SampleSizeRegression

Version 1.0, July 2019

1. Introduction

The program **SampleSizeRegression** --- available for Windows only --- was developed to estimate sample size requirements in the context of Bayesian linear and logistic regression parameter estimation with possible covariate measurement error, along the lines of our paper

Bayesian Sample Size Criteria for Linear and Logistic Regression in the Presence of Confounding and Measurement Error
Lawrence Joseph and Patrick Bélisle
Unpublished (see link below)

We recommend that you read the above paper carefully before using this software; this paper is available from
http://www.medicine.mcgill.ca/epidemiology/Joseph/publications/Methodological/SSReg.pdf

You are free to use this program, for non-commercial purposes only, under two conditions:

- This note is not to be removed;
- Publications using **SampleSizeRegression** results should reference the manuscript mentioned above;
- While we have done our best to ensure the program works as described in this manual, the user acknowledges that this program is not necessarily bug-free. We assume no liability for any errors or consequences that may arise from the use of this program. The use of this software is at the exclusive risk of the user.

If you have not installed **SampleSizeRegression** yet, please read the Installation Instructions (InstallInstructions.html) first.

The easiest way to open this program[1] is to use the shortcut found in Programs list from the Start menu. Once opened, you will be prompted by a graphical user interface (GUI) to describe the problem, that is:

- choose between sample size calculations or outcome prediction for fixed sample size(s)

[1] You can start SampleSizeRegression by browsing through the User's Programs menu (available by clicking the Start button and then Programs) and selecting SampleSizeRegression. You can also start SampleSizeRegression by opening Windows Explorer, browsing to this package's location (c:\Users\user name\Documents\Bayesian Software\ SampleSizeRegression or c:\Documents and Settings\user name\My Documents\ Bayesian Software\ SampleSizeRegression by default, depending on your platform) and clicking on SampleSizeRegression.vbs.

- choose between linear and logistic regression
- fill in your prior information about regression parameters for each independent variable included in your regression model
- select a sample size criterion
- select an output file (where you want the results to be saved)
- and a few more technical questions (number of Gibbs iterations, where to start the seach for the optimal sample size, etc.).

Once the GUI has collected all of the inputs required for the problem, it will be closed and the program will continue almost invisibly; the only thing that you will see on your screen is a WinBUGS window, which you can minimize.

When the program has finished (running time can vary, and could be many hours if you are running sample size calculations) another GUI will appear announcing program completion and giving you the opportunity to view the output immediately. This GUI will not appear when **SampleSizeRegression** is called from a script (see section 3.1).

When started from the .vbs file (for example, when run from the Start menu), **SampleSizeRegression** will always run at low priority, allowing your system to use more CPU for higher priority tasks when needed. Thus, you can continue to work comfortably as this program runs in the background.

2. Problem description

Suppose P independent variables $X_i = (X_{i1}, X_{i2}, \ldots, X_{iP})$ are to be collected, with or without **measurement error**, on N subjects as well as the outcome (or response) variable Y_i, $i = 1, 2, \ldots, N$, and that the relationship between the outcome and the P independent variables will be modeled through either

a) a linear model, that is,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_P X_{iP} + \varepsilon_i,$$

where the ε_i are independent normally distributed random error variables with mean 0 and common variance σ^2, when the outcome is a continuous variable or

b) a logistic model, that is,

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

where $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_P X_{iP}$

when the outcome is binary.

In both linear and logistic model, we assume a multivariate normal prior distribution on the regression parameters $\boldsymbol{\beta} = (\beta_0,\ \beta_1,\ \beta_2, \ldots, \beta_P)'$. For variables X_k that come with measurement error, we assume that the measured value is a random value centered about the true value X_k^*, that is, $X_k \sim N(X_k^*, \sigma_k^2)$ where $\sigma_k \sim U(a_k, b_k)$ for given constants a_k and b_k.

In the linear model, we assume a uniform prior distribution on σ, that is, $\sigma \sim U(a, b)$ for given constants a and b. Note that the presence of the intercept β_0 in the linear model is optional.

The posterior density for $\boldsymbol{\beta}$ in the context of a **linear regression** is given by

$$f(\boldsymbol{\beta} \mid \boldsymbol{Y}) \ \propto \int_{\sigma} f(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{\sigma})\ f(\boldsymbol{\beta})\, f(\sigma) \ \prod_{k \in S} \prod_{i=1}^{N} \int_{\sigma_k} f(X_{ik} | X_{ik}^*, \sigma_k)\, f(\sigma_k)\ d\sigma_k\ d\sigma$$

where

$\boldsymbol{Y}$	is the response vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)'$
$\boldsymbol{X}$	is the design matrix
$f(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{\sigma})$	is the likelihood of the data $\boldsymbol{Y}$ – that is, a multivariate normal distribution with mean $\boldsymbol{X\beta}$ and covariance matrix $\sigma^2\ I_q$, where I_q is the $q\ \times q$ identity matrix, $q\ =\ P + 1$ if the model has an intercept, $q = P$ if not
$f(\boldsymbol{\beta})$	is the analysis prior distribution for $\boldsymbol{\beta}$
$f(\sigma)$	is the analysis prior distribution for σ
S	is the set of variables measured with error, that is, $S = \{k : X_k \text{ is measured with error}\}$; if S is null – that is, if all variables are measured exactly – then $\prod_{k \in S} \prod_{i=1}^{N} \int_{\sigma_k} f(X_{ik} \mid X_{ik}^*, \sigma_k)\, f(\sigma_k)\ d\sigma_k$ = 1
X_{ik}^*	is the measured/observed value for variable X_k in subject i
X_{ik}	is the true value for variable X_k in subject i (unobserved for variables with measurement error).

SampleSizeRegression was developed to compute the minimal sample size N such that a regression parameter (say β_1, without loss of generality) is estimated within a given pre-specified accuracy.

2.1 How SampleSizeRegression works

For a fixed sample size N, a large number M (M is called the preposterior sample size) of data points $X_i, i = 1, 2, \dots, N$ is sampled.

The distribution of $X_i = (X_{i1}, X_{i2}, \dots, X_{iP})$ can be assumed to be a multivariate normal distribution or any multivariate distribution F; in the latter case, the user must provide R code to generate random values for $X_i, i = 1, 2, \dots, N$: an example will be provided in section 4.2.

For each of the M samples of $\{X_1, X_2, \dots, X_N\}$, a set of regression parameters $(\beta_0, \beta_1, \beta_2, \dots, \beta_P)$ is sampled from the $\boldsymbol{\beta}$ design prior distribution, and finally a set of response variables $\{Y_1, Y_2, \dots, Y_N\}$ is generated along the linear or logistic regression model parameters. Response and independent variables are then saved and analyzed through a WinBUGS model taking into account the measurement error around independent variables (if any) and the uncertainty around the regression parameters (through the $\boldsymbol{\beta}$ analysis prior distribution). The WinBUGS Markov Chain Monte Carlo process leads to the approximation of the posterior distribution of the regression parameters $\boldsymbol{\beta}$.

The coverage or length of the HPD interval of a predetermined regression parameter is then calculated for each of the M samples and the sample size N is then ranked as being sufficient or not depending on whether or not the selected sample size criterion is met.

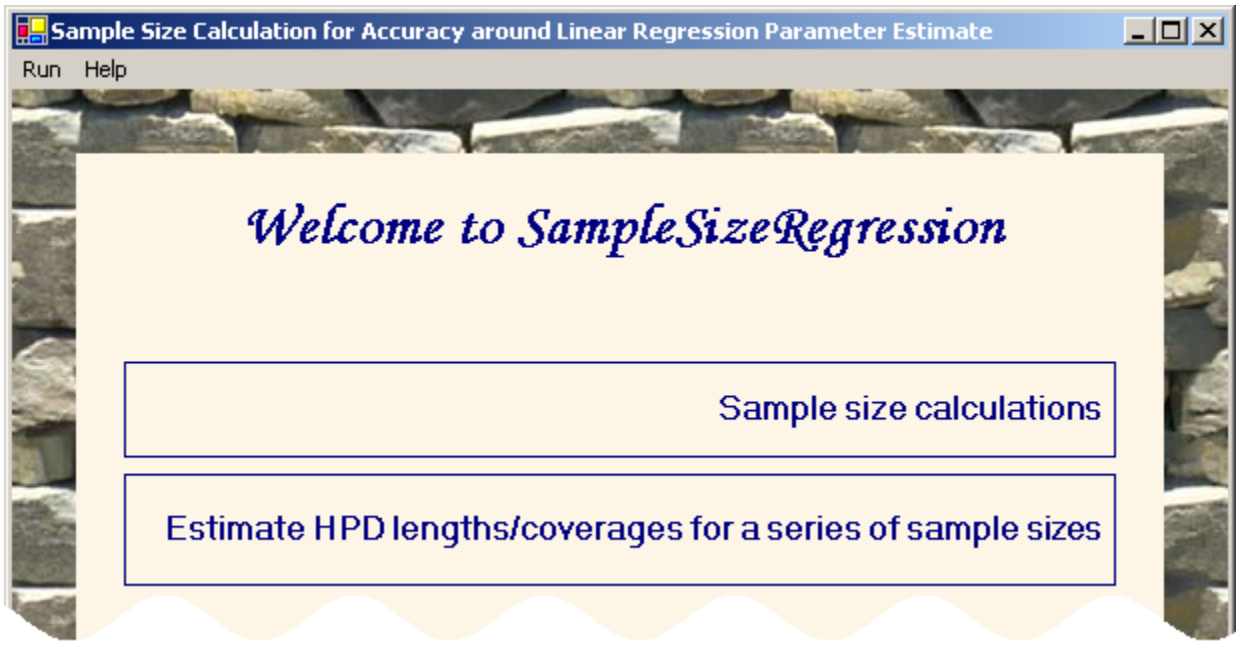
SampleSizeRegression iterates over N until

