ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 141, 1997

*Wiesław Wagner**

USEFULLNESS OF DURBIN METHOD FOR INVESTIGATING NORMALITY IN ONE-DIMENSIONAL LINEAR MODEL

Abstract. In this paper the verification of hypotheses of univariate normality by Durbin randomized method is presented.

The method of elimination of the nuisance parameters by calculating the residual vector and connected residual vector is presented, too.

Key words: linear model, Durbin randomized method, tests for normality of residuals.

1. INTRODUCTION

Investigating the normality of one-dimensional simple sample has been done by many statisticians. Vast literature on the subject is given by Mardia (1980). Most frequently the problem may be reduced to constructing tests statistics which are functions of sample estimators of unknown expectation μ and variance σ^2, which in the problem of testing for normality are nuisance parameters. We eliminate them by transforming observable random variables. One of the ways is given by Durbin (1961). We will extend it to the case of investigating the normal distribution of random errors in a linear model, which performed on the transformed residuals obtained from the LSM in order to get a simple sample. For such samples which comprise independent random variables with identical distributions we apply normality tests.

In the paper the Durbin method for investigating normality of random errors in a linear model is given. Its characteristic is supplemented with assisting results.

* Department of Mathematical Methods Application. Academy of Agriculture, Poznań.

2. LINEAR MODEL AND RESIDUALS

Let y: nx1 be a vector of n observable random variables given by a linear model

$$y = X\beta + e, \tag{2.1}$$

where: X: nxq is a known matrix of order $r(X) = m \leqslant q < n$,

β: qx1 – a vector of unknown constant parameters,

while e: nx1 is a vector of unobserved random variables called random errors. The linear model is expressed by a triple $(y, X\beta, \sigma^2 I)$, where $\sigma^2 > 0$ is unknown and I: $n \times n$ is the identity matrix. This means that E(e) = 0 and $D(e) = \sigma^2 I$.

The vector of residuals from the LSM is expressed by $r = \psi y$, where $\psi = I - (X(X'X)^- X'$ and $(X'X)^-$ is the g-converse matrix of $X'X$.

For the vector of residuals r we have properties which proofs can be found in many monographies concerning linear models: $r'1 = 0$, $r'X = 0$, $r'r = y'\psi y = e'\psi e$, $E(r) = 0$, $D(r) = \sigma^2 \psi$, $E(rr') = Cov(y, r) = Cov(e, r) = \sigma^2 \psi$ and $s^2 = r'r/(n-m)$ is the BLUE estimator of σ^2. The property $D(r) = \sigma^2 \psi$ states that the components of vector r are correlated and (usually) have different variances. For presenting the issue of investigating normality of the distribution of vector e we give some assisting results which are used in the applications of the Durbin method.

3. ASSISSTING RESULTS

Let $X_1, ..., X_n$ be a simple random sample, being a sequence of n independent values of random variable X which is assumed to have distribution $N(\mu, \sigma^2)$. By $\bar{X}$ and S^2 we denote the sample mean and unbiased estimator of variance from this sample, for which some properties hold: $\bar{X} \sim N(\mu, \sigma^2/n)$, $(n-1)S^2 \sim \sigma^2 \chi^2_{n-1}$, $Cov(\bar{X}, S^2) = 0$, $\bar{X}$ and an arbitrary function $g(X_1 - \bar{X}, ..., X_n - \bar{X})$ are independent $(n-1)S^2$ and are independent and $(X_i - \bar{X})/(n-1)S$ are independent.

