Test (2013) 22:432–436
DOI 10.1007/s11749-013-0332-8

DISCUSSION

Comments on: An updated review of Goodness-of-Fit tests for regression models

Simos G. Meintanis

Published online: 25 July 2013

First I would like to thank the authors for providing a thorough review of Goodness-of-Fit (GoF) tests for regression models. It is truly an achievement to cover such a wide spectrum of methods of GoF within a paper of reasonable length. With their expertise in GoF, Professors González-Manteiga and Crujeiras (GMC) have succeeded in providing an up-to-date review of methods and corresponding theory of tests dealing mostly with the structure of the regression function, and in doing so manage to cover the basic parametric regression model as well as alternative models which may be generally described as nonparametric or semiparametric. They also consider, apart from the typical scenario of independent observations, GoF tests with dependence, and tests under more complex data structures.

In what follows I will try to discuss in more detail specific alternative aspects of the regression model. For these aspects one may formulate corresponding hypotheses that may also be included under the general heading 'GoF for regression'. In addition, I will try to provide some insight on methods of GoF which utilize the characteristic function (CF).

S.G. Meintanis was on sabbatical leave from the University of Athens to NWU during the preparation of this work.

This comment refers to the invited paper available at doi:10.1007/s11749-013-0327-5.

S.G. Meintanis (✉)
Department of Economics, National and Kapodistrian University of Athens, 8 Pesmazoglou Street, 105 59 Athens, Greece
e-mail: simosmei@econ.uoa.gr

S.G. Meintanis
Unit for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa

1 Aspects of GoF for regression

Assume the general model

$$Y = m(X) + \varepsilon, \tag{1.1}$$

where $(X, Y) \in \mathbb{R}^{p+1}$ are observed, ε denotes the unobserved error, and $m(\cdot)$ is the unspecified regression function.

On the basis of n independent observations on (Y, X), one wishes to test hypotheses which refer to the different 'ingredients' involved in Eq. (1.1). Predominantly, the literature in general as well as the literature reviewed by GMC, is concerned with hypotheses referring to the structure of the regression function. However, some other aspects of the regression model are also of interest. Such an aspect, which is already mentioned in Sect. 2.4 of GMC, is the type of the error distribution F_ε in model (1.1). The corresponding hypothesis of interest is commonly formulated as $H_{0\theta} : F_\varepsilon \in \mathcal{F}_\theta$, where $\mathcal{F}_\theta$ denotes a specific parametric family of distributions indexed by a parameter θ.

Moving away from the completely parametric null hypothesis $H_{0\theta}$ it is also of interest to consider the hypothesis of symmetry around the origin

$$H_{0S} : F_\varepsilon(\cdot) + F_\varepsilon(-\cdot) \equiv 1. \tag{1.2}$$

Typically the identity involved in H_{0S}, or a variation thereof, is properly estimated and the test statistic is constructed as a distance-type measure based on the estimated version of (1.2). For instance, Fan and Gencay (1995) and Dette et al. (2002) use kernel density estimation to construct their test statistic. In the same spirit, classical Kolmogorov–Smirnov and Cramér–von Mises statistics for the null hypothesis H_{0S} have been proposed by Neumeyer et al. (2005) and Neumeyer and Dette (2007). On the other hand Hušková and Meintanis (2012) use the equivalent formulation of (1.2) as $\mathcal{H}_{0S} : \Im(\varphi_\varepsilon) \equiv 0$, where $\Im(\cdot)$ denotes the imaginary part in complex arithmetic and $\varphi_\cdot$ is a generic symbol for the CF of an arbitrary distribution. Based on this formulation, the authors employ the empirical CF and proposed Cramér–von Mises type statistics for testing symmetry in the nonparametric setting of Eq. (1.1) as well as for the linear regression model. The hypothesis of symmetry (1.2) for the error (or innovation) distribution has also been considered in the context of regression-type models involving dependence. See for instance the papers by Bai and Ng (2001), Delgado and Escanciano (2007), Pérez-Alonzo (2007), Klar et al. (2012), Ngatchou-Wandji (2009), and the review article by Meintanis and Ngatchou-Wandji (2012).

