A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 269, 2012

Grzegorz Kończak[*]

ON TESTING THE SIGNIFICANCE OF THE COEFFICIENTS IN THE MULTIPLE REGRESSION ANALYSIS

Abstract. The multiple regression analysis is a statistical tool for the investigation relationships between the dependent and independent variables. There are some procedures for selecting a subset of given predictors. These procedures are widely available in statistical computer packages. The most often used are forward selection, backward selection and stepwise selection. In these procedures testing the significance of parameters is used. If some assumptions such as normality errors are not fulfilled, the results of testing significance of the parameters may not be trustworthy. The main goal of this paper is to present a permutation test for testing the significance of the coefficients in the regression analysis. Permutation tests can be used even if the normality assumption is not fulfilled. The properties of this test were analyzed in the Monte Carlo study.

Key words: linear regression model, permutation test, Monte Carlo.

I. INTRODUCTION AND BASIC NOTATIONS

A.C. Rencher (2002) considered multiple linear regression models for fixed and random *x*'s. The errors in these models can be normally or non-normally distributed.

Let us consider the multiple linear regression model for fixed x's given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q + \varepsilon \tag{1}$$

where

y is the dependent variable

$x_1, x_2, \ldots, x_q$ represent q different variables (fixed)

β_0 is the intercept

$\beta_1, \beta_2, ..., \beta_q$ represent the corresponding q regression coefficients

ε is the random error where

[*] Ph.D., Associate Professor, Department of Statistics, Katowice University of Economics, grzegorz.konczak@ue.katowice.pl

$$E(\varepsilon) = 0 \text{ and } D^2(\varepsilon) = \sigma^2. \tag{2}$$

In this model each y (y_1, y_2, ..., y_n) in the sample of n observations can be expressed as a linear function of x's plus random error ε. The model (1) can be rewritten as follows

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{3}$$

or equally

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{4}$$

The assumptions (2) can be rewritten as follows

1. $E(\boldsymbol{\varepsilon}_i) = 0,$ for all $i = 1, 2, ..., n$.
2. $D^2(\varepsilon_i) = \sigma^2$, ,for all $i = 1, 2, ..., n$.
3. $Cov(\varepsilon_i, \varepsilon_k) = 0,$ for all $i \neq k$.

The hypothesis statements to test the significance of a particular regression coefficient β_j ($j = 1, 2, \ldots, q$) can be written as follows:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \tag{5}$$

For testing the significance of the individual regression coefficient β_j the t test is used. The t test statistics is based on the t distribution and has the form

$$t = \frac{\hat{\beta}_j}{S(\hat{\beta}_j)} \tag{6}$$

where $\hat{\beta}_j$ is the least square estimator of the parameter β_j ($j = 1, 2, ..., q$) and $S(\hat{\beta}_j)$ is the estimated standard error of $\hat{\beta}_j$. The standard error of each parameter $\hat{\beta}_j$ is given by the square root of the diagonal elements of the matrix $Var(\hat{\beta})$ where

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

The *t* test could be used if errors are independent and normally distributed. The statistic (6) under H_0 has the *t* distribution with *n* - *q* - 1 degrees of freedom. The null hypothesis (5) is rejected if $|t| > t_{\alpha/2, n-q-1}$. If the null hypothesis (5) is not rejected, this indicates that the regressor x_j could be removed from the model. If this hypothesis is rejected, then the regressor x_j could be added to the model.

The multiple linear regression model for random *x*'s has the same form as (1) but it is assumed that x_1, x_2, ..., x_q are not under control of experimenter. Many regression applications involve *x*'s that are random variables. If we assume that the vector (y, x_1, x_2, ..., x_q) has a multivariate normal distribution (L. Godfrey, 2009 and D.J. Sheskin, 2004), we can proceed with testing in the same way as in the fixed *x*'s case.

The statistic (6) could be used if the following assumptions are fulfilled

a) the *x*'s are fixed
b) errors are independent and normally distributed

or

a) the *x*'s are random and normally distributed
b) errors are independent and normally distributed

If these assumptions are not fulfilled then testing the significance of parameters in the regression analysis can't be performed.

