A review of random effects models in EGRET for Windows (Version 2.0.3)

Min Yang
Centre for Multilevel Modelling
Institute of Education, University of London
m.yang@ioe.ac.uk

1. Introduction

1.1 Background

EGRET was originally developed at the School of Public Health of University of Washington USA (Mauritsen, R.H., 1984). Designed for analysing data from Biomedical and Epidemiology studies, EGRET stands for Epidemiological GRaphics Estimation Testing. It fits generalised linear models with and without random effects and survival models. It concentrates on models for categorical data collected from Epidemiology and Biomedical studies including cohort data, cross-sectional data, case-control data, clinical trial data and survival data. It is widely used by Epidemiologists and Biostatisticians.

EGRET for Windows was developed based from an early MS-DOS platform. Released in 1999, the current Window version was developed by a team in CYTEL Software Corporation of Cambridge, MA in the USA.

1.2 Software and hardware requirements

The recommended hardware and software requirements for the window version 2.0.3 (October 2000) include:

- A system running MS Windows 95/98 or MS Windows NT.
- A Pentium-II 200MHz processor.
- 32 MB or more of RAM.
- A hard disk with at least 32 MB of available disk space.

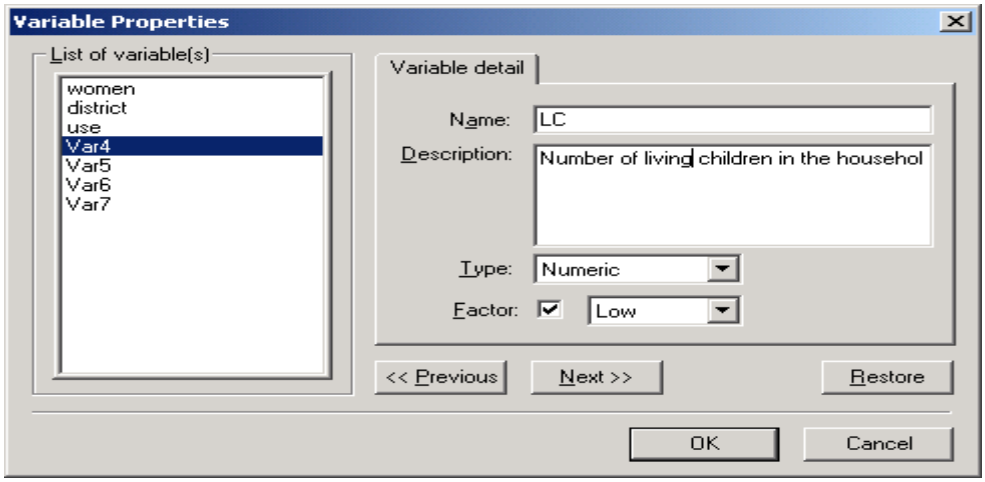
The on line user's manual can be browsed using Acrobat Reader.

1.3 Data input/output functionality

EGRET for Windows can read data files as follows: Dos Egret (*.hdr), LOGXACT (*.cy1), STATXACT (*.cy3), ASCII (*.txt), Text data (*.dat), Excel 5.0 or 7.0 (*.xls), Excel CSV, SPSS (*.scv), SYSTAT (*.syd) and SAS (*.xpt). The same types of data files can be exported from EGRET for Windows except for SPSS, SYSTAT and SAS files.

Data file input is by means of the Import option in the dropdown list of the File window, and output is by the Save / Save As options in the same list. For opening a saved EGRET system file (*.cy1) or (*.hdr), the Open option should be used.

Once a dataset is imported in EGRET, there is a standard procedure for naming, transforming and defining variables – see window below. For a categorical covariate, the *factor* box should be ticked. The reference category, either at the *low* end of the *high* end of the category range is selected in the box next to the *factor* box.



Copying a segment of the data into another file can be done by highlighting it in the Case Editor window, then clicking on copy button and paste in other files.

