Young Researchers' Day

23 September, 2011

9^{00} **Daniel Koch**   Large Portfolio Optimization by Wavelet Thresholding

9^{30} **Federico Rotolo**   A Copula-Based Simulation Method for Clustered Multi-State Survival Data

10^{00} **Fabian Bocart**   The Puzzling Volatility of Art Prices

Coffee Break

11^{00} **Cedric Taverne**   A Model Based on the Beta Distribution to Deal with Rating Scales

11^{30} **Marco Munda**   Adjusting for centre effects in the analysis of survival data from multicentre clinical trials

12^{00} **Bernard Francq** How to Accept the Equivalence of Two Measurement Methods? Comparison and Improvements of the Bland and Altman's Approach and Errors-in-Variables Regression

The seminar is followed by a sandwich lunch in the cafeteria.

The Young Researchers' Day is held in room c.115 of the Institut de statistique, biostatistique et sciences actuarielles, Voie du Roman Pays 20, Louvain-la-Neuve.

Large Portfolio Optimization by Wavelet Thresholding

Daniel Koch   (daniel.koch@uclouvain.be)

The mean-variance portfolio optimization framework takes the form of a quadratic programming problem which uses the mean return vector and the cross-covariance matrix of a N-dimensional stationary process as input. Due to the fact that the process cannot be observed directly, the mean return vector and the covariance matrix need to be replaced by estimates. In high-dimensional settings, e.g. when the number of assets is large relative to the sample size, the sample covariance matrix is badly conditioned. This leads to an estimate of the optimal solution vector characterized by a poor out-of-sample performance. Hence, building a regularized version of the sample covariance matrix helps improve the prediction power of the estimate.

Unbalanced Haar wavelets represent a powerful tool to design appropriate regularization methods. However, standard non-linear thresholding methods deal with each wavelet coefficient individually and hence do not take into account the structure which exists among the wavelet coefficients.

In order to overcome full non-linearity, we propose an iterative thresholding algorithm submitted to constraints which are derived from the hierarchical structure between the wavelet coefficients.

Simulation studies show the good performance of this new form of thresholding compared to existing regularization methods.

A Copula-Based Simulation Method for Clustered Multi-State Survival Data

Federico Rotolo   (federico.rotolo@stat.unipd.it)

Generating survival data with a clustered and multi-state structure is useful to study Multi-State Models [5], Competing Risks Models [4] and Frailty Models [2]. The simulation of such kind of data is not straightforward as one needs to introduce dependence between times of different transitions while taking under control the probability of each competing event, the median sojourn time in each state, the effect of covariates and the type and magnitude of heterogeneity.

Here we propose a simulation procedure based on Clayton copulas [1] for the joint distribution of times of each competing events block. It allows to specify the marginal distributions of time variables, while their dependence is induced by the copula. Furthermore, even though a dependence is obtained between all the time variables, only some joint distributions have to be handled.

The choice of simulation parameters is done by numerical minimization of a criterion function based on the ratio of target and observed values of median times and of probabilities of competing events.

The proposed method further allows to simulate discrete and continuous covariates and to specify their effect on each transition in a proportional hazards way. A frailty term can be added, too, in order to provide clustering. No particular restriction is needed on covariates distributions, frailty distribution, number and sizes of clusters.

An example is provided simulating data mimicking those from an Italian multi-center study on head and neck cancer [3]. The multi-state structure of these data arises from the interest in studying both time to local relapses and to distant metastases before death.

We show that our proposed method reaches very good convergence to the target values.

References

[1] D. G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

[2] L. Duchateau and P. Janssen. *The frailty model.* Springer, 2008.

[3] Filippo Grillo Ruggieri, M.P. Pace, F. Bunkeila, F. Cartei, B.M. Panizza, L. Fabbietti, G. Moroni, S. Cammelli, P. Api, C. Giorgetti, and E. Barbieri. Subcutaneous amifostine in head and neck cancer radiotherapy. *I supplementi di Tumori*, 4(1), 2005.

[4] M. Pintilie. *Competing risks: a practical perspective.* Wiley, 2006.

[5] H Putter, M Fiocco, and R B Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*, 26(11):2389–430, 2007.

The Puzzling Volatility of Art Prices

FABIAN BOCART　(fabian.bocart@uclouvain.be)

We suggest a new heteroskedastic hedonic regression model that takes into account time-varying volatility and is applied to a blue chips art market. Furthermore, we use a nonparametric local likelihood estimator that is more precise than the of ten used dummy variables method. The empirical analysis reveals that errors are considerably non-Gaussian, and that a student distribution with time-varying scale and degrees of freedom does well in explaining deviations of prices from their expectation.

