FACTA UNIVERSITATIS
Series: **Economics and Organization** Vol. 1, N^o 10, 2002, pp. 25 - 29

TESTING IN MULTIPLE REGRESSION ANALYSIS

UDC 519.233.5

Vera Djordjević

Faculty of Economics, University of Niš, 18000 Niš, Serbia and Montenegro

Abstract. *Multiple regression analysis takes great place in the statistical science and practice. F. Galton, English statistician and geneticist, has written that the aim of statistical science is "to find out methods in collecting large groups of data in short, condensend terms convenient for discussion". This thought, although it was told more than one century ago, can be applied to the subject and method of multiple linear regression. The significance of this method is in its usage to predict and estimate the value of this phenomenon (it has been identified as dependent variable). We have to test the significance of given results if we want to use the statistics of the F-test and statistics of the Student's t-test.*

Key words: *testing, multiple linear regression, regression model, residual, regression coefficient, standard error of regression, coefficient of multiple determination, zero hypothesis, alternative hypothesis, analysis of variance, F-test, t-test.*

INTRODUCTION

In simple regression analysis we exemine dependence between variation of two phenomena. Due to the fact that many factors influence some phenomena, especially social, it's necessary to calcualte the interraction among phenomena. In order to attain this we can use multiple regression methods.

Multiple regression has taken a very significant place in statistical science. Not interfering into reasons of this interest, according to the analytician's opinion we can consider the method of regression convenient in one of two cases:

a) in order to control dependent variable according to independent variables, and
b) in order to predict dependent variable. In both cases it is necessary to make researches.

Received February 15, 2003

MATHEMATICAL MODEL

In case of researching relationship between two phenomena and in case of prediction of the value of dependent variable, first we are going to identify variables and then to find out random sample *n*-size for the chosen values of dependent variables.

Suppose that k phenomenon is identified as independent variable (predictor), or X_i, i = 1, 2, …, k, and Y as dependent random variable. The whole multiple linear model can be presented as one equation for volontarily dependent variable Y_i:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_i \tag{1}$$

where:

Y_i– dependent random variable,
$x_1, x_2, \ldots, x_k$ – values of independent variable,
$\beta_0, \beta_1, \ldots, \beta_k$ – model parameters (regression coefficient)
ε_i – a supporting element, or a random error which has normal distribution, zero mean and constant variance.

Multiple linear regression model (1) consists of two parts:

– determined (Y_i')

$$Y_i' = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{2}$$

– stohastics (ε_i), so that from (1) we can get:

$$\varepsilon_i = Y_i - Y_i' \tag{3}$$

Determined part of the linear regression model is an average value of dependent variable (Y_i) for the given values of independent variables:

$$Y_i' = E(Y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{4}$$

and other values of Y_i show average values $E(Y_i)$.

The whole regression model (1) was estimated by the sample regression model:

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \tag{5}$$

where we have:

$\hat{y}_i$ – adjustable or foreseen value of dependent variableY_i,
$x_1, x_2, \ldots, x_k$– values of independent variables,
$b_0, b_1, \ldots, b_k$ – estimations of unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$.

We should choose the multiple linear regression model which presents in the most suitable way the relationship between observed phenomena. It can be acheived by minimizing a sum of square equations of empirical points from the regression model (for example: regression plane when k=2), or:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \min\ , \tag{6}$$

where e_i –is random error in sample.

Multiple linear regression model as statistical model does not mean only mathematical expression but also assumptions which supply the optimal estimation of parameters $\beta_0, \beta_1, \ldots, \beta_k$. These assumptions are usually connected with random error:

– the random error has normal distribution,

– it is equal zero (on the average)

– supporting elements have equal variances.

In case when k=2, multiple linear regression model is regression plane equation in samples (the easiest example of multiple linear regression):

$$\hat{y}_i = b_0 + b_1x_1 + b_2x_2 \tag{7}$$

In order to establish adaptation of the estimated regression model by empirical data we use standard error of the sample regression which represents the estimation of standard deviation of the random error σ_ε. It is market by S_ε, and it is presented as square root of repetition, or:

$$S_\varepsilon = \sqrt{\sigma^2} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}} \tag{8}$$

where SSE is a sum of square root aberration of the empirical points of regression model (Error Sum of Squares).

The standard error of regression as absolute measure of the unexplained variability is not convenient for comparison. That's the reason why we use relative indicator – coefficient of multiple determination R^2. It is presented as a measure of explained variability and it is calculated by this equation:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SSR}{SSy} \tag{9}$$

where SSR presents Regression Sum of Squares (explained variability) and SSy presents the total Sum of Squares (total variability).