Data types EGRET accepts can be String, Numeric and Date.

1.4 Other interface features

For keeping information on data analysis and model fitting, EGRET has four windows for review and output: Log, Result, Desc. Stats, and Current Model Info.

Log window stores the history of analysis or operations and fitted models including estimation procedure, deviance value by iteration, final parameter estimates and their standard errors in table form. They can be either copied partly and pasted into other files or saved wholly as a text file using Save Log option in the File drop down list.

Result workbook window keeps details of the current model definition, number of observations, number of parameters (fixed + random coefficients), deviance, parameter estimates in table form, and timings. The whole window can be output as an Excel (*.xls) or *.htm file via the Save Output option. One can also use the Scratch tool in the same window to organise any information piece by piece into an Excel spreadsheet for presentation or graphing.

Other results such as Fitted values/proportion and Residuals can be copied into the Results window to be saved via the Save Output option. A residual here refers to the difference between the observed and the Fitted value at a data point.

Desc. Stats. window stores histogram of single variable or scatter plot of two variables as part of the descriptive statistics. They can be output as Windows Bitmap file (*.bmp) or JPEG file (*.jpg) using Export Graph option.

Current Model Info. window shows text information including model name, fixed and random terms by variable names, weighting variable, denominator, response variable and level 2 identifier. This is a read-only window.

In the original MS-DOS version of EGRET, simple commands in two groups called DEF and PECAN were required. Commands under DEF were for data manipulation and defining models, and commands under PECAN were for fitting models and model diagnostics. In the window version these commands are made redundant and replaced by Window options with dialog boxes.

2 Standard modelling tools for multilevel analysis

2.1 A brief check of model list

EGRET has window screens with dialog boxes for defining and fitting Logistic Regression, Poisson Regression, Cox Proportional Hazards Regression, and Parametric Regressions for Failure Time. However, as it is designed for analysing categorical data, it has no suitable tool for modelling Normal response data.

For binary and binomial outcomes with a (2-level) clustering structure, EGRET fits logistic models with random. The estimation algorithms are Modified-Newton (default), or Newton-Raphson or Quasi-Newton or Nelder-Mead method with marginal maximum likelihood estimates provided. Covariates are allowed in the fixed part of the models. The random effect at level 2 can be modelled as a linear function of level 2 covariates. Random effect models are also known as two-parameter models in EGRET. We review here models for binary and binomial outcomes.

EGRET terms data with covariates varying within cluster as *distinguishable* data, and data with constant covariates within cluster as *indistinguishable.* In a two-level structure of individuals nested within cluster, if the response is binary with individual level covariates, the data are distinguishable. For a proportion response, if it is nested within a higher level cluster with covariates at the response level, the data are also distinguishable. Where the proportion response is at the cluster level, the term *indistinguishable* is used.

Some other terms in EGRET different from the conventional ones are as follows.

Conventional description	EGRET terms
Level 2	cluster
Level 2 identifier	Match variable
Denominator	Group size variable
Weight	Repetition count variable
Response variable - numerator/denominator	Outcome variable - numerator
Intercept	%GM, Grand mean
Random effect associated with the intercept	%SCL, Scalar

2.2 Tools for statistical inference and model diagnostics

After fitting each new model, EGRET reports in the Result window an overall deviance and the tail probability of χ^2 distribution using a Wald test statistic for each parameter estimate except for the variance. For each fixed parameter estimate, a 95% confidence interval (CI) for the estimated odds ratio is reported too. When adding further variables to an existing model, the Extend option in the DefineModel window can be used. After the model runs, a likelihood ratio test statistic is reported in the Result window to enable one to evaluate the significance of the new variable(s) added in any part of the existing model. However, this tool does not apply to the situation when one wants to remove parameter(s) from the existing model.