a) the desired parameter accuracy is met for sample size N but not for $N - 1$ or
b) in a series of six consecutive sample sizes, the larger three satisfy the sample size criterion while the smaller three do not, and these six consecutive sample sizes do not span more than 2% of their midpoint value.

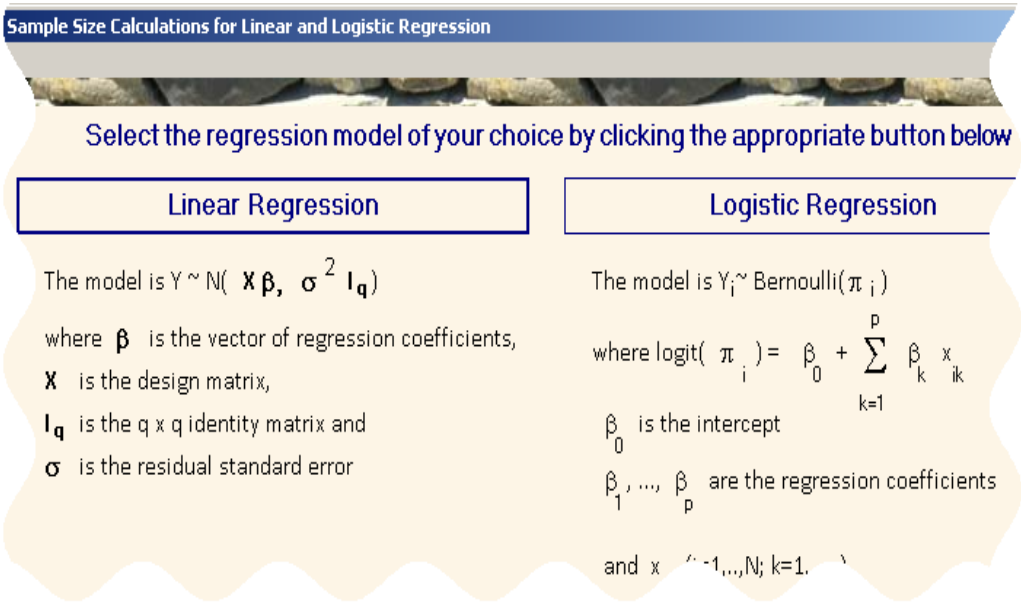
Stopping criterion (b) proves useful when the final sample size is large (e.g. more than a thousand).

3. How to use SampleSizeRegression

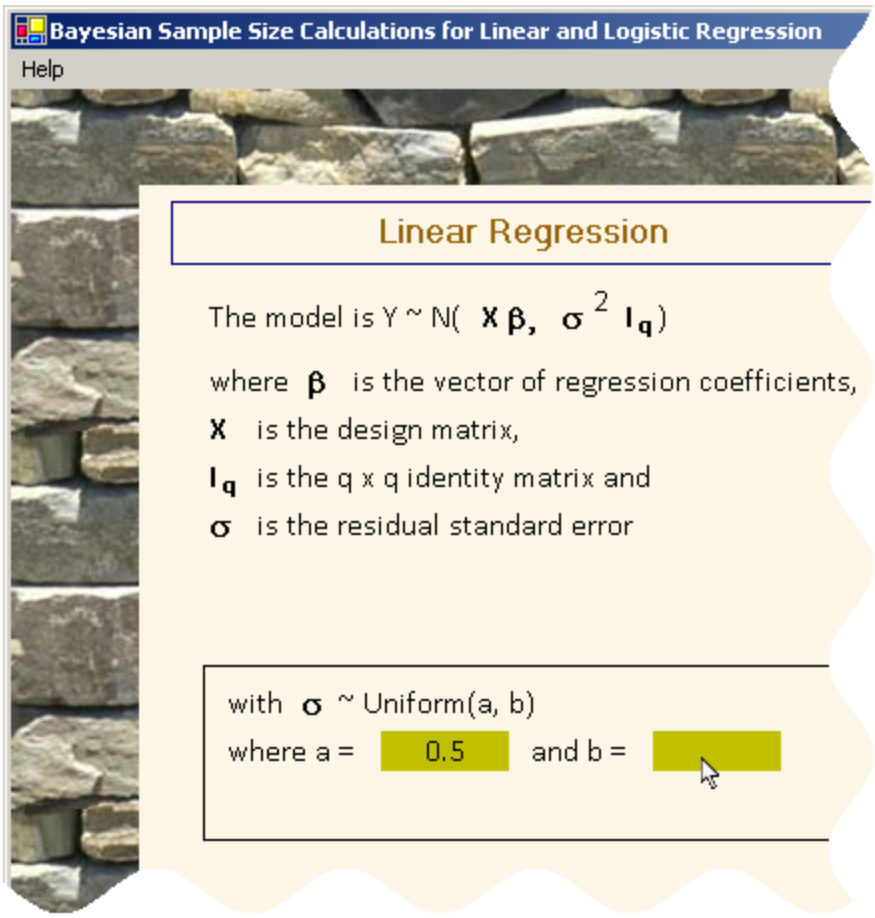
The initial window (below) is used to choose between sample size calculations, or the estimation of average or percentile of HPD lengths or coverages for a series of predetermined sample sizes.



The next form will first be used to select between linear or logistic regression by clicking the corresponding label.



In the context of linear regression, the form will then be used to enter the prior distribution on the standard deviation of then (independent) error terms.



The remaining of this section illustrates the entry form in the context of linear regression. This entry form and the following forms will be very similar in the context of logistic regression; the form entries specific to the logistic model will be introduced in section 3.2.

You will next be asked whether or not you wish to include an intercept in your model and whether the independent variables $X_i = (X_{i1}, X_{i2}, \ldots, X_{iP})$ come from a multivariate normal distribution or not. In epidemiological studies, independent variables will often include dichotomous (e.g., gender) or class variables (e.g., socio-economic status, race) which obviously cannot be modeled through a multivariate normal distribution.

If the distribution for X_i is assumed to be multivariate normal, then the next step (below) will be to enter the number of independent variables, the variable names (second form below) and the parameters associated with the X_i distribution (third form below).

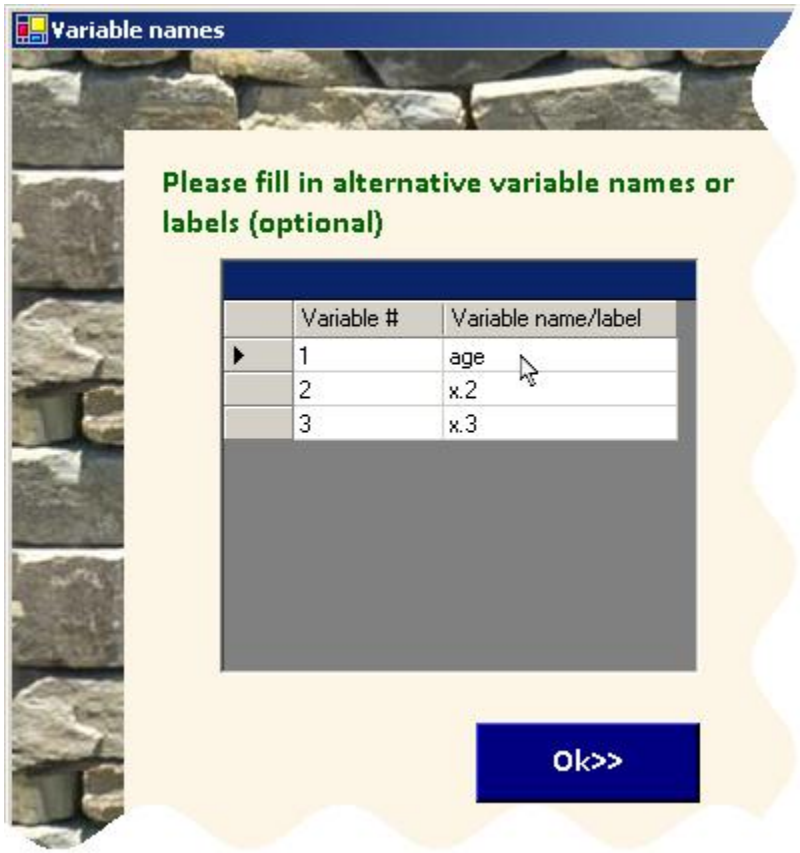
Select a distribution for the i.i.d. vectors