Another aspect involved in model (1.1) is the potential dependence between the regressor (or covariate) X and the error ε. Homoscedasticity is often postulated for the errors ε in model (1.1), and several tests for violation of this hypothesis have been considered. However, instead of looking at the conditional variance only, one could consider the full conditional distribution of ε given X. The benchmark hypothesis then is the hypothesis of independence. In other words, and in obvious notation, one wishes to test the null hypothesis

$$H_{0I} : F_{X,\varepsilon} - F_X F_\varepsilon \equiv 0. \tag{1.3}$$

Alternatives of interest are those for which some feature of the error distribution (mean, variance, kurtosis, etc.) depends on the observed regressor. Einmahl and

van Keilegom (2008a, 2008b) consider modifications of the classical Kolmogorov–Smirnov, Cramér–von Mises and Anderson–Darling statistics for the null hypothesis H_{0I} in model (1.1), both under homoscedasticity and heteroscedasticity. Neumeyer (2009) employs as a test statistic an L2-distance between kernel estimators of the conditional distribution and the unconditional distribution of the covariates. While theory for the earlier tests is developed for univariate regressor, Neumeyer's test is applicable for covariates in arbitrary dimension. Yet another approach utilizes the equivalent formulation of the null hypothesis H_{0I} in terms of CFs as $\mathcal{H}_{0I} : \varphi_{X,\varepsilon} - \varphi_X \varphi_\varepsilon \equiv 0$. This approach is followed by Hlávka et al. (2011) in which the authors develop weighted L2-type statistics involving an estimated version of the identity in $\mathcal{H}_{0I}$.

2 Interpretations of the characteristic function approach

While methods based on CFs have by now been set on a firm theoretical basis and are often found to outperform classical approaches, it is always an issue to give a clear intuitive interpretation to these methods. In this connection notice that, as statisticians naturally think in terms of distributions, densities, or even (low order) moments, interpretations expressed in terms of these classical quantities are most appropriate for intuitive understanding. The interpretations given below are not at any rate new (see for instance Hušková and Meintanis 2012), but it seems fitting to mention them here on the occasion of the synopsis provided by GMC.

I will start with the test statistic in Eq. (15) of GMC and consider a specific form for the weight function ω. In particular, let Z be a random variable which is symmetric around zero with density $K(\cdot)$ and CF $\varphi > 0$. Note that this implies that $\varphi(\cdot)$ is real and even. Then by setting $\omega = \varphi^2$ in Eq. (15) of GMC we have

$$\frac{T_n}{n} = \int \left|\hat{\phi}_\varepsilon(t)\varphi(t) - \hat{\phi}_{\varepsilon_0}(t)\varphi(t)\right|^2 dt.$$

Clearly $\hat{\phi}_\varepsilon \varphi$ (resp. $\hat{\phi}_{\varepsilon_0}\varphi$) is the CF of the convolution $\mathcal{P}_n \star Z$ (resp. $\mathcal{P}_{n0} \star Z$), where $\mathcal{P}_n$ (resp. $\mathcal{P}_{n0}$) is the empirical distribution putting mass $1/n$ on the residuals $\hat{\varepsilon}_j$ (resp. $\hat{\varepsilon}_{j0}$), $j = 1, \ldots, n$. Then by the Parseval–Plancherel theorem

$$\frac{T_n}{2\pi n} = \int \left|f(x) - f_0(x)\right|^2 dx,$$

with f and f_0 denoting the densities of $\mathcal{P}_n \star Z$ and $\mathcal{P}_{n0} \star Z$, respectively. Notice, however, that

$$f(x) = \frac{1}{n}\sum_{j=1}^{n} K(x - \hat{\varepsilon}_j), \qquad f_0(x) = \frac{1}{n}\sum_{j=1}^{n} K(x - \hat{\varepsilon}_{j0}),$$

which shows that T_n with weight function φ^2 is equivalent to an L2-type distance between a pair of nonparametric density estimators with kernel equal to the density corresponding to the CF $\varphi(t)$, and bandwidth equal to one. A standard choice for the law of Z is the zero-mean normal distribution with variance h^2, $h > 0$. Then $\varphi(t) = e^{-(1/2)h^2t^2}$, and by simple algebra the test statistic reduces to an L2-distance between

two kernel density estimators employing the standard normal density as kernel, with bandwidth equal to h.

An alternative interpretation of T_n in terms of moments is also possible. To this end assume that the weight function satisfies $\int t^a \omega(t)\,dt < \infty$, for each $a > 0$. Then by straightforward algebra we have from Eq. (15) of GMC that

$$\frac{T_n}{n} = \int g^2(t)\omega(t)\,dt, \tag{2.1}$$

where $g(t) = C_n(t) + S_n(t) - C_{n0}(t) - S_{n0}(t)$, with $C_n(t) = n^{-1}\sum_{j=1}^n \cos t\hat{\varepsilon}_j$ and $S_n(t) = n^{-1}\sum_{j=1}^n \sin t\hat{\varepsilon}_j$, and likewise for $C_{n0}(t)$ and $S_{n0}(t)$. In turn, simple Taylor expansions of the $\sin(\cdot)$ and $\cos(\cdot)$ functions lead to the expansion

$$g(t) = \sum_{k=1}^{K} \frac{\gamma_k t^k}{k!}\big(\overline{\hat{\varepsilon}^{(k)}} - \overline{\hat{\varepsilon}_0^{(k)}}\big) + o\big(t^K\big), \quad t \to 0,\ K = 1, 2, \ldots, \tag{2.2}$$

