Bulletin of Statistics & Economics,
Spring 2008; Volume 2; Number S08; Bull. Stat. Econ.
ISSN 0973-7022; Copyright © 2008 by BSE, CESER

Power Function for Tests of Null Hypotheses on Mutual Linear Regression Functions' Relations

Jiri Knizek[1], Jan Sindelar[2],Ladislav Beranek[3], Borivoj Vojtesek[4], Rudolf Nenutil[4], Kristyna Brozkova[4], Viktor Drazan[5], Martin Hubalek[6], Lubomir Kubacek[7]

[1]Charles University in Prague, Faculty of Medicine in Hradec Kralove:
Department of Medical Biophysics, Simkova 870, 500 38 Hradec Kralove, Czech Republic
Email: *knizekj@lfhk.cuni.cz*

[2]Academy of Sciences of the Czech Republic, Institute of Information Theory and Automation:
Department of Stochastic Informatics, Czech Republic

[3]University of South Bohemia in Ceske Budejovice, Faculty of Education:
Department of Computer Science, Czech Republic

[4]Masaryk Memorial Cancer Institute in Brno: Experimental oncology division, Czech Republic

[5]Academy of Sciences of the Czech Republic: Institute of Biophysics in Brno, Czech Republic

[6]University of Defense in Brno, Faculty of Military Health Sciences in Hradec Kralove: Institute of Molecular Pathology, Czech Republic

[7]Palacky University in Olomouc, Natural science Faculty: Department of Mathematical Analysis and Mathematical Applications, Czech Republic

ABSTRACT

The main purpose of this work is the derivation of relations for calculation of „power function for tests of null hypotheses on mutual relation of linear regression functions". The normality of error disturbances of processed data is the condition for power function validity. Presented regression model can be widely applied for sophisticated data processing, e.g. in genomics and proteomics.

Key words: Biostatistics; Regression model; Decision-making; Power of test; Power function

2000 Mathematics Subject Classification: 62F03, 62J99

1. INTRODUCTION

Thanks to the development in technologies for genome and proteome analysis, the accumulation of large amounts of data has taken place. So that these data can provide answers to important questions, it is necessary to find efficient biostatistics methods for their treatment (Benevides and Overman, 2005; Malini and Ventkatakrishna, 2006; Schrader and Bougeard, 1995; Smith and Dent, 2004; Tibshirani et al., 2004; Wu et al. 2003). Advanced statistical algorithms which can give more detailed structured information are inevitable in this case.

It seems[1] that the regression model "*disturbance-related sets of regression equtions,*

[1] Project *"Theoretical basis of new methodology of mathematical-statistical and fuzzy-logical identification and decision making in biomarkers from mass spectra"*, No. 301/06/0267, Grant Agency of the Czech Republic.

the case with contemporaneously correlated disturbances" or alternatively *"the case of set of equations with autocorrelations of the first-order*", see (Judge et al., 1985), is particularly proper for the evaluation of data in these cases. One regression function can represent e.g. the course of mass (Knizek et al., 2004a; Tibshirani et al., 2004; Wu et al., 2003) or Raman (Benevides and Overman, 2005; Malini Ventkatakrishna, 2006; Schrader and Bougeard, 1995; Smith and Dent, 2004) spectrum etc. Aim of this work is derivation of powerfunction for tests in these regression models.

The interpretive termination of this methodology results from classical statistical decision-making based on *p*-value and reached power of test $1-\beta(\alpha)$ at a chosen significance level of test α in the sense of conventional rules of statistical decision-making, see (Cohen, 1988; Daly and Bourke, 2000).

The power of a statistical test is used in biological and experimental medical science research applications for statistic decision making, unfortunately, only rarely. Scientific decision making isn't entire herewith. Often *power of test* serves as *an indicator of „saturation" of an trial by experiments*, i.e. as indicator whether results of statistical decision making changes very little when other experiments (from the same data source) are subsequently added.

