Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 11 (18): 3328-3332, 2011
ISSN 1812-5654 / DOI: 10.3923/jas.2011.3328.3332

Significant Tests of Coefficient Multiple Regressions by using Permutation Methods

Ali Shadrokh
Department of Statistics, University of Payame Noor, 19395-4697,
Islamic Republic of Iran

Abstract: Tests of significance of a single partial regression coefficient in a multiple regression model are often made in situations where the standard assumptions underlying the probability calculation (for example assumption of normally of random error term) do not hold. When the random error term fails to fulfill some of these assumptions, one need resort to some other nonparametric methods to carry out statistical inferences. Permutation methods are a branch of nonparametric methods. This study compared empirical type one error of different permutation strategies that proposed for testing nullity of a partial regression coefficient in a multiple regression model, using simulation and show that the type one error of Freedman and Lane's strategy is lower to than the other methods.

Key words: Permutation test, exchangeable observation, regression coefficient, empirical type one error, per mutative matrix

INTRODUCTION

The first descriptions of permutation tests for linear statistical models, including analysis of variance and regression, can be traced back to the work of Fisher (1935).

Permutation tests were first introduced by Fisher (1935) and Bizhannia *et al*. (2010). But they were not pleased so much because they needed time consuming calculations. But nowadays these calculations by production of fast and powerful computers that calculating of a p-value is faster than finding an amount of charts for a parametric test. Parametrical tests take place in the area of free- distribution. As we know for doing the question of testing of hypothesis, in addition to hypothesizes like independency of error term, stability of their variance and random sampling, we need to the hypothesis of normal distribution of error term, while for doing permutation tests we don't need to basic hypothesizes. In fact we for doing a non parametrical test, we use series of simple hypothesizes that leaves the researcher. One of these hypothesizes which in fact is the base for permutation tests is hypothesis of exchange ability of observations which defined like this.

Definition 1: Let's realize exchange ability of random vector (n) dimensional, $x = (x_1, x_2,..., x_n)$ with joint distribution of $f\ x_1, x_2,..., x_n\ (x_1, x_2,..., x_n)$ X is called exchangeable if joint density of observations per each permutation vector is suitable. So for each permutation vector which is shown by X we have:

$$f_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n) = f_{X_1^*,X_2^*,...,X_n^*}(x_1^*,x_2^*,...,x_n^*)$$

Various permutational strategies have been proposed for testing nullity of a partial regression coefficient in a multiple regression model (AL-Salihi *et al*., 2010; Bughio *et al*., 2002; Edriss *et al*., 2008; Kandhro *et al*., 2002; Laghari *et al*., 2003; Rashid *et al*., 2002; Serhat Odabas *et al*., 2007; Tariq *et al*., 2003; Alam, 2004). The proposed permutation methods for such tests have different bases in term of their philosophies and have been proposed in different contexts.

In Anderson and Legendre (1999) point of view just four cases of these strategies are suitable. These four methods are: Manly method (Manly, 1991), Kennedy method (Kennedy and Cade, 1996) Freedman-Lane method (Freedman and Lane, 1983) and Ter braak (Ter-Braak, 1992) but in this article we just compare Kennedy's method and Freedman and Lane's method.

KENNEDY'S METHOD

Suppose variable answer Y and variable $x_1, x_2,..., x_p$ and assume we have n observation for regression. Then we have this equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1} + \beta_p x_{ip} + \varepsilon_i \quad (1)$$

where, ε is (sentence wrong) error term and has an indeterminate distribution F with by a mean zero variance σ^2. We can write the Eq. 1 as the following vector-matrix:

$$Y_{n\times 1} = X_{1_{n\times p}} \beta_{1_{p\times 1}} + x_{p_{n\times 1}} \beta_p + \varepsilon, \quad \varepsilon \sim 0, \sigma^2 I_n \quad (2)$$

Where:

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X_1 = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}, x_p = \begin{pmatrix} x_{1p} \\ \vdots \\ x_{np} \end{pmatrix}, \beta_1 = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

We want to test hypothesis H_0: $\beta_p = 0$ in comparison with, H_2: $\beta_p \neq 0$. For doing this test by permutation method we follow Kennedy algorithm like this: we define matrix:

$$H_1 = X_1 (X_1^t X_1)^{-1} X_1^t$$

and then multiple correspondents of Eq. 2 in (I_n H_1) to solve equation shown here:

$$\tilde{y} = \tilde{x}_p \beta_p + \tilde{\varepsilon} \quad (3)$$

Where:

$$\tilde{y} = (I_n - H_1) y, \ \tilde{x}_{p_{n\times 1}} = (I_n - H_1) x_p, \ \tilde{\varepsilon}_{n\times 1} = (I_n - H_1)\varepsilon$$