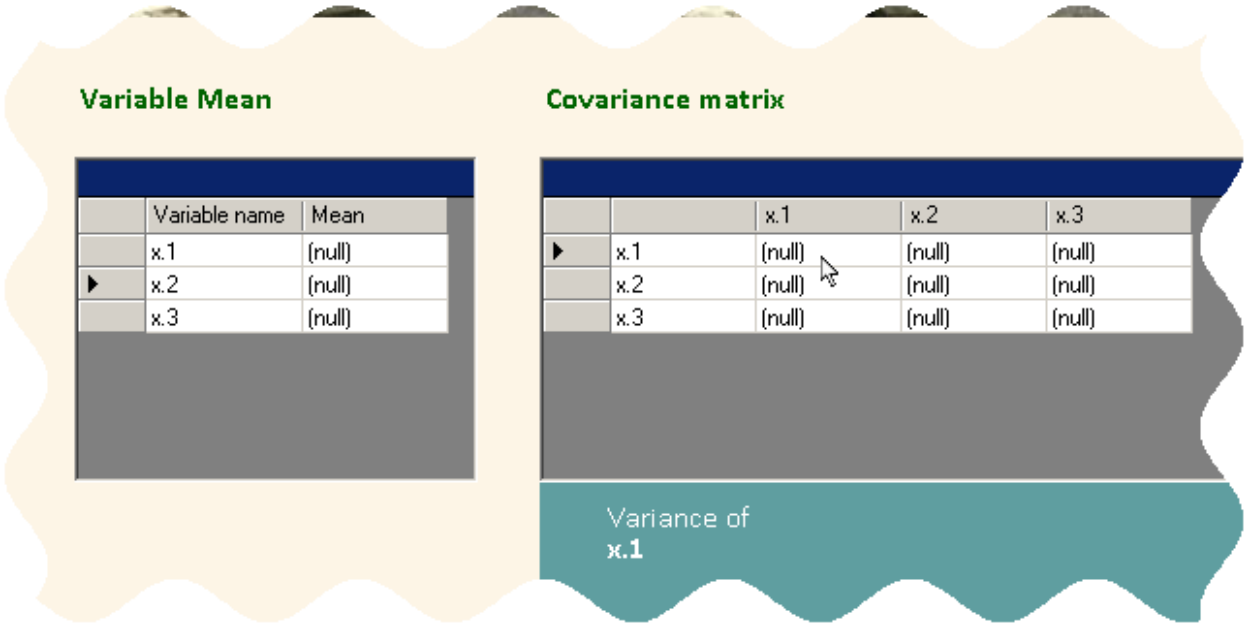
$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $p = q - 1$

- (•) $x_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- ( ) other (to be defined through your own R code, next form)

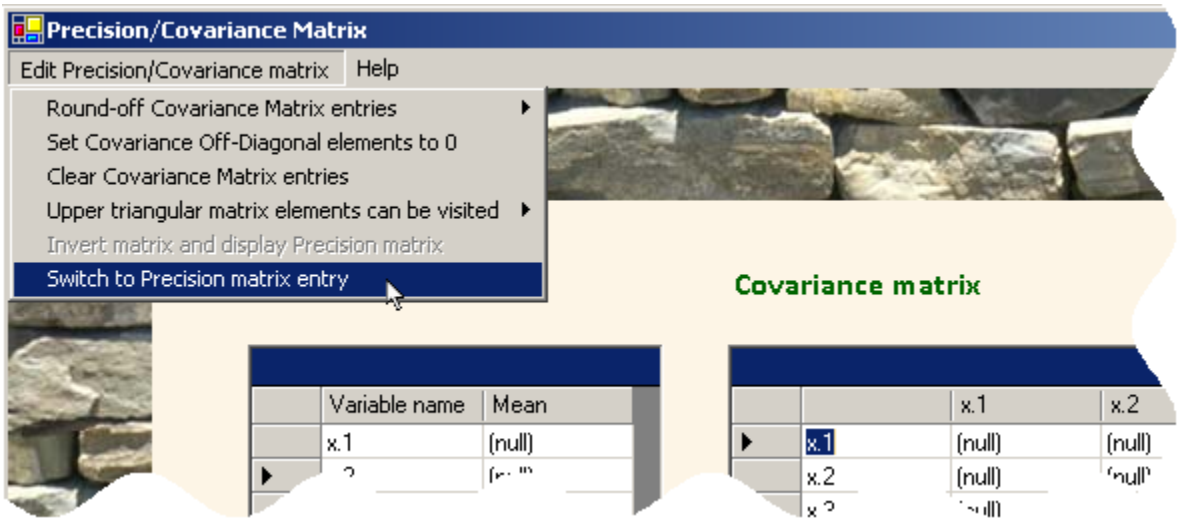
How many independent variables do you wish to include in the model?

1
2
3
4
5
6

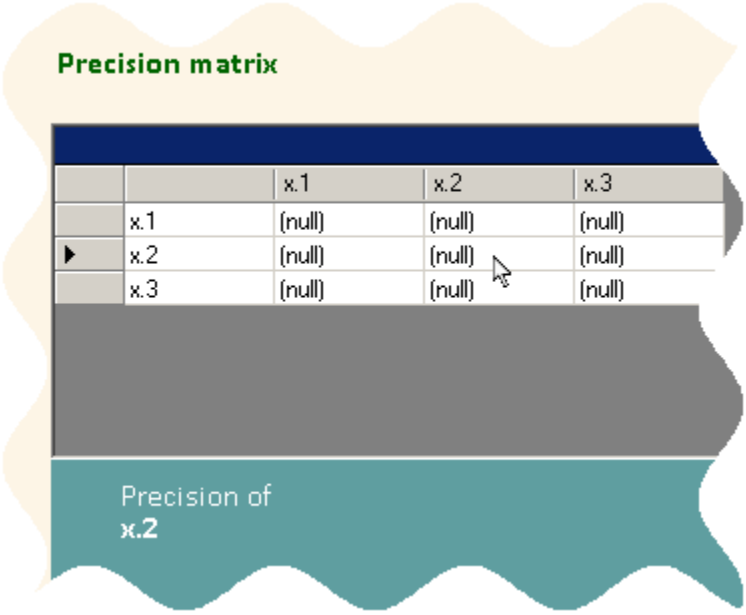




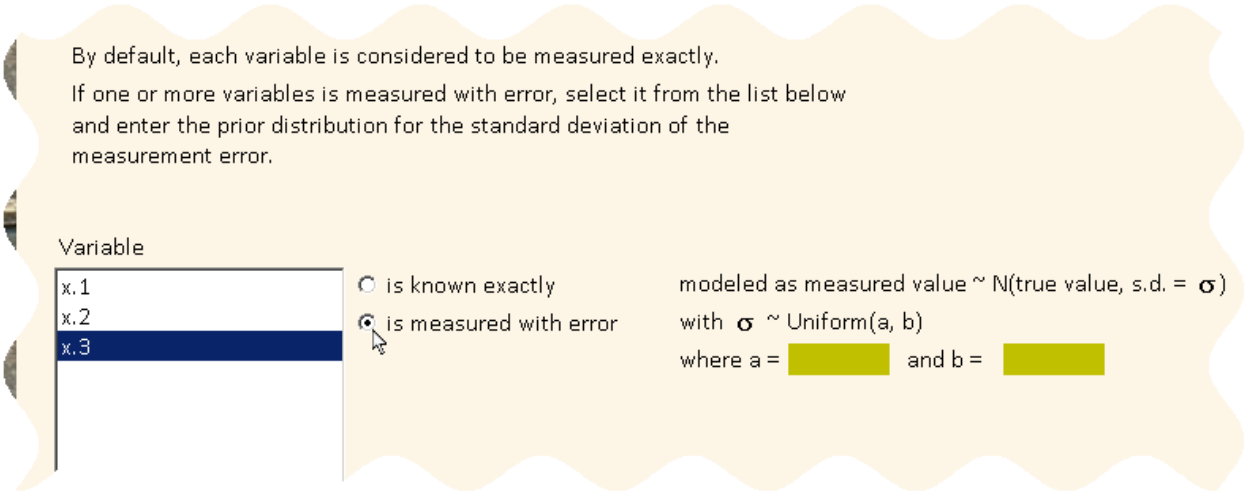
The parameters of the prior density for X_i are the P means and the $P \times P$ covariance matrix, by default; you can alternate between covariance and precision matrices through the top menu item *Switch to Precision (Covariance) matrix entry*.



