A Beginner’s Guide to
Stat-JR’s TREE interface
version 1.0.0

Programming and Documentation by

William J. Browne, Christopher M.J. Charlton, Danius T. Michaelides*, Richard M.A. Parker, Bruce Cameron, Camille Szmaragd, Huanjia Yang*, Zhengzheng Zhang, Harvey Goldstein, Kelvyn Jones, George Leckie and Luc Moreau*

Centre for Multilevel Modelling,
University of Bristol.

*Electronics and Computer Science,
University of Southampton.

December 2013
A Beginner’s Guide to Stat-JR’s TREE interface version 1.0.0

© 2013. William J. Browne, Christopher M.J. Charlton, Danius T. Michaelides, Richard M.A. Parker, Bruce Cameron, Camille Szmaragd, Huanjia Yang, Zhengzheng Zhang, Harvey Goldstein, Kelvyn Jones, George Leckie and Luc Moreau.

No part of this document may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, for any purpose other than the owner’s personal use, without the prior written permission of one of the copyright holders.

ISBN: To be confirmed

Printed in the United Kingdom
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>About Stat-JR</td>
<td>2</td>
</tr>
<tr>
<td>1.1</td>
<td>Stat-JR: software for scaling statistical heights</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>About the Beginner’s guide</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Installing and Starting Stat-JR</td>
<td>4</td>
</tr>
<tr>
<td>2.1</td>
<td>Installing Stat-JR</td>
<td>4</td>
</tr>
<tr>
<td>2.2</td>
<td>The use of third party software and licenses</td>
<td>4</td>
</tr>
<tr>
<td>2.3</td>
<td>Starting up TREE</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>Quick-start guide</td>
<td>6</td>
</tr>
<tr>
<td>3.1</td>
<td>Stage 1: Selecting a template &amp; dataset</td>
<td>7</td>
</tr>
<tr>
<td>3.2</td>
<td>Stage 2: Providing template-specific input</td>
<td>9</td>
</tr>
<tr>
<td>3.3</td>
<td>Stage 3: Outputting the files to run the desired execution</td>
<td>12</td>
</tr>
<tr>
<td>3.4</td>
<td>Stage 4: Running the execution</td>
<td>12</td>
</tr>
<tr>
<td>3.5</td>
<td>Stage 5: The results</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>A detailed guide with worked examples</td>
<td>15</td>
</tr>
<tr>
<td>4.1</td>
<td>The structure and layout of the TREE interface</td>
<td>15</td>
</tr>
<tr>
<td>4.2</td>
<td>Application 1: Analysis of the tutorial dataset using the eStat engine</td>
<td>24</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Summarising the dataset and graphs</td>
<td>24</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Single-level Regression</td>
<td>28</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Multiple chains</td>
<td>35</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Adding gender to the model</td>
<td>37</td>
</tr>
<tr>
<td>4.2.5</td>
<td>Including school effects</td>
<td>39</td>
</tr>
<tr>
<td>4.2.6</td>
<td>Caterpillar plot</td>
<td>41</td>
</tr>
<tr>
<td>4.3</td>
<td>Interoperability – a brief introduction</td>
<td>43</td>
</tr>
<tr>
<td>4.3.1</td>
<td>So why are we offering Interoperability?</td>
<td>43</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Regression in eStat revisited</td>
<td>43</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Interoperability with WinBUGS</td>
<td>45</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Interoperability with OpenBUGS</td>
<td>48</td>
</tr>
<tr>
<td>4.3.5</td>
<td>Interoperability with JAGS</td>
<td>50</td>
</tr>
<tr>
<td>4.3.6</td>
<td>Interoperability with MLwiN</td>
<td>52</td>
</tr>
<tr>
<td>4.3.7</td>
<td>Interoperability with R</td>
<td>55</td>
</tr>
<tr>
<td>4.3.8</td>
<td>Interoperability with AML</td>
<td>60</td>
</tr>
<tr>
<td>4.4</td>
<td>Application 2: Analysis of the Bangladeshi Fertility Survey dataset</td>
<td>61</td>
</tr>
<tr>
<td>4.4.1</td>
<td>The Bangladeshi Fertility Survey dataset</td>
<td>61</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Modelling the data using logistic regression</td>
<td>62</td>
</tr>
</tbody>
</table>
4.4.3 Multilevel modelling of the data.................................................................66
4.4.4 Comparison between software packages..................................................69
4.4.5 Orthogonal parameterisation. ..................................................................73
4.4.6 Predictions from the model ......................................................................75
4.5 Miscellaneous other topics e.g. Data Input/Export........................................77
5 References ........................................................................................................78
6 Appendix: List of Third Party Software that are used by Stat-JR .....................79
Acknowledgements

The Stat-JR software is very much a team effort and is the result of work funded under three ESRC grants: the LEMMA 2 and LEMMA 3 programme nodes (Grant: RES-576-25-0003 & Grant:RES-576-25-0032) as part of the National Centre for Research Methods programme, and the e-STAT node (Grant: RES-149-25-1084) as part of the Digital Social Research programme. We are therefore grateful to the ESRC for financial support to allow us to produce this software.

All nodes have many staff that, for brevity, we have not included in the list on the cover. We acknowledge therefore the contributions of:

Fiona Steele, Rebecca Pillinger, Paul Clarke, Mark Lyons-Amos, Liz Washbrook, Sophie Pollard, Robert French, Nikki Hicks, Mary Takahama and Hilary Browne from the LEMMA nodes at the Centre for Multilevel Modelling.

David De Roure, Tao Guan, Alex Fraser, Toni Price, Mac McDonald, Ian Plewis, Mark Tranmer, Pierre Walthery, Paul Lambert, Emma Housley, Kristina Lupton and Antonina Timofejeva from the e-STAT node.

A final acknowledgement to Jon Rasbash who was instrumental in the concept and initial work of this project. We miss you and hope that the finished product is worthy of your initials.

WJB  November 2013.
1. About Stat-JR

1.1 Stat-JR: software for scaling statistical heights.

The use of statistical modelling by researchers in all disciplines is growing in prominence. There is an increase in the availability and complexity of data sources, and an increase in the sophistication of statistical methods that can be used. For the novice practitioner of statistical modelling it can seem like you are stuck at the bottom of a mountain, and current statistical software allows you to progress slowly up certain specific paths depending on the software used. Our aim in the Stat-JR package is to assist practitioners in making their initial steps up the mountain, but also to cater for more advanced practitioners who have already journeyed high up the path, but want to assist their novice colleagues in making their ascent as well.

One issue with complex statistical modelling is that using the latest techniques can involve having to learn new pieces of software. This is a little like taking a particular path up a mountain with one piece of software, spotting a nearby area of interest on the mountainside (e.g. a different type of statistical model), and then having to descend again and take another path, with another piece of software, all the way up again to eventually get there, when ideally you’d just jump across! In Stat-JR we aim to circumvent this problem via our interoperability features so that the same user interface can sit on top of several software packages thus removing the need to learn multiple packages. To aid understanding, the interface will allow the curious user to look at the syntax files for each package to learn directly how each package fits their specific problem.

To complete the picture, the final group of users to be targeted by Stat-JR are the statistical algorithm writers. These individuals are experts at creating new algorithms for fitting new models, or better algorithms for existing models, and can be viewed as sitting high on the peaks with limited links to the applied researchers who might benefit from their expertise. Stat-JR will build links by incorporating tools to allow this group to connect their algorithmic code to the interface through template-writing, and hence allow it to be exposed to practitioners. They can also share their code with other algorithm developers, and compare their algorithms with other algorithms for the same problem. A template is a pre-specified form that has to be completed for each task: some run models, others plot graphs, or provide summary statistics; we supply a number of commonly used templates and advanced users can use their own – see the Advanced User’s Guide. It is the use of templates that allows a building block, modular approach to analysis and model specification.

At the outset it is worth stressing that there a number of other features of the software that should persuade you to adopt it, in addition to interoperability. The first is flexibility – it is possible to fit a very large and growing number of different types of model. Second, we have paid particular attention to speed of estimation and therefore in comparison tests, we have found that the package compares well with alternatives. Third it is possible to embed the software’s templates inside an e-book which is exceedingly helpful for training and learning, and also for replication. Fourth, it provides a very powerful, yet easy to use environment for accessing state-of-the-art Markov Chain Monte Carlo procedures for calculating model estimates and functions of model estimates, via its eStat engine. The eStat engine is a newly-developed estimation engine with the advantage of being transparent in that all the algebra, and even the program code, is available for inspection.
While this is a beginner’s guide – it is a beginner’s guide to the software. We presume that you have a good understanding of statistical models which can be gained from for example the LEMMA online course (http://www.bristol.ac.uk/cmm/learning/online-course/index.html). It also pre-supposes familiarity with MCMC estimation and Bayesian modelling – the early chapters of Browne (2012) available at http://www.bristol.ac.uk/cmm/software/mlwin/download/2-26/mcmc-web.pdf provide a practical introduction to this material.

Many of the ideas within the Stat-JR system were the brainchild of Jon Rasbash (hence the “JR” in Stat-JR). Sadly, Jon died suddenly just as we began developing the system, and so we dedicate this software to his memory. We hope that you enjoy using Stat-JR and are inspired to become part of the Stat-JR community: either through the creation of your own templates that can be shared with others, or simply by providing feedback on existing templates.

Happy Modelling,

The Stat-JR team.
1.2 About the Beginner’s guide

We have written three initial guides to go with the software: this Beginner’s Guide will cover how to start up and run the software, with a particular focus on the TREE (Template Reading and Execution Environment) interface. It will provide some simple examples and is designed for the researcher who wishes to be able to use the software package without worrying too much about how the mathematics behind the modelling works. As such, it does not go into detail on how users can contribute to extending the software themselves: that is covered in the second, Advanced User’s, guide, designed for those who want to understand in greater detail how the system works. There is also a third, E-book User’s, guide which deals with the software’s alternative DEEP (Documents with Embedded Execution and Provenance) E-book interface.

As well as these three Guides, we also publish support, such as answers to frequently asked questions, on our website (http://www.bristol.ac.uk/cmm/software/statjr), where you can also find our forum in which users can discuss the software.

In this Beginner’s Guide we first describe how to install Stat-JR, and then provide a ‘Quick-start’ guide as a quick visual overview, with brief notes, of the basics of how to work with Stat-JR via TREE. There then follows more detailed sections which provide further explanation, together with point-and-click examples for you to work through.

We look at an example application taken from education research, fitting a Normal response model for a continuous outcome. Here our aim is more to illustrate how to use the software than primarily how to do the best analysis of the dataset in question, and we will demonstrate the interoperability features with the other software packages that link to Stat-JR as well. We will then look at a second example from demography that illustrates binomial response models for a discrete outcome.

