

# 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<sup>1</sup>, [David.V.Conesa@uv.es](mailto:David.V.Conesa@uv.es)  
Anabel Forte<sup>2</sup>, [forte@eco.uji.es](mailto:forte@eco.uji.es)  
Antonio López-Quílez<sup>1</sup>, [Antonio.Lopez@uv.es](mailto:Antonio.Lopez@uv.es)  
Facundo Muñoz<sup>1</sup>, [Facundo.Munoz@uv.es](mailto:Facundo.Munoz@uv.es)

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

<sup>2</sup> 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 Komárek  
*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. On the one hand, the five invited papers show new advances in theoretical research but always keeping an eye in their applied interest. On 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

<b>Berger et al.</b> <i>Risk Assessment for Pyroclastic Flows: Combining Deterministic and Statistical Modeling</i> .....	3
<b>Firth</b> <i>Quasi-variances and extensions</i> .....	10
<b>Gómez</b> <i>Some theoretical thoughts when using a composite endpoint to prove the efficacy of a treatment</i> .....	14
<b>Green et al.</b> <i>Identifying influential model choices in Bayesian hierarchical models</i> .....	22
<b>Jørgensen et al.</b> <i>The Ecological Footprint of Taylor's Universal Power Law</i> .....	27

## Part 2. Contributed papers

<b>Aerts et al.</b> <i>Incomplete Clustered Data and Non-Ignorable Cluster Size</i> .....	35
<b>Alvaro-Meca et al.</b> <i>Bayesian Lee-Carter Model: A Spatio-Temporal Approach.</i> .....	41
<b>Andrés-Ferrer and Ney</b> <i>From Empirical Bayes to Leaving-One-Out</i> .....	45
<b>Aregay et al.</b> <i>Model Based Estimates of Long-Term Persistence of Induced HPV Antibodies: A Flexible Subject-Specific Approach</i> .....	49
<b>Armero et al.</b> <i>Bayesian model selection for assessing the progression of chronic kidney disease in transplanted children.</i> .....	53
<b>Badiella et al.</b> <i>Area under the ROC curve using logistic regression with random effects: Estimation and Inference</i> .....	57
<b>Barber et al.</b> <i>Optical properties of fresh date palm in different stages of maturity</i> .....	63
<b>Bárcena et al.</b> <i>Measuring the real estate bubble: a house price index for Bilbao.</i> .....	67

<b>Baxter et al.</b> <i>Missing data, multiple imputation and the UK National Vascular Database</i> .....	71
<b>Belgrave et al.</b> <i>A Comparison of Frequentist and Bayesian Approaches to Latent Class Modelling of Susceptibility to Asthma and Patterns of Antibiotic Prescriptions in Early Life</i> .....	75
<b>Boixadera et al.</b> <i>Who uses Complementary and Alternative Medicine? An analysis for cancer patients</i> .....	79
<b>Bowman and Crujeiras</b> <i>Assessing isotropy with the variogram</i>	83
<b>Brechmann et al.</b> <i>Simplified regular vines for modeling high-dimensional financial risk data</i> .....	87
<b>Brewer et al.</b> <i>Climate Envelopes for Species Distribution Models</i>	93
<b>Burke and MacKenzie</b> <i>XD survival regression models with frailty</i> .....	99
<b>Caballero-Águila et al.</b> <i>Least-squares signal estimation using correlated delayed observations transmitted by different sensors</i> .	105
<b>Caballero-Águila et al.</b> <i>Filtering algorithm for fractional order discrete systems with uncertain observations</i> .....	109
<b>Carrasco et al.</b> <i>The Log-Generalized Modified Weibull Regression Model</i> .....	113
<b>Castillo and Serra</b> <i>An exponential dispersion family to modelling critical phenomenon</i> .....	117
<b>Catelan and Biggeri</b> <i>Hierarchical Bayesian modelling to assess divergence in disease mapping</i> .....	121
<b>Conde and MacKenzie</b> <i>LASSO Penalised Likelihood in High-Dimensional Contingency Tables</i> .....	127
<b>Conesa et al.</b> <i>Describing the geography of Spanish bank branching.</i> .....	133
<b>Corberán-Vallet and Lawson</b> <i>Spatio-temporal disease modeling and surveillance with Bayesian hierarchical Poisson models</i> ...	137
<b>Corberán-Vallet et al.</b> <i>Time series modeling and Bayesian forecasting with exponential smoothing models</i> .....	141
<b>Costa and Dias</b> <i>Assessment of e-government maturity in Portuguese municipalities using regression and clustering approaches</i> .....	146