Let us define standardized variables $U_i = (X_i - \bar{X})/S$ from sample $X_1, ..., X_n$. The variables U_i, $i = 1, ..., n$, are not independent but we have:

$$Cov(X_1, U_i) = 0, \quad Cov(\bar{X}, U_i) = 0, \quad Cov(S^2, U_i) = 0.$$

The density function $f(u_i) = f(u)$ for random variable U_i is (Pearson, Sekar 1936, Cramer 1958, p. 273).

$$f(u) = \frac{\sqrt{n}}{(n-1)\sqrt{\Pi}} \cdot \frac{g_{n-1}}{g_{n-2}} \left[1 - \frac{n}{(n-1)^2} u^2\right]^{(n-4)/2} \tag{3.1}$$

for $-(n-1)/\sqrt{n} \leqslant u \leqslant (n-1)/\sqrt{n}$, where $g_n = \Gamma(n/2)$.

Density (3.1) is a particular case of random variable U with the generalized beta distribution

$$f(u) = \frac{1}{(b-a)^{p+q-1}} \cdot \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} (u-a)^{p-1}(b-u)^{q-1}, \quad a < u < b,$$

when and $a = -(n-1)/\sqrt{n}$, $b = (n-1)/\sqrt{n}$ and $p = q = (n-2)/2$

Density (3.1) is a symmetric function, so all moments about the origin of odd degree are equal 0, while the ones of even degrees are given by the formula

$$E(U^k) = \frac{(n-1)^k}{\sqrt{\Pi}\, n^{k/2}} \frac{g_{k+1} \cdot g_{n-1}}{g_{n+k-1}}. \tag{3.2}$$

Moments about the origin of degree k of standard deviation S are given by (e.g. Pawłowski 1976, p. 46)

$$E(S^k) = \frac{2^{k/2}}{(n-1)^{k/2}} \frac{g_{n+k-1}}{g_{n-1}} \sigma^k, \quad k = 0, 1, 2,... \tag{3.3}$$

4. DURBIN METHOD FOR A SIMPLE SAMPLE

Let us suppose that we verify a composite hypothesis HC that $X_1,..., X_n$ is a simple sample from the distribution $N(\mu, \sigma^2)$, with unknown μ and σ^2. Let $\bar{X}$ and S^2 be the sample mean and unbised variance estimator. Moreover, let us denote by $\bar{Y}$ and S'^2, the sample mean and variance from the population with the standardized normal distribution $N(0, 1)$. Thus, we have $\bar{Y} \sim N(0, 1/n)$, $(n-1)S'^2 \sim \chi^2_{n-1}$ and $\bar{Y}$ and S'^2 are independent. What's more, after replacing σ by 1 and k by 2 we get $E(S'^2) = 1$.

As we have already mentioned in the first paragraph the parameters μ and σ^2 in the problem of investigating normality are nuisance parameters. Durbin (1961) suggests a randomization process to eliminate them. The idea of it is to consider two further random variables $\bar{Y}$ and S'^2 which have the distributions mentioned earlier. According to this formula, we determine such a random sample $Y_1,..., Y_n$ that

$$(Y_i - \bar{Y})/S' = (X_i - \bar{X})/S, \quad i = 1,..., n. \tag{4.1}$$

We will show that sequence $Y_1,\ldots, Y_n$ is a simple sample from the population with the distribution $N(0, 1)$. The relation (4.1) we write as

$$Y_i = \bar{Y} + S'U_i, \quad i = 1,\ldots, n$$

where U_i is given in point 3 and $\bar{Y}$ and S' are random variables generated independently of $X_1,\ldots, X_n$.

Lemma 4.1. Random variables $\bar{X}$, S^2, U_i, $\bar{Y}$, $S'^{2\prime}$ are pairwise independent.

Proof. The independence of $\bar{X}$, S^2, U_i follows from the results given in point 3. Other independences follow from assumption that $\bar{X}'$ and S'^2 are independent of $X_1,\ldots, X_n$.