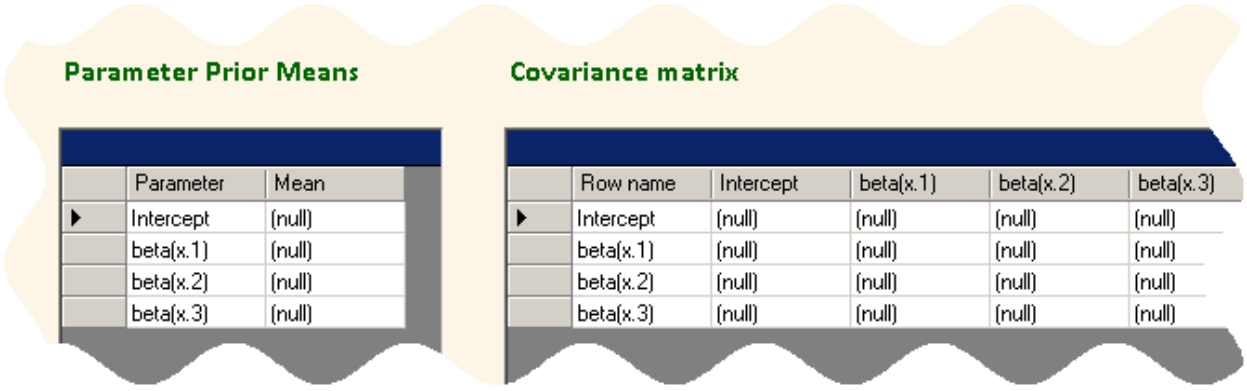
Scale parameters can be equally entered through a covariance or a precision matrix.



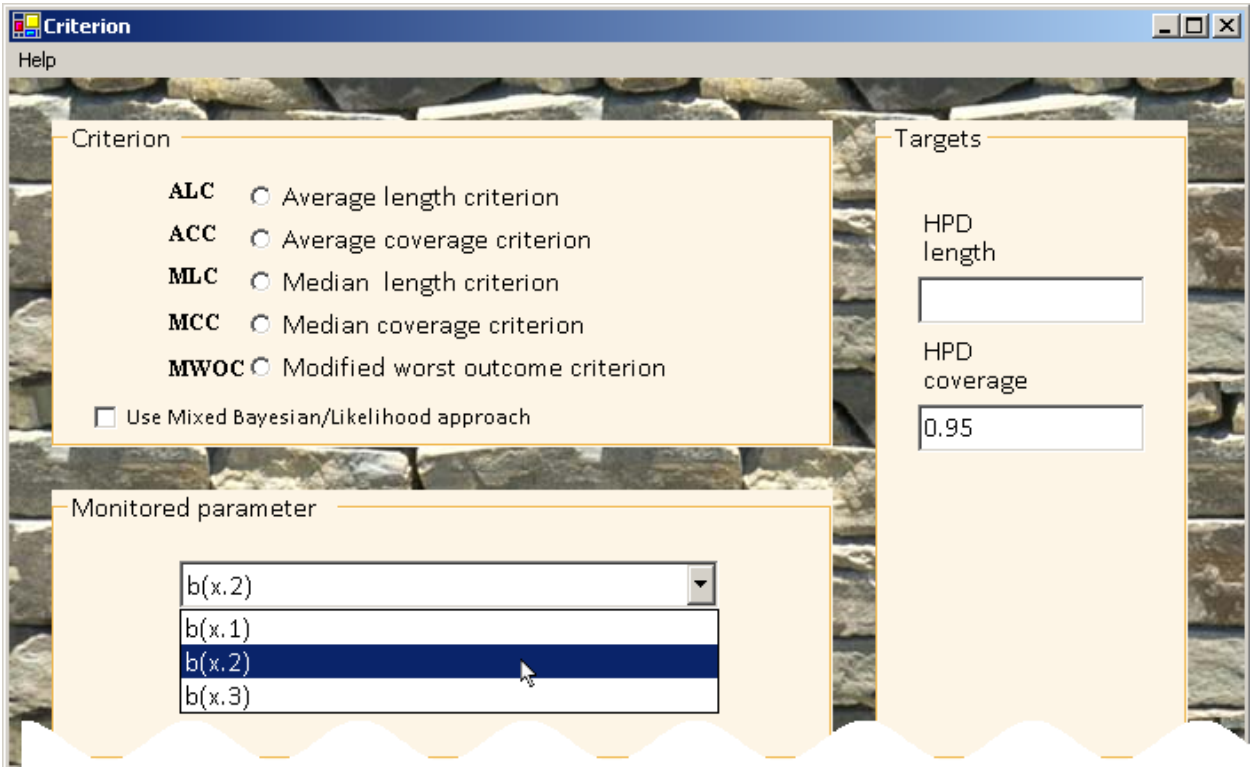
The next form (below) will be used to enter a prior distribution for uncertainty around each independent variable that is measured with some error. Note that all can also be considered as being measured exactly.



The next form (below) is for entry of the parameters of the multivariate normal prior distribution for the $\boldsymbol{\beta}$ regression parameters. It is similar to the entry form for the X_i multivariate distribution and also allows the choice of entry through a covariance or a precision matrix.

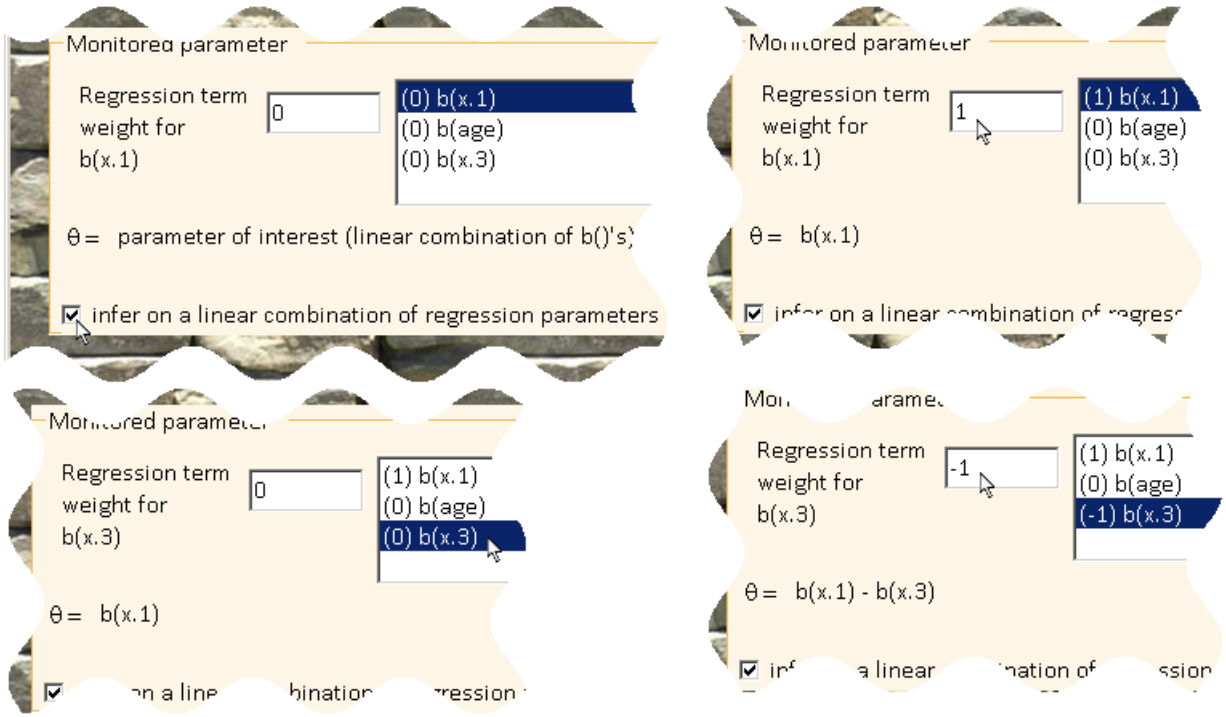


If sample size determination was selected in the initial window, the next form (below) allows selecting one of the ten sample size criteria available. This window is also used to specify the fixed or target (depending on criterion selected) HPD length and coverage for the parameter of interest, also selected from this form.