<b>Creemers et al.</b> <i>Joint Modeling Longitudinal Health Care Costs and Time-to-Event Data in Matched Pairs</i> .....	150
<b>Cysneiros</b> <i>Bartlett-type Correction in Heteroscedastic Symmetric Nonlinear Models</i> .....	156
<b>Cysneiros et al.</b> <i>A Symbolic Robust Regression Model</i> .....	160
<b>Czado et al.</b> <i>Bayesian inference for copula based GARCH models</i>	164
<b>Dejardin et al.</b> <i>Bayesian Dose Escalation in phase I studies of Combinations of Drugs with Control</i> .....	169
<b>De Rooi and Eilers</b> <i>Using text mining tools to compose structure priors for inferring gene networks.</i> .....	173
<b>Djennad et al.</b> <i>Markov-Switching Multifractal models within GAMLSS</i> .....	178
<b>Djeundje and Currie</b> <i>Smooth mixed models for nested curves</i> ..	183
<b>Dondelinger et al.</b> <i>A Bayesian regression and multiple changepoint model for systems biology</i> .....	189
<b>Dooley et al.</b> <i>Analysis of an Observational Study</i> .....	195
<b>Eilers et al.</b> <i>Sea Level Trend Estimation by Seemingly Unrelated Penalized Regressions</i> .....	200
<b>Fabio et al.</b> <i>Generalized random intercept log-gamma exponential family models</i> .....	206
<b>Faria and Gonçalves</b> <i>Modelling Financial Data using Poisson Mixture Approach</i> .....	210
<b>Finazzi et al.</b> <i>A multivariate space-time model for heterogeneous air quality networks</i> .....	214
<b>Fonseca et al.</b> <i>Predictive distributions for non-regular parametric models</i> .....	220
<b>Forte et al.</b> <i>Objective Bayes Criteria for Variable Selection.</i> ....	224
<b>Franco-Villoria et al.</b> <i>Conditional Probability of Flood Risk in Scotland</i> .....	228
<b>Fried et al.</b> <i>Outliers and interventions in INGARCH time series</i>	234
<b>Furche et al.</b> <i>Bivariate Ordinal Regression Models for the Analysis of Neural Data</i> .....	240

<b>Gallego et al.</b> <i>Modelling endocytosis by means of non-homogeneous temporal Boolean models.</i> .....	244
<b>García-Donato et al.</b> <i>A Prior for multiplicity control and closed-form Bayes factors in variable selection</i> .....	248
<b>García-Mora et al.</b> <i>Approximated Survival function in the Sum of Two Independent Homogeneous Markov Processes: Application to Bladder Carcinoma.</i> .....	249
<b>Gargoum</b> <i>On using the Hellinger distance in checking the validity of approximations based on dynamic generalized linear models</i>	253
<b>George and Ünlü</b> <i>Parameter Estimation in Skills-based Knowledge Space Theory and Cognitive Diagnosis Models: A Comparison</i>	258
<b>Gilchrist et al.</b> <i>Forecasting film revenues using GAMLSS</i> .....	263
<b>Gilthorpe et al.</b> <i>Importance of correctly specifying the random structure in growth mixture models</i> .....	269
<b>Gomes et al.</b> <i>Modeling swimming marks through Blocks and POT methods</i> .....	273
<b>Gonçalves and Costa</b> <i>Improvement of surface water quality variables modelling that incorporates a hydro-meteorological factor: a state-space approach</i> .....	276
<b>Gottard et al.</b> <i>Modelling fertility and education in Italy in the presence of time-varying frailty component</i> .....	281
<b>Grisotto et al.</b> <i>Empirical Bayes models to estimate contextual effects</i> .....	287
<b>Habteab Ghebretinsae et al.</b> <i>Generalized Frailty Model for Comet Assays</i> .....	292
<b>Ha et al.</b> <i>Interval Estimation of Random Effects in Frailty Models</i>	298
<b>Haggarty et al.</b> <i>Functional Clustering of Water Quality Data in Scotland</i> .....	303
<b>Hasso and Matawie</b> <i>Using Probability Models to Classify Software Patterns</i> .....	308
<b>Hernandez et al.</b> <i>Linear Model comparison with structured mean and dispersion parameters</i> .....	312
<b>Huertas et al.</b> <i>Joint Modelling of Two Sequential Times to Events With Longitudinal Information</i> .....	316