2. Regression model

We express the common regression model according to Judge (Judge et al., 1985) (for example for M spectral courses) in the form

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_M \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{X}_M \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_M \end{pmatrix}$$

or alternatively in the brief form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \,,$$

where $\boldsymbol{y}$, $\boldsymbol{X}$, β and $\boldsymbol{e}$ are $(MT \times 1)$, $(MT \times K)$, $(K \times 1)$ a $(MT \times 1)$ dimensioned vectors of sub-vectors and matrices of sub-matrices, where $K = \Sigma_{i=1}^{M}(K_i + 1)$. It is assumed, nevertheless, that the mean values of sub-vectors $E[\boldsymbol{e}_i] = \boldsymbol{0}$, $i = 1,2,\ldots,M$, and $E(\boldsymbol{e}_i\boldsymbol{e}_j') = \sigma_{ij}\boldsymbol{I}_T$, where $i,j = 1,2,\ldots,M$. The covariance matrix of the vector of sub-vectors of error disturbances[2] $\boldsymbol{e}$ is given by the relation $E[\boldsymbol{ee}'] = \boldsymbol{\Omega}_{(MT\times MT)} = \boldsymbol{\Sigma}_{(M\times M)} \otimes \boldsymbol{I}_{(T\times T)}$.

This regression model has proved very well in the past in solution of related problems of proteomics, see (Knizek et al., 2004a; Knizek et al., 2004b).

3. Statistic decision making on mutual linear relations of linear regression functions

Spectral methods are plentifully used in molecular biology and also in many other fields of biology and biochemistry often in different, more or less rigorous, quantitative studies and analyses. Spectral dependence can be generally expressed with the help of linear regression function, e.g. as a superposition of orthogonal polynomials, see (Forsythe, 1957; Ralston, 1973; Golub and Van Loan, 1994). If we want to scientifically perform some statistic decision making *(i. e. use statistical test) on miscellaneous mutual linear relations among spectral dependences*, then this is enabled just by the below mentioned theorems. One of the many such problems is the *identification of biomarker areas* in mass or Raman spectrometry.

3.1 *p*-Value

Provided that the vector of sub vectors of error disturbances $\boldsymbol{e}$ has normal distribution, it is possible to test *the null hypothesis*

$$\left.\begin{array}{lcccc} H_0: & \boldsymbol{R}(\xi) & \beta & = & \boldsymbol{r}(\xi) \\ & (J\times K) & (K\times 1) & & (J\times 1) \end{array}\right\} \qquad (3.1)$$

[2] As a rule, these error disturbances dominantly consist of the influence of the so-called biological variability, further of the influences of laboratory experimental errors, variability at biological material preparation and influence of random disorders of prime mass spectra in consequence of technical limits of measuring apparatus (Knizek et al., 2007)

on any J-equational determined (e. g. biophysical) mutual linear relations of M *regression equations, where* $J \leq M$ and $K = \Sigma_{i=1}^{M}(K_i + 1)$. The null hypothesis (3.1) is then tested against its *two-sided alternative* H_1 stating that from J equations *at least one is not valid*[3].

We calculate the *p*-value for the null hypothesis (3.1) test by means of the relation

$$p = 1 - F_{J,\,MT-K}(\hat{\lambda}) \,. \tag{3.2}$$

$F_{J,\,MT-K}(\hat{\lambda})$ in (3.2) is the cumulative distribution function of Fisher-Snedecor distribution with J and $MT - K$ degrees of freedom and the test characteristic

$$\hat{\lambda} = \frac{(\boldsymbol{r} - \boldsymbol{R}\hat{\hat{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{B}}\boldsymbol{R}')^{-1}(\boldsymbol{r} - \boldsymbol{R}\hat{\hat{\beta}}) \Big/ J}{(\boldsymbol{y} - \boldsymbol{X}\hat{\hat{\beta}})'(\hat{\Sigma}_{(M\times M)}^{-1} \otimes \boldsymbol{I}_{(T\times T)})(\boldsymbol{y} - \boldsymbol{X}\hat{\hat{\beta}}) \Big/ (MT - K)} \,, \tag{3.3}$$

where $\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_{(K\times K)} = (\boldsymbol{X}'_{(K\times MT)}\,\hat{\boldsymbol{\Omega}}_{(MT\times MT)}^{-1}\,\boldsymbol{X}_{(MT\times K)})^{-1}$, where the matrix of estimation of covariance $\hat{\boldsymbol{\Omega}}_{(MT\times MT)}^{-1} = \hat{\Sigma}_{(M\times M)}^{-1} \otimes \boldsymbol{I}_{(T\times T)}$. Mixed variances matrix $\hat{\Sigma}_{(M\times M)}$ elements are estimated by means of the relations $\hat{\sigma}_{ij} = T^{-1}\hat{\boldsymbol{e}}_i'\hat{\boldsymbol{e}}_j$, $i, j = 1, 2, \ldots, M$. At the same time the so-called EGLS-estimation[4] of the vector of sub-vectors of the regression coefficients $\hat{\hat{\beta}}$ is a result of the two-stage solving of the set of normal equations $\hat{\hat{\beta}} = (\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{y}$, where in the first stage (the first approximation) the matrix of covariances $\boldsymbol{\Omega}$ estimation is approximated by unit matrix, i.e. $\hat{\boldsymbol{\Omega}}_{(MT\times MT)} \approx \boldsymbol{I}_{(MT\times MT)}$.