2 Installing and Starting Stat-JR

2.1 Installing Stat-JR

Stat-JR has a dedicated website (http://www.bristol.ac.uk/cmm/software/statjr) from which you can request a copy of the software, and which contains instructions for installation.

2.2 The use of third party software and licenses

Stat-JR is written primarily in the Python package but also makes use of many other third party software packages. We are grateful to the developers of these programs for allowing us to use their products within our package. When you have installed Stat-JR you will find a directory entitled licences in which you can find subdirectories for each package detailing the licensing agreement for each. The list of software packages that we are using can be found in the Appendix to this document.
2.3 Starting up TREE

Stat-JR’s interface is viewed and operated via a web browser, but it is started by running an executable file.

To start Stat-JR select the Stat-JR TREE link from the Centre for Multilevel Modelling suite on the start up menu. This action opens a command prompt window in the background to which commands are printed out. This window is useful for viewing what the system is doing: for example, on the machine on which we have run TREE, you can see the following commands:

WARNING:root:Failed to load package GenStat_model (GenStat not found)
WARNING:root:Failed to load package gretl_model (Gretl not found)
WARNING:root:Failed to load package MATLAB_script (Matlab not found)
WARNING:root:Failed to load package Minitab_model (Minitab not found)
WARNING:root:Failed to load package Minitab_script (Minitab not found)
WARNING:root:Failed to load package MIXREGLS (MIXREGLS not found)
WARNING:root:Failed to load package Octave_script (Octave not found)
WARNING:root:Failed to load package OpenBUGS (OpenBUGS not found)
WARNING:root:Failed to load package SABRE (Sabre not found)
WARNING:root:Failed to load package SAS_model (SAS not found)
WARNING:root:Failed to load package SAS_script (SAS not found)
WARNING:root:Failed to load package SPSS_model (SPSS not found)
WARNING:root:Failed to load package SPSS_script (SPSS not found)
WARNING:root:Failed to load package Stata_MLwiN (Stata not found)
WARNING:root:Failed to load package Stata_model (Stata not found)
WARNING:root:Failed to load package Stata_script (Stata not found)
WARNING:root:Failed to load package SuperMix (SuperMIX not found)
INFO:root:Trying to locate and open Firefox
ERROR:root:could not locate runnable browser
INFO:root: Trying to locate and open Chrome
DEBUG:root:could not locate runnable browser
INFO:root:Trying to locate and open default web browser
http://0.0.0.0:50215/

The most important command when starting up is the final line (the precise five-digit number written out towards the end of the line will likely differ, though). This only appears when the program has successfully performed all its initial set-up routines. This may take a while, particularly the first time you use the program. You should then be able to view the start page of TREE in your browser; if you can’t, then try refreshing the browser window, or typing localhost:50215 (in this example) into the address bar. The lines such as WARNING:root:Failed to load package GenStat model (GenStat not found) are not necessarily problematic but are warning you that the Genstat statistical package has not be found and loaded on your particular machine.
Quick-start guide

This section provides an overview ‘quick-start’ guide to using Stat-JR, via the TREE application; for more detailed instructions, together with worked point-and-click examples, see later sections. We’re assuming you’ve installed Stat-JR, and can see the opening page of the TREE application in your browser (see Section 2).

When operating Stat-JR through TREE, you generally proceed through the following five stages:

**Stage 1.** Firstly, choose the dataset you want to analyse / plot / summarise / etc., and the template you want to use to do so. Each template contains commands to perform certain functions: some run models, others plot graphs, or provide summary statistics, and so on...

**Stage 2.** You will be asked for further template-specific input: e.g. which variables from your dataset you would like to include in your model / which variables you would like to plot / summarise / etc.

**Stage 3.** Once you’ve answered all the input queries, Stat-JR generates all the commands, scripts, macros, equations, and instructions necessary to perform, or describe, the function you’ve requested. You can view these within TREE, and can download them too...

**Stage 4.** Stat-JR then runs these commands / scripts / macros, employing externally-authored software (e.g. R, MLwiN, WinBUGS, SPSS, Stata, etc.), or in-house software (such as the eSTAT engine), as appropriate.

**Stage 5.** Finally, the results are returned; depending on the template these can include model estimates, graphs, summary tables, and so on. Again, these can be viewed within TREE, and are also downloadable. The output may also include datasets (e.g. MCMC chains), which you can then feed back into the system by matching them up with a template back in Stage 1.
Below we briefly highlight the main features, with screenshots, of each of these five stages.

3.1 Stage 1: Selecting a template & dataset
- On opening Stat-JR, the page below, containing introductory information, will be displayed in a web browser. To proceed to choosing a template and dataset, click on the Begin button.
- Having pressed Begin, the page below will be displayed. Note that here, and on other screens, wherever you see the question mark symbols, context-specific help is revealed if you hover your cursor over them. Hover-over help can appear elsewhere too: e.g. describing the options along the top navigation bar.
- Here you can specify the template and dataset you want to use, and then begin to specify your inputs.
- Selecting Dataset > Choose or Template > Choose from the top bar will reveal lists of available datasets and templates. For each, find the one you want from the list, and then press Use.
- Note, when choosing a template, you can use the cloud terms to help your search: the blue tags describe functional aspects of the templates, whilst the red terms describe which engines / packages the templates support (you can combine search terms by clicking on more than one, and cancel your selections by pressing [reset]).

This link, available on all subsequent pages of TREE too, allows you to change settings such as paths to data and template folders, paths to interoperating software, and optimisation settings for generated code.

Click here to progress to the next screen where you can choose a template & dataset, and can start specifying your inputs...

If you have modified any files in the templates, datasets or packages folders, then you can reload their contents into the current session via the Debug menu here, which is also available on all subsequent pages of TREE.

A link to the Stat-JR webpages which contain further support, including frequently asked questions & a user forum.
Wherever you see these (question mark) symbols, context specific help is revealed if you hover your cursor over them. Hover-over help can appear elsewhere too: e.g. describing the options along the top navigation bar.

You can select one or more of these terms to help you find relevant templates; the blue tags describe the functional aspects of a template function, whilst the red terms describe the engines/packages supported by a template. To unselect terms, press [reset].

Here you can see which dataset and template are currently selected. Hovering your cursor over these names will reveal a description of each (if available).

Clicking on the down arrow symbol just to the right of the Dataset heading in the top bar will bring up a menu. Select Choose to bring up the window, below, allowing you to nominate a dataset other than that currently selected...

...and likewise for the Template...

Clicking the 'label' symbol brings up a list of tags, whilst the clicking the 'cog' symbol brings up a list of supported engines/packages.
3.2 Stage 2: Providing template-specific input

- Once your desired Dataset and Template is selected, you can start answering the input questions back on the main page. These are required by Stat-JR to allow the template to perform the appropriate executions with your dataset; these inputs vary between templates, and also within templates too, depending on your earlier choices as you progress through the screens.

- For multi-choice lists you can de-select variables by simply clicking on their name in the list of selected items.

- Press Next each time you’ve completed the input questions on the current page.

- Then, if applicable, more inputs will be revealed, and those you have already selected will be greyed-out. However, you can still change each input via the remove button which you’ll see next to each one. Alternatively, to re-specify all your inputs, press Start again (in the top bar).

- When asked for the Name of the output results, this will be the name given to any outputted dataset which results from running the template (see Stage 5).
For multi-select lists, you can de-select variables by clicking on their name here.

If, at any point, you want to re-specify all your inputs, then press Start again.

As you progress through the screens, you can see your choices reflected in the input string and the RunStatJR command, at the bottom; if you copy the former (up to & including the curly brackets), and paste in the box below and press Set, Stat-JR will automatically populate the inputs boxes with your choices; the latter (i.e. the RunStatJR command) can be used to call Stat-JR via a command line.

You can remove specific inputs via these buttons here.

Again, once you're happy with your inputs, press Next.
This is the name given to any outputted dataset (e.g. MCMC chains produced by the model run).

We've now completed all the inputs, and so we press Next for the final time...

(We've skipped a screen or two where we were asked about this input ...)}
3.3 Stage 3: Outputting the files to run the desired execution

- Once you’re pressed **Next** after the final input, Stat-JR returns a number of initial outputs which you can view in the output pane at the bottom of the window.

- Note that Stat-JR hasn’t done everything you want it to do yet: it’s just producing preliminary files telling you what it’s going to do, and how it’s going to do it.

- To select particular content to view in the output pane, use the drop-down menu just above it.

- The **Popout** button, just above the output pane, allows you to view its contents in a new browser tab.

- Pressing **Run** performs the executions described by the scripts, etc, returned in the output pane.

3.4 Stage 4: Running the execution

- Once you’re pressed **Run**, the executions specified by you are performed.

- Depending on your choices, this may take anything from a second or two (e.g. to produce a simple plot, fit a model using a non-iterative method of estimation, produce summary data, etc.), to many minutes (e.g. to run MCMC chains for a large number of iterations).

- If appropriate (e.g. if the template supports inter-operability, and if you have chosen to employ it when prompted), externally-authored software packages (e.g. R, MLwiN, WinBUGS, SPSS, etc.) are opened, run, then closed, and the results are returned to Stat-JR.

- Whilst the execution runs, you may see a lot of activity in the black command window, which can help you keep a track of progress.
You may see a lot of activity in the command window as the execution is performed...

Whilst it performs these executions, the progress gauge indicates that Stat-JR is still working...
3.5 Stage 5: The results

- Once the executions have run, the progress gauge, in the top-right corner, will change from “Working” to “Ready”, and the drop-down list, just above the output pane, will now be populated with more results.

- Depending on the template, a range of buttons / boxes appear above the output pane allowing you to e.g. Download the results, Add to ebook, and run chains for Extra iterations.

- If applicable, an outputted dataset now appears in the list of datasets (see Dataset > Choose, via the top bar).

This ends the quick start guide. In the next chapter we describe the operation of TREE in more detail, and work through examples.
4 A detailed guide with worked examples

4.1 The structure and layout of the *TREE* interface

Stat-JR can be thought of as a system that manages the use of a set of templates written either by the developers, and supplied with the software, or by users themselves. Each template will perform a specific function: for example, fitting a specific family of models, summarising a dataset, or plotting a graph. The Stat-JR system therefore allows the user to select and use specific templates with their datasets, and to capture and display the outputs that result.

Returning to our start-up of the software, when the line `http://0.0.0.0:50215/` appears, and after refreshing the web browser, the browser window should appear as follows:

![Start screen for Stat-JR interface](image)

This is the start screen for the *TREE* interface to Stat-JR, and contains information on funders, authors, and a link to the Stat-JR website which contains further guidance, such as answers to frequently asked questions, and a user forum.

