Out-of-Sample Bootstrap Tests for Non-Nested Models

Patrick Mair **Achim Zeileis**
Department of Statistics and Mathematics, Wirtschaftsuniversität Wien

Abstract

When testing non-nested models, the asymptotic distribution theory of the ordinary likelihood ratio statistic is not valid anymore. Several test statistics, some of them based on information criteria, have been proposed in order to test such non-nested hypotheses. Concerning bootstrap approaches to simulate goodness-of-fit measures such as the likelihood ratio value, have been elaborated as well. Based on these methods, we extend existing bootstrap simulations towards out-of-sample bootstrap evaluation. As an application, a parametric bootstrap on simulated regression data is provided.

Keywords: non-nested models, out-of-sample bootstrap.

1. Introduction

In standard statistical theory, nested models are usually compared by using a likelihood ratio (LR) statistic which is asymptotically χ^2-distributed with the corresponding difference in the degrees of freedom. However, sometimes it is desired to make a decision between models which are non-nested. In this case, typically AIC or BIC are used to compare the models on a "descriptive" level: The model which minimizes the corresponding IC is chosen. However, this comparison does not allow for a statement that one model fits significantly better than the other one.

Basically, statistical models can be non-nested in the likelihoods, in the regressors, and in the link-function (e.g., linear against log-linear regressions). The estimated parameters optimize a certain target function, typically the maximization of the likelihood. An in-sample evaluation of the target function may be too "optimistic"; with respect to the generalizability of the results, an out-of-sample evaluation is more feasible. We present corresponding out-of-sample bootstrap (OOB) evaluations on the differences in the log-likelihoods as target function which is applicable for nested as well as for non-nested model hypotheses.

For such non-nested models, a broad application spectrum has been shown, ranging from econometrics to psychometrics. As a consequence, numerous test procedures have been developed. An brief overview is given in the following section.

2. General Problem Formulation

Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ denote a random vector with the corresponding observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. When testing non-nested hypotheses, it is not obvious, which one should be considered as H_0 and H_1, respectively. Thus, for model M_f we state the hypothesis H_f that M_f fits the data, and that model M_g fits the data is imposed in H_g. Pertaining to parameter notation, $\boldsymbol{\theta}_f \in \Theta_f$ is the parameter vector for M_f and $\boldsymbol{\theta}_g \in \Theta_g$ is the parameter vector for M_g. It follows that the LR statistic which tests the appropriateness of M_f can be expressed by

$$T_f = \left(L_f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_f) - L_g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_g)\right) - E_{\widehat{\boldsymbol{\theta}}_f}\left(L_f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_f) - L_g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_g)\right). \quad (1)$$

This is a preprint of an article published in the Proceedings of the 22nd Workshop on Statistical Modelling, 2007, 427–430.

In order to test the fit of M_g, T_g can be established straightforwardly by switching the hypotheses. Cox (1962) gives the proof that T_f and T_g, respectively, are asymptotically normally distributed with mean zero. However, the two problematic steps in constructing the test statistic are computing the expectation term and the variance of both statistics. Further developments and generalizations emanating from Cox's test statistic are numerous. A recent overview of existing testing approaches can be found in Golden (2003) and Watnik, Johnson, and Bedrick (2001).

3. Out-of-Sample Bootstrap Approach

3.1. Simulation Setting and DGP

Hinde (1992) accomplished a bootstrap simulation based on the LR-criterion in order to make a decision between two non-nested GLM's M_f and M_g. We extend this approach with respect to a parametric OOB simulation following Hothorn, Leisch, Zeileis, and Hornik (2005). A repeated subsampling is performed where the differences in the log-likelihoods, i.e., $\Delta L = L_f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_f) - L_g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_g)$, are computed by using the training data (50% of the sample) and furthermore evaluated on the test data (remaining 50% of the sample).

As an application, we simulate data which are consistent with the following regression model M_f (see also Watnik et al., 2001):

$$\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{\beta}_v + \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{\epsilon}_x \tag{2}$$

$\boldsymbol{Y}$ is the $n \times 1$ response vector of the simulated observations, $\boldsymbol{V}$ is an $n \times 2$ matrix with the intercept vector and one predictor variable. The remaining k predictors are defined in the $n \times k$ matrix $\boldsymbol{X}$. The joint vector of regression parameters is $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_v, \boldsymbol{\beta}_x)$.

The alternative model M_g is expressed as

$$\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{\beta}_v + \boldsymbol{Z}\boldsymbol{\beta}_z + \boldsymbol{\epsilon}_z. \tag{3}$$

$\boldsymbol{Z}$ are k different regressors with respect to $\boldsymbol{X}$. Thus, $\boldsymbol{V}$ is the vector of common regressors, $\boldsymbol{\epsilon}_x$ and $\boldsymbol{\epsilon}_z$ are i.i.d. with mean $\mu = 0$ and variance $\sigma^2 = 1$, whereas $\boldsymbol{\beta}_z = \boldsymbol{\beta}_x$.