<b>Ibacache Pulgar and Paula</b> <i>Elliptical semiparametric mixed models</i> .....	322
<b>Kelly</b> <i>The change-point problem in regression with correlated data and change in variance</i> .....	326
<b>Komárek</b> <i>Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data</i> .....	330
<b>Lambert</b> <i>Additive location-scale model when the response and some covariates are interval censored</i> .....	334
<b>Letón and Molanes-López</b> <i>Second order delta method for estimating the Youden index and optimal threshold</i> .....	338
<b>Little et al.</b> <i>Modeling growth patterns of the swift tern using nonlinear mixed effect models</i> .....	342
<b>Loquiha et al.</b> <i>Zero-Inflated Poisson and Negative Binomial Models Applied to Maternal Mortality Rate in Mozambique</i> .....	346
<b>Lynch and MacKenzie</b> <i>On Bivariate Survival Regression Models</i>	352
<b>Marchetti et al.</b> <i>Regression graph models: an application to joint modelling of fertility intentions among childless couples</i> .....	358
<b>Martínez-Beneito et al.</b> <i>A spatio-temporal monitoring system for Influenza-Like Illness incidence</i> .....	364
<b>Martínez-Coscollà et al.</b> <i>Bayesian hierarchical modelling for analyzing the efficiency in the European banking system.</i> .....	368
<b>Marx et al.</b> <i>Multidimensional Single-Index Signal Regression</i> ...	372
<b>Mauff and Little</b> <i>Multivariate Nonlinear Multi-Level Mixed Effect Models: Techniques and Application to Pharmacokinetic Data</i>	378
<b>Mayr et al.</b> <i>Boosting Generalized Additive Models for Location, Scale and Shape</i> .....	384
<b>Menten et al.</b> <i>Estimation of Infection Rates from Repeated ELISA Optical Density Data using Hidden Markov Models</i> .....	390
<b>Mirkov and Friedl</b> <i>Nonlinear and Spline Regression Models for Forecasting Gas Flow on Exits of Gas Transmission Networks</i>	394
<b>Mohd Din et al.</b> <i>Prediction of the rheumatoid arthritis activity score: a joint modeling approach</i> .....	400
<b>Molanes-López et al.</b> <i>Covariate-adjusted inference for the Youden index and associated classification threshold</i> .....	404

<b>Moreira and Machado</b> <i>An R Package for the Estimation of the Bivariate Distribution for Censored Gap Times</i> .....	410
<b>Muggeo and Lovison</b> <i>Testing for a breakpoint in segmented regression: a pseudo-score approach</i> .....	415
<b>Muñoz and López-Quílez</b> <i>Geostatistical modelling with non-Euclidean distances</i> .....	419
<b>Murawska et al.</b> <i>Multi-state models for non Markov process</i> ..	423
<b>Mutsvari et al.</b> <i>Some approaches to correct for misclassification in the absence of an internal validation data set</i> .....	427
<b>Nicholls and Ryder</b> <i>Phylogenetic models for Semitic vocabulary.</i>	431
<b>Nicholls and Watt</b> <i>Partial Order Models for Episcopal Social Status in 12th Century England</i> .....	437
<b>Nysen et al.</b> <i>Testing Goodness-of-Fit of Parametric Models for Censored Data</i> .....	441
<b>Oller and Gómez</b> <i>Testing against ordered alternatives with interval-censored data</i> .....	445
<b>Palarea-Albaladejo and Martín-Fernández</b> <i>Examining distance-based grouping on the simplex sample space: the fuzzy clustering case</i> .....	450
<b>Pardo and Pérez</b> <i>The use of GEE for analyzing housing prices</i> .	454
<b>Peng and MacKenzie</b> <i>Precision of estimators in interval censored parametric survival models</i> .....	458
<b>Pennino et al.</b> <i>A Bayesian spatial approach to modelling fish species occurrence.</i> .....	464
<b>Pereira et al.</b> <i>The truncated inflated beta regression</i> .....	468
<b>Perra et al.</b> <i>A Bayesian analysis of survival times for stage IV non-small cells lung cancer</i> .....	472
<b>Pfeifer</b> <i>On probabilities of avalanches triggered by alpine skiers. Models with random effects taking the stratified data into account.</i> .....	476
<b>Pita-Fernández et al.</b> <i>Cancer incidence in kidney transplant recipients</i> .....	480
<b>Pomann et al.</b> <i>Evaluating Change Detection in Data Streams</i> .	486

