Assessing the Fit of Conditional Distributions derived by Bayesian Structured Additive Distributional Regression

Alexander Silbersdorff*[1]

[1]Chair of Statistics, University of Göttingen

Version 0.3
Last changes: 7th March 2017

Abstract

This paper considers testing the adequacy of parametric specifications of conditional distributions for Structured Additive Distributional Regression estimated in a Bayesian framework.

Keywords: Kolmogorov-Smirnov test; Structured Additive Distributional Regression

1 Introduction

One quintessential assumption made by Structured Additive Distributional Regression (SADR) is the use of a parametric form for the conditional distributions. This assumption may be tested by considering the following null-hypothesis $\mathcal{H}_0$ against the alternative hypothesis $\mathcal{H}_1$:

$\mathcal{H}_0$: The conditional distributions can be modelled by parametric form, $p(y \mid \boldsymbol{\theta})$, for all observed values of y and some values of $\boldsymbol{\theta}$ derived for the corresponding covariates, $\mathbf{x}$.

vs.

$\mathcal{H}_1$: The conditional distributions cannot be modelled by parametric form, $p(y \mid \boldsymbol{\theta})$, for all observed values of y and any values of $\boldsymbol{\theta}$ derived for the corresponding covariates, $\mathbf{x}$.

2 Theory and Algorithm

In order to test these hypotheses one can use an adaptation of the well known Kolmogorov-Smirnov test. Our adaptation is based on the work of Andrews (1997) and Rothe and Wied (2013) who

*Corresponding author: asilbersdorff@uni-goettingen.de.

proposed a frequentist framework for the testing of conditional distributions. Using their idea to transform the conditional moment restrictions imposed by the parametric specification of our structured additive distributional regression model into unconditional ones (see Rothe and Wied, 2013), we are able to specify the test statistic T_n as

$$T_n = \sqrt{n} \sup_{(y,\mathbf{x})} \mid \hat{H}_n(y,\mathbf{x}) - \hat{H}_n^0(y,\mathbf{x}) \mid, \tag{1}$$

where $\hat{H}_n(y,\mathbf{x})$ and $\hat{H}_n^0(y,\mathbf{x})$ constitute estimates of the joint cumulative distribution function of both dependent and independent variable for n observations integrated up with respect to the marginal distribution of the conditioning variables:

$$\hat{H}_n(y,\mathbf{x}) = n^{-1} \sum \mathbb{1}_{\{Y_i \leq y\}} \mathbb{1}_{\{\mathbf{X}_i \leq \mathbf{x}\}}$$

and

$$\hat{H}_n^0(y,\mathbf{x}) = n^{-1} \sum \hat{P}_n \mathbb{1}_{\{\mathbf{X}_i \leq \mathbf{x}\}},$$

where $\hat{P}_n$ denotes the estimated cumulative density function based on n samples using structured additive regression while $\mathbb{1}$ again denotes an indicator function.

As the asymptotic distribution of T_n under the null hypothesis depends on the data-generating process in a complex fashion we propose to use a bootstrap procedure to simulate it. In order to incorporate the uncertainty attached to the parameter estimates we use draws from the MCMC realisations and contrast it with simulated realisations of Y for a set of randomly selected covariate combinations of $\mathbf{X}$. Our bootstrap algorithm thus as follows:

Step 1 Draw a bootstrap sample of covariates $\{\mathbf{X}_{b,i}; 1 \leq i \leq n\}$ with replacement from the obtained values in the sample $\{\mathbf{X}_i; 1 \leq i \leq n\}$.

Step 2 Randomly select the $\mathbf{m}$-th MCMC draw from set $\mathcal{M}$ for the parameter estimates, yielding $\theta_{b,1}^{(\mathbf{m})}(\mathbf{x}), \ldots, \theta_{b,K}^{(\mathbf{m})}(\mathbf{x})$.

Step 3 Use $\{\theta_{b,k}(\mathbf{x}); 1 \leq k \leq K\}$ and $\{\mathbf{X}_{b,i}; 1 \leq i \leq n\}$ to simulate $\{Y_{b,i}; 1 \leq i \leq n\}$ in accordance with the parametrically specified conditional distributions.