The coefficient of multiple determination shows the procentage of variations of dependent variable Y which is described by common influence of independent variables which are involved in this model. During its calculation we should take care of the number of independent variables and of sample size. It is achieved by calculation of the adjusted coefficient of multiple determination:

$$R^2_{adj.} = 1 - \frac{n-1}{n-k-1} \cdot (1 - R^2) \tag{10}$$

where: n – is the sample size and k – number of independent variables.

MODEL USAGE TESTING

In order to use the estimated regression equation we firstly have to test the significance of given estimates.

This is zero and alternative hypothesis:

H_0: $\beta_0 = \beta_1 = \beta_2 = \ldots + \beta_k = 0$

H_A: at least one $\beta_i \neq 0$

According to this, we have laid, zero hypothesis in that way that a linear connection between observed phenomena variations does not exist, or that x_1, x_2, …, x_k has not influence on Y.

If we start from the assumption that the total variability of dependent variable is conditioned by the variability of independent variables involved in the model and by the unexplained variability, we can write:

$$SSy = SSR + SSE \qquad (11)$$

where SSy – presents the total Sum of Squares (total variability),
SSR – Regression Sum of Squares (explained variability),
SSE – Error Sum of Squares (unexplained variability),

We apply F-test, and test and test the possibility of the regression model usage by analysis of variance. The table of this analysis is presented here:

sources of variation	degrees of freedom	sum of squares	mean squares	F-ratio
Regression	*k*	SSR	$MSR=\frac{SSR}{k}$	$F=\frac{MSR}{MSE}$
Error	*n*–*k*–1	SSE	$MSE = \frac{SSE}{n-k-1}$	
Total	*n* –1	SSy		

The decisions rules:

if $F \geq F_{\alpha;k;n\text{-}k\text{-}1,}$ we reject null hypothesis,
if $F < F_{\alpha;k;n\text{-}k\text{-}1,}$ we accept null hypothesis.

According to this, if the realized value of the F-test is lesser than theoretical, or we accept null hypothesis, we come to a conclusion that the linear influence of independent variables on dependent variable doesn't exist.

REGRESSION COEFFICIENTS TESTING

If we use estimated regression model to estimate and predict the values of dependent variable Y, we have to test the significance of estimates of each parameter apart (β_i, i = 1, 2, …, k).

In the easiest case of the multiple linear regression k = 2 we test the estimate significance of two parameters β_1 and β_2. The null and alternative hypothesis are presented here:

I H_0: $\beta_1 = 0$ II H_0: $\beta_2 = 0$
H_A: $\beta_1 \neq 0$ H_A: $\beta_2 \neq 0$

Test statistics:

$$t_1 = \frac{b_1}{S_{b_1}} \qquad t_2 = \frac{b_2}{S_{b_2}}$$

has the Student's t-distribution with n – k – 1 degrees of freedom.

If $|t_i| < t_{\alpha/2,\ i=1,2}$, the value of the test statistics has fallen in the field of accepting null hypothesis. In that case we accept null hypothesis that independent variables (X_1, i.e X_2) does not influence dependent variable Y.

Generally in multiple regression model we apply testing:

H_0: $\beta_i = 0$
H_A: $\beta_i \neq 0$

(for i = 1, 2, …, k); the test statistics

$$t_i = \frac{b_i}{S_{b_i}}$$

has the Student's t-distribution with degrees of freedom $n - k - 1$.

We accept null hypothesis if $|t_i| < t_{\alpha/2.}$

We can also mention that when testing the significance greater realized value of the t-test statistics does not mean that the variable which corresponds has greater relative influence on dependent variable.

In order to determine mathematical model of multiple regression and to test it we have to do many calculations. It's necessary to use special calculation programme for these calculations. The application of some current package of calculation programme is significant not only for ananalyst but also for statistical practice.

REFERENCES

1. Mc Clave J.; Benson G.; Sincich T. Statistics for business and economiCS, Prentice Hall, 2001.
2. Aczel A. Complete business statistics, Homewood, 1986.
3. Phaffenberger P., Paterson J.. Statistical methods, Homewood, 1987.

TESTIRANJE MODELA LINEARNE VIŠESTRUKE REGRESIJE

Vera Djordjević

Upotrebljivost modela linearne višestruke regresije za predviđanje i ocenjivanje vrednosti zavisne promenljive proverava se testiranjem. Za testiranje možemo koristiti F-test i t-test. I jednim i drugim testom proverava se da li oblik zavisnosti utvrđen na osnovu raspoloživih podataka (podataka uzorka) važi za osnovni skup. Samo u tom slučaju regresioni model može poslužiti analitičaru za dalju statističku analizu i korišćenje.