Pressing **Begin** returns the following screen:

![Screen with explanatory variables](image)

At the top you’ll see a black **title bar**. From left to right, this contains:

- a link (**Stat-JR:TREE**) back to the opening page;
• an option (Start again) to clear all inputs the user has chosen for the current template;

• a Dataset menu allowing the user to choose, drop (from temporary memory cache), return summary statistics for the current dataset, view (the entire dataset; see below), return a list of datasets, upload / download (see Section 4.5) datasets. For example, selecting Dataset > Choose returns a scrollable list of all the datasets that the system is aware of: i.e. those which appear in the datasets subdirectory of this installation of Stat-JR. This pane can be used to change the selected dataset via the Use button; for inputting your own data set you can use the Upload button.

• the name of the currently-selected dataset (in the grey box) – if you hover your cursor over this name, it returns a textual description of the dataset;

• a Template menu allowing the user to choose, list (described below) or upload individual templates not already uploaded in the current session. If you select Template > Choose, a box appears which contains a scrollable list of all the templates that the system is aware of: i.e. those which appear in the templates subdirectory of this installation of Stat-JR. This can be used to change the selected template via the Use button. As we anticipate there being many templates, each template has defined ‘tags’ which are terms to describe what it does. These appear as blue phrases in the ‘cloud’ above the list of templates, whereas the estimation engines supported by each template appear in the cloud in red. When you select a template, its name and description appear to the right of the list. Clicking on the symbol that looks like a baggage label returns the tags for that template, whereas clicking on the ‘cog’ symbol returns a list of engines that particular template supports;

• the name of the currently-selected template (in the grey box) – again, if you hover your cursor over this name, it also returns a description of the template;

• a progress gauge indicating whether Stat-JR is “Ready”, “Working”, “Idle” or whether it has encountered an “Error”;

• a link to a page containing options to change a variety of Settings (discussed further below);

• a Debug button; this produces a drop-down list from which one can choose to reload the templates, datasets or packages, allowing users upload changes to files they make outside the TREE interface, without having to start-up Stat-JR again. For example, a user could paste a new dataset into the datasets directory, or modify a template in the templates directory, and reload them so that they appear in their lists in the browser window.

The Settings window, accessible via the black title bar, displays a number of settings that the program uses with each possible software package: some of these are relatively straightforward, such as where the executables for each package are found, and some are relatively advanced, such as for the eStat engine, optimisation, starting values and standalone code options.
We will now look at the **View dataset** window:

Select **Dataset > Choose** from the menu in the black **title bar**.

Scroll down the dataset list, towards the bottom, and click on **rats**; its name and description will appear to the right of the list.

Click on the **Use** button, and the name of the current dataset (in the grey box in the black title bar at the top) should have changed accordingly.

Select **Dataset > View**; this will open a new tab in your browser: if you click on this you will be able to see the dataset we have just selected, as follows:

![Image of the View dataset window](image)

The **rats** dataset is a small, longitudinal animal growth dataset which contains the weights of 30 laboratory rats on 5 weekly occasions from 8 days of age (see Gelfand et al (1990) for more details). The five measurements are labelled \( y_8, y_{15}, y_{22}, y_{29} \) and \( y_{36} \), respectively, and the dataset also contains a constant column – a set of 1’s, named **cons**, and a rat identifier column, **rat**. Initially, we are going to perform a regression analysis of the initial weight (\( y_8 \)) on the final weight (\( y_{36} \)), including an intercept (**cons**). The boxes above the data allow the user to quickly add a new variable or delete an existing variable from the dataset. We can also view a summary of the dataset:

To view a summary of the dataset, click on the first tab in the browser window containing the main menu screen.

Select **Dataset > Summary** button via the black title bar at the top.

Click on the new tab and the screen will look as follows:
Here we get a very short summary of the dataset, giving, for each variable, the minimum value, maximum value, mean and standard deviation. If the dataset has had descriptions added or has categorical variables then they will appear in the last two columns. More extensive summaries are available by using specific templates to summarise datasets, as we will see later.

Let’s now look at the Template menu:

Back on the main page, if you click on Template > List the following screen will appear in a new tab:

This rather busy screen (we don’t reproduce it all here, due to its length) contains, in the two columns on the left, a tabular list of all the templates that are available with a short description of what each template does. The other three columns are of more interest to advanced users, but contain a list of functions in the template code, tags that identify the template type, and the engines that are supported by the template.

We will next demonstrate running a template, using the default Regression1 template that fits a 1-level Normal response regression model, this is appropriate as the response, the weights of the rats, is a continuous measure.

Return to the main menu screen, which should look as follows:
In the middle of the screen you can see the inputs required for this template (these are template-specific, and may change when you use a different template). Since some inputs are conditional (i.e. are only required when earlier inputs take specific values), the opportunity to specify inputs proceeds through sequential steps. Here we see the two initially-required inputs for the Regression1 template are the Response variable and Explanatory variables. Since this template only allows for 1 response variable to be specified, a pull-down list is displayed, but since it allows for several explanatory variables to be specified, a multiple selection list is displayed for that input. In the case of the latter, variables are selected by clicking on their name in the left-hand list; to de-select them, click on their name in the right-hand list.

The Start again link (in the top black bar) will clear any inputs the user has already selected and return you to the first template input screen (i.e. the current screen, in this case), whilst the Next button will allow the user to move on and specify further inputs once those on the current screen have all been chosen.

Use the input controls and the Next button(s) to fill in the screen as follows:
Note that an option to **remove** appears next to each input previously submitted; this will remove the current input, but keep the other inputs you have specified (as far as it can; if they are conditional on the input you have removed, then they will be, out of necessity, removed too).

So, here we are performing a regression of the initial weight \( (y_8) \) on the final weight \( (y_{36}) \), including an intercept \( (\text{cons}) \).

The other inputs refer to the Monte Carlo Markov chain (MCMC) estimation procedures in Stat-JR. MCMC estimation methods are simulation-based, and so require certain parameters to be set. The methods involve taking a series of random (dependent) draws from the posterior distribution of the model parameters in order to summarise each parameter. The inputs required here are as follows:

- **the number of chains**: this is the number of starting points from which we will take random draws;
- **random seed**: the value from which random numbers are initially drawn. This allows repeatability, as a run using the same starting values and random seed will give the same answers. When multiple chains are used this seed is generally multiplied by the chain number to give a unique seed for each chain;
- **length of the burnin**: the initial length of the chain (i.e. the number of iterations at the start) which are excluded from the parameter summaries (the rationale for this is explained a little further in the example, below, with the tutorial dataset);
- **number of iterations**: the length of chain following the burnin, from which the parameter summaries are drawn;
- **thinning**: this determines how often the values are stored: i.e. store every \( n \)th iteration.
By answering ‘Yes’ to the question *use default algorithm settings*, we have used defaults for other settings for which we will therefore not be prompted to complete. By answering ‘No’ to *generate prediction dataset* we have chosen not to generate a dataset of predictions from our model. By answering ‘Yes’ to *use default starting values* we have chosen not to start the chain at values of our choosing, instead accepting Stat-JR’s defaults. We will discuss MCMC estimation in slightly more detail in the applications in the next section. The final input we’re asked for is the *name of output results*: this is the name (here we’ve chosen *out*) given to any dataset of parameter chains that results from running the template.

You will notice, towards the bottom of the window, a box for an **Input string** to be entered, next to a **Set** button, with a rather long text string above it and another labelled as **Command** below it. The input string allows the user to specify all the inputs directly, without having to point-and-click through the list as we have done. These have to be formatted in a certain way, however, as illustrated by the current (*Input String*) text string which Stat-JR has written for us as a result of our inputs. This (i.e. the string between, and including, the curly brackets: in this example `{‘burnin’:’1000’... ‘makepred’: ‘No’}`) can be copied and pasted into the box, and the Set button pressed (following any edits you would like to make to the input values), in order to specify inputs directly. The second text string (labelled **Command**) can be used by the command driven version of Stat-JR to perform the same operations, although we will not discuss this further here.

**Clicking on the Next button** will now pre-process the template inputs, and will result in the following new pane at the bottom of the window:

![Image](image-url)

The object currently specified in the pull-down list (*equation.txt* is selected by default here) appears in the pane below it. These objects are any outputs constructed by Stat-JR before and during the execution of the template, so here we see a nice mathematical description of the model. If we now select the object *model.txt* from the list we see a description of the regression model that we wish to fit in the language that is used by the eStat engine:
At this point we haven’t actually run the template, and so the objects that can be selected from the pull down list are those present pre-model run, and include computer code to actually fit the model.

Click the **Run** button to run the template.

Once the progress gauge, towards the right of the black title bar, has changed from “Working” (blue) to “Ready” (green), select **ModelResults** from the pull-down list.

The screen will then look as follows:
Here we see parameter estimates, along with standard deviations (SDs), as a measure of precision for each parameter. We will explain these further in the next section. At the top of the screen shot above (which is in fact the middle of the full window, vertically-speaking) we now have a few additional buttons. The Extra Iterations box, along with the More button, will allow us to run for longer (i.e. for a number of iterations additional to those we have already run for). The Download button will produce a zipped file that contains a folder with files for many of the objects contained in the two pull-down lists whilst the Add to ebook button can be used if one wants to construct an ebook to be used with the DEEP eBook interface into Stat-JR.

You’ll recall that we earlier named the output results ‘out’, so if we choose this from the pull-down list just above the output pane, we’ll be able to view it, as follows:

Here we see columns containing the chains of values for each parameter in the model. As well as being able to view this file here, it is also a dataset (stored in temporary memory) and so will appear in the dataset list (at least for the duration of our current session using the software) accessible via the Dataset menu in the top title bar. This means that we can string templates together, as we can select out as a dataset and perform operations on it using another template.

This ends our whistle-stop tour of many of the windows in Stat-JR. We will next look at a practical application.
4.2 Application 1: Analysis of the tutorial dataset using the eStat engine

4.2.1 Summarising the dataset and graphs

In this section we will look at performing some analysis of an example dataset from education. The dataset in question is known as the tutorial dataset, and is used as an example in the MLwiN software manuals (see, for example, Browne 2012). In fact, much of the material here owes a lot to Browne (2012), which employs similar analysis but using MLwiN.

Let us start by looking at the tutorial dataset.

Select tutorial via Dataset > Choose (see the title bar), then click Use.

If you then select Dataset > Summary, a new tab should appear in the browser window containing summary information, as follows:

The tutorial dataset contains data on exam scores of 4059 secondary school children from 65 schools at age 16. These exam scores have been normalised to have a mean of zero and a standard deviation of one and are named normexam. There are several predictor variables, including a (standardised) reading test (standlrt) taken at age 11, the pupils gender (girl), and the school’s gender (schgend) which takes values 1 for mixed, 2 for boys and 3 for girls. Each variable is described in the Description column and if you hover over any of the variables that say “Yes” in the value labels column, the category labels will be displayed.

