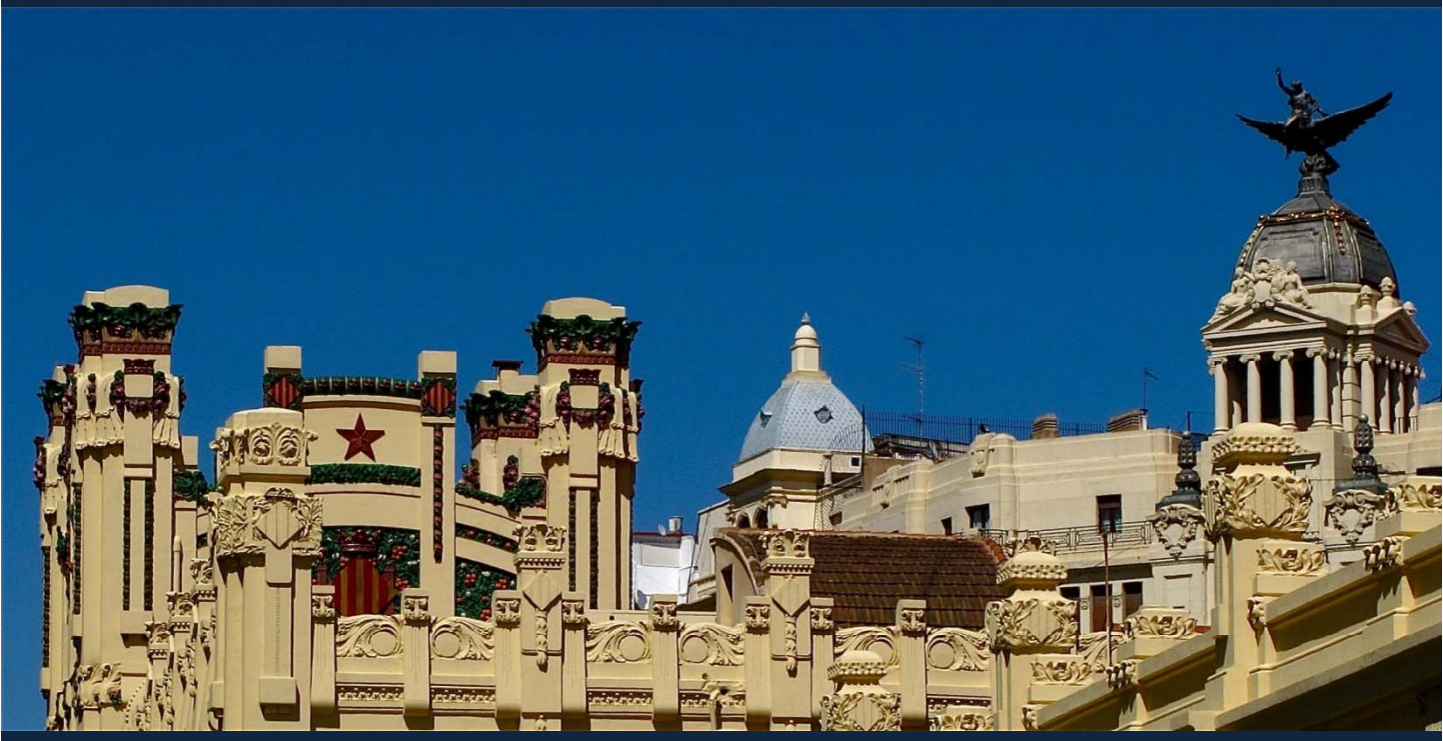
IWSM 2011

Proceedings of the

26th International Workshop on Statistical Modelling

Valencia (Spain), July 11-15, 2011



Editors:
David Conesa
Anabel Forte
Antonio López-Quílez
Facundo Muñoz

Proceedings of the 26th International Workshop on Statistical Modelling

July 11-15, 2011

València

David Conesa, Anabel Forte,
Antonio López-Quílez, Facundo Muñoz
(editors)

Proceedings of the 26th International Workshop on Statistical Modelling.
València, July 11-15, 2011
David Conesa, Anabel Forte, Antonio López-Quílez, Facundo Muñoz, eds.
València 2011.
ISBN 978-84-694-5129-8

Editors:

David Conesa[1], David.V.Conesa@uv.es
Anabel Forte[2], forte@eco.uji.es
Antonio López-Quílez[1], Antonio.Lopez@uv.es
Facundo Muñoz[1], Facundo.Munoz@uv.es

[1] Departament d'Estadística i Investigació Operativa
Universitat de València (Estudi General)
Facultat de Matemàtiques
Dr. Moliner 50, 46100 Burjassot, Spain.