For model diagnostics, EGRET has Post-Fit tools in a table and a graph to report predicted numerator (Fitted values) and predicted proportion (Fitted proportion) as well as residuals for each observed unit. A graph of fitted values against case number for each unit by $y_{ij} = 1$ (the case in Epidemiology) and $y_{ij} = 0$ (control) is presented. One can change settings of the graph, for example, a graph of fitted proportion against residual or against other covariate. A click on any point in the graph will highlight the case record in the table, and vice versa.

A summary of these tools is given in Table 1.

3 Model specifications — Basic models

3.1 Two-level Logistic-binomial models for distinguishable data

Using the standard notation for the i^{th} case in the j^{th} cluster, the probability of response π_{ij} for the i^{th} covariate pattern is related to the covariates by

$$\log it(\pi_{ij}) = x_{ij}\beta + \sigma u_j , \qquad (1)$$

where $x_{ij}\beta = x_{ij1}\beta_1 + ... + x_{ijp}\beta_p$ is the linear predictor in the fixed part, σ is a positive scalar, and is used to model over-dispersion in the data. The distributional assumptions for the random effects are that u_j is a standard Normal random variable. Note that EGRET also allows for the response proportion to have a Beta-binomial distribution with mean given by the fixed part and a scale parameter, or alternatively a binomial-binomial distribution.

The level 2 variation is effectively modelled by the scalar σ which can be a function of cluster-specific covariates as

$$\sigma_j = z_{1j}\theta_1 + ... + z_{qj}\theta_q \qquad (2)$$

where $Z_j = (z_{j1}, ..., z_{jq})$ represent a vector of q cluster-specific covariates, θ_q is the coefficient for z_q. This allows complex variation at cluster level to be fitted.

The example data for illustration purpose are from the 1988 Bangladesh Fertility Survey (Steele, et ct). A sub-sample of 1,934 women grouped in 60 districts had response on the contraceptive use status at time of survey: $y_{ij} = 1$ (using) or $y_{ij} = 0$ (not using). Three background variables of each woman are: number of living children at time of survey (LC) coded as none, 1, 2, 3+ with three dummy variables (x_{1ij}, x_{2ij}, x_{3ij}); age of woman in years (AGE, x_{4ij}) centred at the mean age; and type of region of residence (URBAN, x_{5ij}, Urban=1 and Rural=0). The data are distinguishable and only Logistic-binomial random effect models can be fitted. (See web site data set descriptions)

For a single level logistic model with all three covariates fitted we have,

$$\log it(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} \qquad (4)$$

The model expression in EGRET is

use ~ %GM + LC + AGE + Urban,

where ‘LC’ and ‘Urban’ are two category variables or factors termed in EGRET.

For a variance component or random intercept model, there is

$$\log it(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \sigma u_{0j} , \qquad (5)$$

and the expression in EGRET is

use ~Fixed + random

Fixed = %GM + LC + AGE + Urban

Random = %SCL

For a model with different random effects for urban and rural districts, the model is

$$\log it(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \sigma u_{0j} \qquad (6)$$

and $\sigma = \theta_1 + \theta_2 x_{5ij}$

EGRET uses the model

use ~ Fixed + random

Fixed = %GM + LC + AGE + Urban

Random = %SCL + Urban

In (6) σ is fitted as a function of the intercept and the Urban group.

In fitting the three models above, EGRET produces a Likelihood Ratio Test (LRT) statistic to test the significance of the random intercept term %SCL in Model (5) compared to the single level model (4), and a LRT for the random effect associated with the variable Urban in (6) compared to (5).

Two steps are required to fit a model: defining the model and analysing it. Options DefineModel and Analyze in the menu carry out the steps, as shown below.

