Statistical Modelling of Skewed Data in Clinical Trials Using Transformations and Two-Part Models

(Griffin JM, Lall R, Warwick J, Lamb SE)

Warwick Medical School
CLINICAL TRIALS UNIT

WARWICK
THE UNIVERSITY OF WARWICK

INTRODUCTION

• It is common to encounter skewed outcome data in clinical trials (e.g. resource data, recovery time, pain scores). Such data may can also be characterised by a distribution with a mass at one or more points (i.e. semi-continuous).

• A number of approaches often used include log transformed OLS, non-parametric, Bayesian analysis; but each have their limitations.

• The two part model (Duan;1983; Mullahy;1998) was developed for healthcare expenditure data. It has received attention in clinical trials for analysing skewed outcome data.

OBJECTIVE

• To compare two part models perform against commonly used models when analysing the Roland Morris Questionnaire (RMQ) scores collected from a large RCT of back pain (Back Skills Training trial (BEST)) and outline the purpose and direction of future simulation studies.

METHODS

STUDY DESIGN

-BEST compared the clinical effectiveness of active management (AM) in general practice versus AM plus a group-based, professionally led cognitive behavioural approach (CBA) for sub acute and chronic low back pain (LBP)

-Follow up was at 3, 6 and 12 months post randomisation.

-The primary outcome for the BEST trial was the RMQ score.

-It is the most extensively used outcome measure in back pain studies, with scores ranging from 0-24.

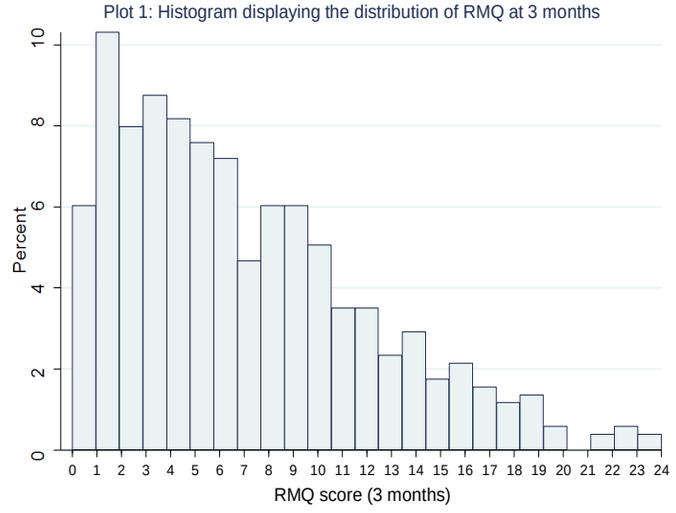
- It is known to have ceiling effects (as shown in Plot 1) low scores indicate less disability.

STATISTICAL METHODS

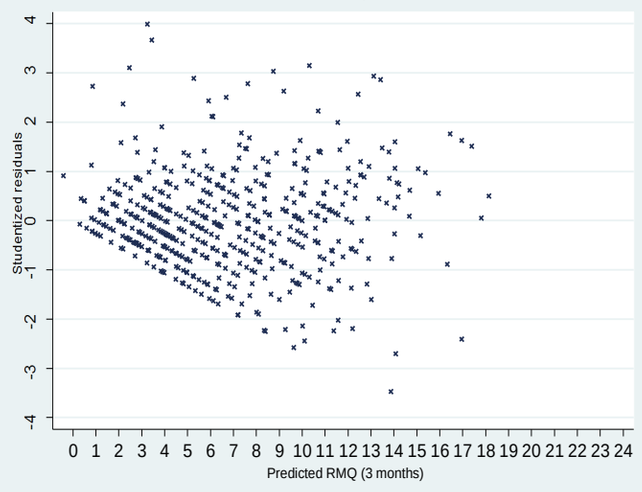
•For each model under study (described in Table 1) the main covariate of interest was treatment – CBA versus AM (adjusting for age (continuous), sex (female/male) and baseline RMQ score).

• Exploratory analysis consisted of histograms, Q-Q plots and residual plots.

Plot 1: Histogram displaying the distribution of RMQ at 3 months



Plot 2: Studentised residuals versus predicted values at 3 months



METHODS

To select the 'best' fit model we calculated root mean squared error & mean absolute prediction error. These statistics were ranked and summed for each model. The sum of ranks indicate the fit of the model – lowest rank was the best fit model.

