UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

INTERNATIONAL YEAR OF STATISTICS

CLADAG 2013

9 th SCIENTIFIC MEETING OF THE CLASSIFICATION AND DATA ANALYSIS GROUP OF THE ITALIAN STATISTICAL SOCIETY

September 18 - 20, 2013
University of Modena and Reggio Emilia
San Geminiano Complex - Modena, Italy

Book of Abstracts

Editors: Tommaso Minerva, Isabella Morlini, Francesco Palumbo

CLEUP
ISBN: 9788867871179

Patronage

Dipartimento di Economia Marco Biagi

Dipartimento di Comunicazione ed Economia

Dipartimento di Scienze Fisiche, Informatiche e Matematiche

Table of Contents

An overview on multiple regression models based on permutation tests

Massimiliano Giacalone[1] and Angela Alibrandi[2]

Abstract
When the population, from which the samples are extracted, is not normally distributed, or if the sample size is particularly reduced, it becomes preferable to use nonparametric statistical tests. Within the regression models, it is possible to use permutation tests, considering their ownerships of optimality, when the normal distribution of the response variables is not guaranteed especially in the multivariate context. In the literature there are numerous permutation tests applicable to the estimation of regression models. In this paper we focused our attention on the permutation tests of the independent variables, proposed by Oja, and other methods of nonparametric inference, in the regression models context.

1 Permutation test in regression models

In many cases, when the classical conditions of regression models are not respected, it's possible to use the permutation tests, considering their ownerships of optimality, especially in the multivariate context. The evaluation of the parameters' significance is an inferential procedure, based on randomization tests (if the same experimental plan justifies them) or permutation tests (if the observed samples are random, so that the analyzed samples justify the calculations) (Kempthorne and Doerfler, 1969). Through the use of permutation tests, we assess the null hypothesis of casualness: in fact, it suggests that, if the examined phenomenon has a certain tendency, confirmed by a model that appears as gives, it is a purely accidental effect of the observations in casual order. We proceed choosing a useful S statistic test to

1 Massimiliano Giacalone, lecturer with annual appointment, Bologna University, email: massimilia.giacalone@unibo.it

2 Angela Alibrandi, Department of Economical, Business, Environmental Sciences and Quantitative Methods, Messina University, email; aalibrandi@unime.it

measure the entity of the phenomenon of interest in relation to the observed data and we compare the observed s statistic test value of S and the distribution of S, obtained by casually rearranging the data. The test is based on the following principle: if the null hypothesis were true, then all the possible arrangements of the observations would have equal probability to verify, that is the order of the observed data is one of the possible equally probable arrangements and s appears as one of the possible values of the randomization distribution of S. If s is a significant value, then the null hypothesis is rejected, then, for implication, the alternative hypothesis is considered more reasonable. The significance level of s is, the percentage of the values that are great or equal to s in the randomization distribution. It represents a measure of evidence strength against the null hypothesis.

2 The Oja permutation test of the independent variables

The experimental plan presented by Oja (1987) considers n subjects to which a treatment variable x is assigned in order to study their effects on a response variable Y. In addition, for each k subject, further Z explanatory variables (covariates) are considered. The non-parametric permutation tests proposed by Oja are relative to a completely permuted plane: in fact, they are based on the assumption that the treatment values are randomly assigned to the subjects. Therefore, the permutation distribution used to verify the significance of a relationship between X and Y, taking into account the effects of the Z covariates, is obtained by permuting the X values to the n statistical units. In formal terms, this is a regression plan model where the results can be generalized to multiple regression. The model can be expressed as:

$$Y_i=\alpha + \beta X_i + \gamma Z_i + \varepsilon_i \tag{1}$$

with i=1,...,n, where α, β and γ are unknown parameters, X is the explicative variable of the plane such that Σ_i^n=0, z is the explanatory covariable and $\varepsilon_1,\ldots,\varepsilon_n$ are independent and identically distributed random errors with zero mean. The attention is focused on the β parameter; therefore the null hypothesis is expressed by H_0: β=0; α and γ are nuisance parameters. Let's suppose that Y, X and Z are given for all i: the X variable is considered as a realization of the random permutation x^* of x. Then, the corresponding y values, which have not been realized, are $y^*=y+\beta(x^*-x)$, from which we can easily obtain $y^*-\beta x^*=y-\beta x$. The test statistic proposed by OJA to assess the null hypothesis is:

$$T=\sum_{i<j<k}\Delta_{ijk}{}^{y}\Delta_{ijk}x^{*} \quad \text{where} \quad \Delta_{ijk}{}^{y}=\begin{Vmatrix}1 & 1 & 1\\ y_i & y_j & y_k\\ z_i & z_j & z_k\end{Vmatrix} \tag{2}$$

with $i<j<k$ and similarly for $\Delta_{ijk}x^{*}$.

This statistic is not easily calculable; so, Oja proposed an alternative form of this test in order to facilitate the calculations:

$$T=\sum_{i}\hat{y}_i x_i^{*} \quad \text{where} \quad \hat{y}_i=\sum_{j<k}\Delta_{ijk}{}^{y}\delta_{jk}{}^{x} \quad \text{with} \quad \Delta_{jk}{}^{x}=\begin{Vmatrix}1 & 1\\ z_j & z_k\end{Vmatrix} \tag{3}$$

where $j<k$. Collins (1987) tried to approximate the permutation distributions proposed by Oja, with other distributions. In particular Oja has suggested a standardized normal approximation or, equivalently, to square the test statistic proposal and compare the

result with the critical value of a χ^2 . Of course there is no certainty that these distributions provide adequate approximation to the corresponding distributions of permutation. Collins (1987) has proposed, in his work, a reformulation of the Oja statistics, using easier methods of calculation, to obtain the explicit formulas of the moments of permutation and especially to have the advantage of being able to recognize a beta distribution as an approximation of the exact distribution of null permutation. These procedures presented by Oja and Collins have not had much success and development because, by permuting the independent variables, they violate the principle of ancillarity, according to which the plan should be subject to maintain the collinearity between the explanatory variables (Kennedy, 1995).