3.2 Power of test

Test statistics (3.3) has in the case of null hypothesis validity (3.1) the Fisher-Snedecor distribution of probability with J and $MT - K$ degrees of freedom. In case the null hypothesis (3.1) is not valid, i.e. $\boldsymbol{r} - \boldsymbol{R}\beta = \varsigma \neq \boldsymbol{0}$, then the distribution of test statistics (3.3) is non-central Fisher-Snedecor distribution with the following non-centrality parameter

[3] This property at the mentioned testing relations guarantees significant sensitivity of the so called “first catching” e.g. of otherwise a difficult-to-identify trend. In other words: only a small trace in data suffices for the algorithm to be able to recognize it.

[4] EGLS = estimated generalized least squares (Judge et al., 1985)

$$\delta = (\boldsymbol{r} - \boldsymbol{R}\beta)'(\boldsymbol{R}\boldsymbol{B}\boldsymbol{R}')^{-1}(\boldsymbol{r} - \boldsymbol{R}\beta).$$

Power of test in the alternative ς (i.e. in the case of unsatisfactory null hypothesis (3.1)) can be expressed by the relation

$$P\left\{F_{J,\,MT-K}(\delta) \geq F_{J,MT-K}(0;1-\alpha)\right\} = \text{power},$$

where $F_{J,\,MT-K}(0;1-\alpha)$ is α-critical value of Fisher-Snedecor central distribution and $F_{J,\,MT-K}(\delta)$ is a random quantity with the probability density ($k = J$, $l = MT - K$)

$$f_{k,l,\delta}(z) = \begin{cases} \exp\left(-\dfrac{\delta}{2}\right)\displaystyle\sum_{r=0}^{\infty}\frac{1}{r!}\left(\frac{\delta}{2}\right)^{r} \dfrac{\Gamma\left(\dfrac{k+l}{2}+r\right)\left(\dfrac{k}{l}\right)^{\frac{k}{2}+r} z^{\frac{l}{2}+r-1}}{\Gamma\left(\dfrac{k}{2}+r\right)\Gamma\left(\dfrac{l}{2}\right)\left(1+\dfrac{k}{l}z\right)^{\frac{k+l}{2}+r}} & , \quad z > 0, \\ 0 & , \quad z \leq 0. \end{cases}$$

Hence the power function

$$\text{power}(\boldsymbol{\beta}) = \int_{F_{k,\,l}\,(0\,;1-\alpha)}^{\infty} f_{k,l,\delta}(z)\,dz,$$

see fig. 1.

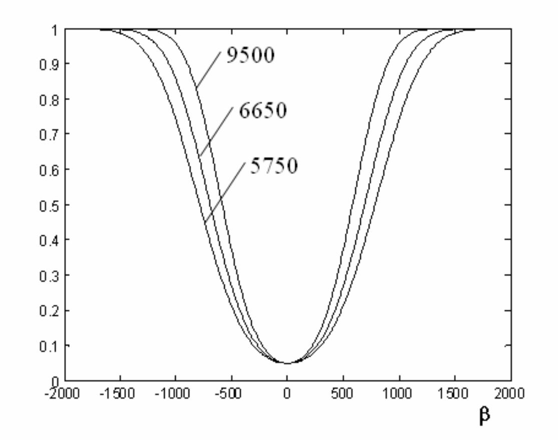

Figure 1: A typical power function course in case that all elements of the vector of sub-vector β match each other, for various ranges of sampling, i. e. for different M and/or T , in the concrete for $J = 5\,; MT - K = 9500$, $J = 5\,; MT - K = 6650$ and $J = 5\,; MT - K = 5750$.