A Model Based on the Beta Distribution to Deal with Rating Scales

CEDRIC TAVERNE　(cedric.taverne@uclouvain.be)

Rating scales are used everywhere in questionnaires whatever the subject of interest. It could be a Likert scale from *Strongly disagree* to *Strongly agree* or a scoring scale from 0 to 10. In modeling, these kind of scale are very often used as response variable in linear regression with normal errors. With this approach, the fitted values might exceed the original bounds of the scale. This is quite problematic even if the model fits the data very well. Furthermore, responses on rating scales are generally skewed and often inflated at the bounds of the scale.

In response to that problem we develop a model which can be seen as an adaptation of the beta regression proposed by Ferrari and Cribari-Neto in 2004. They proposed a regression model where the response is conditionally beta distributed. The relation between the mean response and the regressors is modeled through a logit link. A precision parameter is also estimated. In our work, we add three novelties: first, we deal with the inflation at the lower bound by coupling the beta regression model with a simple binary logit model. Second, we have adapted the expression of the likelihood to fit discrete rating scales. Third, we jointly model the dispersion treated as a constant in Ferrari and Cribari-Neto (2004). Our goal is to fit this extended model into the scheme of stated choice experiment wherein complex correlations are linking the observations used for estimation.

References

[1] Silvia Ferrari and Francisco Cribari-Neto. Beta Regression for Modelling Rates and Proportions *Journal of Applied Statistics*, Taylor and Francis Journals, vol. 31(7), pages 799-815, 2004.

Adjusting for centre effects in the analysis of survival data from multicentre clinical trials

Marco Munda   (`marco.munda@uclouvain.be`)

In multicentre clinical trials, patients are recruited at several hospital centres. This primarily allows the enrolment of an adequate number of patients within the established time frame. It also deepens the generalisability of its findings.

Stratification or inclusion of fixed effects are traditional methods for taking clustered structures into account. In survival analysis, another possibility is the shared frailty model that has been introduced as an extension of the Cox model for clustered time-to-event data. Heterogeneity in baseline risks across centres is encompassed into a random quantity (the so-called frailty) that accounts for the association between event times.

It is still troublesome for the statistical practitioner to use the frailty model in a real-world application. On one hand, the choice of a particular distribution for the frailties is commonly driven by mathematical reasons rather than by clinical ones or by the data themselves. On the other hand, different distributions model different association structures in the data. Few results are available with regard to the misspecification of the frailty distribution. As a result, there is a lack of guidance on how to adjust for centre effects in a proportional hazards model. All in all, should the frailty model methodology be recommended to account for centre effects given that the frailty distribution might well be misspecified?

Concentrating on the treatment effect, a simulation study in the context of a cancer clinical trial is conducted in order to assess the performance of the frailty model over its competitors. A special attention is given to the misspecification of the frailty distribution.

How to Accept the Equivalence of Two Measurement Methods? Comparison and Improvements of the Bland and Altman's Approach and Errors-in-Variables Regression

Bernard Francq   (`bernard.francq@uclouvain.be`)

The needs of the industries to quickly assess the quality of products and the performance of the manufacturing methods leads to the development and improvement of alternative analytical methods sometimes faster, easier to handle, less expensive or even more accurate than the reference method. These so-called alternative methods should ideally lead to results comparable or equivalent to those obtained by a standard method known as a reference.

To compare two measurement methods, a certain characteristic of a sample can

be measured by the two methods in the experimental domain of interest. Firstly, to statistically test the equivalence of measurement methods, the pairs of points (X_i, Y_i), representing the measures given by the reference method and the alternative one can be modelled by an errors-in-variables regression (a straight line). The estimated parameters are very useful to test the equivalence. Indeed, an intercept significantly different from zero indicates a systematic analytical bias between the two methods and a slope significantly different from one indicates a proportional bias. A joint confidence interval can also be used to test the equivalence. Secondly, the differences between paired measures X_i-Y_i can be plotted against their averages and analyzed to assess the degree of agreement between the two measurement methods by computing the limits of agreement. This is the very well known and widely used Bland and Altman's approach.

We review and compare the two methodologies, the Bland and Altman's approach and the errors-in-variables regression, with simulations and real data. Then, we'll propose some improvements like the tolerance interval on the Bland and Altman's approach and how to accept the equivalence by specifying a practical difference threshold on the results given by two measurement methods in the approach of errors-in-variables regression.