Step 1: Then by using least error squares, we estimate β_p like this:

$$\tilde{\beta}_p = \frac{\tilde{x}_p^t \tilde{y}}{\tilde{x}_p^t \tilde{x}_p} = \frac{x_p^t (I_n - H_1) y}{x_p^t (I_n - H_1) x_p}$$

After that we estimate β_p, amount of statistic test $t = \tilde{\beta}_p / s(\tilde{\beta}_p)$ calculate and call it referential t.

Step 2: We do permutation for $\tilde{y}$ amounts and show it as $\tilde{y}^*$

Step 3: We make a regression equation on $\tilde{x}_p$, model of Eq. 3, from found vector. Then we estimate β_p by least sum of square error, like one given here:

$$\tilde{\beta}_p^* = \frac{\tilde{x}_p^t \tilde{y}^*}{\tilde{x}_p^t \tilde{x}_p}$$

Step 4: By repeating second and third step and finding t^*'s we also find permutation distribution t's, next, by using that we find p-value which is the relation of amount of permutation statistics that their absolute value is larger than their referential E absolute value and at least we admit or reject zero hypothesis.

PROBLEM WITH KENNEDY METHOD

In permutation methods, permutating amounts of y, causes amounts of ε permutate. Thus a question comes to mind is that whether this obligatory permutation changes its parametrical distributions or not? So we first give this definition.

Permutative matrix: Definition 2 permutative matrix is a square matrix that there are numbers of zero and only a one each row and the position of number one in each row is different from other rows. This matrix is shown by sample P and has a special feature of $P^t P = P P^t$.

Now, with respect to the above definition, we analyze the mentioned problem: We know that:

$$\varepsilon \overset{iid}{\sim} F(0, \sigma^2 I_n)$$

so, considering the above definition, we can put the permutated vector ε which is known by the notation of ε^* this way ε^* pε. As a result:

$$E(\varepsilon^*) = E(P\varepsilon) = 0, \ Var(\varepsilon^*) = Var(P\varepsilon) = \sigma^2 I_n$$

It is clear that by the permutation of values of *y* does not change the distribution of ε. Now, we study this in case of the reduced equation of which Kennedy has made use Eq. 3.

It can easily be proved that:

$$\tilde{\varepsilon} \overset{iid}{\sim} F\sigma^2 (I_n - H_1)$$

Regarding the definition 2, we can easily prove that:

$$E(\tilde{\varepsilon}^*) = E(P\tilde{\varepsilon}) = 0, \ Var(\tilde{\varepsilon}^*) = Var(P\tilde{\varepsilon}) = \sigma^2 (I_n - H_1^*)$$

Which shows the variance of $\tilde{\varepsilon}^*$ is dependent on $\tilde{H}_1^*$ thus by any permutation of $\tilde{H}_1, \sigma^2$, is multiplied by a new number that changes the variance. Therefore, by any permutation of $\tilde{y}$, the parameters of distribution of $\tilde{\varepsilon}$ changes. So due to the null hypothesis H_0: $\beta_P = 0$. The distribution of $\tilde{y}$ also changes by any permutation.

MODIFIED KENNEDY'S METHOD

Huh and Jhun (2001) improve this problem, with regard to this point that the matrix (I_n H_i) is the one to a power of its own value and with the rank of n' = n-p, added the following steps to kennedy's algorithm:

Step 1: First, we work the eigenvectors and rank of the matrix (I_n H_i) out

Step 2: We save the eigenvectors equivalent to the particular measures of one which equals the rank of in the form of (I_n H_i) the matrix

Step 3: Through the orthogonalizing process of Gram Schmith, we change/turn this matrix to an orthogonal one

Step 4: Divide each column of this matrix by the norm of that column so that an orthogonal unit vector is resulted. We come to realize this new matrix by V_1 the dimensions of which are ($n\times n'$)