Lemma 4.2. $E(\bar{Y} + S'U_i) = 0$ and $D^2(\bar{Y} + S'U_i) = 1$

Proof. From Lemma 4.1 we get

$$E(\bar{Y} + S'U_i) = E(\bar{Y}) + E(S'U_i) = E(\bar{Y}) + E(S'U_i) = E(S')E(U_i) = 0$$

which follows from disappearing of the moments of odd orders of variable U_i. From the fact, that $Cov(\bar{Y} + S'U_i) = 0$, we get for the variance

$$D^2(\bar{Y} + S'U_i) = D^2(\bar{Y}) + D^2(SU) = \frac{1}{n} + E(S'^2U_i^2) - [E(S')E(U_i)]^2 =$$
$$= \frac{1}{n} + \frac{n-1}{n} = 1$$

where we used $E(S'^2) = 1$.

We have shown that the first two moments of the left side expression $Y_i = Y + S'U_i$ are identical with those for the variable with distribution $N(0, 1)$. Now, we will give a lemma in which we will prove that variable $S'U$ is normally distributed and the first one has the chi distribution and the second has the symmetric beta distribution.

Lemma 4.3. Random variable $Z = S'U$ has the normal distribution $N(0, (n-1)/n)$.

Proof. We use the result given by Fisz (1967, p. 71). If S' and U are independent random variables with densities $f_1(s')$ and $f_2(u)$ then the distribution of $Z = S'U$ is given by the density

$$f(z) = \int_{-\infty}^{\infty} \frac{1}{s'}, \cdot f_1(s')f_2\left(\frac{z}{s'},\right)ds'$$

Let us denote by C_1 and C_2 the constants

$$C_1 = \frac{2(n-1)^{(n-1)/2}}{2^{(n-1)/2} g_{n-1}} \quad \text{and} \quad C_2 = \frac{\sqrt{n}\, g_{n-1}}{(n-1)\sqrt{\Pi}\, g_{n-2}}$$

Using the density $f_1(s')$ given, among others, by Pawłowski (1976, p. 45) and density (3.1) we get

$$f(z) = C_1 C_2 \int_0^\infty \frac{1}{s},s'^{n-2} \exp\left\{-\frac{(n-1)s'^2}{2}\right\}\left[1 - \frac{n}{(n-1)^2}\frac{z^2}{s'^2}\right]^{(n-4)/2} ds' =$$

$$= C_1 C_2 \int_0^\infty s' \exp\left\{-\frac{(n-1)s'^2}{2}\right\}\left[s'^2 - \frac{n}{(n-1)^2} z^2\right]^{(n-4)/2} ds'.$$

We change the variable s' by consecutive substitutions $t = (n-1)s'^2/2$, $v = 2(n-1)t - nz^2$, $w = /2(n-1)$. Then we get

$$f(z) = \frac{C_1 C_2}{(n-1)^{n-3}} (2(n-1)^{(n-4)/2} \exp\left[-\frac{nz^2}{2(n-1)}\right] \int_0^\infty e^{-w} w^{(n-2)/2-1} dw.$$

Coming back to the constants C_1 and C_2, from the definition of the gamma function, we get

$$f(z) = \frac{2(n-1)^{(n-1)/2}}{2(n-1)/2 g_{n-1}} \frac{\sqrt{n}}{(n-1)} \frac{g_{n-1}}{g_{n-2}} \frac{2^{(n-4)/2}(n-1)^{(n-4)/2}}{(n-1)^{n-3}}$$

$$x g_{n-2} \exp\left\{-\frac{nz^2}{2(n-1)}\right\} = \frac{1}{\sqrt{2\Pi}} \sqrt{\frac{n}{n-1}} \exp\left\{\frac{nz^2}{2(n-1)}\right\}.$$

Finally, we get the density function of the distribution $N(0, (n-1)/n)$ i.e. random variable Z has the normal distribution.

5. DURBIN METHOD FOR A LINEAR MODEL

In point 2 we gave the vector of residuals from LSM for model (2.1). We are transforming it further, using Durbin's randomization procedure. Its use is connected with the elimination of the nuisance parameter σ^2 on which the covariance matrix of the vector of residuals depends. It requires the use of a transformation which expresses the quotient of two random variables with the chi-square distribution. On the other hand, the transformation of the vector of residuals should give such random variables which are uncorrelated. Both problems can be solved in two ways.