We can explore the dataset in more detail, prior to fitting any models, by using the many data manipulation templates available in Stat-JR. We will first look at some plots of the data:

Select Template > Choose and then select Histogram from the template list that appears and click Use.
Fill in the inputs as shown below and click **Next** and then **Run** and select *histogram.svg* from the output list.

Here you will see, in the output pane, a histogram plot that shows that the response variable we will model, *normexam*, appears Normally-distributed.

**Select Template > Choose** and this time select **XYPlot** from the template list, then click **Use**.

Fill in the inputs as shown below and click **Next** and then **Run** and select *graphxy.svg* from the list.
Here we see that there appears to be a positive relationship between *normexam* and *standlrt*, with pupils that have higher intake scores performing better, on average, at age 16.

We can display the graph in a separate tab in the browser window by clicking on the **Popout** button next to the pull down list:
For the sake of brevity, for the remainder of this documentation we will assume you now know how to change template/dataset, and also how to display output in separate tabs, so we’ll refrain from repeating this information in detail again.

Next, we might like to examine how correlated the two variables, *normexam* and *standlrt*, actually are:

Select **Average and Correlation** as the template, and complete the inputs as follows before clicking on **Next** and **Run** and selecting **table** from the outputs:
Here we see that the correlation is 0.59, so fairly strong and positive. We might also like to look at how exam score varies by gender:

**Select Tabulate as the template, and complete the inputs as follows, before clicking on Next and Run and selecting table from the output list:**

![Tabulate template](image)

We have to enter variables for column values and row values, and so here we have specified column values as *girl* (taking value 1 for girls and 0 for boys) and row values as *cons* (which is a constant), and then we get 2 columns in the output labelled 0 and 1 for boys and girls, respectively. Looking at the means, it appears that girls do slightly better than boys, and looking at the standard deviations (sd's) they are slightly less variable than boys in their scores. Let us now consider performing some statistical modelling on the dataset.

### 4.2.2 Single-level Regression

As in the last chapter, with the rats dataset, we will start by fitting a simple linear regression model to the tutorial dataset. Here we will regress normexam on standlrt by using a modelling template.

**Select Regression1 as the template and fill it in as follows:**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Here we are fitting a linear regression, and so have `standlrt` as an explanatory variable, but also `cons` (which is a column of 1s) as we would like to include an intercept as well. For now we have set-up the MCMC estimation options as we did for the `rats` dataset, and we will overwrite the output file `out`.

Clicking on the **Next** button will populate a pull down list of objects created by Stat-JR at the bottom of the screen and by default we see the object `equation.tex`:

In the pane we find a mathematical representation of the chosen model. Note that the file is a LaTeX file that is being rendered in the browser by a piece of software called MathJaX (v2.3, 2013), so if you are a LaTeX-user you can copy this file straight into a document. If we instead choose `model.txt` from the list we see the following:
Here we see the text file that represents the model we wish to fit in the language that the algebra system used by the built-in eStat engine requires. The Regression1 template only uses the eStat MCMC-based estimation engine, so as you can see in the mathematical formulae in equation.tex we are fitting a Bayesian version of a linear regression, and the last four lines of the output are prior distributions for the unknown parameters, $\beta_0$, $\beta_1$ and the precision $\tau$ (where $\tau=1/\sigma^2$).

Whilst we will keep our description of Bayesian statistics and MCMC estimation to a minimum, and recommend Chapter 1 of Browne (2012) for more details, in brief we are interested in the joint posterior distribution of all unknown parameters given the data (and the prior distributions specified). In practice, in complex models, this distribution has many dimensions (in our simple regression we have 3 dimensions) and is hard to evaluate analytically. Instead, MCMC algorithms work by simulating random draws from a series of conditional posterior distributions (which can be evaluated). It can then be shown (by some mathematics) that after a period of time (required for the simulations to move from their possibly arbitrary starting point) that the draws will be a dependent sample from the joint posterior distribution of interest. It is common, therefore, to throw away the first $n$ draws which are deemed a burn-in period.

For the simple linear regression, it is a mathematical exercise to show that the conditional posterior distributions have standard forms and are Normal (for the fixed effect) and Gamma (for the precision $= 1/\text{variance}$). The eStat engine has a built in algebra system which takes the text file (model.txt) in the left-hand pane and returns the conditional posterior distributions; you can view these as follows:

Select algorithm.tex from the list and click on the Popout button and the algebra steps will appear in a new tab as follows:
The eStat engine then takes these posterior distributions and wraps them up into computer code (C++) which it will compile and run for the model. By default this will be several pieces of code that are linked together by Stat-JR, although the Settings screen (accessible via a link towards the top of the main menu screen, as we saw earlier) has an option to output completely standalone code that can be taken away and run separately from the Stat-JR system; this is, however, a topic for more advanced users.

Returning to the tab, in the browser window, containing the model template, click on the Run button and wait for the model to run.

Then select ModelResults from the pull down list and pop it out into a separate tab.
Here the model results can be split into two parts:

The first part of the results (under the heading ‘Parameters’) contains the actual parameter estimates. Here, for each parameter, we get 3 numbers: a posterior mean estimate (mean), a posterior standard deviation (sd), and an effective sample size (ESS).

Here we see that beta_0 has a mean estimate of approximately 0, which we would expect as both the response and predictor have been normalised, or standardised. The slope beta_1 has mean 0.595 with standard deviation 0.013, and is highly significant, as it’s mean estimate is many times it’s standard deviation (a Bayesian equivalent of a standard error) The value 0.595 represents the average increase in the normexam score for a 1-point (1 sd, due to standardising) increase in standlrt. The residual variance, sigma2, has value 0.649 showing that, as the initial response variance was 1.0, standlrt has explained 35.1% of the variability.

The ESS is a diagnostic which reflects the simulation-based (stochastic) nature of the MCMC estimation procedure: we have based our results on the 5,000 iterations post burn-in, but we know that the method produces dependent samples, and so the ESS gives an equivalent number of independent samples for the parameters involved; in effect a measure of the information content of the chain. In this case, all parameters have ESS of > 4000, and so the chains are almost independent.

The second part (under the heading ‘Model’) refers to the model fit for this particular model and the DIC diagnostic (Spiegelhalter et al. 2002). The DIC diagnostic is an information criterion which is a measure of how good a specific model is, consisting of a combination of how well the model fits the

---

Here the model results can be split into two parts:

The first part of the results (under the heading ‘Parameters’) contains the actual parameter estimates. Here, for each parameter, we get 3 numbers: a posterior mean estimate (mean), a posterior standard deviation (sd), and an effective sample size (ESS).

Here we see that beta_0 has a mean estimate of approximately 0, which we would expect as both the response and predictor have been normalised, or standardised. The slope beta_1 has mean 0.595 with standard deviation 0.013, and is highly significant, as it’s mean estimate is many times it’s standard deviation (a Bayesian equivalent of a standard error) The value 0.595 represents the average increase in the normexam score for a 1-point (1 sd, due to standardising) increase in standlrt. The residual variance, sigma2, has value 0.649 showing that, as the initial response variance was 1.0, standlrt has explained 35.1% of the variability.

The ESS is a diagnostic which reflects the simulation-based (stochastic) nature of the MCMC estimation procedure: we have based our results on the 5,000 iterations post burn-in, but we know that the method produces dependent samples, and so the ESS gives an equivalent number of independent samples for the parameters involved; in effect a measure of the information content of the chain. In this case, all parameters have ESS of > 4000, and so the chains are almost independent.

The second part (under the heading ‘Model’) refers to the model fit for this particular model and the DIC diagnostic (Spiegelhalter et al. 2002). The DIC diagnostic is an information criterion which is a measure of how good a specific model is, consisting of a combination of how well the model fits the
data (here defined by the model deviance) and how complex the model is (here defined by pD: the effective number of parameters). Basically the better fitting the model is, the better the model is, but it has to be penalised by how complex it is. The DIC statistic is defined as the deviance of the mean + 2pD. In this example the deviance at the mean (D(\theta_bar)) is 9760.5 and pD is \approx 3 (reflecting the three parameters of the model that are being estimated) and so we have a DIC value of 9766.4. This number is not particularly interesting in isolation but it is when we compare values for several models.

We can also get more information from the diagnostic plots that are available in the list of objects

Return to the model run tab in the browser window, and select beta_1.svg from the pull-down list above the output pane and view in a separate tab.
This “sixway” plot gives several graphs that are constructed from the chain of 5,000 values produced for $\beta_1$. The top-left graph shows the values plotted against iteration number, and is useful to confirm that the chain is ‘mixing well’, meaning that it visits most of the posterior distribution in few iterations. The top-right graph contains a kernel density plot which is like a smoothed histogram and represents the posterior distribution for this parameter. Here the shape is symmetric and looks like a Normal distribution which we expect given theory for fixed effects in a normal model.

The two graphs in the middle row are time series plots known as the autocorrelation (ACF) and partial autocorrelation (PACF) functions. The ACF indicates the level of correlation within the chain; this is calculated by moving the chain by a number of iterations (called the lag) and looking at the correlation between this shifted chain and the original. In this case, the autocorrelation is very small for all lags. The PACF picks up the degree of auto-regression in the chain. By definition a Markov chain should act like an autoregressive process of order 1, as the Markov definition is that the future state of the chain is independent of all the past states of the chain given the current value. If, for example, in reality the chain had additional dependence on the past 2 values, then we would see a significant PACF at lag 2. In this case all PACF values are negligible. All of this suggests that we have good mixing and it would be appropriate to proceed to the interpretation of the parameters.

The bottom-left plot is the estimated Monte Carlo standard error (MCSE) plot for the posterior estimate of the mean. As MCMC is a simulation-based approach this induces (Monte Carlo) uncertainty due to the random numbers it uses. This uncertainty reduces with more iterations, and is measured by the MCSE, and so this graph details how long the chain needs to be run to achieve a specific MCSE. The sixth (bottom-right) plot is a multiple chains diagnostic and doesn’t make much sense when we have run only one chain, and we will therefore consider running multiple chains in the next section.

We can also get some other diagnostics and summary statistics for the model as follows:

Click on the Template pull down list at the top of the screen and select Choose and SummaryStats as the template.

Next click on the Dataset pull down list and select Choose and out as the dataset.