Power of test for the test of the null hypothesis (3.1) is then calculated for a concrete EGLS-estimation of regression coefficients $\hat{\hat{\beta}}$ with the help of the formula

$$\textit{power of test} = \mathrm{power}(\hat{\hat{\beta}}) \ .$$

Power of test is the probability of rejection of the null hypothesis (3.1) when reality does not coincide with the null hypothesis, i. e. when $\boldsymbol{r} - \boldsymbol{R}\beta = \varsigma \neq \mathbf{0}$.

3.3 Practical calculation of power of test

We calculate in the MATLAB language quantity $\textit{power of test}$ with the help of the program structure (schematically outlined)

$$\mathrm{power\_of\_test} = 1 - \mathrm{ncfcdf}(\mathrm{finv}(1 - \alpha, J, MT - K), J, MT - K,\ \delta) \ .$$

4. Discussion and conclusions

There is not any doubt nowadays that proteomics needs new approaches. This work deals with a key algorithmic relation of one of these new approaches: by derivation of the formula for calculation of *„power function for tests of null hypotheses on mutual linear regression functions' relations"*.

With the help of variation of second number of degree of freedom $MT - K$, *the statistical analysis of power of test will enable* to extract these quite exclusive information:

- *Statistical estimation of the number of subsequent regression function (e.g. spectrum dependences)* that are to be additionally experimentally measured so as to fulfill the conventional requirements on *power of test* (Cohen, 1988; Daly and Bourke, 2000),

- *statistical estimation of the need for higher density of sampling* (i. e. the independent variable sampling); such, that it fulfills the conventional requirements on *power of test* (Cohen, 1988; Daly and Bourke, 2000). At the same time it generally holds that *these two estimations can influence each other*.
- I. e. the increase of sampling density can reduce the need for dependences that are to be additionally experimentally measured. And vice-versa, the increase of the number of dependences can reduce the requirement for higher sampling density.

References

1. Benevides, J.M., Overman, S.A., 2005, *Raman, polarized Raman and ultraviolet resonance Raman spectroscopy of nucleic acids and their complexes*, J. Raman Spect., 36, 279-299.

2. Cohen, J, 1988, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Lawrence Erlbaum, Mahwah, New Jersey.

3. Daly, L.E., Bourke, G.J., 2000, *Interpretation and Uses of Medical Statistics*, 5th edn. Blackwell Science, Oxford.

4. Forsythe, G.E., 1957, *Generation and Use of Orthogonal Polynomials for Data-fitting on a Digital Computer*. J. Soc. Indust. Appl. Math., 5, 74-88.

5. Golub, G.H., Van Loan, C.F., 1996, *Matrix Computations*. The Johns Hopkins University Press, Baltimore.

6. Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H., Tsoung-Chao, L., 1985, *The Theory and Practice of Econometrics*, J. Wiley, New York.

7. Knizek, J., 2004a, *Theoretical basis of new methodology of mathematical-statistical decision making with the help of biomarkers from mass spectra*, Acta Medica (Hradec Kralove), 47, 291-298.

8. Knizek, J., Bergmann, M., Sindelar, J., Kovarova, H., 2004b *MIAPS - programovy system pro odhadovani poradi vzajemne podobnosti/nepodobnosti chovani proteinu v case (in Czech)*. Acta Medica (Hradec Kralove) Supplementum; 47, 131-135.

9. Knizek, J., Pulpan, Z., Hubalek, M., Beranek, L., Pokorny, P., 2007, *Stochastic Model of Mass Spectrum Random Disturbances and its Simulation*. In publication process.

10. Malini, R., Venkatakrishna, K., 2006, *Discrimination of normal, inflammatory, premalignant, and malignant oral tissue*: A Raman spectroscopy study, Biopolymers, 81, 79-193.

11. Ralston, A., 1973, *A First Course in Numerical Analysis*. McGraw Hill Book Company, New York.

12. Schrader, B., Bougeard, D., 1995, *Infrared and Raman Spectroscopy*. Weinheim, Wiley-VCH Verlag GmbH, Berlin.

13. Smith, E., Dent, G., 2004, *Modern Raman Spectroscopy*: A Practical Approach, Wiley, New York.

14. Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., Le, Q.T., 2004, *Sample classification from protein mass spectrometry by "peak probability contrasts*", Bioinformatics - Bioinformatics Advance Access, Oxford University Press, 1-34.

15. Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., Zhao, H., 2003, *Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data*, Bioinformatics, 19, 1636-1643.