II. VARIABLES SELECTION METHODS IN THE LINEAR MODEL

One of the most important problems in the multiple regression analysis is the selection of variables. The methods of selecting variables are a way of selecting a particular set of independent variables to be used in the regression model. There is a large number of commonly used methods which are called stepwise techniques. The most often used are forward selection, backward selection and stepwise selection:

– *forward selection* starts with no variables selected. Next we add the most significant variable. At each step we add the most significant variable until there are no variables that meet the criterion set by the user,

– *backward selection* starts with all variables selected. At each step the least significant variable is removed from the model until none of them meets the criterion set by the user,

– *stepwise selection* is a method that is a combination of two previous methods, testing at each stage for including or excluding variables.

The methods described above (and some other) are included in statistical packages such as SPSS, Statistica, MiniTab, Statgraphics, R. For each of these methods the significance of parameters is assessed at each step of the procedure. The *t* test is used many times at each step. If one of the following conditions is fulfilled

- the errors are not normally distributed
- the errors are not independent
- the errors are not homoscedastic
- the *x*'s are random and not normally distributed

then the *t* test shouldn't be performed. In this case the permutation test can be used instead. Permutation tests can be performed even if the normality assumption is not fulfilled.

III. PROBLEMS FOR NORMALITY TESTING

In many regression applications *x*'s are not fixed. If the methods of selecting predicting variables described above are used then normality of independent random variables has to be tested. The normality hypothesis can be tested using normality test (for example Shapiro-Wilk's test, Lilliforse's test or chi square goodness of fit test). Even if the null hypothesis in normality testing is not rejected, it is possible that the sample is taken from non-normal distribution. Normality testing for small sample sizes was analyzed in the Monte Carlo study.

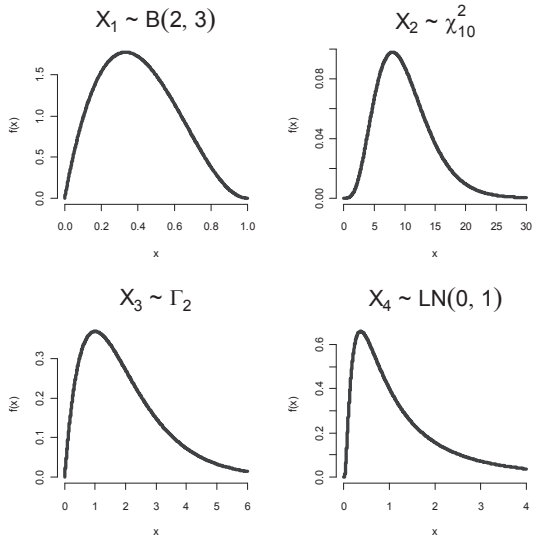

Fig. 1. The densities of random variables under study

There 4 non-normal distributions were considered: beta (random variable X_1), chi-square (X_2), gamma (X_3) and log-normal (X_4). The details of the random variables X_1, X_2, X_3 and X_4 are given in table 1. The density functions of analyzed random variables are presented in Fig. 1. The sample of the size n = 15 was generated from these distributions. Then the normality hypotheses were tested using Lillieforse's test and Shapiro-Wilk's test. This procedure was repeated 1 000 times and the probabilities of the rejection of the null hypothesis were estimated. The simulation study was performed using R (http://www.r-project.org) procedures. The results of computer simulation are presented in table 1.

Table 1. Estimated probabilities of failing to reject the normality hypothesis

Details of the random variable	R function	Lilliforses test	Shapiro-Wilk's test
$X_1 : B(2,3)$ $E(X_1) = 0{,}4 \qquad D^2(X_1) = 0{,}04$	rbeta(n, 2, 3)	0.9454	0.9392
$X_2 : \chi^2_{10}$ $E(X_2) = 10 \qquad D^2(X_2) = 20$	rchisq(n, 10)	0.8737	0.8181
$X_3 : \Gamma(2)$ $E(X_3) = 2 \qquad D^2(X_3) = 2$	rgamma(n, 2)	0.8597	0.7914
$X_4 : LN(0,1)$ $E(X_4) = \sqrt{e} \qquad D^2(X_4) = e(e-1)$	rlrnorm(n, 0, 1)	0.7420	0.6123

Source: Monte Carlo study.

It is easy to notice (table 1) that for the analyzed non-normal random variables testing the hypothesis often leads to "no reject the null hypothesis". The standard error of estimation of probabilities included in table 1 is less than 0.016.

IV. PERMUTATION TEST VERSUS *t* TEST – MONTE CARLO STUDY

Variables selection methods in the multiple regression analysis are based on the *t* test. This test can be performed only if the assumptions mentioned above are fulfilled. It is possible that the *t* test is performed, due to the results of normality testing, even for non-normally distributed *x*'s. The results of the use of the *t* test and the permutation test were compared.