Run the SummaryStats template and select the inputs as follows before clicking on Run:

Now select table from the drop-down list of outputs, and display it in a separate tab:
Here we see a more extensive summary of the three parameters of interest. This summary table includes various quantiles of the distribution which are calculated by sorting the chain and picking the values that lie x% into the sorted chain (where x is 2.5, 5, 50 etc.). These allow for accurate interval estimates that do not rely on a Normal distribution assumption. The inter-quartile range (IQR) is similarly calculated by picking the values that lie 25% and 75% through the sorted list and calculating the distance between them.

The final statistic is an MCMC diagnostic designed to suggest a length of chain to be run. The Brooks-Draper diagnostic is based on measuring the mean estimate to a particular accuracy (number of significant figures set to 2 by default). For example, it states that to quote \( \sigma^2 \) as 0.65 with some desired accuracy only requires 32 iterations. The anomaly here is \( \beta_0 \), however, since the true value is 0 we have difficulty quoting such a value to 2 significant figures!

### 4.2.3 Multiple chains

MCMC methods are more complicated to deal with than classical methods as we have to specify many estimation parameters, including how long to run the MCMC chains for. The idea of running chains for a longer period is to counteract the fact that the chains are serially-correlated, and therefore are not independent samples from the distribution. Another issue that might cause problems is that the posterior distribution of interest may have several possible maxima (i.e. may be multimodal). This is not usually an issue in the models we cover in this book, but it is still a good idea to start off the estimation procedure from several places, or with several runs with different random number seeds, to confirm we get the same answers.

From the top bar change Template and Dataset using the respective pull down lists and Choose so you have Regression1 as the template and tutorial as the dataset.

This time fill in the screen as follows:
Click on the **Next** and **Run** buttons.

When the model has run select *beta_1.svg* from the outputs list and pop it out to view it in a new tab.

Here we see the three chains superimposed on each other in the top-left pane – note the chain looks primarily red simply because this chain (chain 3) has been plotted on top of the other two, and due
to good mixing obscures them. Each chain has its own kernel plot in the top-right pane and this also suggests that, by the similarity of shape and position, the chains are mixing well.

We have previously described what all the graphs here mean in Section 4.2.2, apart from the Brooks-Gelman-Rubin diagnostic plot (BGRD; Brooks and Gelman, 1998) in the bottom-right corner. This plot looks at mixing across the chains: the green and blue lines measure variability between and within the chains, and the red is their ratio. For good convergence this red line should be close to 1.0, and here the values get close to 1.0 fairly quickly. We can have a lot of faith in the estimates of our model.

### 4.2.4 Adding gender to the model

We have so far been more focused on understanding the MCMC methods but now we will return to modelling. We next wish to look at whether gender has an additional effect on normexam on top of that we have observed for intake score (standlrt).

To do this, click on the remove link next to explanatory variables in the browser window, and fill-in the template as follows:

Click on **Next** and then **Run** to run the model.

After the model finishes running select **ModelResults** from the drop-down list of outputs, and display in a new tab.
This new model has one additional fixed effect parameter (beta_2) associated with gender, and we see it has a positive effect (0.170) which appears highly-significant (at least twice its sd, which is 0.025). Note that in our earlier tabulation we saw that the difference in gender means was 0.093 - (-0.140) = 0.233 and so the effect here is somewhat smaller, probably due to correlation between gender and intake score.

Looking at the DIC diagnostic to assess whether this model is better we see this has dropped from 9766.4 to 9724.9, which is a big drop, and so the model with gender is indeed much better.

Finally we see that the ESS for two of the parameters is lower (beta_0 and beta_2), at around 1600, so the model doesn’t mix quite as well; however, these ESS are still large enough not to require further iterations. Here is the graph for **beta_2.svg**, displayed in a new tab:
We see reasonable mixing, and can clearly see the significance of the effect as well (as the kernel density plot in the top-right corner indicates that 0 is nowhere near the posterior distribution). From a modelling perspective we have thus far ignored the fact that our data is a two-stage sample and that we should account for the clustering of the pupils within secondary schools. To do this we need to fit a 2-level model, and use a different template.

### 4.2.5 Including school effects

Stat-JR contains many different model-fitting templates some of which can fit whole families of models and some of which can fit just one or two specific models. We have thus far looked at the rather restrictive Regression1 template that only fits single level normal response models. To include school effects we will now look at the 2LevelMod template, which not only includes a set of random effects but also supports different response types and estimation engines, features that we will look at later.

On the Template pull down list at the top of the screen select Choose and select 2LevelMod as the template and stick with tutorial for the dataset.

Set-up the inputs as shown below:
Press **Next** and then **Run** to fit the model. Note that running will take a while as we are storing all 65 school effects and so for each one the software needs to construct diagnostic plots.

When the model finishes select **ModelResults**, from the output list, and show the results in a separate tab.
Here if you scroll down we see that the DIC value for the two-level model is 9245, compared with 9725 for the simpler model, showing that it is important to account for the two levels in the data. If you scroll down to the beta fixed effect parameters, as shown in the table below, you will find that their mean estimates have changed little.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Single level</th>
<th>Single level</th>
<th>2level</th>
<th>2level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean(sd)</td>
<td>ESS</td>
<td>Mean(sd)</td>
<td>ESS</td>
</tr>
<tr>
<td>beta_0</td>
<td>-0.103 (0.0196)</td>
<td>1615</td>
<td>-0.091 (0.0429)</td>
<td>320</td>
</tr>
<tr>
<td>beta_1</td>
<td>0.590 (0.0126)</td>
<td>5482</td>
<td>0.560 (0.0126)</td>
<td>4956</td>
</tr>
<tr>
<td>beta_2</td>
<td>0.170 (0.0254)</td>
<td>1623</td>
<td>0.170 (0.0330)</td>
<td>776</td>
</tr>
</tbody>
</table>

The standard deviations for $\beta_0$ and $\beta_2$ have increased due to taking account of the clustering, and the ESS values have reduced due to correlation in estimating the fixed effects and level 2 residuals.

4.2.6 Caterpillar plot
The random effects in the 2-level model are also interesting to look at, and one graph that is often used is a caterpillar plot. This can be produced in Stat-JR using a template specifically designed for producing this plot. This template requires the user to select all the ‘u’s to be displayed in the plot, which can be time-consuming if there are many of them:
From the top bar we need to select **Choose** for **Template** and **Dataset**.

Choose **CaterpillarPlot95** as the template and **out2level** as the dataset.

You need now to select all the u’s from $u0$ to $u64$ which is best done by clicking on $u0$ and holding down the mouse and scrolling down to multiselect all the u’s together.

Once all are selected press the **Run** button.

Select **caterpillar.svg** in the pull down list and view in a new tab as follows:

This graph shows the schools in order of ascending mean whilst the bars give a 95% confidence interval around each mean. The school in the middle with the wide confidence interval (i.e. very large bars) has only 2 pupils and so there is much greater uncertainty in the estimate.

In this chapter we have explored fitting three models to the tutorial dataset. This has illustrated how the Stat-JR system works, how to interpret the output from MCMC and eStat, and how to compare models via the DIC diagnostic. There are better models that can be fitted to the dataset: for example, we could look at treating the effect of intake score ($standlrt$) as random, and fit a random slopes model using the template **2LevelRS**; in the future we may add material on this subject to this manual, but for now we leave this as an exercise for the reader. Next we turn to the interoperability features of Stat-JR.
4.3 Interoperability – a brief introduction

In this section we look at interoperability with other software packages. In order to run this section you will need to have installed the other packages and told Stat-JR where they are. For more details look at the Stat-JR website (www.bristol.ac.uk/cmm/software/statjr/).

4.3.1 So why are we offering Interoperability?
There are many motivations that could be given for the benefits of having an interoperability interface. First and foremost it opens up functionality in other software packages through a common interface.

One important feature that the template, Regression1AML, which we cover at the end of this chapter, shows is that not all model templates need to use the built-in eStat engine. It would be perfectly reasonable for a user to construct a template that fitted a specific family of models in the WinBUGS software and then novice users would have access to a user-friendly interface to such models without having to understand the subtleties of writing WinBUGS code; it can thus play an important role introducing packages, such as WinBUGS, to new users. This follows earlier work: for example the MLwiN-WinBUGS interface that we developed 10 years ago.

It also offers an easy way of comparing different software packages for a multitude of examples, and we will return to this in Section 4.4.4. Finally it can be thought of as a teaching tool, so that a user familiar with one package can use Stat-JR and directly compare the script files, etc., required for the package with which they are familiar to those required for an alternative package.

4.3.2 Regression in eStat revisited
In Section 4.2 we looked at fitting a few models to the tutorial dataset using the built-in eStat engine: a newly-developed estimation engine with the advantage of being transparent in that all the algebra, and even the program code, is available for inspection. It is an MCMC-based estimation method, but is also rather quick. In this chapter we will stick with one simple example, the initial linear regression model that we fitted to the ‘tutorial’ dataset that we considered in Section 4.2. We will need to use a new template, Regression2, as the Regression1 template only supports the eStat engine.

We will begin by setting-up the model and running it in eStat:

From the top bar select Regression2 as the template, and tutorial as the dataset using the Choose options on the pull down lists for templates and datasets and set-up the inputs as follows:
Click on **Next** and **Run** to fit the model.

Select **ModelResults** from the pull down list, and show this output in a new tab which should look as follows:

These results are identical to those we obtained using **Regression1** earlier, although we only looked at the plot for $\beta_1$ in Section 4.2.3. We will use this as a benchmark, contrasting these results with those we obtain from the other packages, although it is worth noting that all packages will have
good mixing and converge quickly for this simple linear regression model. You might like to explore differences between engines / packages for other models yourself after reading this chapter.

4.3.3 Interoperability with WinBUGS
WinBUGS (Lunn et al., 2000) is an MCMC-based package developed (as BUGS – Bayesian inference Using Gibbs Sampling) originally in the early 1990s by a team of researchers at the MRC Biostatistics Unit in Cambridge. It is a very flexible package and can fit, in a Bayesian framework, most statistical models, provided you can describe them in its model specification language. In Stat-JR we have borrowed much of this language for our own algebra system, and so many templates feature interoperability with WinBUGS.

To fit the current model using WinBUGS we can click on remove next to the Choose estimation engine input and set up the template inputs as follows:

When we press Next the Stat-JR software will construct all the files required to run WinBUGS so for example we can choose model.txt from the list:
Here we see the model defined in the WinBUGS model specification language in the output pane. This file is almost identical to that used by eStat aside from the expression \( \text{length}(	ext{normexam}) \) being replaced here by its value 4059. Selecting \textit{script.txt} from the list and popping out to a new tab gives the following:

Here we see a list of the commands to be run in the WinBUGS language to fit the model. Note that this is done using a temporary directory and so this pathname appears in many commands.