Pertaining to the data generation process (DGP), the regressor matrices are drawn from a multivariate normal distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and the similarity of M_f and M_g is steered by the correlation $r(\boldsymbol{X}, \boldsymbol{Z})$ in the VC-matrix $\boldsymbol{\Sigma}$. The lower $r(\boldsymbol{X}, \boldsymbol{Z})$, the larger difference in model fit between M_f and M_g. As a consequence, $\Delta L = L_f(\boldsymbol{y}|\boldsymbol{\beta}_v, \boldsymbol{\beta}_x) - L_g(\boldsymbol{y}|\boldsymbol{\beta}_v, \boldsymbol{\beta}_z)$ becomes large and the OOB test should detect this deviation.

3.2. Simulation Conditions and Bootstrap Results

The correlation parameters for model similarity vary from .20 to .99, the sample size n from 20 to 100. In order to examine effects with respect to the variations in goodness-of-fit of M_f, different $\boldsymbol{\beta}$-constellations from .20 up to 2.00 are provided. Within each combination of these parameters, 1000 data sets for $k = 2$ x- and z-predictors are simulated. On each of these data sets $B = 250$ bootstrap samples are drawn and evaluated out-of-sample. Finally the rejection rate of the hypothesis that $\Delta \log L = 0$ is determined.

By inspecting the resulting rejection rates in Table 1, several issues of test behavior of the OOB test become obvious: For low β-values, which implies that neither of the models fits the data, the test is not able to detect differences even for low model correlations. With growing β-values, it is capable to discriminate between M_f and M_g with respect to the goodness-of-fit. For large β's and large n-values, the OOB test becomes significant also for negligible model differences (i.e., high correlations). By inspecting the sample size n for medium ranged regression coefficients, the decrease in test power over the model correlations is striking. For $n = 20$, the rejection rate is considerably low, whereas an increasing n leads to a noticeable gain in testing power.

Table 1: Rejection rates for non-nested regression models

			$r(\boldsymbol{X}, \boldsymbol{Y})$									
k	$\boldsymbol{\beta}$	n	.20	.50	.70	.80	.85	.90	.93	.95	.97	.99
2	.2	20	.31	.35	.32	.27	.27	.27	.29	.34	.26	.28
2	.2	30	.39	.44	.30	.37	.35	.33	.34	.35	.34	.34
2	.2	50	.52	.46	.47	.46	.46	.46	.48	.48	.45	.44
2	.2	100	.71	.67	.70	.70	.57	.65	.57	.60	.67	.54
2	.5	20	.40	.33	.36	.34	.32	.36	.29	.31	.27	.36
2	.5	30	67	.69	.59	.52	.53	.46	.47	.46	.46	.34
2	.5	50	.96	.92	.82	.79	.73	.72	.67	.67	.62	.56
2	.5	100	1.00	.99	.99	.95	.94	.91	.92	.80	.78	.73
2	.8	20	.69	.48	.51	.44	.47	.43	.40	.34	.35	.32
2	.8	30	.94	.88	.81	.71	.66	.64	.58	.56	.54	.52
2	.8	50	1.00	1.00	.98	.94	.95	.85	.83	.79	.71	.64
2	.8	100	1.00	1.00	1.00	1.00	.99	.99	.96	.92	.88	.75
2	1.0	20	.67	.57	.54	.49	.47	.45	.36	.49	.36	.41
2	1.0	30	.95	.97	.90	.85	.78	.70	.67	.60	.58	.49
2	1.0	50	1.00	1.00	1.00	.98	.98	.95	.88	.81	.78	.68
2	1.0	100	1.00	1.00	1.00	1.00	1.00	.99	.99	.96	.94	.81
2	1.5	20	.84	.78	.76	.61	.65	.54	.53	.43	.43	.38
2	1.5	30	1.00	.99	.98	.94	.95	.86	.81	.75	.64	.52
2	1.5	50	1.00	1.00	1.00	1.00	1.00	.99	.99	.95	.82	.79
2	1.5	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.98	.89
2	2.0	20	.94	.90	.89	.79	.76	.72	.59	.56	.42	.35
2	2.0	30	1.00	1.00	1.00	.98	.98	.98	.91	.86	.80	.58
2	2.0	50	1.00	1.00	1.00	1.00	1.00	1.00	.99	.96	.93	.81
2	2.0	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.95

References

Cox, D. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B*, **24**, 406-424.

Golden, R.M. (2003) Discrepancy risk model selection test theory for comparing possibly misspecified or nonnested models. *Psychometrika*, **68**, 229-249.

Hinde, J. (1992) Choosing between non-nested models: A simulation approach. In: *Advances in GLIM and Statistical Modelling, Lecture note in Statistics 78.* 119-124, New York: Springer.

Hothorn, T., Leisch, F., Zeileis, A., and Hornik, K. (2005) The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, **14**, 675-699.

R **Development Core Team** (2006) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. `http://www.R-project.org`

Watnik, M., Johnson, J., Bedrick, E. J. (2001) Nonnested linear model selection revisited. *Communications in Statistics - Theory and Methods*, **30**, 1-20.