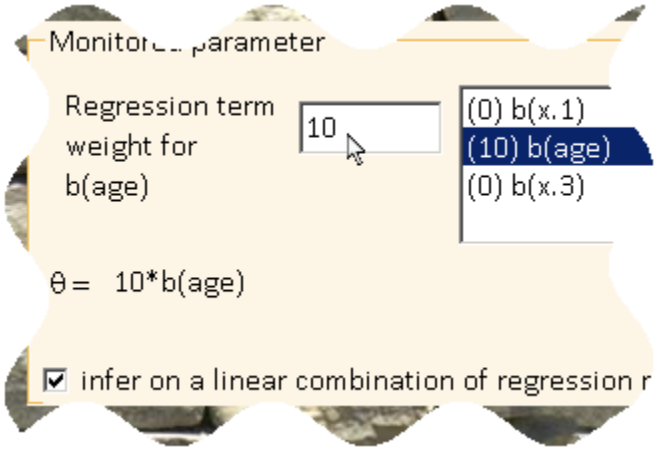


When the Mixed Bayesian/Likelihood approach is chosen in form above, the prior distributions for $\boldsymbol{\beta}$ and σ will be used at the design stage (to generate data) but different prior distributions (called analysis prior distributions) will be used at the analysis stage (that is, in the WinBUGS model written to estimate the posterior distribution for $\boldsymbol{\beta}$): these analysis prior distributions would be collected later through forms similar to those already presented.

Note that it is also possible to base sample size calculation on a linear combination of regression parameters rather than on a single regression parameter as illustrated above. Indeed, clicking the tick box labeled *infer on a linear combination of regression parameters* will modify the elements displayed in *Monitored parameter* frame: you can then click on the regression parameters of your choice in the list box displayed an assign a weight to the selected regression parameter by entering it in the text box to its left. In the example below, we first click b(x.1) and assign it a +1 weight, and then click b(x.3) and assign it a -1 weight to build the difference b(x.1) – b(x.3). Note that the text to the right of θ gives the expression of the regression parameters linear combination such defined.



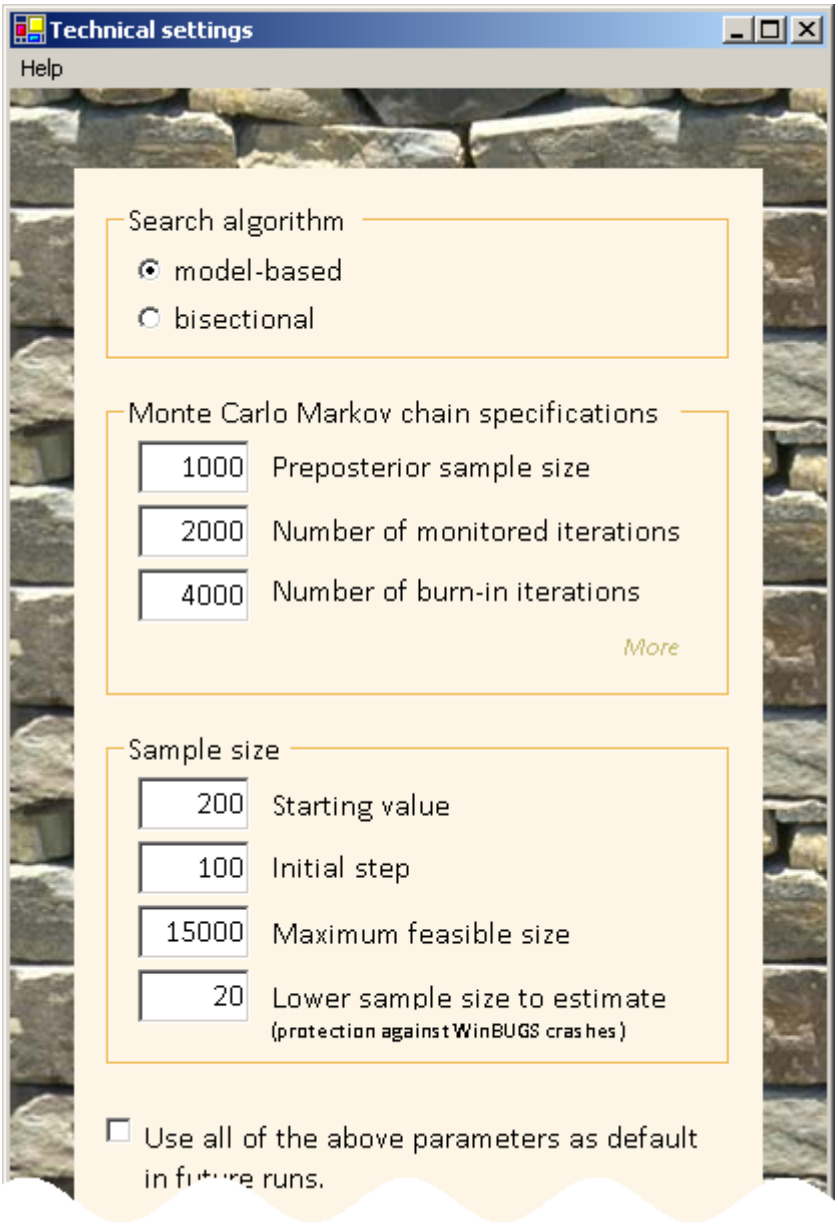
In the context of logistic regression, that feature could also be used to base sample size calculation on the Odds Ratio of a variable on a different scale than the default one-unit OR. The opposite image illustrates an example where the interest would be on the Odds Ratio for age expressed in terms of a 10-years difference.



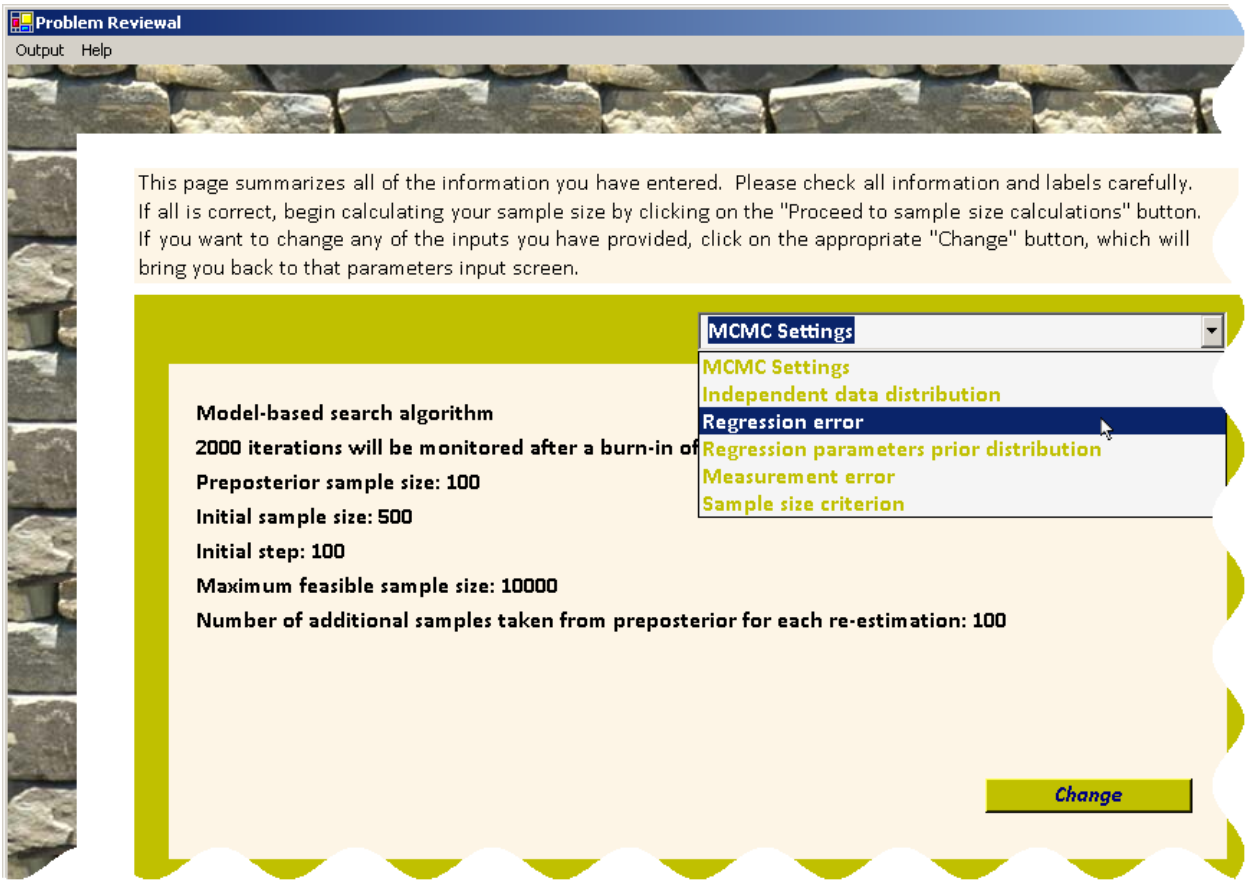
The next window allows the user to specify the preposterior sample size and the number of burn-in and monitored iterations of the Gibbs sampler algorithm that is used throughout. Changing these values is optional, the default values will usually provide reasonable estimates.