Return to the tab containing the main page and click on the Run button.
The WinBUGS package then pops up in its own window, runs the above script, and returns control to Stat-JR when it has finished estimating the model. If we look at the `ModelResults` output from the list and pop it out to its own tab we will see the following:

![Stat-JR: Model Results](image)

These estimates, as one might expect, are very close to those from eStat, and again all ESS values are around 5,000-6,000. We can also look at the log file from WinBUGS:

Return to the template tab and choose `log.txt` in the outputs list.

Scroll the `log.txt` file down to the bottom, and the screen should look as follows:
Here we see that the estimates and the DIC diagnostic are embedded in the log file, and take a similar value to eStat. WinBUGS required initial value files for each run (and these are stored in 3 text files beginning with *inits* and the chain number), together with a data file as well as the model and script files already shown. All of these are available to view and to use again, thus Stat-JR is useful for learning how these other packages, such as WinBUGS, work.

### 4.3.4 Interoperability with OpenBUGS

Our next package to consider is OpenBUGS (Lunn et al., 2009). OpenBUGS was developed by members of the same team who developed WinBUGS, but differs in that it is open source so other coders may get access to the source code, and in theory develop new features in the software.

To run OpenBUGS via Stat-JR click on the word *remove* next to the **Choose Estimation engine** input, set up the template as follows, and then click on **Next**:
This will have set-up the files required for OpenBUGS; these are similar, but not identical, to WinBUGS: the script file, in particular, is somewhat different and is split into 3 parts called *initscript.txt*, *runscript.txt* (shown below) and *resultscript.txt*, (you can access this from the objects list):

OpenBUGS allows us to change the working directory, and so there is no need for other commands to include the temporary directory path. Unlike WinBUGS, OpenBUGS will run in the background, and so will not appear when we click run.
Clicking on Run and selecting ModelResults in its own tab gives the following:

Again, these results are very similar in terms of parameter estimates and ESS values to the other software packages.

4.3.5 Interoperability with JAGS
The third standalone MCMC estimation engine available, via Stat-JR, is JAGS (Just Another Gibbs Sampler), developed by Martyn Plummer (Plummer, 2003). JAGS also uses WinBUGS model language, but has a few differences in terms of script files and data files.

To run JAGS via Stat-JR click on the remove text next to Choose estimation engine and set-up the template as follows, before clicking on Next:
This will set-up the files required for JAGS; for example, here you can see the script file (script.txt) which show some differences to those for WinBUGS (as to the initial value file formats):

Like OpenBUGS, JAGS will run in the background (i.e. it will not open as a window on your screen).

Clicking on Run, and placing ModelResults in a new tab, gives the following:
As you can see, we have similar estimates and effective sample sizes to the other estimation methods we’ve used. Whilst JAGS can be faster than WinBUGS and OpenBUGS, it fits a slightly smaller number of models.

4.3.6 Interoperability with MLwiN
MLwiN (Rasbash et al. 2009) is a software package specifically written to fit multilevel statistical models. It features two estimation engines (for MCMC and likelihood-based (IGLS) methods, respectively) with a menu-driven, point-and-click user interface. It also has an underlying macro language, however, and this is what we use to interoperate with Stat-JR. We will first consider the MCMC engine. As it is limited in the scope of models it fits, this means it is generally quicker than the other MCMC packages. MLwiN is a single chain program, but can be made into a multiple chain engine with Stat-JR, since the latter can start-up three separate instances of MLwiN. At present these are given different random number seeds, but the same starting values, however we will try and change this in future.

To run MCMC in MLwiN, via Stat-JR, click on the **remove text by Choose estimation engine** input and set-up the template as follows before clicking on **Next** :
You can see, in the pulldown list the dataset (in .dta format) that is used by MLwiN. There are also several MLwiN script files for the multiple chains and the several stages of model fitting.

Clicking on the **Run** button will set off three instances of MLwiN (in the background) and Stat-JR will then collate the results together. Choosing **ModelResults**, and displaying them in a new tab, gives the following:

Once again here we have similar estimates, although the naming convention is slightly different for MLwiN. To show that we have multiple chains we can examine the chains for the slope (**beta3**), as shown below:
Stat-JR also offers the option of using the likelihood-based IGLS estimation engine in MLwiN.

To do this in MLwiN, via Stat-JR, click once again on the remove text next to the Choose estimation engine input and set-up the template as follows, before clicking on Next:

Again the dataset will appears in the output pane, and this time pressing Run will give the following in the ModelResults output:
Here we get the Deviance (-2*Loglikelihood) value, together with parameter estimates with standard errors. The likelihood-based methods are far faster to run than the MCMC-based methods.

4.3.7 Interoperability with R

R (R Development Core Team, 2011) is another more general purpose package that can be used to fit many statistical models. R has many parallels with Stat-JR in that users can supply functions (like Stat-JR templates) which are then added to the library of R packages. We have thus far implemented interoperability features with R for several of these R functions; for example, for the template Regression2, we have implemented two R engines: R_MCMCglmm, which is MCMC-based, and R_glm, which is a standard regression modelling function. We will firstly demonstrate MCMCglmm.

To run MCMC in R, via Stat-JR, click on the remove text by the Choose estimation engine input and set-up the template as follows, and click on Next:
After pressing **Next**, if we look at the script file, *script.R*, which we can select from the outputs list, we see the following:

---

**MCMCglmm** can fit all forms of generalised linear mixed models, of which a linear regression is a rather trivial case. You will see that the script file contains some setup code which will actually
download and install the \textit{MCMCglmm} library the first time you execute the script (so ensure your machine is connected to the internet) before calling the \textit{MCMCglmm} command and then producing summaries.

Clicking on \textbf{Run} in the main window will create several outputs.

The \textit{ModelResults} are similar to other software but we can also look at diagnostics plots that are specific to R by selecting \textit{DiagPlots1.png}:

![DiagPlots1.png](image)

Here R gives trace plots and kernel density plots for both the intercept and the slope parameter.

Turning next to the \textit{glm} package we can click on the \textbf{remove} text by \textbf{Choose estimation engine} and set-up the template as follows, before clicking on \textbf{Next}:

![Choose estimation engine](image)
Clicking on Run will this time run the MASS package and give results in ModelResults as usual. There are additional graphical plots that come back from R; for example, below is a plot of residuals of the model fit against fitted values (ResivsFitted.svg).

Before finishing with R, we will also demonstrate a non-model template developed with R called PlotsViaR that gives the Stat-JR user access to R’s lattice graphics package through the Stat-JR interface.

Click on Choose from the Template pull down list at the top of the screen to get a list of all the templates. Note that the search cloud is useful with interoperability as it can be used to show which templates offer interoperability with a particular package (the engines are in red).

Click on Plots and also R_script in the blue tag cloud. You’ll see that the list of templates, underneath, is accordingly reduced to just those that draw plots using R.

Select PlotsViaR from the list, and click Use.

Set up the template inputs as shown below:
These options will display kernel plots for the exam scores of pupils grouped by gender, with separate (panelled or trellise) plots for each school gender type. We can now press Run and show the plot (Plot1.svg) in a separate tab:

Here (by coincidence) we have blue for boys and pink for girls!
4.3.8 Interoperability with AML

We will next look at another software package that can fit many statistical models via likelihood-based estimation. AML (Lillard & Panis, 2003) is very useful for fitting multi-process models, but as with other software packages can fit a simple regression as a special case. In our development work on Stat-JR we have written special templates for interoperability with AML as opposed to incorporating interoperability in the standard templates. We therefore need to do the following:

Click on the Choose option from the Template pull down list.
Select Regression1AML from the template list and click on Use, and stick with the tutorial dataset.
Note that if you have earlier clicked on Plots and R_script in the cloud of terms you will need to either unselect them or click on [reset] to see the required template.
Fill in the inputs as follows, and press Next:

Now click on Run to run the model in AML and select ModelResults from the list:

Here we see the model results are similar to other packages. AML has three input datasets (amlfit.raw, amlfit.aml and amlfit.r2a). There are also three additional output files from AML:
amlfit.out, amlfit.tab and amlfit.sum. For more information on how AML works we recommend looking at the reference manual for the software.

We also have interoperability support for a variety of other packages, including GenStat, MATLAB, Minitab, Octave, Sabre, SAS, SPSS, Stan (via RStan) and Stata. These packages are either not installed on the machine we are currently using, or are not supported by the Regression2 template that is being demonstrated, and so Stat-JR realises this and does not offer them.

4.4 Application 2: Analysis of the Bangladeshi Fertility Survey dataset

4.4.1 The Bangladeshi Fertility Survey dataset

The Bangladeshi dataset (bang1) is an example dataset from the 1988 Bangladeshi Fertility Survey. It contains records from 1934 women based in 60 districts in Bangladesh, and we are planning to investigate variables that predict whether the women were using contraception or not at the time of the survey. Let us first look at the data and the variables we will consider.

Select Choose and pick bang1 from the Dataset list and click on Use. Click on View from the Dataset list to view the data as follows:
Here we see records for the first 24 women in district 1 displayed. The response variable use takes value 1 if the woman was using contraceptives during the time of the survey, and 0 if she was not. There are then several predictor variables, both woman-level and district-level. Here we will focus on just two: the number of living children (lc), which is a categorical variable with four categories (no kids, one kid, two kids, three+kids), and the respondents’ age, which is measured to the nearest year and has been centred around its grand mean. We will now consider modelling the dataset.

### 4.4.2 Modelling the data using logistic regression

We will firstly consider a simple linear regression model relating contraception use to the age of the woman.

Choose the template 1LevelMod from the Template list and click on Use. Then setup the model with inputs as below.

Clicking on Next and choosing equation.tex in the pull down list and we see the following:

\[
\begin{align*}
\text{use}_i & \sim \text{Binomial}(\text{cons}_i, \pi_i) \\
\logit(\pi_i) & = \beta_0 \text{cons}_i + \beta_1 \text{age}_i \\
\beta_0 & \sim 1 \\
\beta_1 & \sim 1
\end{align*}
\]
Here we the logistic regression model, in LaTeX, in the output pane. If we select `model.txt` we can then see the model code that the algebra system will interpret:

```
model{
    for (i in 1:length(uss)) {
        u[i] = dbin(p[i], cons[i])
        logit(p[i]) <- cons[i] * beta0 + age[i] * beta1
    }
    # Prior
    beta0 ~ dflat()
    beta1 ~ dflat()
}
```

Now choosing `algorithm.tex` from the right-hand pane, and placing it in its own tab in the browser window, gives the following:

Here we see that the eStat engine uses a different MCMC method, random walk Metropolis, for the steps for the fixed effects (`beta0` and `beta1`) when fitting logistic regression models. We will come back to this modelling decision in Section 4.4.4 when we compare different software packages.