The model analyzed in the Monte Carlo study had the form of:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \tag{7}$$

where $\varepsilon \sim N(0,1)$ and $\beta_j \in \{0,1\}$ for $j = 1, 2, 3, 4$.

There were 15 variants of the model (7) with one, two, three or four significant variables. The details of the analyzed models are presented in table 2.

Table 2. The details of the analyzed variants of the multiple regression linear model

Parameter	Symbol					
	M_1	M_2	...	M_{12}	...	M_{1234}
β_1	1	0		1		1
β_2	0	1		1		1
β_3	0	0		0		1
β_4	0	0		0		1

Permutation tests are computer-intensive statistical methods. These tests were introduced by R.A. Fisher in 1930's (P. Good, 2005 and W.J. Welch, 1990). In the permutation test instead of comparing the observed value of the test statistic to a standard distribution, the reference distribution is generated from the data. These tests can give results that are more accurate than those obtained with the use of traditional statistical methods. The concept of these tests is simpler than of the tests based on normal distribution. The main application of these tests is a two-sample problem (B. Efron, R. Tibshirani, 1993). Permutation tests were used for determining the significance of the linear regression model coefficients. These results were compared to the *t* test results.

Simulation procedure

There 15 models were analyzed in the Monte Carlo study. The details of the analyzed models are described in tables 1 and 2. For each model a set of data (y, x_1, x_2, ..., x_q) was generated 1 000 times.

The Monte Carlo study was performed for the model (7) where random variables' details are described in table 2. The significance level = 0.05 in the testing procedures in the Monte Carlo study was assumed. The steps for each model of this study were as follows:

1. A sample of the size $n = 15$ from the considered model M_x (see table 2) was generated.

2. Parameters of the linear model were estimated (least squares method) and the significance of each parameter using the *t* test was calculated. Then a set of significance parameters was denoted by S_1.
3. The significance of the parameters was determined with the use of the permutation test (for $L = 1\,000$ randomly shuffled *x*'s – see Fig.2). Then a set of significance parameters was denoted by S_2.
4. Steps 1-3 were repeated $N = 1\,000$ times.
5. The number of consistent results ($S_1 = S_2$) was calculated and the estimated probabilities of achieving the consistent results in the *t* test and the permutation test were estimated.
6. The whole procedure (steps 1–6) was repeated for each model (see table 2).

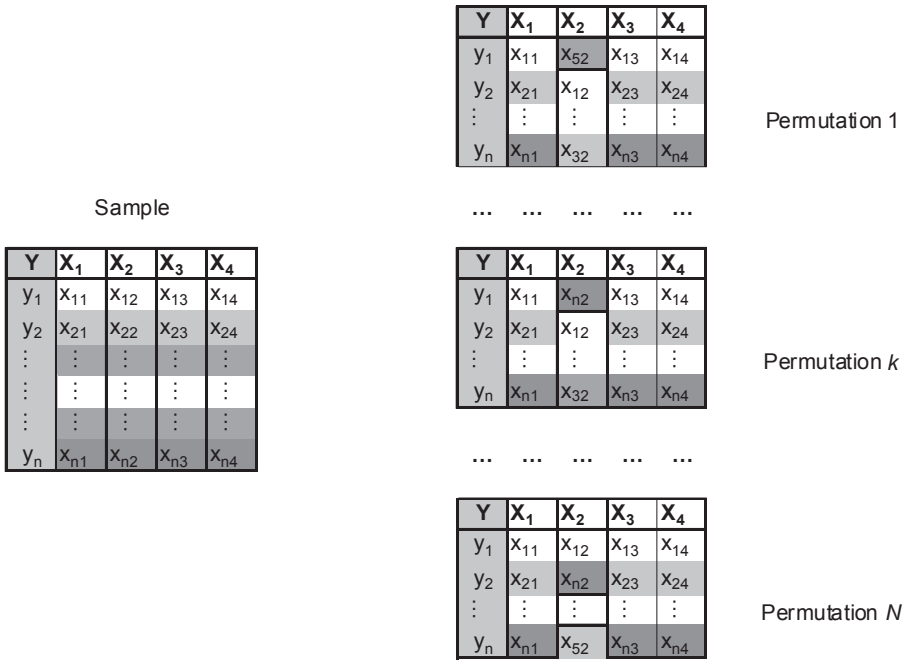

Fig. 2. The scheme of permutation *x*'s

The estimated probabilities of consistent indications of the *t* test and the permutation test are presented in table 3. The standard error of estimated probabilities is less than 0.016. The results from table 3 are presented in Fig. 3. It can be noticed that the *t* test should not be performed, but due to the result of normality testing (table 1) researchers often use this test.