Figure 1 The DefineModel window

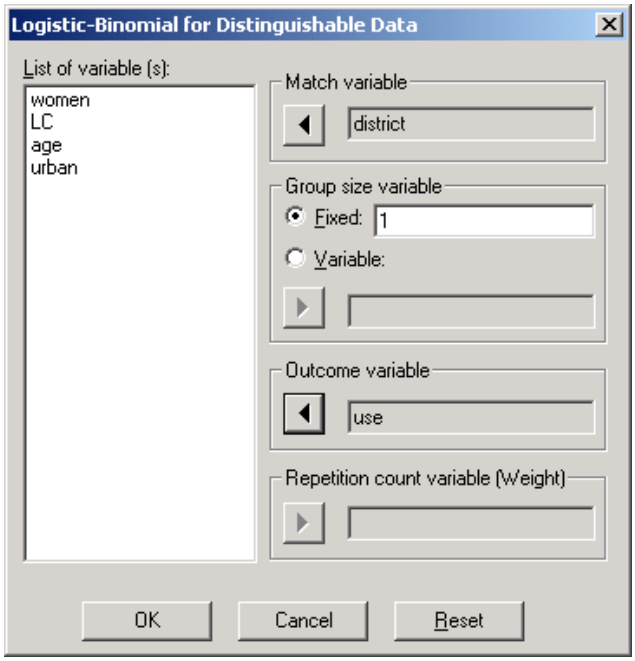
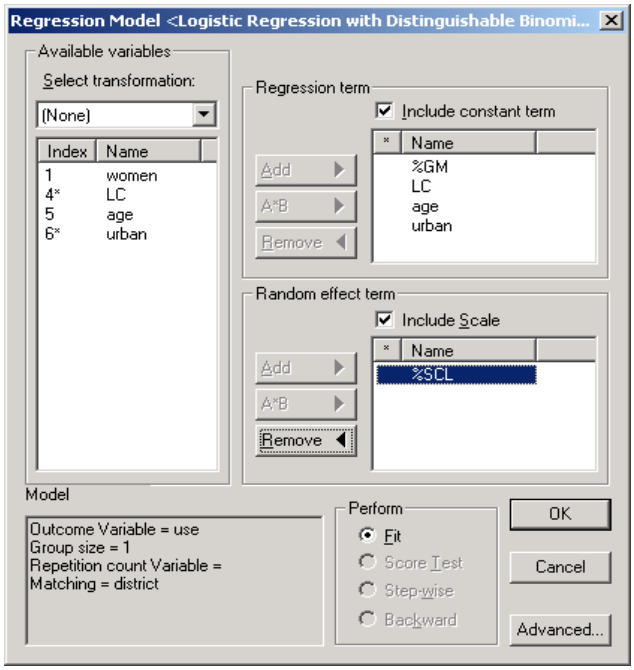


Figure 2 The Analyze window for Model (5)



From the DefineModel window, the information about the cluster, response, denominator and weight are specified. In the Analyze window, Model (5) is set up. Model (4) is fitted by excluding the Scale from the Random effect term. In the Advanced dialog box, one out of four estimation algorithms can be selected.

Estimates for models (4)~(6) are presented in Table 2. The run time has been converted for a Pentium II 433 Mhz processor under Windows 2000.

3.2 *Two-level Logistic models for indistinguishable data*

We first fit a model with different random effects for urban and rural

$$\log it(y_j) = \beta_0 + \beta_5 x_{5j} + (\theta_1 + \theta_2 x_{5j}) \qquad (7)$$

In this review we fit a logistic-Normal model. The model specification and results are in Tables 1 and 2.

