Test (2013) 22:414–418
DOI 10.1007/s11749-013-0329-3

DISCUSSION

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

Jacobo de Uña-Álvarez

Published online: 25 July 2013

This paper presents a review on Goodness-of-Fit (GoF) tests, with the focus on regression models. First of all, I want to thank Wenceslao and Rosa for their enormous effort in bringing us all this existing material on a topic of the highest relevance. As a mathematician, I have enjoyed a lot reviewing how all the challenges involved in this complicated area of research have found many interesting answers and proposals, materialized in clever statistical methods and deep asymptotic results. As a teacher of Statistics courses for graduate students in Biology and other applied sciences at my home University, I must say that the main question underlying this review ('does this regression model fit my data well?') is already present in the mind of the non-mathematicians. Partial answers provided by the classical F-test (for nested models) are not always enough to satisfy my students' curiosity, and they ask for more sophisticated methods as those reviewed in this work. With this in mind, I would like to ask the authors about the level of implementation of the existing GoF tests for regression in user friendly software. I am specially concerned with R software, but I am open to any other package (with emphasis on free software!)

Going now to methodological issues, it is quoted by the authors (Sect. 1.2) that selecting the smoothing parameter in smooth tests is a matter with 'serious gaps' (p. 7, line 47). I basically agree. Even when some theory has been developed (e.g. Gao and Gijbels 2008), it is still unclear what to do in practice. Given a smooth test $T(h)$ depending on a bandwidth h, one would like to choose h such that $P_1(T(h) > c_{\alpha,h})$ is maximum while satisfying $P_0(T(h) > c_{\alpha,h}) \leq \alpha$, where P_0 and P_1 stand for the probability law under the null and the alternative, respectively, and $c_{\alpha,h}$ is the critical

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

J. de Uña-Álvarez (✉)
Department of Statistics and OR, Facultad de CC. Económicas y Empresariales, Universidad de Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain
e-mail: jacobo@uvigo.es

point of $T(h)$ at level α. Martínez-Camblor and de Uña-Álvarez (2013a) introduced different algorithms to find such a bandwidth in practice. Although their ideas were introduced in the context of k-sample smooth tests, they can be applied here in their original form.

Just to give some details, define $h_{\max} = \arg\max_h P_1(T(h) > c_{\alpha,h})$, that is, the bandwidth leading to the largest power. In practice, since P_1 is unknown, it is estimated by bootstrapping the alternative hypothesis. Here, by 'bootstrapping the alternative hypothesis' it is meant to draw B_1 bootstrap resamples from the model $Y^* = m_{nh}(X^*) + \varepsilon^*$, where ε^* is a random variate following the distribution function of the nonparametric residuals $Y_i - m_{nh}(X_i)$ (naive bootstrap) or an alternative wild bootstrap or smoothed bootstrap resampling plan. If $T^b(h)$ denotes the statistic based on the bth bootstrap resample, $P_1(T(h) > c_{\alpha,h})$ is approximated by

$$P_1^{B_1}\left(T(h) > c_{\alpha,h}^{B_0}\right) = \frac{1}{B_1}\sum_{b=1}^{B_1} I\left(T^b(h) > c_{\alpha,h}^{B_0}\right),$$

where $c_{\alpha,h}^{B_0}$ is an approximation of $c_{\alpha,h}$ based on an independent set of B_0 resamples under the null model (as discussed in Sect. 3.1 of the review). Consequently,

$$h_{\mathrm{DB}} = \arg\max_{h\in\mathcal{H}} P_1^{B_1}\left(T(h) > c_{\alpha,h}^{B_0}\right)$$

approximates the 'optimal' smoothing factor $h_{\max}$, where here $\mathcal{H}$ stands for a grid of possible bandwidths. This algorithm is termed the 'double bootstrap' (DB) algorithm in Martínez-Camblor and de Uña-Álvarez (2013a), and it is connected to the one suggested in Cao and Van Keilegom (2006). Alternatively, Martínez-Camblor and de Uña-Álvarez (2013a) suggest as another optimal bandwidth the one corresponding to the minimum p-value; this is $h_{\min} = \arg\min_h \pi(t_h)$, where $\pi(t_h) = P_0(T(h) > t_h)$ and where t_h denotes the actual value of the test statistic $T(h)$ on the given sample. Unlike $h_{\max}$, the criterion $h_{\min}$ is not influenced by the particular level α under which the test is performed. This appealing alternative idea of finding the minimum p-value has, however, one criticism; one may argue that, when the null model is true, the value t_h will probably fall in an area without much interest, since ability to reject the null is desired under the alternative. Therefore, Martínez-Camblor and de Uña-Álvarez (2013a) introduced a slightly modified minimum p-value criterion, namely