[2] Departamento de Economía
Universitat Jaume I
Facultad de Ciencias Jurídicas y Económicas
Campus del Riu Sec, E-12071 Castelló de la Plana, Spain.

Cover photo: Victor Roda

Printed by Copiformes S.L.

Scientific Programme Committee

- Susie Bayarri (Chair)
  Universitat de València, Spain
- Carmen Armero
  Universitat de València, Spain
- Adrian Bowman
  University of Glasgow, UK
- Charmaine B. Dean
  Simon Fraser University, Canada
- María Durbán
  Universidad Carlos III de Madrid, Spain
- Claire Ferguson
  University of Glasgow, UK
- Herwig Friedl
  Graz University of Technology, Austria
- Gillian Heller
  Macquarie University, Australia
- John Hinde
  University of Galway, Ireland
- Thomas Kneib
  Carl von Ossietzky Universität Oldenburg, Germany
- Arnost Komàrek
  Charles University in Prague, Czech Republic
- Antonio López-Quílez
  Universitat de València, Spain
- Brian Marx
  Louisiana State University, USA
- Pere Puig
  Universitat Autònoma de Barcelona, Spain

Preface

This volume contains all the papers of the 26th International Workshop on Statistical Modelling. Many things have changed since in 1986 an enthusiastic group of statisticians interested in statistical modelling started these series of workshops within a friendly and supportive academic atmosphere. New technologies, more attendants, but always with the same initial spirit: to promote and develop the use of statistical modelling in research and applications.

We are glad to present you these Proceedings, which clearly reflect the aliveness of that spirit. One the one hand, the five invited papers show new advances in theoretical research but always keeping an eye in their applied interest. One the other hand, the great amount of contributions (a total of 140) and their quality demonstrate that the workshop is in good shape. Authors should receive most of the credit for the quality of these Proceedings. Nevertheless, all submissions were carefully reviewed by the members of the Scientific Committee. Their detailed work has been reflected in a big improvement of the preliminary versions jointly with the final selection of contributions.

This 26th edition of the IWSM will be held in Valencia (Spain) in an informal environment (ADEIT- FUNDACIÓ UNIVERSITAT-EMPRESA of the Universitat de València) to encourage discussion and exchange of ideas which could result in future research. Valencia has a great tradition in Statistics and in particular in Bayesian Statistics. This why we are so happy to see that this way of thinking and doing statistics is quite present in these Proceedings reflecting its important role in the Society. We will also like to comment, that many of the contributions in these Proceedings are due to students, which clearly have the future in their hands.

Finally, we wish to acknowledge Carmen Armero, the chair of the local Committee for putting together all the pieces needed in the process of organising this event. Without her interest and passion it would have been impossible.

So welcome to Valencia. Enjoy the city and surroundings and have a great conference.

David Conesa, Anabel Forte, Antonio López-Quílez, Facundo Muñoz
Valencia, June 2011

Contents

Part 1. Invited papers

Part 2. Contributed papers

Testing for a breakpoint in segmented regression: a pseudo-score approach

Vito M. R. Muggeo, Gianfranco Lovison

[1] Dipartimento Scienze Statistiche e Matematiche 'S. Vianelli', Università di Palermo, ITALY - email: vito.muggeo@unipa.it, lovison@unipa.it

Abstract: To overcome the well known oddities in testing for the existence of a breakpoint in segmented regression models, we discuss a novel approach based on the generalized Pearson X^2 statistic, which can be considered as an approximation of the Score statistic. We describe the method and present results from some simulations.

Keywords: segmented regression; break-point; hypothesis testing; Pearson chi-squared; non-standard inference.

