Test (2013) 22:437–441
DOI 10.1007/s11749-013-0333-7

DISCUSSION

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

Holger Dette

Published online: 25 July 2013

A survey on the new developments in goodness-of-fit testing in regression models is very welcome and the authors have done an excellent job reviewing the recent literature. I would like to complement this nearly exhaustive exposition by focusing on two further issues in the context of testing model assumptions for regression models. The first refers to hypothesis testing in nonparametric quantile regression models, while the second discusses goodness-of-fit tests in inverse regression models.

1 Recent goodness-of-fit tests in quantile regression

Consider a d-dimensional vector $X_1 = (X_{11}, \ldots, X_{1d})^T$ and denote by $F(y|x)$ the conditional distribution function of Y given $X = x = (x_1, \ldots, x_d)^T$ and by $Q(\tau|x) = F^{-1}(y|x)$ the corresponding conditional quantile function. In the following we fix some quantile $\tau \in (0, 1)$. The recent literature on goodness-of-fit testing in quantile regression models discusses the problem of significance testing and testing for additivity. Both testing problems are motivated by the fact that in practical applications nonparametric quantile regression methods suffer from the curse of dimensionality and therefore do not yield precise estimates of conditional quantile surfaces for realistic sample sizes. Structural assumptions can improve the performance of estimators substantially. For example, under the additional assumption of additivity

$$H_0 : Q(\tau|x) = Q(\tau|x_1, \ldots, x_d) = \sum_{k=1}^{d} Q_k(\tau|x_k) + c(\tau) \tag{1}$$

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

H. Dette (✉)
Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany
e-mail: holger.dette@rub.de

the quantile regression function can be estimated at a one dimensional rate (Doksum and Koo 2000; De Gooijer and Zerom 2003; Horowitz and Lee 2005; Yu and Lu 2004; Lee et al. 2010; Dette and Scheder 2011). Dette et al. (2012) used the approach of Zheng (1996) and consider the statistic

$$T_n = \frac{1}{n(n-1)g_n^d} \sum_{i=1}^{n} \sum_{j \neq i}^{n} L\big((X_i - X_j)/g_n\big) \widehat{R}_i \widehat{R}_j. \tag{2}$$

where L denotes a kernel, g_n a bandwidth and $\widehat{R}_j$ are "residuals" from an additive quantile regression estimate, that is,

$$\widehat{R}_i = I\big\{Y_i \leq \widehat{Q}_{\text{add}}^{-i}(\tau|X_i)\big\} - \tau. \tag{3}$$

Here $\widehat{Q}_{\text{add}}^{-i}(\tau|X_i)$ denotes an additive estimate of the quantile regression function Q without the ith observation (for fixed τ). For a specific class of estimators (see Dette and Volgushev 2008 and Chernozhukov et al. 2010) they showed the asymptotic normality of a standardized version of T_n under the null hypothesis, local alternatives and fixed alternatives. Because the asymptotic distribution depends in a complicated way on the underlying distributions a bootstrap test is implemented to obtain critical values.

An alternative way of reducing the dimensionality is to exclude components from the predictor which are not significant. Although variable selection in the framework of linear quantile regression models has been recently considered by Zou and Yuan (2008), Wu and Liu (2009) and Belloni and Chernozhukov (2011) the problem of testing significance in quantile regression has found much less attention. Jeong et al. (2012) proposed a test for significance of the variable $Z = (X_{11}, \ldots, X_{1d_1})$ in a multivariate quantile regression model, which is motivated by Zheng's method. Their test is based on the statistic

$$J_n = \frac{1}{n(n-1)g_n^{d-d_1}} \times \sum_{\substack{i,j \\ i \neq j}} L\big((Z_i - Z_j)/g_n\big)\big(I\big\{Y_i \leq \hat{q}_\tau(W_i)\big\} - \tau\big)\big(I\big\{Y_j \leq \hat{q}_\tau(W_j)\big\} - \tau\big)$$

where $W_i = (X_{\text{id}_1+1}, \ldots, X_{\text{id}})$, L is a kernel and g_n is a bandwidth converging to 0 with increasing sampling size. These authors state that under the null hypothesis a normalized version of this test statistic converges weakly to a normal distribution (with rate $n^{-1/2}g_n^{-d/4}$). It should be pointed out here that the proof in this paper is not correct (see Dette et al. 2012 for more details).