Figure 3 The DefineModel window

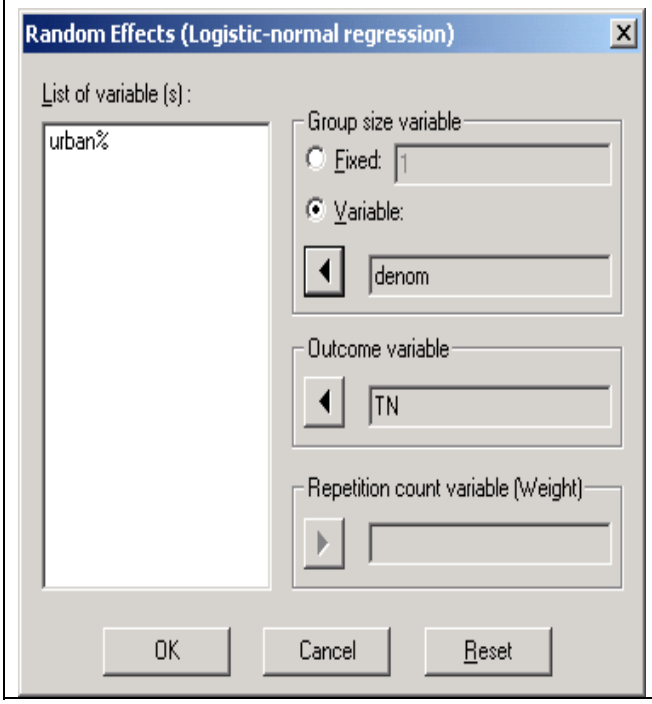
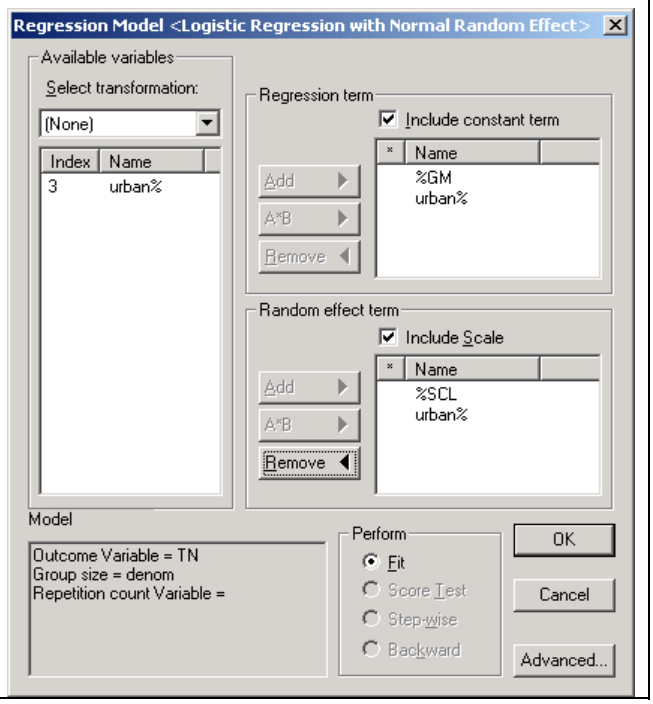


Figure 4 The Analyze window for Model (7)



4 Documentation and user support

The User Manual EGRET for Windows is well organised and well written with clear detail. Part I deals with installation, data input/output, menus, windows and tutorial. Part II describes how to define and run regressions. Part III is about nonparametric procedures, IV is about assessing goodness of fit and V has appendices on special topic such as modelling strategies and troubleshooting, example datasets and program limits. It has a chapter bridging EGRET (DOS) use with the Windows version, and has a list of Beta version testers.

The full document is available also through the on-line help in the program. In the package is included free technical support and new product announcements.

Summary

EGRET for Windows is very easy to use with a user friendly environment for data handling, model definition and fitting and reporting. Being dedicated to binomial and count data as well as survival data, and being unable to fit Normal response models, it has limited functionality and those from outside the medical sciences some of the terminology may be unfamiliar. Furthermore, Poisson models with random effects cannot be fitted. EGRET for Windows is nevertheless good at what it does and a useful package for teaching Epidemiological modelling.

The current prices of EGRET for Windows are $395 per copy for academic users and $795 for commercial. Shipping and handling costs are $75 to abroad and $15 domestic. The program is currently distributed by

Cytel Statistical Software
CYTEL Software Corporation
675 Massachusetts Avenue
Cambridge. MA 02139
USA
www.cytel.com/products/egret

Email: Sales@cytel.com

References

Mauritsen, R.H. (1984). Logistic regression with random effects. Ph.D. thesis, Department of Biostatistics, University of Washington.