The idea of the first one is to use some random variables, exactly as many as there are unknown nuisance parameters, in the problems of investigating normality. The problem was developed in that direction by Sarkadi (1960, 1967), Stormer (1964), Theil (1968) and Sally and Sarkadi (1982).

The other way considered is to extend the set of random variables by the number of them equal exactly to the number of nuisance parameters. These additional variables are treated as generated random variables, independent of the observed ones. Such reasoning was already presented in point 4. When the elements of the sample are correlated, one should still generate a random vector with the n-dimensional $N_n(0, I)$ distribution in order to use it to eliminate correlated variables. This idea was presented by Golub et al. (1973) and Wagner (1982, 1990).

In our considerations we apply the Durbin formula. We have one nuisance parameter (variance σ^2) and n correlated random variables being the components of the random vector r of the residuals. We generate a random variable $(n-1)S'^2 \sim \chi^2_{n-1}$ and independently of it random vector v with an arbitrary distribution with the moments $E(v) = 0$ and $D(v) = I$. We create a corrected vector of LSM residuals

$$r^* = \frac{S'}{S} r + (I - \psi)v, \tag{5.1}$$

where $S^2 = r'r/(n-m)$.

To prove that the components of vector r^* are uncorrelated we use the result given in points 2, 3 and 4.

Lemma 5.1. $E(r^*) = 0$ and $D(r^*) = I$.

Proof. From independence of S'^2 and vector v of vector y, we get $Cov(S', r) = Cov(S', v) = Cov(S, r) = Cov(S, v) = 0$ and $Cov(r, v)$. For the expectation we have $E(r^*) = E\left(\frac{S'}{S} r\right) + (I - \psi)E(v) = E(S')E\left(\frac{r}{S}\right)$. But every component of vector r/S has the same distribution with expectation equal 0 according to Lemma 4.3, i.e. $E(r^*) = 0$. Further, due to the earlier mentioned covariances, we get

$$D(r^*) = D\left(\frac{S'}{S} r\right) + (I - \psi)D(v)(I - \psi)' = E\left[\frac{S'^2}{S^2} rr'\right] -$$

$$- E\left(\frac{S'}{S} r\right)\left[E\left(\frac{S'}{S} r\right)'\right] + I - \psi = \frac{E(S'^2)}{E(S^2)} E(rr') + I - \psi = \frac{1}{2}\sigma^2\psi + I - \psi = I,$$

what follows from $E(S'^2) = 1$ and (3.3) at $k = 2$.

Given lemma shows that the components of vector r^* are uncorrelated and get a simple sample.

6. TESTING FOR NORMALITY OF RANDOM ERRORS

The result of Lemma 5.1 will be used to test normality of random errors in model (2.1). Let $N = \{N_n(\mu, \sigma^2 I): \mu = X\beta, \sigma^2 > 0\}$ be a class of n-dimensional normal distributions with the given parameters. We set the null hypothesis that the distribution of vector e belongs to class N, which we write as $H_0\!:e \in N$ against the alternative $H_1\!:e \notin N$.

We verify the hypothesis H_0 with vector of r^* of corrected residuals. It is the sum of two vectors. The first is created from the transformation of observable random vector y and generated random variable with the chi-square distribution with n-1 degrees of freedom. If the hypothesis H_0 is true then, according to Lemma 4.3 each of its components is normally distributed. About the second vector we can assume that, in particular, it is a random vector with the distribution $N_n(0, I)$. Thus, the sum of the two independent vectors, each normally distributed, gives a random vector normally distributed. And conversly, with the help of Cramer Lemma, (see e.g. Rao 1982, p. 525) assuming that vector r^* is normally distributed and, at the same time, it is composed of the two earlier mentioned independent random vectors, then each of them is normally distributed. It implies that the vector of random errors is normally distributed.