Step 4 Use bootstrapped data $\{Y_{b,i}; 1 \leq i \leq n\}$, $\{\mathbf{X}_{b,i}; 1 \leq i \leq n\}$ and $\{\theta_{b,k}(\mathbf{x}); 1 \leq k \leq K\}$ to compute estimates $\hat{H}_{b,n}$ and $\hat{H}_{b,n}^0$ yielding the bootstrap realisation of the test statistic: $T_{b,n} = \sqrt{n} \sup_{(y,\mathbf{x})} \mid \hat{H}_{b,n}(y,\mathbf{x}) - \hat{H}_{b,n}^0(y,\mathbf{x}) \mid$.

Using the simulated distribution of T_n we can then derive the corresponding p-value and or critical

values to assess the test statistic.

3 An Application Scenario

As an application let us consider earnings distributions, which are conditioned on a set of covariates (like education, age, etc.). These distributions are often assumed to follow a Dagum Type II distribution, as is done in Sohn (2016). The Dagum Type II distribution is given by

$$p(y \mid \pi_0, a, b, c) = \pi_0 \mathbb{1}_{\{y=0\}} + (1 - \pi_0) p_+(y \mid a, b, c),$$

where π_0 yields a point mass for zero earnings, while the positive domain is modelled by a Type I Dagum distribution that takes the following form:

$$p_+(y \mid a, b, c) = \frac{acy^{ac-1}}{b^{ac}(1 + (y/b)^a)^{p+1}}, \quad a \in \mathbb{R}_{>0},\ b \in \mathbb{R}_{>0},\ c \in \mathbb{R}_{>0}.$$

The adequacy of modelling the conditional earnings distributions may thus be judged on the basis of assessing the following hypotheses:

$\mathcal{H}_0$: The conditional earnings distributions can be modelled by a Dagum Type II distribution, $p(y \mid \pi_0, a, b, c)$, for all observed earnings y and some values of π_0, a, b, c derived for the corresponding covariates, $\mathbf{x}$.

vs.

$\mathcal{H}_1$: The conditional earnings distributions cannot be modelled by a Dagum Type II distribution, $p(y \mid \pi_0, a, b, c)$, for all observed earnings y and any values of π_0, a, b, c derived for the corresponding covariates, $\mathbf{x}$.

Following the algorithm described in the previous section, this hypothesis may be tested as follows: The Dagum Type II distribution usually features two independent MCMC smaples, one for π_0 and one for a, b and c. Supposing that we have 1,000 MCMC draws each, we have a two dimensional sample set $\mathcal{M} = (1, \ldots, M_1) \times (1, \ldots, M_2)$, with $M_1 = M_2 = 1000$. A given random bootstrap sample b thus takes a random set of realisations from the MCMC output $\{m_1; m_2\} = \mathbf{m} \in \mathcal{M}$ to yield $\pi_b^{m_1}$, $a_b^{m_2}$, $b_b^{m_2}$ and $c_b^{m_2}$.

4 Simulation Study

In order to assess the misspecification test we consider three simple simulation studies in order to validate its performance. In each simulation study, we consider a simple framework with one explanatory variable which has a linear effect on all the predictors of the Dagum distribution, i.e.

$$g(\eta^{\theta_k}) = \beta_0^{\theta_k} + \beta_1^{\theta_k} x, \tag{2}$$

where x is an integer from the interval $[1, 10]$ and g is the log-link.

For the simulations we use 1,000 observations and 1000 bootstrap repetitions and contrast the results for a true specification, as specified above, with a misspecified parametric model. For the misspecification we use a log-normal distribution with mean and coefficient of variation equivalent to that of the Dagum specification.

The result from our simulation studies are displayed in Table 1. The p-value for each simulation run as well as the mean of the p-values of all simulation runs (μ), their standard deviation (σ) as well as the three quartiles (Q_1, Q_2, Q_3).

In Simulation Study 1 entails a for the a scenario where we have negligible parameter uncertainty such that the standard deviation of the posterior distribution is 1% of its expectation. The results are displayed in the first and second column. As can be observed the p-values in the first column roughly follow a uniform distribution, as we would expect, while the second column repeatedly rejects the null.