Cytel Software Corporation (2000), EGRET for Windows, User manual, Cambridge, USA.

Steele, F., Diamond, I. & Amin, S. (1996). Immunization uptake in rural Bangladesh: a multilevel analysis. Journal of the Royal Statistical Society, Series A, **159**, 289-299.

Table 1 EGRET specifications for single level and 2-level models (distinguishable data)

Model	Fixed parameter and random effects Ests.	Machine steps to run models	Estimates (SE) & Deviance	Seconds to convergence
Single level model	*Fixed* %GM, β_0 LC='2', β_1 LC='3', β_2 LC='4', β_3 Age, β_4 Urban, β_5 *Random* None	*Choose from the menu*: DefineModel Logistic regression with random effects. Logistic-binomial regression for distinguishable data… *In the dialog box, select:* **1. District** as Match variable **2.** Group size variable fixed as 1 **3. Use** as Outcome variable **4.** click on OK. *Choose from the menu* Analyze New… *In the dialog box:* 1. Add **LC**, **Age** and **Urban** to the Model Terms 2. Tick off the random effect term, **%SCL** 3. Click on OK.	β_0 = -1.568 (0.126) β_1 = 1.059 (0.152) β_2 = 1.288 (0.167) β_3 = 1.216 (0.171) β_4 =-0.024 (0.0075) β_5 = 0.797 (0.105) D = 2,456.73 (df=1,928)	7
Random intercept	*Fixed* %GM, β_0 LC='2', β_1 LC='3', β_2 LC='4', β_3 Age, β_4 Urban, β_5 *Random* %SCL, σ	*Choose from the menu* Analyze Extend… *In the dialog box of Random effect term:* 1. Tick the box of Include Scale 2. Click on OK.	β_0 = -1.694 (0.148) β_1 = 1.109 (0.158) β_2 = 1.378 (0.175) β_3 = 1.347 (0.180) β_4 =-0.027 (0.0079) β_5 = 0.730 (0.120) σ = 0.455 (0.071) D = 2,412.9 (df=1,927) LRT (Special) = 43.77 (df=1)	10
Random effect different between Urban and rural groups	*Fixed* %GM, β_0 LC='2', β_1 LC='3', β_2 LC='4', β_3 Age, β_4 Urban, β_5 *Random* %SCL, θ_1 Urban, θ_2	*Choose from the menu* Analyze Extend… *In the dialog box:* 1. Add **Urban** to the Random effect term 2. Click on OK.	β_0 = -1.706 (0.156) β_1 = 1.103 (0.158) β_2 = 1.372 (0.175) β_3 = 1.346 (0.180) β_4 =-0.027 (0.0079) β_5 = 0.805 (0.126) θ_1 = 0.582 (0.099) θ_2 = -0.314 (0.138) D = 2,407.69 (df = 1,926) LRT = 5.28 (df=1)	10

Table 2 EGRET specifications for 2-level Logistic models (indistinguishable data)

Model	Fixed & random effects Ests.	Machine steps to run models	Estimates (SE)	Seconds to convergence
Logistic-normal Quasi-Newton Raphson algorithm	*Fixed* %GM, β_0 Urban%, β_5 *Random* %SCL, θ_1 Urban%, θ_2	*Choose from the menu*: DefineModel Logistic regression with random effects. Logistic-Normal regression… *In the dialog box, select:* 1. Group size variable **Denom** 2. Outcome variable **TN** 3. Click on OK. *Choose from the menu* Analyze New… *In the dialog box:* 1. Add **Urban%** to the Model Terms 2. Add **Urban%** to the Random effect term 3. Click on OK.	β_0 = -0.843 (0.118) β_5 = 1.224 (0.328) θ_1 = 0.434 (0.132) θ_2 =-0.078(0.363) D=113.3 (df=56)	3