Application of Smooth Tests of Goodness of Fit to Generalized Linear Models

Paul Rippon

B Eng (Chem, Hons1)

Grad Dip Comp Sci

Grad Dip Math Stud

Grad Cert Prac Ter Teach

Thesis submitted to satisfy requirements for the degree of Doctor of Philosophy in Statistics.

October, 2012.

The thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying subject to the provisions of the Copyright Act 1968.

Paul Rippon

Acknowledgements

I would like to acknowledge and thank my supervisors Professor John Rayner and Dr Frank Tuyl for their support and guidance in completing this research. John, in particular, agreed to take over as my supervisor several years ago and provided the possibility of a new research area, interesting ideas and a path forward when I had lost my way in the dark PhD forest. In addition to his knowledge and insight, his unfailingly positive approach and enthusiasm has been of immense value over the long and often painful process of completing this thesis. More recently, Frank has provided a different perspective in helping to critically review the writing and suggest improvements in how certain ideas were presented.

I would also to acknowledge the patience of my examiners, whoever they may be, who have been waiting for some time now to receive this thesis.

Finally, to a long list of family and friends, you are now permitted to ask that dreaded question: "How's the thesis?"

Contents

Abstract

Statistical models are an essential part of data analysis across many diverse fields. They are used to test research hypotheses, aid decision making, estimate effect sizes and/or improve understanding of the underlying processes generating the data of interest. However it is essential to critically assess any fitted model, confirming that the model really is compatible with the data, before meaningful conclusions are possible.

Generalized linear models (GLMs) provide a flexible modelling framework encompassing many commonly used models including the normal linear model, logistic regression model and Poisson regression model. This thesis explores how **the smooth testing concept** – originally proposed by Neyman (1937) and further developed by Rayner et al. (2009) among others – can be **used to test the distributional assumption in a GLM**. However sensible interpretation of this test, or any other test used to assess the fit of a GLM, must recognize that:

- the stochastic, deterministic and link components that make up a GLM should all be considered when assessing model validity,
- the validity of any one of these three components cannot be sensibly considered in isolation as it is confounded by the validity of the other two.

It is therefore important to consider how the smooth test developed in this thesis might be usefully incorporated into an overall model development strategy for GLMs, either replacing or supplementing existing diagnostic tools. Simulation studies demonstrate that the **power of the smooth test** is competitive with other existing tests. However, it also offers the possibility of **improved diagnostic ability**

through the breakdown of the overall smooth test statistic into a sum of squares of interpretable components. The `SmoothGLM` package has been developed which **implements the smooth test** in a form that can be easily applied to models fitted using the standard `glm()` function **within the R statistical computing environment**.