Table 3. Estimated probabilities of consistent indications of the *t* test and the permutation test

Model	Estimated probability		Model	Estimated probability		Model	Estimated probability
M1	0.884		M13	0.903		M123	0.931
M2	0.920		M14	0.818		M124	0.850
M3	0.913		M23	0.929		M134	0.851
M4	0.874		M24	0.872		M234	0.878
M12	0.919		M34	0.856		M1234	0.876

Source: Monte Carlo study

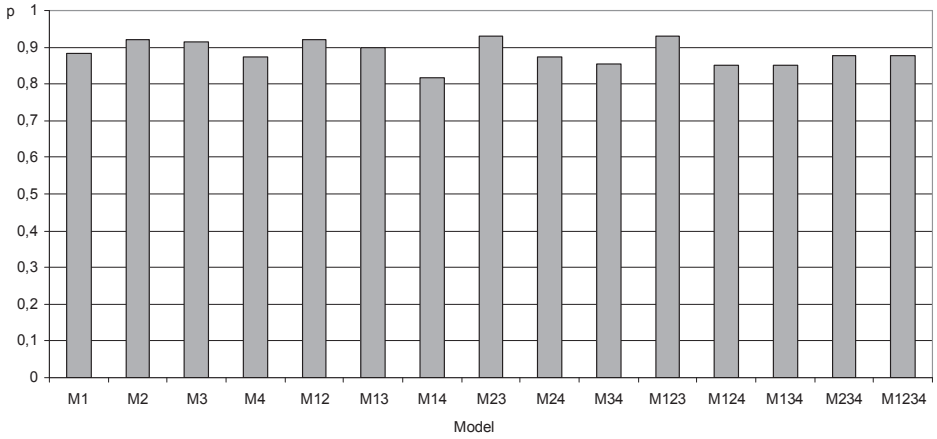

Fig. 3. Estimated probabilities of consistent indications of the *t* test and the permutation test

V. CONCLUDING REMARKS

Regression analysis is an important issue in different scientific areas. Many studies are carried out by investigating the regression parameter of the independent variable before adding or removing the predictor in the regression analysis. In these procedures the *t* test is performed.

The procedure of testing the significance of parameters in the linear regression analysis using the permutation test was proposed in the paper. The properties of this procedure were analyzed in the Monte Carlo study. Permutation tests can be used even if the normality assumption is not fulfilled. The Monte Carlo study has shown that it is a good replacement for the *t* test in case where independent variables are non-normally distributed.

REFERENCES

Efron B., Tibshirani R. (1993) *An Introduction to the Bootstrap*, Chapman & Hall. New York.
Godfrey L. (2009) *Bootstrap Tests for Regression Models*, Palgrave Text in Econometrics, London.
Good P. (2005) *Permutation, Parametric and Bootstrap Tests of Hypotheses*, Springer Science Business M edia, Inc., New York.
Rencher A.C. (2002) *Methods of Multivariate Analysis*, Wiley-Interscience, New York.
Sheskin D.J. (2004) *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton.
Welch W.J. (1990) *Construction of Permutation Tests*, Journal of the American Statistical Association, vol. 85, No. 411, Theory and Methods.

Grzegorz Kończak

O TESTOWANIU ISTOTNOŚCI WSPÓŁCZYNNIKÓW W MODELU REGRESJI WIELORAKIEJ

Model regresji liniowej pozwala na badanie i opis powiązań pomiędzy zmienną zależną i zmiennymi niezależnymi. W analizach dotyczących modelu regresji liniowej zakłada się m.in. normalność rozkładu reszt oraz jednorodność wariancji. Jeżeli wspomniane założenia nie są spełnione, to rezultaty testowania istotności modelu regresji mogą nie być wiarygodne. W opracowaniu zaproponowano wykorzystanie testu permutacyjnego do weryfikacji istotności modelu regresji liniowej. Testy permutacyjne mogą być stosowane bez zakładania postaci rozkładu zmiennej. Analizę własności proponowanego testu przeprowadzono z wykorzystaniem symulacji Monte Carlo.