Returning to the main pane and clicking on **Run** will now run the model. Once it has finished, if we select `ModelResults` from the list, and look at it in a new tab, we get the following:
Perhaps disappointedly and surprisingly, age doesn’t appear to have a significant effect (its estimate (0.0064) is similar in magnitude to its standard error (0.0051)). To see this more clearly we can look at the graph beta_1.svg in its own browser tab:

Here, whilst the values on the x-axis overlap and therefore aren’t particularly clear, we can see that all three chains show strong support for the value 0.00 in the kernel density plot (i.e. it’s comfortably
within the distribution). It might be the case, however, that contraceptive use has a non-linear relationship with age (possibly quadratic) and this could also be confounded by how far through their own family-formation process the woman is, which we will model via the variable \( lc \). We might also be interested in accounting for any clustering effects of having women nested within districts.

In order to fit a quadratic function to age we will need to construct the variable \( \text{age}^2 \) which we can easily do by viewing the Dataset and use the variable creation tool.

Select View from the Dataset pull-down list at the top of the page
In the New Variable name box type the following:

![Variable creation tool](image)

Here we are going to overwrite the existing dataset (at least in temporary memory) with a version in which we have appended an additional column to it. Clicking on Create and looking at the data below gives the following:

![Data view](image)

Here you see \( \text{age}^2 \) \((\text{age}^2)\) appearing in the column on the far right. Whilst we could explore adding further explanatory variables to this 1-level model, we are going to move straight into fitting a 2-level model to account for districts.
4.4.3 Multilevel modelling of the data
We will now require a template that will fit a 2-level logistic regression model to our dataset. In the earlier sections we looked at the template 2LevelMod and we will once again use it here and also illustrate how to fit categorical predictor variables.

On the main tab, click on Choose in the Template pull down list and select 2LevelMod and click on Use button to run this template.

Fill in the template inputs as follows:

Here we need to specify several extra inputs, including an input for the level 2 identifiers and also to let the software know which predictor variables are categorical. Continue with the inputs as follows:
Clicking on **Next** will run the algebra system and set up code to fit the model. If we select `model.txt` in the output list we will see the following:

Here we see the more complicated model code for this 2-level model in the left-hand pane. Note that the `lc` predictor is treated as categorical and thus appears as 3 dummy variables (`lc_1 – lc_3`)

If we select `tau_u.xml` in the output list we will see the following:

Here we see the algorithm step for the parameter `tau_u`. Although most parameters in this model are updated by Random Walk Metropolis sampling, this parameter is updated by Gibbs Sampling as its conditional posterior distribution has a standard form.

If we now click on **Run** then after 130s (including around 100s of compiling time) the model will have run and if we select `sigma2_u.svg` we will see the following:
Here we can see that convergence and mixing, for this parameter at least, are reasonable. In fact, if we look at the diagnostic plots for the other parameters, we see similar convergence there as well. Next we can look at ModelResults in its own tab to see the parameter estimates:
Here we see that $\beta_2$ is significant and negative (and larger than $\beta_1$) suggesting a quadratic fit to the age predictor. As the data is centred around its mean, this implies that contraceptive use is reduced the further from the mean age the woman is. We will look at this in more detail at the end of the chapter.

The parameters $\beta_3$-$\beta_5$ are all significant, and positive (and of similar magnitude), which suggests that women with children are more likely to use contraceptives than those without. The parameter $\sigma^2_u$ is fairly large, suggesting there are differences between districts in terms of contraceptive use.

What is slightly disappointing here are the ESS values for all the fixed parameters. We have run each chain, after burnin, for 2,500 iterations resulting in a total of 7,500 actual iterations (i.e. from 3 chains) but the effective sample sizes are of the order of 100-350. As this indicates, the default algorithm in eStat – random walk Metropolis – is not very efficient for this example. We will look at two possible solutions in the next two sections.

### 4.4.4 Comparison between software packages

Not all software packages fit the same MCMC algorithm for this model. So here we will show how to fit the same model in another package, OpenBUGS, which uses a different method: namely multivariate updating for the fixed effects in a GLMM, as developed by Gamerman (1997). This method results in slower estimation, but, as we will see, far better ESS. We will then look at a table comparing all the possible MCMC algorithms in the different packages for this model, which you can verify for yourselves.

To fit the model in OpenBUGS click on the remove text next to Choose estimation engine and set-up the model as follows:
Clicking on **Next** and **Run** will (after 2 min 50s on my machine) give the following, having selected *ModelResults* from the drop-down box above the output pane, and opening it in a new tab:
Here we see far better effective sample size values, with runs of 7,500 iterations translating into ESS values of between 2,500 and 5,500 for the beta parameters.

We can repeat this analysis using WinBUGS, JAGS and MLwiN with the same run lengths. Note for JAGS you will need to edit the initial value files or it will not run. To do this view each in the output window and click on the Edit button. If you change the value for beta_2 (the fixed effect associated with age2) from 0.1 to 0.0 in all three initial values files and click Set each time then JAGS should run.

We could also fit the model using the MCMCglmm package in R, although here we would need to run a single chain and logistic regression models for binary data are the one GLMM where the answers can be a little different as it assumes over-dispersion which is inappropriate in this case.

The table overleaf details the results of fitting many of these options:
In summary we see that MLwiN is by far the fastest of the packages, with eStat quicker than the other three as well. Both MLwiN and eStat use the simple random walk Metropolis algorithm, which is not the best method for this model and gives fairly poor ESS. Interestingly, both WinBUGS and OpenBUGS use the Gamerman method, but in this case OpenBUGS performs better in terms of time taken and ESS. This is somewhat puzzling as when each is run with a single chain, their performance is almost identical. Finally, for this example, JAGS is the slowest and its performance is disappointingly in the middle with regard ESS; however, there have been many comparisons between JAGS and WinBUGS for different models, and which method is better varies from model to model, so we shouldn’t dismiss it based on just this one example. The final column shows another eStat method which we will discuss next.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>eStat</th>
<th>WinBUGS</th>
<th>OpenBUGS</th>
<th>JAGS</th>
<th>MLwiN</th>
<th>eStat orthogonal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta0</td>
<td>-0.761(0.183)</td>
<td>-0.789(0.170)</td>
<td>-0.790(0.173)</td>
<td>-0.776(0.177)</td>
<td>-0.833(0.175)</td>
<td>-0.784(0.180)</td>
</tr>
<tr>
<td>Beta0 ESS</td>
<td>92</td>
<td>396</td>
<td>2595</td>
<td>255</td>
<td>94</td>
<td>980</td>
</tr>
<tr>
<td>Beta1</td>
<td>0.0076(0.0097)</td>
<td>0.0068(0.0090)</td>
<td>0.0066(0.0091)</td>
<td>0.0069(0.0090)</td>
<td>0.0051(0.0091)</td>
<td>0.0068(0.0096)</td>
</tr>
<tr>
<td>Beta1 ESS</td>
<td>183</td>
<td>952</td>
<td>5033</td>
<td>550</td>
<td>244</td>
<td>1793</td>
</tr>
<tr>
<td>Beta2</td>
<td>-0.0049(0.00074)</td>
<td>-0.0048(0.00072)</td>
<td>-0.0048(0.00073)</td>
<td>-0.0048(0.00071)</td>
<td>-0.0047(0.00073)</td>
<td>-0.0048(0.00073)</td>
</tr>
<tr>
<td>Beta2 ESS</td>
<td>319</td>
<td>1288</td>
<td>5062</td>
<td>926</td>
<td>332</td>
<td>1800</td>
</tr>
<tr>
<td>Beta3</td>
<td>0.761(0.164)</td>
<td>0.779(0.160)</td>
<td>0.782(0.162)</td>
<td>0.778(0.163)</td>
<td>0.797(0.163)</td>
<td>0.779(0.165)</td>
</tr>
<tr>
<td>Beta3 ESS</td>
<td>227</td>
<td>1117</td>
<td>5296</td>
<td>631</td>
<td>250</td>
<td>1686</td>
</tr>
<tr>
<td>Beta4</td>
<td>0.809(0.192)</td>
<td>0.822(0.181)</td>
<td>0.826(0.186)</td>
<td>0.818(0.188)</td>
<td>0.855(0.187)</td>
<td>0.823(0.190)</td>
</tr>
<tr>
<td>Beta4 ESS</td>
<td>170</td>
<td>780</td>
<td>5177</td>
<td>477</td>
<td>192</td>
<td>1725</td>
</tr>
<tr>
<td>Beta5</td>
<td>0.805(0.192)</td>
<td>0.824(0.180)</td>
<td>0.828(0.184)</td>
<td>0.817(0.187)</td>
<td>0.861(0.183)</td>
<td>0.823(0.191)</td>
</tr>
<tr>
<td>Beta5 ESS</td>
<td>114</td>
<td>547</td>
<td>4442</td>
<td>330</td>
<td>128</td>
<td>1666</td>
</tr>
<tr>
<td>Sigma2u</td>
<td>0.320(0.099)</td>
<td>0.318(0.100)</td>
<td>0.317(0.100)</td>
<td>0.317(0.100)</td>
<td>0.317(0.099)</td>
<td>0.322(0.101)</td>
</tr>
<tr>
<td>Sigma2u ESS</td>
<td>806</td>
<td>1764</td>
<td>1754</td>
<td>1446</td>
<td>770</td>
<td>756</td>
</tr>
<tr>
<td>Pd</td>
<td>42.96</td>
<td>42.44</td>
<td>42.66</td>
<td>42.16</td>
<td>42.49</td>
<td>43.21</td>
</tr>
<tr>
<td>DIC</td>
<td>2394.03</td>
<td>2393.36</td>
<td>2394.0</td>
<td>2393.39</td>
<td>2393.65</td>
<td>2394.51</td>
</tr>
<tr>
<td>Time (s)</td>
<td>130 (30)</td>
<td>219</td>
<td>170</td>
<td>284</td>
<td>14</td>
<td>132 (31)</td>
</tr>
</tbody>
</table>
4.4.5 Orthogonal parameterisation.

The reason eStat (and MLwiN) perform badly in this instance is that they are performing single-site updating, and the parameters are correlated. So here we will consider a reparameterisation method that aims to fit parameters that are less correlated, and then translates them back to the original parameters. For this we construct a set of orthogonal vectors from the original predictor variables (see Browne et al. 2009 for details).

We will therefore now look at the NLevelOrthogParamRS template in order to use orthogonalisation on our model. This template actually fits a larger family of models: those with any number of higher levels/classifications (hence ‘NLevel’), allowing for the possibility of random slopes at each of these levels (hence ‘RS’), and so our 2-level random intercept model is perhaps the simplest case that the template fits.

Click on the Template pull down list and click Choose then select NLevelOrthogParamRS from the template list.