When **SampleSizeRegression** is used to calculate sample sizes, this form also allows the user to specify whether the optimal sample size should be found via a bisectional search or with a so called model-based algorithm. The latter, the default choice, usually converges to the optimal sample size neighbourhood with fewer steps.

In either case, the first three sample sizes for which the outcome of interest (e.g. average HPD length) will be estimated are based on a bisectional search, after which this option comes into effect.



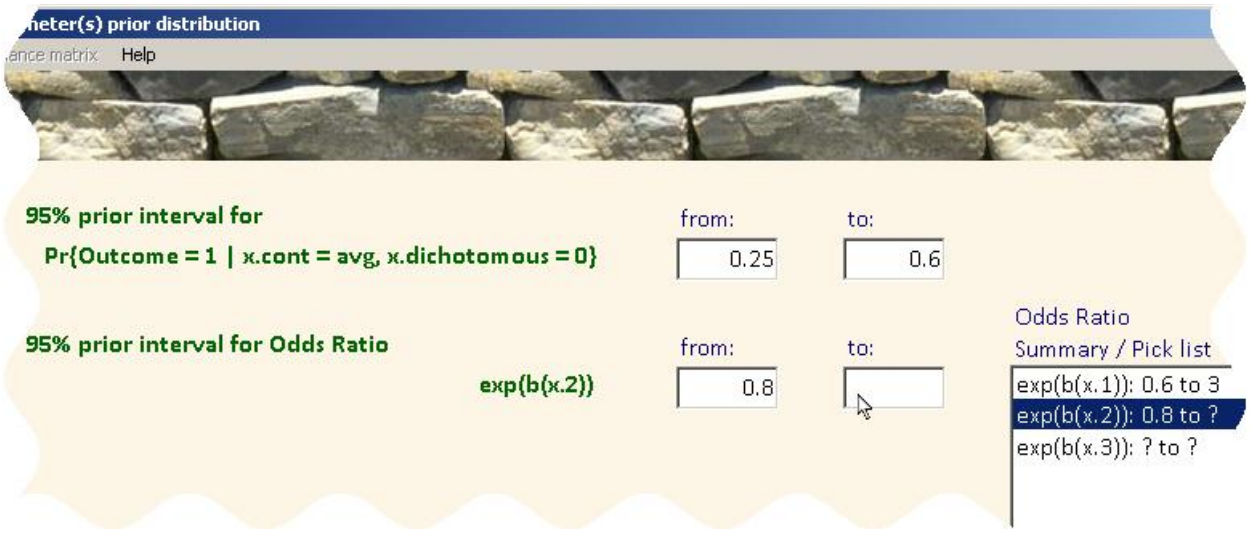
Finally, a Problem Reviewal form (below) allows the user to review each parameter entered through the different forms and modify any, if necessary, by clicking the appropriate *Change* button.



The above form is also used to select the output file location, either by selecting the top-left menu item *File/Save as…* or by clicking the Output location link in the lower left portion of the form.

3. 1 Entering prior distribution for regression parameters in the context of logistic regression

In the context of logistic regression, the prior distribution for the β regression parameters can be entered through its mean vector and covariance or precision matrix as illustrated in the context of linear regression in section above, or through the 95% limits of prior intervals for the Odds Ratio of each independent variable included in the regression model, as shown in figure below.

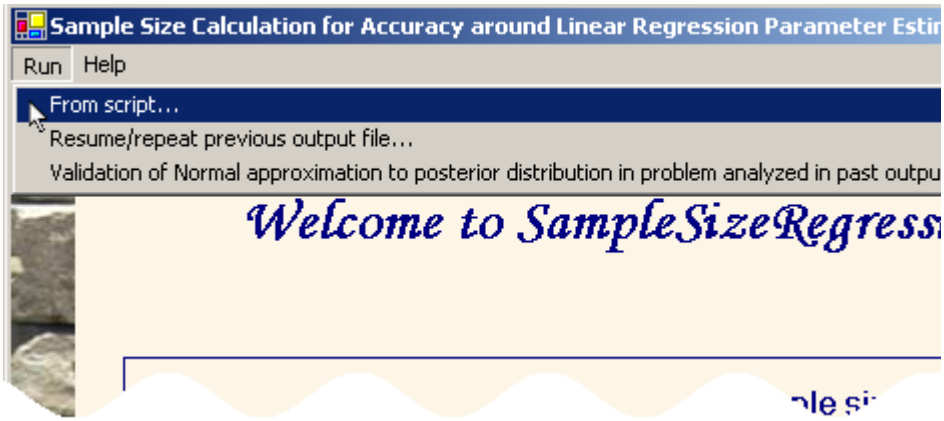


The top 2 boxes are used to enter the 95% interval for the prior probability of a positive outcome ($Y_i = 1$) when the continuous independent variables are equal to their average and the dichotomous independent variables are 0's. The other values to enter are the 95% prior intervals for each of the independent variables' Odds Ratios (per one unit change): this is done by first clicking a X variable name in the right box labeled *Odds Ratio Summary / Pick list*, and then entering the lower and upper limits of its 95% Odds Ratio prior interval. The prior information contained in the above 95% limits on OR's is turned into a multivariate normal distribution with independent components, with mean and sd (μ_i, σ_i) for regression parameter β_i such that $\exp(\mu_i + z_{0.025}\sigma_i)$ are equal to the two endpoints of the 95% prior interval for β_i, i=1, 2,..., P, and (μ_0, σ_0) such that $\exp(\mu_0 + z_{0.025}\sigma_0)$ are equal to the $\text{logit}()$ value of the two endpoints of the 95% prior interval for Pr{Y = 1 | continuous X variables = their average and dichotomous X variables = 0}.

3. 2 Saving and running scripts

Once a problem is fully described by completing the appropriate forms, the actual computations can be launched right away by clicking the *[Run now] Proceed to Sample Size calculations* or saved for future submission by clicking the *[Run later] Register Problem Description* button, both found in the lower right corner of the Problem Reviewal form. Problems saved for future computation will be saved as *script* files, identified by a label entered by the user in the next form.

The initial form of **SampleSizeRegression** allows the user to run one or more previously registered scripts through the *Run/From script…* top-left menu item.

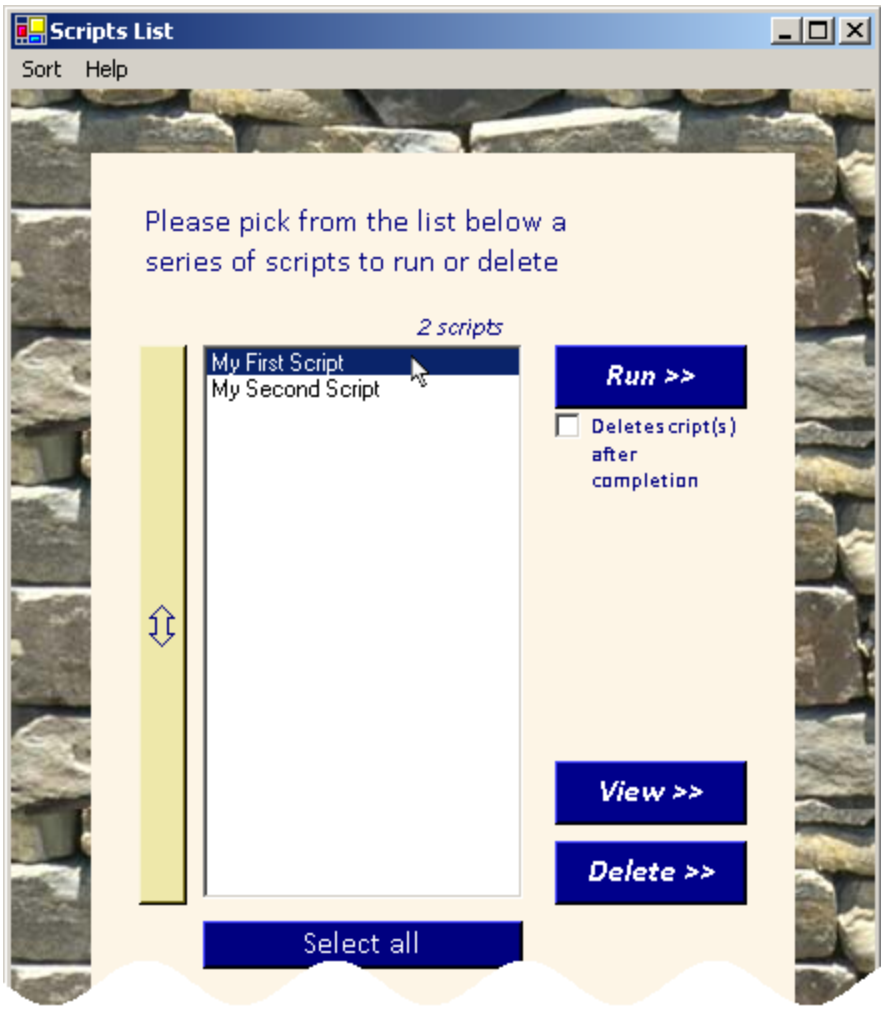


Running and submitting a script is useful when computing sample sizes for a number of variants (e.g., with different criteria or with different prior distributions), in that it eliminates delays between each run.