Step 5: Using the spectrographic analysis and knowing the point that particular measures of matrix V_1 is one, we reduce the matrix (I_n H_i) this way:

$$(I_n - H_1) = V_1 I V_1^t = V_1 V_1^t$$

Step 6: After reduction of this matrix, we multiply two sides of the Eq. 3 by the matrix V_i^t. So that a new equation is resulted in this form:

$$\tilde{\tilde{y}}_{n'\times 1} = \tilde{\tilde{x}}_{Pn'\times 1}\beta_p + \tilde{\tilde{\varepsilon}}_{n'\times 1}, \quad \tilde{\tilde{\varepsilon}} \overset{iid}{\sim} F(0, \sigma^2 I_{n'})$$

Where:

$$\tilde{\tilde{y}} = V_1\tilde{y}, \quad \tilde{\tilde{x}}_p = V_1^t\tilde{x}_p, \quad \tilde{\tilde{\varepsilon}} = V_1^t\tilde{\varepsilon}$$

Now, we turn to the distribution of the permutated vector of $\tilde{\tilde{\varepsilon}}$. To do so, we again refer to the definition 2:

$$E(\tilde{\tilde{\varepsilon}}^*) = E(P\tilde{\tilde{\varepsilon}}) = 0, \; Var(\tilde{\tilde{\varepsilon}}^*) = Var(P\tilde{\tilde{\varepsilon}}) = \sigma^2 I_{n'}$$

So it is pinned down that $\tilde{\tilde{\varepsilon}}^* \sim \tilde{\tilde{\varepsilon}}$ and this gives clue to the fact that this new method keeps the distribution of the error expressions constant and stable even after the permutation.

After the modification/improvement of the Eq. 3 and turning it into the form (4), we apply the steps contributed by Kennedy to the new resulted equation.

SIMULATION

Anderson and Legendre (1999), having an extent simulation done (ignoring the error in Kennedy's method), showed that in Freedman-Lawn method, a Type I error probability is less than that the Kennedy's methods. Shadrokh and d'Aubigny (2010), Shadrokh (2011) analytically show that the Type I error of Freedman and Lane method is lower than that of Kennedy's approach.

Here, with the aid of a simulation, we intend to check out whether the claim Anderson and Legendre (1999) and Shadrokh (2011) have made remains valid after modifying the error in the method of Kennedy or not? To do so, we calculate the empirical probability of type I error in all the three permutating methods on the basis of a double regression model $y = \beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$ considering the four following factors:

- The sample size {10, 20, 30}
- The correlation between the two variances x_1, x_2, ρ = {0.1, 0.9}
- The quantity/value of the coefficient of regression which is not tested β = {0.5, 1.5}
- The distribution of error expressions: the exponential with the parameter 2

The focused simulation is done employing the software S-plus and the results are drawn as the following charts.

In all four case (showed in four chart) the probability of type one error of Freedman and Lane method is lower than that of Kennedy's method and modified Kennedy's method (Huh and Jhun's method).

The certain considerable point is that, considering Fig. 1a-b, once β equals a small quantity, ignoring the fact that the correlation is high or low and when sample size reaches 30, Freedman-Lane's method and modified Kennedy's method which is depicted in the charts as the Huh and Jhun method, both, show a convex quantity.

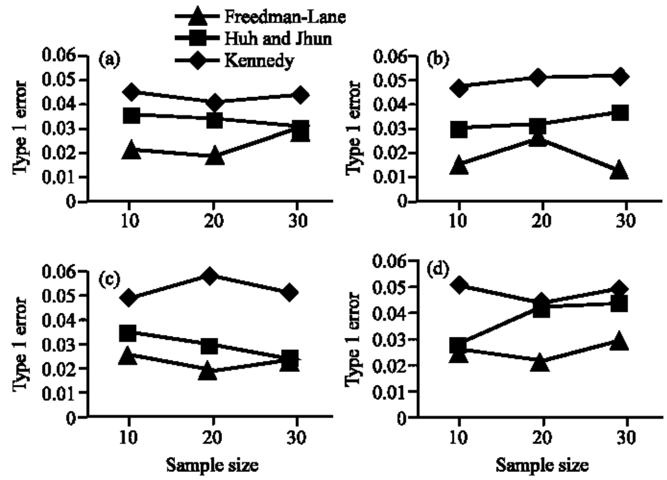

Fig. 1: The consistent views in one view chart; (a) Beta 1 = 0.5, r = 0.1, (b) Beta 1 = 1.5, r = 0.1, (c) Beta 1 = 0.5, r = 0.9 and (d) Beta 1 = 1.5, r = 0.9

CONCLUSION

The objective of this article was to select the best test of significance of a single partial regression coefficient in a multiple regression model. Hence the Kennedy's method, modified Kennedy's method and Freedman and Lane's methods compared by simulation. Then, with accordance to the results of the simulation the best method was selected. This led us to the fact that the results Shadrokh's analytical and results Anderson's simulation still remain unquestionable after the modification of the problem with Kennedy's method and it can be asserted that when selecting a method for testing one of the coefficients of multiple regressions, Freedman and Lane's method is to be the first choice.