1 Introduction

The segmented regression model for a response variable Y and a covariate X postulates that the relationship between X and the conditional mean $E[Y|x] = \mu$ is piecewise linear, i.e. two straight lines connected at an unknown point to be estimated. More broadly we can assume the response belongs to the exponential family with link function $g(\cdot)$ leading to the regression equation

$$g(\mu_i) = z_i^T\gamma + \beta(x_i - \psi)_+ \qquad i = 1, 2, \ldots, n \tag{1}$$

where $(x_i - \psi)_+ = (x_i - \psi)I(x_i > \psi)$ and $z_i^T\gamma$ may include additional linear terms, such as other covariates, the model intercept, and the linear term for the segmented variable that represents the 'left slope' of the piecewise relationship. The choice of a variance function $V[Y|x_i] = \phi v(\mu_i)$ completes the specification of the GLM. This paper deals with testing for the existence of ψ in model (1). When ψ does not exist, model (1) reduces to a 'simple' GLM with linear effects. Roughly speaking, estimation and inference in the segmented regression model are difficult and challenging for several reasons. In particular, testing for the existence of a breakpoint is a non-regular problem which makes the usual statistical tests invalid and involves a lot of theoretical issues, see Feder (1975) for an early work on the topic. The traditional tests are far from being helpful in this context: for instance, the null distribution of the likelihood ratio statistic is bimodal with a zero mean, but its analytical density is unknown. At the best of our

knowledge two approaches have been suggested in the literature. Davies (1987) proposed an approach based on the theory of stochastic processes; the test is currently implemented by the `davies.test()` function in the R package segmented (Muggeo, 2008). The other approach by Kim et al. (2000) uses permutations to obtain the null distribution and to compute the p-value accordingly. However, both approaches provide sub-optimal solutions in some contexts: the permutation test has been discussed only for continuous responses using permutations of the residuals of the null fit and therefore generalizations to other responses, e.g. binary, are not immediate; moreover this approach may become computationally cumbersome for large samples. The Davies test may also be hard to use with large datasets, as several fits (about ten) are needed; furthermore it does not generalize to multiple breakpoints. We discuss a simple and very intuitive approach based on a Pearson-type statistic which performs reasonably well under different models and is simple to implement.

2 Methods

We are interested in testing for the existence of the breakpoint in model (1). Without loss of generality, let $\hat{\mu}_{0i}$ be the fitted values for the 'null' (i.e. no breakpoints) model and $\hat{\mu}_i$ the fitted values under the alternative, namely for the segmented regression fit. The link function $g(\cdot)$ and possible presence of additional covariates do not matter. A generalized form of the Pearson statistic which can be used to compare the two models is

$$X^2_{1|0} = \sum_{i=1}^{n} \frac{(\hat{\mu}_i - \hat{\mu}_{0i})^2}{\phi v(\hat{\mu}_{0i})}, \tag{2}$$

where the dispersion parameter, if unknown, is usually replaced by a corresponding consistent estimate. Notice that, when the alternative model is the saturated model, i.e. $y_i = \hat{\mu}_i$, $X^2_{1|0}$ is the usual Pearson goodness of fit statistic which is equivalent to the score statistic for any GLM. Lovison (2005) showed that for canonical GLM $X^2_{1|0}$ is greater than the equivalent score statistic and he also gave an X^2-like formula for the score statistic. Motivated by these connections, the Pearson-type statistic (2) is referred to as *pseudo*-score statistic, and Agresti and Ryu (2010) used it to build confidence intervals in discrete statistical models. Here we use it for testing for a breakpoint in segmented GLMs, where the usual asymptotic tests fail and current proposals do not appear to be fully satisfactory.
To perform hypothesis testing we need to known the null distribution of $X^2_{1|0}$. With respect to the null linear model, the segmented 'alternative' model has two additional parameters, the difference in slope parameter and the breakpoint, therefore it seems reasonable that under H_0 $X^2_{1|0} \stackrel{d}{\rightarrow} \chi^2_2$. Like for interval estimation problems in Agresti and Ryu (2010), we do not

yet have formal arguments to show that the chi-squared distribution holds under H_0, but we show its performance via simulations.
Table 1 reports the actual sizes of the pseudo score statistic $X^2_{1|0}$ to test for the existence of the breakpoint. We consider different scenarios, with four sample sizes and three densities for the responses: Gaussian, $Y_i \sim \mathcal{N}(\mu_i = 0.15x_i, 0.01^2)$; Poisson, $Y_i \sim \mathcal{P}(\mu_i = e^{2+0.5x_i})$; Negative Binomial, $Y_i \sim \mathcal{NB}(\mu_i = e^{2+0.5x_i}, \mu_i + \mu_i^2/2)$, where $x_i = i/n$ in every scenario. The Negative Binomial family has been selected to assess the performance of the pseudo-score statistic when the model is estimated via a quasi-likelihood approach; for this and the Gaussian example, the dispersion parameter is assumed unknown and it is replaced by a corresponding method-of-moments estimate under the null hypothesis (i.e. from the linear model).