Select the scripts you wish to run now by clicking the appropriate script label(s) from the list and click the *Run>>* button.

Note the tick box below the *Run>>* button which can be ticked if you wish to delete the script files when the calculation is completed.

By default, the scripts are listed in order of entry date and time. Clicking the two-sided arrow button to the left of the list will reverse the order of the list.



4. Examples of running SampleSizeRegression

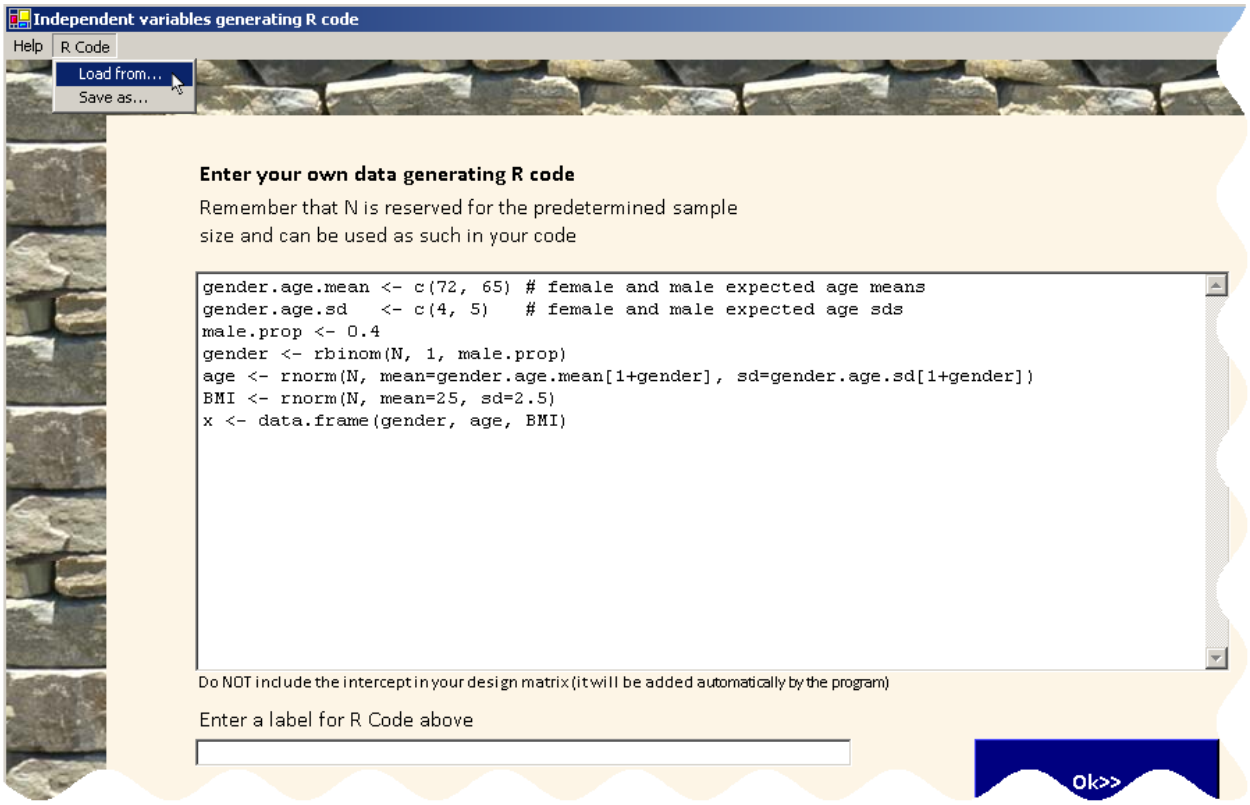
4.1 Sample size calculation

ICI on va illustrer l'utilisation du GUI pour répéter un exemple tiré de l'article.

4.2 Running sample size calculations on problems where X is not multivariate normally distributed

The independent variables $X_i = (X_{i1}, X_{i2}, \ldots, X_{iP})$ do not necessarily come from a multivariate normal distribution, as discussed in Section 2.1. In many applications to medicine and other fields of application, categorical variables will be present, meaning that the multivariate normal density is not appropriate. As the algorithm needs to sample from the X_i distribution (again, see Section 2.1), you will be asked to provide R code to sample from the X_i distribution.

Following the second form, where you indicate that X_i does not come from a multivariate normal distribution, you will need to enter/load your R data-generating code: you can load your R code through the top-left menu item or enter (or cut and paste) your R code into the main text box of the form illustrated below.



The generating-data R code from example above is reproduced below to ease readability:

```
gender.age.mean <- c(72, 65) # female and male expected age means
gender.age.sd   <- c(4, 5)   # female and male expected age sds
male.prop <- 0.4
gender <- rbinom(N, 1, male.prop)
age <- rnorm(N, mean=gender.age.mean[1+gender], sd=gender.age.sd[1+gender])
BMI <- rnorm(N, mean=25, sd=2.5)
x <- data.frame(gender, age, BMI)
```

The above code will sample values for three independent variables, namely gender, age and BMI. The proportion of males in the study is expected to be 40% and the age for women is expected to be 72 years old on average, with an SD of 4, while men's age is expected to be slightly lower with mean 65 years old and with an SD of 5. BMI is expected to be normally distributed (with mean 25 and s.d. 2.5) for both men and women.

As can be seen in the R code above, the fourth line uses the reserved variable name N, used for the sample size. N is a reserved variable name in the sense that it should NOT be used in your code for any variable other than sample size. The object defined on the last line of your code (x, in the above example) should be a data frame containing each variable to be used in the linear regression model. Obviously, your R code should be thoroughly tested before you use it in **SampleSizeRegression** as no debugging for this code is done by the program itself.

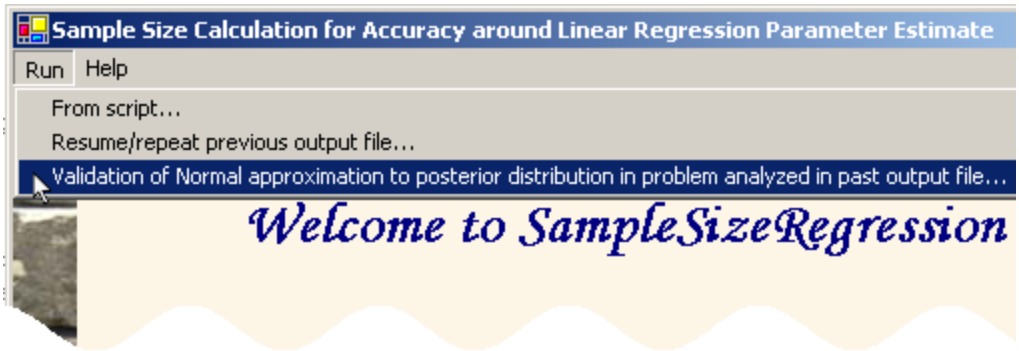
We suggest that you enter a label in the text box under the label *Enter a label for R code above*: doing so will make the use of your R code only one click away the next time you run **SampleSizeRegression**.

Note that the use of class variables with $K > 2$ classes in **SampleSizeRegression** is possible only through the use/definition of $K - 1$ dummy variables in the R data-generating code. Categorical variables not defined in this way will be treated as continuous variables. This limitation arises from the way WinBUGS treats categorical data.