<b>Porcu et al.</b> <i>Modelling the Timing of Marital Dissolution in Italy: censored quantile regression with additive terms</i> .....	490
<b>Prieto et al.</b> <i>Estimation of the density of the Antarctic Blue whales population using their sequences of sounds</i> .....	494
<b>Ramsey and Futschik</b> <i>Optimal DNA Pooling for the Detection of Single Nucleotide Polymorphisms</i> .....	499
<b>Riebler et al.</b> <i>Modelling seasonal patterns in longitudinal profiles with correlated circular random walks</i> .....	503
<b>Rippe and Eilers</b> <i>Segmented smoothing with an <math>L_0</math> penalty</i> ....	509
<b>Rodríguez-Álvarez et al.</b> <i>Testing for covariate effects in ROC-GAM regression models based on bootstrap methods</i> .....	515
<b>Rodríguez-Díaz et al.</b> <i>D-Optimum designs in random effect logistic regression models</i> .....	519
<b>Rosen et al.</b> <i>Adaptive Spectral Estimation for Nonstationary Time Series</i> .....	523
<b>Rushworth et al.</b> <i>Distributed lag models for hydrological data</i> .	529
<b>Russo et al.</b> <i>Exact and approximate inferences for nonlinear mixed-effects heavy-tailed models</i> .....	534
<b>Sabanés Bové et al.</b> <i>Hyper-<math>g</math> Priors for Generalised Additive Model Selection</i> .....	538
<b>Schnabel et al.</b> <i>Optimal time scaling for plant growth analysis</i> .	544
<b>Sellers</b> <i>Introducing a Model to Determine True Counts via the Conway-Maxwell-Poisson Distribution</i> .....	548
<b>Sikorska et al.</b> <i>Fast genome-wide association analysis in longitudinal studies</i> .....	553
<b>Singh and Huzurbazar</b> <i>Analysis of Gene Duplication Data</i> ....	557
<b>Slaets et al.</b> <i>Flexible Modelling of Functional Data using Continuous Wavelet Dictionaries</i> .....	561
<b>Smith and Bowman</b> <i>Boundary identification in 3D images</i> .....	565
<b>Sobotka et al.</b> <i>Confidence intervals for geoadditive expectile regression models</i> .....	571
<b>Stefanova</b> <i>Measuring Efficiency of Trial Designs with Unreplicated or Partially Replicated Test Lines</i> .....	577

<b>Stöber and Czado</b> <i>A Markov switching model for vine copulas</i> . . . . .	581
<b>Sweeney and Haslett</b> <i>Bayesian residual analysis in Poisson regression models.</i> . . . . .	587
<b>Tamura and Giampaoli</b> <i>Prediction for an observation in a new cluster for Multilevel Logistic Regression considering <math>k</math> random coefficients</i> . . . . .	593
<b>Taylor and Einbeck</b> <i>Multivariate regression smoothing through the “falling net”</i> . . . . .	597
<b>Tharmaratnam and Claeskens</b> <i>Robust model selection in additive penalized regression splines models</i> . . . . .	603
<b>Thompson</b> <i>Statistical modeling of geographic risks for very low birth weights near Texas superfund sites</i> . . . . .	607
<b>Ugarte et al.</b> <i>Spatio-temporal risk smoothing and forecasting with <math>P</math>-splines</i> . . . . .	612
<b>Urbano et al.</b> <i>Bioassays models with natural mortality and random effects</i> . . . . .	616
<b>Usuga et al.</b> <i>A study to compare HGLM and GAMLSS in mixed linear models</i> . . . . .	622
<b>Van den Hout et al.</b> <i>A latent-class semi-parametric change point model for cognitive ability in older age</i> . . . . .	626
<b>Van Oirbeek and Lesaffre</b> <i>Measuring the Brier score for frailty models</i> . . . . .	632
<b>Ventrucci et al.</b> <i>A Dipole Model for MEG Data</i> . . . . .	636
<b>Ventura and Racugno</b> <i>A Bayesian adjustment of the modified profile likelihood</i> . . . . .	642
<b>Waldmann and Kneib</b> <i>Bayesian Structured Additive Quantile Regression</i> . . . . .	648
<b>West et al.</b> <i>Groups within networks</i> . . . . .	652
<b>Worton and Mclellan</b> <i>Robust mixture modelling of telemetry data in wildlife studies of home range</i> . . . . .	656
<b>Yee and Hadi</b> <i>Row-Column Association Models</i> . . . . .	660
<b>Ziegler-Graham and Rohde</b> <i>Use of Marginal Likelihoods in Statistical Inference</i> . . . . .	666

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

Vito M. R. Muggeo, Gianfranco Lovison

<sup>1</sup> Dipartimento Scienze Statistiche e Matematiche ‘S. Vianelli’, Università di Palermo, ITALY - email: [vito.muggeo@unipa.it](mailto:vito.muggeo@unipa.it), [lovison@unipa.it](mailto: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

$$g(\mu_i) = z_i^T \gamma + \beta(x_i - \psi)_+ \quad i = 1, 2, \dots, n \quad (1)$$

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  $\psi$  in model (1). When  $\psi$  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

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

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.  $y_i = \hat{\mu}_i$ ,  $X_{1|0}^2$  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  $X_{1|0}^2$  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 know the null distribution of  $X_{1|0}^2$ . 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  $H_0$   $X_{1|0}^2 \xrightarrow{d} \chi_2^2$ . 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.

$n$	Gaussian			Poisson			Negative Binomial		
	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.

$\psi$	$n$		Family		
			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  $\psi$  is in the middle of the

range of the segmented variable and with larger sample sizes. Although the differences are moderate, generally  $X_{1|0}^2$  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.