The verification of H_0 is done with the help of a simple sample, which is created by the components of vector r^* and with the help of test for normality. At $n \leqslant 50$ the Shapiro, Wilk (1965) test is recommended, and at $50 < n \leqslant \downarrow 100$ the D'Agostino (1971) test. The tests mentioned are omnibus tests i.e. they are both sensitive to departures from the symmetry and curtosis of the normal distribution. They are characterized by power and their critical values are known. The Shapiro–Wilk and Shapiro–Francia tests have left-side critical regions. This means that the big values of the tests statistics do not lead to rejection of H_0. Furthermore, the D'Agostino test has a two-sided critical region. The H_0 hypothesis is not rejected with this test when the value of the test statistic lies between the upper and lower critical values.

REFERENCES

Cramer H. (1958): *Metody matematyczne w statystyce*, PWN, Warszawa.

D'Agostino R. B. (1971): *An omnibus test of normality for moderate and large size samples*, „Biometrika", 58, p. 341–348.

Durbin J. (1961): *Some methods of constructing exact test*, „Biometrika", 48, p. 41–55.

Golub G. H., Gutman I., Duter R. (1973): *Examination of pseudo-residuals of outliers for detecting suporsity in the general univariate linear model*, [in:] (eds) D. G. Kabe, R. P. Gupta, *Multivariate statistical inference*, North-Holland Publishing Company, Amsterdam.

Fisz M. (1967): *Rachunek prawdopodobieństwa i statystyka matematyczna*, PWN, Warszawa.

Mardia K. V. (1980): *Tests of univariate and multivariate normality. Handbook of statistics*, Vol. 1, p. 279–320.

Pawłowski Z. (1976): *Statystyka matematyczna*, PWN, Warszawa.

Pearson E. S., Sekar C. A. (1936): *The efficiency of statistical tools and a criterion for the rejection of outlying observations*, „Biometrika", 28, p. 308–320.

Rao C. R. (1982): *Modele liniowe statystyki matematycznej*, PWN, Warszawa.

Sally L., Sarkadi K. (1982): *Beurteilung der Normaliat an Handmehrerer Stichproben kleinen Umfangs*, „Qualitat und Zuverlassigkeit", 27, p. 197–199.

Sarkadi K. (1960): *On testing for normality*, Publ. Math. Inst. Hungar. Acad. Sci. 5, p. 269–275.

Sarkadi K. (1967): *On testing for normality*, Proc. 5th Berkeley Symp. Math. Statist. Prob. 1, p. 373–387.

Shapiro S. S., Francia R. S. (1972): *Approximate analysis of variance test for normality*, JASA, 67, p. 215–216.

Shapiro S. S., Wilk M. B. (1965): *An analysis of variance test for normality (complete samples)*, „Biometrika", 52, p. 591–611.

Stormer H. (1964): *Ein Test zun Erkennen von Normalverteilungen*, Z. f. Wahrscheinlichkeitstheorie, 2, p. 420–433.

Theil H. (1968): *A simplification of the BLUS procedure for analyzing regression distribuance*, JASA, 63, p. 242–251.

Wagner W. (1982): *Testing for normality of errors in linear models*, Studia Sci. Math. Hungarica, 17, p. 393–401.

Wagner W. (1990): *Test normalności wielowymiarowej Shapiro–Wilka i jego zastosowania w doświadczalnictwie rolniczym*, Roczniki AR w Poznaniu, Rozprawy Naukowe, 197, Poznań.

Wiesław Wagner

PRZYDATNOŚĆ METODY DURBINA DO BADANIA NORMALNOŚCI W JEDNOWYMIAROWYM MODELU LINIOWYM

W pracy przedstawiono metodę randamizacyjną Durbina do testowania normalności błędów losowych.

Zaprezentowano również metodę eliminacji parametrów zakłócenia poprzez obliczanie wektora resztowego i skorygowanego wektora resztowego.