$$h_{e\min} = E_{P_1}\left[\arg\min_h \pi\left(T(h)\right)\right],$$

which is the expectation, under the alternative, of the bandwidth minimizing the p-value of the test. Both $h_{\min}$ and $h_{e\min}$ are implemented by using bootstrap approximations, leading to two selectors termed DM ('double minimum') and BM ('bootstrap minimum'), respectively. The reported simulation results (for the k-sample problem, see also Martínez-Camblor 2010 for paired design) suggest that (a) DB algorithm is too anticonservative (apart from being very computationally intensive); (b) DM is too conservative; and (c) BM is a suitable modification of DM which requires a reasonable computational effort. Finding (a) is not surprising, since no correction for

the automatic choice of h is introduced in the computation of the final p-value of the test, which therefore does not respect the nominal level α (see also simulation results in Cao and Van Keilegom 2006). Finding (c) indicates that the idea of averaging $h_{\min}$ along the alternative distribution of the test statistic leads to a good compromise between level and power. Thus, the application of BM algorithm in the context of GoF for regression seems to deserve some investigation. Explicitly, BM bandwidth is defined as

$$h_{\mathrm{BM}} = \frac{1}{B_1} \sum_{b=1}^{B_1} h_{\mathrm{BM}}^b,$$

where

$$h_{\mathrm{BM}}^b = \arg\min_{h \in \mathcal{H}} \pi^{B_0}\big(T^b(h)\big)$$

is the minimum p-value bandwidth for the bth bootstrap resample drawn under the alternative, $T^b(h)$ denotes the statistic based on that very resample, and π^{B_0} is an obvious bootstrap approximation of the function π obtained from B_0 independent resamples under the null, namely

$$\pi^{B_0}(t) = \frac{1}{B_0} \sum_{b'=1}^{B_0} I\big(T^{b'}(h) > t\big).$$

It would be interesting to see how the several existing smooth tests for regression models work with this automatic selector. See also Martínez-Camblor and de Uña-Álvarez (2013b) for a refinement of the BM algorithm and further discussions.

It is a bit disappointing to see that the test based on the empirical regression process, in spite of its theoretical asymptotic superiority (detecting alternatives at the parametric rate $n^{-1/2}$ regardless the dimension p), may have smaller power than smooth tests for finite sample sizes and $p > 1$; indeed, one would like to trust the empirical regression test statistic since it avoids the issue of bandwidth selection. At this point the authors refer to the simulation study in Miles and Mora (2003) who consider $p = 1, 2$ and $n = 100$ ('small sample size'). I wonder if there is some further evidence on this issue or if the authors have performed simulations on their own to investigate this. In my view, a question of such a big interest should not be left in the hands of a single simulation study. For example, would smooth tests be still superior to the empirical regression test for $p > 2$?

Section 6 is particularly interesting to me, since I have been quite involved (I am still) with the analysis of survival data. In this area, censored, randomly truncated, or length-biased data often appear. The contribution of Cao and González-Manteiga (2008) gives solution to a broad family of testing problems under censoring and truncation, including polynomial regression, proportional hazards, additive risks, and proportional odds model. As the authors say, the key tool here is the generalized conditional Kaplan–Meier estimator of Iglesias-Pérez and González-Manteiga (1999), investigated later in the dependent setting by Liang and de Uña-Álvarez (2011) and Liang et al. (2012). In these papers, the asymptotic properties of the conditional distribution estimator were obtained under the assumption of mutual conditional independence among the censoring, truncating and lifetime variables, given the covariates.

Independence between truncation and censoring times is also assumed in Sánchez-Sellero et al. (2005). In some scenarios, however, the censoring time and the truncation time are dependent, and one rather has a (conditional) independence between the truncation time and the *residual* censoring time ($C - T$, in authors' notation); this typically occurs with cross-sectional survival data, cf. e.g. de Uña-Álvarez (2004), Luo and Tsai (2009), or Huang and Qin (2011). Probably it is worthwhile to investigate how the proposed testing procedures can be adapted to this different scenario. On the other hand, methods for length-biased data reviewed in Sect. 6.4 could be adapted to the situation in which the data suffer from right-censoring besides of the length-biasing. Contributions in regression with length-bias and censoring include Shen (2009), Shen et al. (2009), de Uña-Álvarez and Iglesias-Pérez (2010), Qin et al. (2011), Ning et al. (2011), or Chen and Zhou (2012). So developing GoF tests for regression in this setting seems to be a promising field of research too.