Volgushev et al. (2013) proposed a test for the hypothesis of the significance of the predictor Z in the nonparametric quantile regression model, which can detect local alternatives converging to the null hypothesis at a parametric rate and at the same time does not depend on the dimension of the predictor Z, such that smoothing with respect to the covariate Z can be avoided. Their approach is based on an empirical process $T_n(x, z)$, which estimates the functional

$$\begin{aligned} T(z, w) &= E\big[\big(P\big(Y \leq q_\tau(W)|W, Z\big)\big) - \tau)I\{W \leq w\}I\{Z \leq z\}\big] \\ &= E\big[\big(I\big\{Y \leq q_\tau(W)\big\} - \tau\big)I\{W \leq w\}I\{Z \leq z\}\big] \end{aligned} \tag{4}$$

for all (z, w) in the support of the distribution of the predictor (Z, W), where the inequality $W \leq w$ between the vectors W and w is understood as the vector of inequalities between the corresponding coordinates and $I\{A\}$ denotes the characteristic function of the event A. It is easy to see that the null hypothesis (the predictor Z is not significant) is equivalent to

$$T(z, w) \equiv 0$$

for all (z, w) in the support of the random variable (X, Z), where the functional T is defined in (4). Volgushev et al. (2013) obtained weak convergence of an appropriately scaled and centered version of $T_n(z, w)$ under the null hypothesis, fixed and local alternatives and proposed a Kolmogorov–Smirnov or a Cramer–von Mises type statistic for the hypothesis of the significance of the predictor Z in the nonparametric quantile regression model.

2 Goodness-of-fit tests in inverse regression models

In contrast to "classical" regression models the regression function in inverse models is not directly related to the mean of the response. More precisely, an inverse regression model is defined by

$$Y = Km(x) + \varepsilon \tag{5}$$

where K denotes an operator, which transfers the function m to a new function Km and ε is a centered random error with existing mean. The function m is not directly observable and the problem of estimating m has found considerable interest in the recent statistical literature (see for example Mair and Ruymgaart 1996; Kaipio and Somersalo 2010; Cavalier 2008; Bertero et al. 2009 or Birke et al. 2010 among others). In inverse regression the simplification of the statistical analysis obtained by parametric models is even more substantial than in the direct case and consequently goodness-of-fit testing is particularly relevant in this context. Surprisingly, the problem of testing parametric hypotheses regarding the regression function m in models of the type (5) is not well-studied in the literature. Some results are available for related indirect models. Here tests for the parametric form of the density in deconvolution problems have been discussed in Butucea (2007), and testing parametric model assumptions in the presence of instrumental variables (which is closely related to statistical inverse problems) has been considered in Holzmann (2007).

In order to illustrate the difficulties which appear while testing model assumptions in inverse regression models consider the problem of testing for a parametric form of the function m, where the operator K is a convolution defined by

$$g(x) := K_\psi f(x) = m * \psi(x) = \int_{\mathbb{R}} \psi(x - t) m(t)\, dt, \tag{6}$$

where ψ is a given function, which is defined by the specific application. An estimator of m can be constructed observing the representation

$$m(t) = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{\mathcal{F}g(\omega)}{\mathcal{F}\psi(\omega)} \exp(it\omega)\, d\omega. \tag{7}$$

where $\mathcal{F}g(\omega)$ and $\mathcal{F}\psi(\omega)$ denote the Fourier transforms of $g = Km$ and ψ, respectively. Because ψ is known only an estimator of $\mathcal{F}g(\omega)$ is required, which can be obtained by the empirical Fourier transform, that is,

$$\hat{\mathcal{F}}g(\omega) = \frac{1}{na_n} \sum_{j=-n}^{n} Y_{j,n} \exp(-i\omega x_{j,n}).$$