where $\gamma_k = 1$ or -1, and $\overline{u^{(k)}} = n^{-1}\sum_{j=1}^n u_j^k$, $k = 1, \ldots, K$. Clearly then, matching takes place in $g(t)$ between, on the one hand the empirical moments of $\hat{\varepsilon}_j$ and the empirical moments of $\hat{\varepsilon}_{j0}$, on the other hand. Furthermore by substituting Eq. (2.2) into Eq. (2.1) and by some further algebra we arrive at

$$\frac{T_n}{n} = \sum_{k,\kappa=1}^{K} \frac{\gamma_k\gamma_\kappa}{k!\kappa!}\big(\overline{\hat{\varepsilon}^{(k)}} - \overline{\hat{\varepsilon}_0^{(k)}}\big)\big(\overline{\hat{\varepsilon}^{(\kappa)}} - \overline{\hat{\varepsilon}_0^{(\kappa)}}\big)v_{k,\kappa} + \text{remainder}, \tag{2.3}$$

where $v_{k,\kappa} = \int t^{k+\kappa}\omega(t)\,dt$. Consequently the role of the weight function is to specify the weight according to which each moment equation enters the value of the test statistic. The introduction of parametric weight functions such as $\omega(t) = e^{-h|t|^b}$, and $b = 1, 2$, make things even more clear. In such cases, large values of h lead to increasing rates of decay for the weight functions, which implies that higher order moments enter the calculation of T_n with increasingly reduced weights. In fact, if we set $K = 1$ in Eq. (2.3), a quick calculation leads to

$$\lim_{h\to\infty} h^{3/b}\frac{T_n}{n} = \text{constant} \times \big(\overline{\hat{\varepsilon}^{(1)}} - \overline{\hat{\varepsilon}_0^{(1)}}\big)^2, \tag{2.4}$$

where the limit for $b = 1$ (resp. $b = 2$) corresponds to the weight function $w(t) = e^{-h|t|}$ (resp. $w(t) = e^{-ht^2}$), and the constant depends on $w(t)$. The limits in Eq. (2.4) show that we should be cautious when selecting the value of h, since if h is too large we are led to a test statistic that simply matches the sample means of the residuals. On the other hand, a value of h which is too close to zero (recall the aforementioned connection between h and the bandwidth in density estimation), and since for $h = 0$ the integral in Eq. (2.1) diverges, results in a procedure which is vulnerable to numerical instability. Consequently the choice of the weight function calls for a compromise between functions with a high rate of decay which allow for moment matching of only lower order, and functions with a low rate of decay which also involve higher moments but are prone to instabilities.

References

Bai J, Ng S (2001) A consistent test for conditional symmetry in time series models. J Econom 103:225–258

Delgado MA, Escanciano JC (2007) Nonparametric tests for conditional symmetry in dynamic models. J Econom 141:652–682

Dette H, Kusi–Appiah S, Neumeyer N (2002) Testing symmetry in non-parametric regression models. J Nonparametr Stat 14:477–494

Einmahl J, van Keilegom I (2008a) Tests for independence in nonparametric regression. Stat Sin 18:601–616

Einmahl J, van Keilegom I (2008b) Specification tests in nonparametric regression. J Econom 143:88–102

Fan Y, Gencay R (1995) A consistent nonparametric test for symmetry in linear regression models. J Am Stat Assoc 90:551–557

Hlávka Z, Hušková M, Meintanis SG (2011) Tests for independence in non-parametric heteroscedastic regression models. J Multivar Anal 102:816–827

Hušková M, Meintanis SG (2012) Tests for symmetric error distribution in linear and nonparametric regression models. Commun Stat, Theory Methods 41:833–851

Klar B, Lindner F, Meintanis SG (2012) Specification tests for the error distribution in GARCH models. Comput Stat Data Anal 56:3587–3598

Meintanis SG, Ngatchou–Wandji J (2012) Recent tests for symmetry with multivariate and structured data: a review. In: Jiang J et al (eds) Nonparametric statistical methods and related topics. A Festschrift in honour of Professor PK Bhattacharya. World Scientific, Singapore, pp 35–73

Neumeyer N (2009) Testing independence in nonparametric regression. J Multivar Anal 100:1551–1566

Neumeyer N, Dette H (2007) Testing for symmetric error distribution in nonparametric regression models. Stat Sin 17:775–795

Neumeyer N, Dette H, Nagel E–R (2005) A note on testing symmetry of the error distribution in linear regression models. J Nonparametr Stat 17:697–715

Ngatchou–Wandji J (2009) Testing symmetry of the error distribution in non-linear heteroscedastic models. Commun Stat, Theory Methods 38:1465–1485

Pérez–Alonzo A (2007) A bootstrap approach to test for conditional symmetry in time series models. Comput Stat Data Anal 51:3484–3504