilità, Atti del convegno «La statistica per le imprese» organizzato dal SIS, Torino, 2-4 IV-1997, I, 387-401
Vittadini G. (1999) "Analysis of qualitative variables in structural models with unique solutions", in *Classification and Data Analysis – Theory and Application*, a cura di M. Vichi e O. Opitz, Springer (1999), 203-210.
Young F. W., de Leeuw J., Takane Y (1976) Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features, *Psychometrika*, 41, 4, 505-529.

Nonparametric Estimation Methods for Sparse Contingency Tables

Riccardo Borgoni
Attuariato Rami Danni RAS SpA
Corso Italia, 23 - 20122 Milano, Italy

Corrado Provasi
Dipartimento di Scienze Statistiche, Università di Padova
Via S. Francesco, 33 – 35121 Padova, Italy

Abstract: The problems related with multinomial sparse data analysis have been widely underlined in statistical literature in recent years. Concerning the estimation of the mass distribution, it has been widely spread the usage of nonparametric methods, particularly in the framework of ordinal variables. The aim of this paper is to evaluate the performance of kernel estimators in the framework of sparse contingency tables with ordinal variables comparing them with alternative methodologies. Moreover, an approach to estimate the mass distribution nominal variables based on a kernel estimator is proposed. At the end a case study in actuarial field is presented.

Keywords: Ordinal and nominal variables, Kernel estimator, Sparse data.

1. Introduction

Data in the form of counts occur often in statistical practice. The empirical mass function is a good estimator of the true distribution under usual asymptotic conditions, but it is very poor with small or even moderate cell counts, when sample zeros can appear. In such situation, Bishop, Fiemberg and Holland (1975) introduced the idea of *sparse asymptotic conditions* to give a more realistic asymptotic framework in which the number of classes and the sample size increase together. They also introduced a well known estimator (hereafter *BFH estimator*) obtained by summing a small value to each cell frequency. Even if this estimator is not consistent, it has been proved that it has a better behaviour than empirical distribution in a sparse context. In a sparse asymptotic context kernel estimator has good asymptotic properties (Hall and Titterington, 1987).
In this paper we make a review of the kernel estimator for univariate and multivariate contingency tables with ordered marginals and an evaluation of its performance. Moreover, we make an attempt to apply this approach to nominal variables. In the end we try to underline the utility of smoothing by an application. In section 2 we briefly review some results on the kernel estimator for ordinal data in the unidimensional and multidimensional cases and some useful corrections to improve it. In the multidimensional case, a simulation study is performed to

A possible drawback for the method could appear the choice of the smoother and of the bandwidth parameter, a critical point in nonparametric estimation. Notwithstanding, this choice is somewhat less important for *submodel plots*. In the comparison of two nonparametric estimates, if the same estimator with the same bandwidth parameter is used for both curves, then the pointwise bias in the estimates will be canceled. In a different setting, Bowman and Young (1996) showed that the comparison of nonparametric curves yields a remarkable stability of results over a wide (reasonable) range of the bandwidth parameters.

In addition, we note that the procedure could be time consuming in the presence of many covariates. However, the method can be strongly enhanced by a dynamic graphics device that allows for an easy and quick change of the projection directions.

Further work will involve a general use of the proposed method to compare complex multivariate regression models, not necessarily restricted to parametric models (e.g. generalized additive models, projection pursuit regression models).

References

Bedrick E.J., Tsai C.L. (1994). Model Selection for Multivariate Regression in Small Samples. *Biometrics*, 50, pp. 226-231.

Bowman A., Young S. (1996). Graphical comparison of nonparametric curves. *Applied Statistics*, 45, pp. 83-98.

Brown, B.W. (1980). Prediction analysis for binary data, *Biostatistics Casebook*, Miller, R.J., Efron, B., Brown, B.W. & Moses, L.E. (Eds.), Wiley, New York.

Collett D. (1991). *Modelling binary data*. Chapman and Hall, 1991.

Cook R.D. (1998). *Regression Graphics*. Wiley, New York.

Cook R.D., Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley, New York.

Jovanovic, B.D. and Hosmer, D.W. (1997) A simulation of the performance of Cp in model selection for logistic and Poisson regression, *Computational Statistics and Data Analysis*, 23, pp. 373-379.

On the Use of Multivariate Regression Models in the Context of Multilevel Analysis

Giorgio Vittadini

Department of Statistical Sciences - University of Milan-Bicocca. E-mail: giorgio.vittadini@unimib.it

Abstract: The use of Multivariate Regression Models with mixed data to evalu and decompose relative effectiveness of different social agencies prese numerous problems. The solution proposed is to use the Seemingly Unrela Equations Models (SURE) in the framework of Multilevel Analysis, follow quantification of the response variables by means of simultane Multidimensional Scaling methods. An example is provided.

Keywords: Multivariate Regression Analysis, Mixed Data, Multilevel Analy Interaction Parameters, SURE Analysis, Relative Effectiveness.

1. Study with mixed indicators

In the studies about the effectiveness evaluation of a social service, the outcom defined as a long-term result of the output onto a particular aspect of the u Such an outcome can be measured by an appropriate set of indicators.
Of particular relevance is the relative effectiveness which evaluates diffe social agencies (hospitals, schools, and etc...). In such cases, the outcome depe on the explicative variables connected with the users and on other varia concerning the agencies. Outcomes can be described by a set of mixed indica or by latent variables, obtained with measurement errors from those indicat
To this aim the following models have been utilized:

- Multivariate Models with mixed variables that take into account associat and correlations among outcomes.
- Multilevel Models which clearly reveal the relative effectiveness of e social agency.

This paper proposes an integration of the two families of instruments.

2. Multivariate Models and relative effectiveness

From the interpretative point of view, it is reasonable to calculate simultaneo the value of outcomes which describe different but correlated aspects of the " of well being" of the user. The Multivariate Models are suitable for this purpo