Here $Y_{-n,n}, \ldots, Y_{n,n}$ denote observations at experimental conditions $x_{-n,n}, \ldots, x_{n,n}$, where $x_{j,n} = \frac{j}{na_n}$ and $a_n \to 0$ as $n \to \infty$ (if the predictor is bounded the method is not consistent). If the resulting estimator is denoted by $\hat{m}_n$, Bissantz et al. (2012) proposed to reject the null hypothesis for large values of the statistic

$$T_n = \int_{\mathbb{R}} \left| \hat{m}_n(t) - m(t, \hat{\vartheta}) \right|^2 dt$$

where $m(t, \hat{\vartheta})$ is a parametric estimate of the regression function. These authors established asymptotic normality of a standardized version of T_n under the null hypothesis and fixed alternatives, similarly to the case of direct regression (see Härdle and Mammen 1993). However, the corresponding statement is substantially more complicated. In particular the rates depend on the asymptotic properties of the Fourier transform $\mathcal{F}\psi(\omega)$ (see Bissantz et al. 2012 for more details). In other words, any asymptotic inferences depends sensitively on the properties of the operator K in model (6). For these reasons we expect that an interesting field of future research is the consideration of other testing problems for other hypotheses and different operators in the context of inverse regression models.

Acknowledgements The author thanks Martina Stein, who typed parts of this manuscript with considerable technical expertise. This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt A1, C1) of the German Research Foundation (DFG).

References

Belloni A, Chernozhukov V (2011) ℓ_1-penalized quantile regression in high-dimensional sparse models. Ann Stat 39(1):82–130

Bertero M, Boccacci P, Desiderà G, Vicidomini G (2009) Image deblurring with Poisson data: from cells to galaxies. Inverse Probl 25(12):123006, 26

Birke M, Bissantz N, Holzmann H (2010) Confidence bands for inverse regression models. Inverse Probl 26:115020

Bissantz N, Dette H, Proksch K (2012) Model checks in inverse regression models with convolution-type operators. Scand J Stat 39:305–322

Butucea C (2007) Goodness-of-fit testing and quadratic functional estimation from indirect observations. Ann Stat 35:1907–1930

Cavalier L (2008) Nonparametric statistical inverse problems. Inverse Probl 24(3):034004, 19

Chernozhukov V, Fernandéz-Val I, Galichon A (2010) Quantile and probability curves without crossing. Econometrica 78(3):1093–1125

De Gooijer JG, Zerom D (2003) On additive conditional quantiles with high-dimensional covariates. J Am Stat Assoc 98(461):135–146

Dette H, Scheder R (2011) Estimation of additive quantile regression. Ann Inst Stat Math 63(2):245–265

Dette H, Volgushev S (2008) Non-crossing nonparametric estimates of quantile curves. J R Stat Soc, Ser B 70(3):609–627

Dette H, Guhlich M, Neumeyer N (2012) Testing for additivity in nonparametric quantile regression. http://www.ruhr-uni-bochum.de/mathematik3/research/index.html
Doksum K, Koo JY (2000) On spline estimators and prediction intervals in nonparametric regression. Comput Stat Data Anal 35:67–82
Härdle W, Mammen E (1993) Comparing nonparametric versus parametric regression fits. Ann Stat 21:1926–1947
Holzmann H (2007) Testing parametric models in the presence of instrumental variables. Stat Probab Lett 78:629–636
Horowitz J, Lee S (2005) Nonparametric estimation of an additive quantile regression model. J Am Stat Assoc 100(472):1238–1249
Jeong K, Härdle WK, Song S (2012) A consistent nonparametric test for causality in quantile. Econom Theory 3:1–27
Kaipio J, Somersalo E (2010) Statistical and computational inverse problems. Springer, Berlin
Lee YK, Mammen E, Park BU (2010) Backfitting and smooth backfitting for additive quantile models. Ann Stat 38(5):2857–2883
Mair BA, Ruymgaart FH (1996) Statistical inverse estimation in Hilbert scales. SIAM J Appl Math 56:1424–1444
Volgushev S, Birke M, Dette H, Neumeyer N (2013) Significance testing in quantile regression. Electron J Stat 7:105–145
Wu Y, Liu Y (2009) Variable selection in quantile regression. Stat Sin 19:801–817
Yu K, Lu Z (2004) Local linear additive quantile regression. Scand J Stat 31(3):333–346
Zheng JX (1996) A consistent test of a functional form via nonparametric estimation techniques. J Econom 75:263–289
Zou H, Yuan M (2008) Composite quantile regression and the oracle model selection theory. Ann Stat 36:1108–1126