STATISTICIAL MODELS

1P (one part) Linear regression model - without transformation, with log and square root transformation;

•*1P Generalised linear model* - with the choice of the variance function decided by Park's test;

•*1P Logistic regression model* - dichotomised for disability (score 1-24) and no disability (score 0)

•*Two part (2P) models*- made up of two components:

First part: Uses logistic regression to predict the probability of any disability;

Second part: predicts the amount of disability expressed in terms of the RMQ score conditional on no disability. Four models fitted:

(i) ordinary linear regression with no transformations;
(ii) ordinary linear regression with logistic transformation;
(iii) ordinary linear regression with square root transformation;
(iv) generalised linear regression.

The probabilities from the first part were multiplied by the expected values from part two to obtain the unconditional predicted values (using the twopm command in Stata 13).

•Duan's smearing estimator was used when the error tem was not normal as a variance stabilising transformation. Bootstrapped samples generated from predicted values and the mean treatment difference and 95% confidence intervals were obtained.

RESULTS

• Exploratory analysis showed (a) little normality (Plot 1); (b) the assumption of homoscedasticity (constant variance) was violated (Plot 2); (c) baseline RMQ score appeared to be linearly related to RMQ score at 3 months and age less so. The plots are typical of all time-points.

•Table 1 shows estimate of treatment effects, 95% confidence interval obtained for each of the models, together with the sum of the ranked diagnostic statistics as described in the Statistical Methods section.

Table 1: Sum of ranks from diagnostic test statistics and treatment effect estimates (95% CI's)

		3 months	6 months	12 months
1-P OLS - untransformed	Sum of ranks	**19.5**	**6.0**	**5.5**
	Estimate (95% CI)	1.1 (0.4, 1.7)	1.4 (0.7, 2.1)	1.3 (0.6, 2.1)
1-P OLS lognormal retransformation	Sum of ranks	**23.5**	**26.0**	**25.5**
	Estimate (95% CI)	1.1 (0.3, 1.9)	1.8 (1.0, 2.6)	1.3 (0.4, 2.2)
1-P OLS square root retransformation	Sum of ranks	**17.0**	**20.5**	**19.0**
	Estimate (95% CI)	1.1 (0.4, 1.8)	1.8 (1.0, 2.6)	1.2 (0.4, 2.0)
1-P GLM	Sum of ranks	**14.0**	**17.0**	**18.5**
	Estimate (95% CI)	0.7 (0.01, 1.1)	1.2 (0.5, 1.9)	0.6 (-0.1, 1.4)
2-P BINARY & OLS	Sum of ranks	**11.0**	**8.5**	**7.0**
	Estimate (95% CI)	1.0 (0.4, 1.7)	1.6 (0.9, 2.3)	1.0 (0.3, 1.7)
2-P BINARY & OLS with lognormal retransformation	Sum of ranks	**17.5**	**21.0**	**19.5**
	Estimate (95% CI)	1.0 (0.2, 1.8)	1.8 (1.0, 2.6)	1.4 (0.5, 2.2)
2-P BINARY & OLS with lognormal retransformation (Duan SE)	Sum of ranks	**13.5**	**14.0**	**17.0**
	Estimate (95% CI)	1.2 (0.3, 1.2)	2.2 (1.2,3.2)	1.7 (0.6, 2.7)
2-P BINARY & OLS with square root retransformation	Sum of ranks	**8.5**	**11.0**	**12.0**
	Estimate (95% CI)	1.0 (0.3, 1.7)	1.6 (0.8, 2.4)	1.1 (0.4. 1.8)
2-P BINARY & GLM	Sum of ranks	**10.5**	**11.0**	**11.0**
	Estimate (95% CI)	1.1 (0.4, 1.9)	1.7 (0.9, 2.6)	1.3 (0.4, 2.2)

CONCLUSION

•Two parts models provides an attractive method of analysis, but may be more computationally intensive and may not necessarily correct for constant variance, even after transformation.

•The two part generalised linear model was rated the 'best' model for the 3 month data. There was evidence that the 2-models performed at least as well as conventional methods

• In the case of a constant variance assumption, the linear regression was the best fit model but this is not a valid assumption in the BEST data as Plot 2 demonstrates.

FURTHER WORK

•We are currently exploring the performance of two part models through developing simulation studies. Multiple scenarios will be considered including skewed and bi-modal data.

•Simulated data have been generated from a range of statistical distributions including beta and gamma distributions. Model performance will be assessed, as well as the bias of treatment estimates.

NHS National Institute for Health Research

Birmingham Science City ideasforlife

Advantage West Midlands
www.advantagewm.co.uk

With thanks to Birmingham Science City Translational Medicine, Clinical Research and Infrastructure Trials Platform

Contact : James Griffin T:02476 522480 E:James.Griffin@warwick.ac.uk