Another sampling issue in Survival Analysis is that of random double truncation. Here, the observation of the variable of interest (X, Y) is only possible when $U \leq Y \leq V$, where U and V are two random limits (the left and right truncation times); U and V are also recorded in that case. If (U, V) is independent of (X, Y), the sampled population is a weighted version of that of interest, with weighting function $w(x, y) = P(U \leq y \leq V)$, which is unknown but may be estimated. In Moreira et al. (2012) a nonparametric estimator for the regression function was introduced, following the ideas of Sect. 6.4 but with $w(.,.)$ replaced by its nonparametric maximum-likelihood estimator computed from the sample (U_i, V_i, X_i, Y_i), $i = 1, \ldots, n$. Therefore, it would be useful to investigate possible extensions of the GoF tests referred in Sect. 6.4 to the case of a random weighting function.

Regarding Sect. 6.2 I only want to mention the recent contribution of Dikta et al. (2013) on model checks for regression with missing binary response data. This is a very important problem in Survival Analysis (among other fields), when some of the censoring indicators are missing; in that case, one looks for a suitable imputation-regression model (here the recorded lifetime plays the role of 'covariate') to recover the missing censoring indicators before some Kaplan–Meier type estimation procedure is performed. If the model is miss-specified, then the resulting estimator is systematically biased. Therefore, GoF tests are more than necessary in this context.

Summarizing, I am sure that this review will contribute a lot to the research activity of people concerned with GoF testing in regression. It provides a quick and clever insight into the main ideas, problems, and existing results in the area. I am looking forward to the following contributions of the authors!

Acknowledgements Work supported by the Grant MTM2011-23204 (FEDER support included) of the Spanish Ministerio de Ciencia e Innovación.

References

Cao R, González-Manteiga W (2008) Goodness-of-fit tests for conditional models under censoring and truncation. J Econom 143:166–190

Cao R, Van Keilegom I (2006) Empirical likelihood tests for two-sample problems via nonparametric density estimation. Can J Stat 34:61–77

Chen X-R, Zhou Y (2012) Quantile regression for right-censored and length-biased data. Acta Math Appl Sin (Engl Ser) 28:443–462
de Uña-Álvarez J (2004) Nonparametric estimation under length-biased sampling and type I censoring: a moment-based approach. Ann Inst Stat Math 56:667–681
de Uña-Álvarez J, Iglesias-Pérez MC (2010) Nonparametric estimation of a conditional distribution from length-biased data. Ann Inst Stat Math 62:323–341
Dikta G, Subramanian S, Winkler T (2013) Bootstrap based model checks with missing binary response data. Stat Probab Lett 83:219–226
Gao J, Gijbels I (2008) Bandwidth selection in nonparametric kernel testing. J Am Stat Assoc 103:1584–1594
Huang C-Y, Qin J (2011) Nonparametric estimation for length-biased and right-censored data. Biometrika 98:177–186
Iglesias-Pérez MC, González-Manteiga W (1999) Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. J Nonparametr Stat 10:213–244
Liang H-Y, de Uña-Álvarez J (2011) Wavelet estimation of conditional density with truncated, censored and dependent data. J Multivar Anal 102:448–467
Liang H-Y, de Uña-Álvarez J, Iglesias-Pérez MC (2012) Asymptotic properties of conditional distribution estimator with truncated, censored and dependent data. Test 21:790–810
Luo X, Tsai WY (2009) Nonparametric estimation for right-censored length-biased data: a pseudo-partial likelihood approach. Biometrika 96:873–886
Martínez-Camblor P (2010) Nonparametric k-sample test based on kernel density estimator for paired design. Comput Stat Data Anal 54:2035–2045
Martínez-Camblor P, de Uña-Álvarez J (2013a) Studying the bandwidth in k-sample smooth tests. Comput Stat 28:875–892
Martínez-Camblor P, de Uña-Álvarez J (2013b) Density comparison for independent and right censored samples via kernel smoothing. Comput Stat 28:269–288
Miles D, Mora J (2003) On the performance of nonparametric specification test in regression models. Comput Stat Data Anal 42:477–490
Moreira C, de Uña-Álvarez J, Meira-Machado L (2012) Nonparametric regression with doubly truncated data. Report 12/04. Discussion papers in Statistics and OR, University of Vigo
Ning J, Qin J, Shen Y (2011) Buckley-James-type estimator with right-censored and length-biased data. Biometrics 67:1369–1378
Qin J, Ning J, Liu H, Shen Y (2011) Maximum likelihood estimations and EM algorithms with length-biased data. J Am Stat Assoc 106:1434–1449
Sánchez-Sellero C, González-Manteiga W, Van Keilegom I (2005) Uniform representation of product-limit integrals with applications. Scand J Stat 32:563–581
Shen P-S (2009) Hazards regression for length-biased and right-censored data. Stat Probab Lett 79:457–465
Shen Y, Ning J, Qin J (2009) Analyzing length-biased data with semiparametric transformation and accelerated failure time models. J Am Stat Assoc 104:1192–1202