TABLE 1. Empirical sizes (based on 2,000 replicates) of the pseudo score test testing for a breakpoint in different scenarios.

	Gaussian			Poisson			Negative Binomial		
n	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
50	0.007	0.045	0.099	0.015	0.058	0.104	0.009	0.050	0.106
100	0.012	0.055	0.112	0.013	0.059	0.109	0.016	0.061	0.111
500	0.011	0.056	0.105	0.014	0.057	0.118	0.013	0.050	0.110
1000	0.010	0.047	0.093	0.010	0.057	0.112	0.012	0.061	0.117

We observe that the pseudo score test for a breakpoint in segmented regression performs reasonably well by providing empirical sizes close enough to the corresponding nominal values.
Table 2 shows the power of the proposed X^2-type test and the Davies test in detecting a changepoint: we consider two sample sizes ($n = 50, 100$), $\mu_i = 0.05(x_i - \psi)_+$ for Gaussian responses and $\mu_i = e^{2+(x_i-\psi)_+}$ for Poisson and Negative Binomial responses; we also assess the effect of the location of the breakpoint by considering $\psi = 0.50$ and $\psi = 0.75$.

TABLE 2. Empirical power at level 0.05 (based on 1,000 replicates) of the pseudo score test and the Davies test in different scenarios.

			Family		
ψ	n		Gaussian	Poisson	Neg Binom
0.50	50	X^2	0.593	0.269	0.099
		Davies	0.555	0.227	0.096
	100	X^2	0.910	0.457	0.135
		Davies	0.879	0.374	0.119
0.75	50	X^2	0.303	0.125	0.070
		Davies	0.282	0.100	0.091
	100	X^2	0.568	0.227	0.098
		Davies	0.482	0.170	0.088

As expected both tests perform better when ψ is in the middle of the

range of the segmented variable and with larger sample sizes. Although the differences are moderate, generally $X^2_{1|0}$ outperforms the Davies test, and moreover it is actually much simpler to compute, since it requires only two fits.

3 Conclusions

The estimation problem for GLMs involving segmented relationships appears to have received much attention by several authors in the literature, and different solutions are available; see for instance Muggeo (2003). On the other hand, hypothesis testing problems currently present open research questions. In this paper, we have illustrated a very simple, intuitive, and general approach to the problem of testing for a breakpoint in GLMs. Results from some simulation studies show that the Pearson-type statistic provides satisfactory results, at least in the simple case of testing '1 vs. 0' breakpoints. Possible further uses of the Pearson-type statistic concern testing with multiple breakpoints, e.g. 2 vs. 1 or 0 breakpoints, and testing under model misspecification, e.g. in the presence of heteroscedasticity and autocorrelation with continuous responses. These topics need further investigation.

References

Agresti, A., and Ryu, E. (2010) Pseudo-Score Confidence Intervals for Parameters in Discrete Statistical Models *Biometrika*, **97**, 215-222.

Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**, 33-43.

Feder, P.I. (1975). The log likelihood ratio in segmented regression. *Annals of Statistics* **3**, 84-97.

Kim, H.-J., Fay, M.P., Feuer, E.J., and Midthune, D.N. (2000). Permutation Tests for Joinpoint Regression with Applications to Cancer Rates. *Statistics in Medicine* **19**, 335-351.

Lovison, G. (2005) On Rao score and Pearson X^2 statistics in generalized linear models. *Statistical Papers* **46**, 555-574

Muggeo, V.M.R. (2003). Estimating regression models with unknown breakpoints. *Statistics in Medicine*, **22**, 3055-3071.

Muggeo, V.M.R. (2008). Segmented: an R package to fit regression models with broken-line relationships. *R News* **8**(1), 20-25.