Simulation Study 2 features moderate parameter uncertainty such that the standard deviation of the posterior distribution is 5% of its expectation. The results are displayed in the third and fourth column. As for the first two columns we see that the results are able to clearly distinguish between the correct and the false specification.

For Simulation Study 3, we assume considerable parameter uncertainty with the standard deviation of the posterior distribution is 50% of its expectation. The results are displayed in the last two columns. Given this large uncertainty, the model specification test is less likely to reject the false hypothesis. With higher parameter uncertainty, the test is thus conservative.

Overall, the simulation study indicates that the test generally works although its power is mitigated by large parameter uncertainty.

	Sim. Study 1		Sim. Study 2		Sim. Study 3	
Sim.Run	$\mathcal{H}_0$ TRUE	$\mathcal{H}_0$FALSE	$\mathcal{H}_0$ TRUE	$\mathcal{H}_0$FALSE	$\mathcal{H}_0$ TRUE	$\mathcal{H}_0$FALSE
1	0.30	0.00	0.02	0.00	0.76	0.12
2	0.58	0.00	0.60	0.00	0.66	0.09
3	0.40	0.00	0.60	0.00	0.40	0.03
4	0.21	0.00	0.82	0.00	0.84	0.07
5	0.71	0.00	0.64	0.00	0.45	0.08
6	0.40	0.00	0.92	0.00	0.63	0.07
7	0.07	0.00	0.05	0.00	0.95	0.10
8	0.92	0.00	0.98	0.00	0.81	0.09
9	0.28	0.00	0.45	0.00	0.54	0.10
10	0.10	0.00	0.80	0.00	0.92	0.08
11	0.91	0.00	0.57	0.00	0.62	0.09
12	0.21	0.00	0.50	0.00	0.59	0.09
13	0.47	0.00	0.00	0.00	0.90	0.09
14	0.29	0.00	0.71	0.00	0.37	0.12
15	0.66	0.00	0.15	0.00	0.72	0.10
16	0.15	0.00	0.13	0.00	0.48	0.09
17	0.92	0.00	0.15	0.00	0.48	0.10
18	0.73	0.00	0.23	0.00	0.92	0.12
19	0.49	0.00	0.41	0.00	0.46	0.10
20	0.98	0.00	0.21	0.00	0.67	0.09
21	0.06	0.00	0.31	0.00	0.80	0.09
22	0.29	0.00	0.68	0.00	0.82	0.10
23	0.26	0.00	0.73	0.00	0.84	0.11
24	0.49	0.00	0.82	0.00	0.97	0.10
25	0.39	0.00	0.57	0.00	0.57	0.11
26	0.79	0.00	0.40	0.00	0.60	0.11
27	0.91	0.00	0.06	0.00	0.80	0.10
28	0.20	0.00	0.54	0.00	0.44	0.10
29	0.35	0.00	0.92	0.00	0.56	0.09
30	0.43	0.00	0.05	0.00	0.42	0.16
μ	0.47	0.00	0.47	0.00	0.67	0.10
σ	0.28	0.00	0.31	0.00	0.18	0.02
Q_1	0.26	0.00	0.17	0.00	0.50	0.09
Q_2	0.40	0.00	0.52	0.00	0.64	0.10
Q_3	0.70	0.00	0.70	0.00	0.82	0.10

Table 1: Results from Simulation Studies for Misspecificaiton Test

5 Conclusion

This paper describes and assesses a misspecification test for conditional distributions as estimated by SADR. To this end we outline required modifications to the test by Rothe and Wied (2013). Subsequently, we consider a short simulation study and find that the test generally identifies misspecification although its power is mitigated by large parameter uncertainty.

References

D. W. K. Andrews (1997): A Conditional Kolmogorov Test, in: Econometrica, 65(5), pp. 1097–1128.

C. Rothe and D. Wied (2013): Misspecification Testing in a Class of Conditional Distributional Models, in: Journal of the American Statistical Association, 108(501), pp. 314–324.

A. Sohn (2016): The Gender Earnings Rift: Assessing Hourly Earnings Distributions of Males and Females using Structured Additive Distributional Regression, in: ZfS Working Paper, 07/2016.