Click on Use and fill in the template inputs as follows:

Clicking on Next and selecting equation.tex in the pull down list (we’ve opened it in a new tab) will show the following:
Here we see that the model code is actually fitting a different set of predictors, each with the prefix ‘orth’ and a corresponding set of coefficients. There is then a set of deterministic statements that translate these coefficient values to the coefficient values for the original predictors (again, see Browne et al. (2009) for details).

Clicking on the Run button will run the model (which took 130s on my machine including compiling), after which selecting ModelResults from the pull down list, and popping out into a new tab, gives the following:
The estimates, their ESS, and the time taken to run the model are all added to the end of the software comparison table we looked at above. It indicates that, compared to the other method we employed to fit the model in eStat, there is slightly more of an overhead incurred when performing the orthogonalising algorithm, however it is still faster than OpenBUGS, and the ESS are now much better (if still not as good as OpenBUGS). We therefore have two ways of fitting the model that are reasonably comparable in terms of ESS/s, with OpenBUGS still having a slight edge. This orthogonalising approach is also available in MLwiN: this will be faster again, and should have similar ESS to the method in eStat, and therefore may be the best overall in terms of ESS/s, but we leave this for the reader to investigate.

4.4.6 Predictions from the model

When we ran this model we discussed some interpretation of the fit, but it would be nice to plot some predictions from the model as well. In this latest version of Stat-JR we have added the option to store predictions when fitting the model. So hopefully in the last model fit you will have ticked yes to the generate prediction dataset question. This will generate a new dataset named prediction_datafile which contains the original data and several prediction columns formed from the model fit.

To use this dataset we need to select Choose on the dataset list and select prediction_datafile from the list.

In fact the dataset has a full prediction column called pred_full but this also contains the district random effects. We would here like to simply predict from the fixed part of the model so we can construct the variable pred_fixed as follows:

Click on View from the Dataset menu and input the new variable pred_fixed as indicated below.

Click on Create to create the variable.
This has created a variable on the fixed predictor scale but as we are fitting a logistic regression we need to take an anti-logistic transform to convert these predictions to probabilities. This can be done by creating another column in the dataset as shown below:

In order to plot separate fitted curves for the various numbers of living children we can use the template `XYGroupPlot` as shown below:

Clicking on Run and popping out `graphxygroup.svg` gives the following:
Here we see the four curves (although three of them are very close together) which clearly showing that the women with children have higher probabilities of using contraceptives, and that the peak for each group is around the average age of the sample, as discussed earlier.

Hopefully this section has shown firstly that Stat-JR can fit models other than Normal response models; in fact there are a vast number of model templates which fit lots of other model classes. Secondly, we hope we’ve shown its utility in terms of comparing model-fitting across different software packages for different models, accessing each from a common hub.

### 4.5 Miscellaneous other topics e.g. Data Input/Export

It is very easy to import data into Stat-JR. Data has to be in the .dta dataset format that is used by Stata (and supported by MLwiN). Any .dta files that are in the “datasets” subdirectory of Stat-JR will automatically be loaded in the start-up of the software. As discussed earlier, if you wish to add a dataset to the directory after start-up you can press **Debug > Reload datasets** from the main menu page to reload the whole ‘datasets’ folder again, or you can upload an individual file by finding it via **Browse...** (beneath the scrollable list of datasets on the main menu page) and then click on the **Upload** button. Similarly, any datasets constructed while running the software, and stored in...
temporary memory, can be downloaded for storage elsewhere by selecting them as the current dataset and using the Download button.

5 References


6 Appendix: List of Third Party Software that are used by Stat-JR

Stat-JR makes use of several third party software products that are included within the distributed code or (in the case of MinGW) need to be downloaded separately. These software products each have a license file that can be viewed from the links in the table below and/or in the licences subdirectory of the installed code.

<table>
<thead>
<tr>
<th>Package</th>
<th>Link</th>
<th>Licence terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>MathJax</td>
<td><a href="http://cdn.mathjax.org/mathjax/2.0-latest/LICENSE">http://cdn.mathjax.org/mathjax/2.0-latest/LICENSE</a></td>
<td>Apache</td>
</tr>
<tr>
<td>Bootstrap</td>
<td><a href="https://github.com/twitter/bootstrap/blob/master/LICENSE">https://github.com/twitter/bootstrap/blob/master/LICENSE</a></td>
<td>Apache</td>
</tr>
<tr>
<td>jquery</td>
<td><a href="http://jquery.org/license">http://jquery.org/license</a></td>
<td>MIT</td>
</tr>
<tr>
<td>jquery-ui</td>
<td><a href="http://jquery.org/license">http://jquery.org/license</a></td>
<td>MIT</td>
</tr>
<tr>
<td>jquery-treeview</td>
<td><a href="https://github.com/jzaeffer/jquery-treeview">https://github.com/jzaeffer/jquery-treeview</a></td>
<td>Dual MIT/GPL</td>
</tr>
<tr>
<td>jqgrid</td>
<td><a href="http://www.trirand.com/blog/?page_id=87">http://www.trirand.com/blog/?page_id=87</a></td>
<td>Dual MIT/GPL(v2)</td>
</tr>
<tr>
<td>jQuery File Upload</td>
<td><a href="http://opensource.org/licenses/MIT">http://opensource.org/licenses/MIT</a></td>
<td>MIT</td>
</tr>
<tr>
<td>jQuery text align</td>
<td><a href="http://www.opensource.org/licenses/bsd-license.php">http://www.opensource.org/licenses/bsd-license.php</a></td>
<td>BSD</td>
</tr>
<tr>
<td>cx_freeze</td>
<td><a href="http://cx-freeze.readthedocs.org/en/latest/license.html">http://cx-freeze.readthedocs.org/en/latest/license.html</a></td>
<td>PSF</td>
</tr>
<tr>
<td>numexpr</td>
<td><a href="http://www.opensource.org/licenses/mit-license.php">http://www.opensource.org/licenses/mit-license.php</a></td>
<td>MIT</td>
</tr>
<tr>
<td>mako</td>
<td><a href="http://www.opensource.org/licenses/mit-license.php">http://www.opensource.org/licenses/mit-license.php</a></td>
<td>MIT</td>
</tr>
<tr>
<td>pyparsing</td>
<td><a href="http://www.opensource.org/licenses/mit-license.php">http://www.opensource.org/licenses/mit-license.php</a></td>
<td>MIT</td>
</tr>
<tr>
<td>reset-fonts-grids</td>
<td><a href="http://yUILibrary.com/licenses/">http://yUILibrary.com/licenses/</a></td>
<td>BSD</td>
</tr>
<tr>
<td>markupsafe</td>
<td><a href="http://www.opensource.org/licenses/bsd-license.php">http://www.opensource.org/licenses/bsd-license.php</a></td>
<td>BSD</td>
</tr>
<tr>
<td>pyquery</td>
<td><a href="http://www.opensource.org/licenses/bsd-license.php">http://www.opensource.org/licenses/bsd-license.php</a></td>
<td>BSD</td>
</tr>
<tr>
<td>Package</td>
<td>License Link</td>
<td>License Type</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------------------------------------------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>rdflib</td>
<td><a href="http://www.opensource.org/licenses/bsd-license.php">http://www.opensource.org/licenses/bsd-license.php</a></td>
<td>BSD</td>
</tr>
<tr>
<td>isodate</td>
<td><a href="http://www.opensource.org/licenses/bsd-license.php">http://www.opensource.org/licenses/bsd-license.php</a></td>
<td>BSD</td>
</tr>
<tr>
<td>lxml</td>
<td><a href="http://lxml.de/index.html#license">http://lxml.de/index.html#license</a></td>
<td>BSD</td>
</tr>
<tr>
<td>numpy</td>
<td><a href="http://numpy.scipy.org/license.html">http://numpy.scipy.org/license.html</a></td>
<td>BSD</td>
</tr>
<tr>
<td>scipy</td>
<td><a href="http://www.scipy.org/License_Compatibility">http://www.scipy.org/License_Compatibility</a></td>
<td>BSD</td>
</tr>
<tr>
<td>matplotlib</td>
<td><a href="http://matplotlib.sourceforge.net/users/license.html">http://matplotlib.sourceforge.net/users/license.html</a></td>
<td>Modified BSD</td>
</tr>
<tr>
<td>MinGW</td>
<td><a href="http://www.mingw.org/license">http://www.mingw.org/license</a></td>
<td>Not distributed with software directly</td>
</tr>
<tr>
<td>LAPACK</td>
<td><a href="http://www.netlib.org/lapack/LICENSE.txt">http://www.netlib.org/lapack/LICENSE.txt</a></td>
<td>Modified BSD</td>
</tr>
<tr>
<td>BLAS</td>
<td><a href="http://www.netlib.org/blas/faq.html#2">http://www.netlib.org/blas/faq.html#2</a></td>
<td>Own licence (Netlib)</td>
</tr>
<tr>
<td>web.py</td>
<td><a href="https://github.com/webpy/webpy/blob/master/LICENSE.txt">https://github.com/webpy/webpy/blob/master/LICENSE.txt</a></td>
<td>PSF</td>
</tr>
<tr>
<td>Python</td>
<td><a href="http://docs.python.org/license.html">http://docs.python.org/license.html</a></td>
<td>PSF</td>
</tr>
<tr>
<td>setuptools</td>
<td><a href="http://docs.python.org/license.html/">http://docs.python.org/license.html/</a></td>
<td>PSF</td>
</tr>
<tr>
<td>cssselect</td>
<td><a href="http://www.opensource.org/licenses/bsd-license.php">http://www.opensource.org/licenses/bsd-license.php</a></td>
<td>BSD</td>
</tr>
<tr>
<td>dateutil</td>
<td><a href="http://opensource.org/licenses/bsd-2-Clause">http://opensource.org/licenses/bsd-2-Clause</a></td>
<td>Simplified BSD</td>
</tr>
<tr>
<td>six</td>
<td><a href="https://bitbucket.org/gutworth/six/src/e3da7fd96039a6ed89493f89d121c4f3797e6713/LICENSE?at=default">https://bitbucket.org/gutworth/six/src/e3da7fd96039a6ed89493f89d121c4f3797e6713/LICENSE?at=default</a></td>
<td>MIT</td>
</tr>
<tr>
<td>sparqlwrapper</td>
<td><a href="http://www.w3.org/Consortium/Legal/2002/copyright+software-20021231">http://www.w3.org/Consortium/Legal/2002/copyright+software-20021231</a></td>
<td>W3C</td>
</tr>
<tr>
<td>provpy</td>
<td><a href="http://opensource.org/licenses/BSD-2-Clause">http://opensource.org/licenses/BSD-2-Clause</a></td>
<td>BSD</td>
</tr>
</tbody>
</table>