REFERENCES

Al-Salihi, A.M., M.M. Kadum and A.J. Mohammed, 2010. Estimation of global solar radiation on horizontal surfaces using routine meteorological measurements for different cities in Iraq. Asian J. Sci. Res., 3: 240-248.

Alam, M.Z., 2004. Biosorption of basic dyes using sewage treatment plant biosolids. Biotechnology, 3: 200-204.

Anderson, M.J. and P. Legendre, 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. J. Statistical Comput. Simulation, 62: 271-303.

Bizhannia, A.R., S.Z. Mirhosseini, B. Rabiee and M. Taeb, 2010. The QTLs determination of cocoon shell weight trait in mulberry silkworm (*Bombyx mori* L.). Biotechnology, 9: 41-47.

Bughio, H.R., A.M. Soomro, A.W. Baloch, M.A. Javed and I.A. Khan *et al*., 2002. Yield potential of aromatic rice mutants/varieties in different ecological zones of Sindh. Asian J. Plant Sci., 1: 439-440.

Edriss, M.A., P. Hosseinnia, M. Edrisi, H.R. Rahmani and M.A. Nilforooshan, 2008. Prediction of second parity milk performance of dairy cows from first parity information using artificial neural network and multiple linear regression methods. Asian J. Anim. Vet. Adv., 3: 222-229.

Fisher, R., 1935. Design of Experiments. lst Edn., Oliver and Boyd Ltd., New York.

Freedman, D. and D. Lane, 1983. A nonstochastic interpretation of reported significance levels. J. Bus. Ecom. Statist., 1: 292-298.

Huh, M.H. and M. Jhun, 2001. Random permutation testing in multiple linear regression. Commun. Stat. Theory Meth., 30: 2023-2032.

Kandhro, M.M., S. Laghari, M.A. Sial and G.S. Nizamani, 2002. Performance of early maturing strains of cotton (*Gossypium hirsutum* L.) developed through induced mutation and hybridization. Asian J. Plant Sci., 1: 581-582.

Kennedy, P.E. and B.S. Cade, 1996. Randomization tests for multiple regression. Commun. Statist, 25: 923-936.

Laghari, S., M.M. Kandhro, H.M. Ahmed, M.A. Sial and M.Z. Shad, 2003. Genotype X environment (GxE) interactions in cotton (*Gossypium hirsutum* L.) genotypes. Asian J. Plant Sci., 2: 480-482.

Manly, B.F.J., 1991. Randomization and Monte Carlo Methods in Biology. 2nd Edn., Chapman and Hall, London.

Rashid, A., G.R. Hazara N. Javed, M.S. Nawaz and G.M. Ali, 2002. Genotype×environment interaction and stability analysis in mustard. Asian J. Plant. Sci., 5: 591-592.

Serhat Odabas, M., Sezgin Uzun and G. Ali, 2007. The quantitative effects of temperature and light on growth, development and yield of faba bean (*Vicia faba* L.): I. Growth. Int. J. Agric. Res., 2: 765-775.

Shadrokh, A. and G. d 'Aubigny, 2010. An analytic comparison of permutation methods for tests of partial regression coefficients in the linear model. Applied Math. Sci., 4: 857-878.

Shadrokh, A., 2011. Comparison of permutation methods tests in linear multiple regression model. Int. J. Acad. Res., Vol. 3, No. 2.

Tariq, M., M. Irshad-ul-Haq, A.A. Kiani and N. Kamal, 2003. Phenotypic stability for grain yield in maize genotypes under varied rainfed environments. Asian J. Plant Sci., 2: 80-82.

Ter-Braak, C.J.F., 1992. Permutation Versus Bootstrap Significance Tests in Multiple Regression and ANOVA. In: Bootstrap and Related Techniques, Jockel, K.H., G. Rothe and W. Sendler (Eds.). Springer- Ver1ag, Berlin, pp: 79-86.