3 Some alternative approaches of permutation tests in regression

Other methods of nonparametric inference, in a regression model are:

- **The residual permutation test of the complete model**, proposed by Ter Braak (1992), that is analogous to a bootstrap test and it consists of permuting the residual samples of a multiple regression in order to produce a distribution that can be compared with the value sample of a statistical test. In effects, this test is not a permutation test in the traditional sense, because the data are transformed to get the residues, before their exchange happens. Moreover, it is hybrid between a permutation test and a bootstrap test and its justification can be derived from both value b* around the true b value in the bootstrap samples. Similarly the variability of F_{obs} to test $\beta=\beta_0$ are similar to the variability of F^* to test β =b. These appreciable properties are also justified because the F used statistic test is asymptotically pivotal; whatever the distribution of errors, the F asymptotic distribution doesn't depend on the parameters in the model. This test is well applicable with great samples because the variability of b around the truth β is similar to the variability of the resampled that are not tested (Levin and Robbins, 1983; Gail, Tan and Piantadosi, 1988; Kennedy and Cade, 1996).
- **The permutation test of the dependent variable,** used to verify the null hypothesis H_0: $\beta = 0$. It can be performed by comparing the values of the F test statistic with the distribution obtained by permuting the Y observations, to casually assign to the sets of the observations of X and Z independent variables. Manly (1991) proposes three possible motives to justify this type of permutation approach: in first place, the n observations can be a casual sample from a population of possible observations, where the Y variable could be independent from the X and Z explanatory variables; in according to place, the values of the experimental variables X and Z can casually be assigned to n statistical units and, therefore, the values of the Y response variable can be observed (Y would not to be influenced by the X and Z variables). Besides, if the variable Y and the explanatory variables X and Z are independent, all the possible joining among every value Y and every values X and Z are equally probable in relation to a potential mechanism that generates the data;
- **The exact restricted permutation tests for partial regression models,** fundamentally developed by Brown and Maritz (1982), furnishes an exact permutation test for a partial regression model, within the regression plain. The

proposed scheme, united to a suitable experimental plan, is used for the inference on the regression coefficient β of X, when exists another explanatory Z variable that influences the Y response. The X coefficient is therefore a disturb parameter.

4 Final remarks

In this paper we revisited the use of permutation tests to evaluate non parametric inference in a regression model. Comparing the randomization and permutation tests with the conventional test for inference in a regression model, we can underline some aspects. First of all, the randomization and permutation tests have two important advantages: they are valid and opportunely applicable without casual samples and they allow to select a statistic test appropriated for a particular considered situation. Nevertheless, it's not possible to generalize the conclusions of a randomization test to the whole population of interest. In fact a randomization test identifies the probability that a phenomenon of interest is casual. The concept of population from which to extract samples of observations is not fundamental and this is the reason for which the casual sampling is not required. In the other hand, the generalization of results of the conventional tests to the whole population is based on the assumption, not always verifiable, that the observed samples are equivalent to a casual sample or that the data are available for the whole population of interest (but this last condition is practically unrealizable). So, randomization and permutation tests represent a methodologically adequate solution in a large number of practical experimental contexts in which the samples are not random. These methods seem to be appropriate for particular conditions, alternatively to conventional tests whose assumptions are too restricted.

References

1. Brown, B. M., Maritz, J. S.: Distribution- free methods in regression. In: Australian Journal of Statistics, 24, pp. 318-333 (1982).
2. Collins, M. F.: A permutation test for planar regression. In: Australian Journal of Statistics, 29, pp. 303-308 (1987).
3. Gail, M. H., Tan, W.Y. , Piantadosi, S.: Tests for no treatment effect in randomized clinical trials. In : Biometrika, 75, pp. 57-64 (1988).
4. Kempthorne, O., Dorfler, T.E.: The behavior of some significance tests under experimental randomization. In: Biometrika, 56, pp. 231-248 (1969).
5. Kennedy, P.E.: Randomization tests in econometrics. In: Journal of Business and Economic Statistics, 13, pp. 85-94 (1995).
6. Kennedy, P.E, Cade, B.S.: Randomization tests for multiple regression. In: Communication in Statistics – Simulation and Computation, 25, pp. 923-936 (1996).
7. Levin, B., Robbins, H.: Urn models for regression analysis, with applications to employment discrimination studies. In: Law and Contemporary Problems, 46, pp. 247-267 (1983).
8. Mainly, B. J. F.: Randomization and Monte Carlo methods in biology. Chapman and Hall. London (1991).
9. Oja, H.,: On permutation tests in multiple regression and analysis of covariance problems. In: Australian Journal of Statistics, 29, pp.91-100 (1987).
10. Ter Braak, C. J. F.: Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Bootstrapping and Related Techniques, (K. H. Jockel, G. Rothe and W. Sendler, Eds.) New York, Springer Verlag, pp. 79-85 (1992).