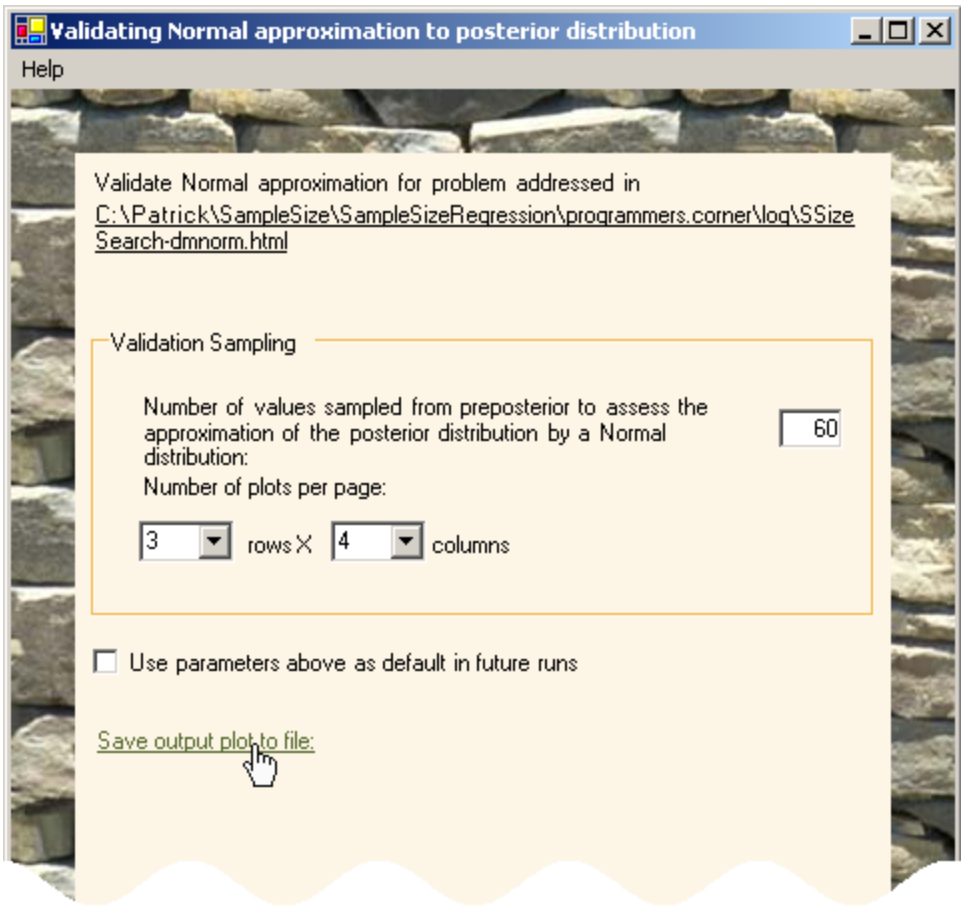
4.3 Validating the normal approximation to the posterior density of the regression parameter of interest

The calculation of HPD interval length or coverage for the parameter of interest (one of the slopes in the regression model) is based on the normal approximation of its marginal posterior distribution. It is hence a good idea to validate the appropriateness of that approximation in the context of your problem for your final sample size.

From the initial form, select the *Run/Validation of Normal approximation…* top-left menu item.



The next form allows the user to enter the number of values sampled from preposterior to assess the normal approximation for the posterior density of the parameter under study.



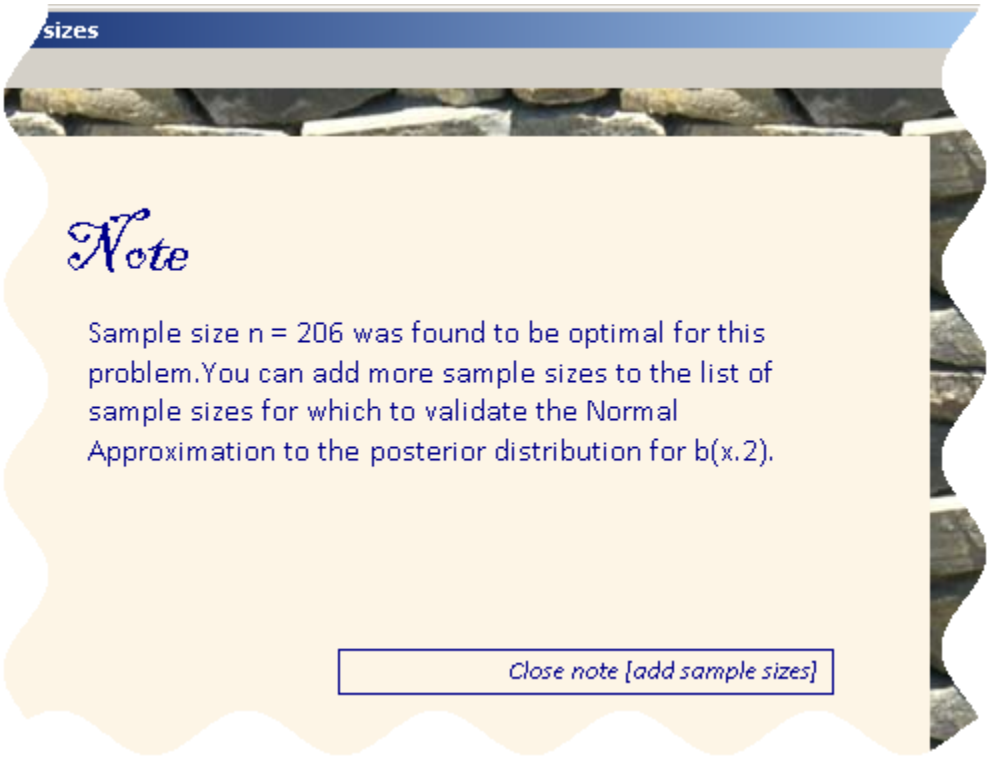
Even though a large preposterior sample size (of the magnitude of thousands of samples) was used in the original problem, this time the sample does not need to be very large to form an opinion on the appropriateness of the normal approximation.

We suggest to run it for 60 samples by default, but this can of course be changed. Remember, however, that a histogram of the values obtained in the MCMC WinBUGS program run will be drawn for each sample, with the distribution function of the best-fitting normal density superimposed. This means we have to monitor and save the values of the parameter of interest for each WinBUGS iteration, which is demanding in terms of both computer time and memory. By default, each page of the pdf output file will display 3 rows and 4 columns of histograms, but this can also be changed in the above form.

The trace OF XX and YY will also be saved, allowing the user to monitor the appropriateness of the chosen number of burn-in and monitored iterations, among other options.

Click the Save output plot to file item in the lower section of the form to select the pdf output file location: make sure not to overwrite an already existing pdf file. When done, click the *Ok>>* button.

The next form displays the optimal sample size obtained for this problem.



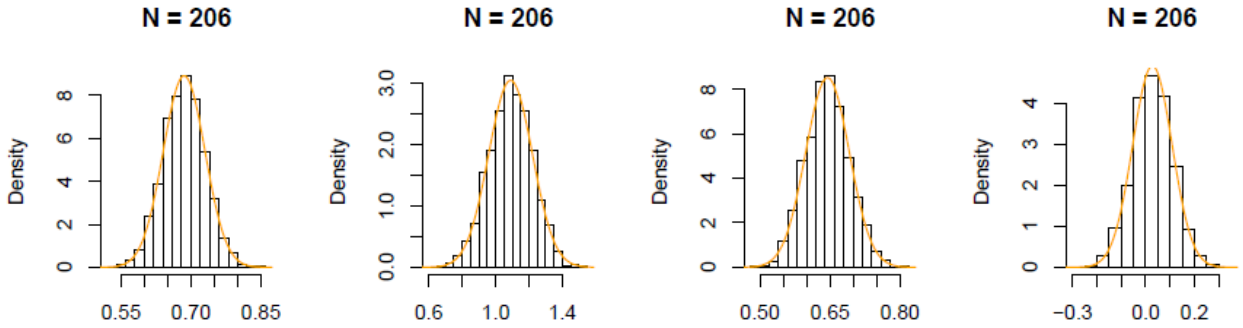
By default, the normal approximation validation check will be run for that optimal sample size only, but you can also add additional sample sizes, should you consider sampling more or less subjects than indicated and wish to validate the normal approximation for these alternative sample sizes as well. Click the *Close note [add sample sizes]* button and then the *Next>>* button to proceed with optimal sample size only.

A Problem Reviewal form will be displayed, allowing you to modify the parameters entered. Some parameters, however – such as number of burn-in and monitored iterations – are not modifiable as it would not make sense to check the convergence and mixing with technical parameters different from those used in the original problem.
Click *[Run now] Proceed>>* when ready.

Upon completion, a form will pop-up, offering you links to the (now modified) main html output file, the .pdf file with the normal approximation superimposed on the histograms and the traces OF XYZ.

A link to the normal approximation validation check pdf output file will be added to the main html output file. Below is an excerpt of the normal approximation validation check pdf output file: the four figures display, respectively, the histogram for β_1 values sampled in the WinBUGS

MCMC run for four samples with $N = 206$: the superimposed orange lines show the distribution function of the best-fitting normal distributions and all show a more than decent fit, which should reassure the user of both the appropriateness of the normal approximation for the posterior distribution for $\beta_1$and for the appropriateness of the optimal sample size returned, given the prior information at hand. Each histogram represents the results from one sample of X_i's, $i = 1, 2, \ldots, N$.

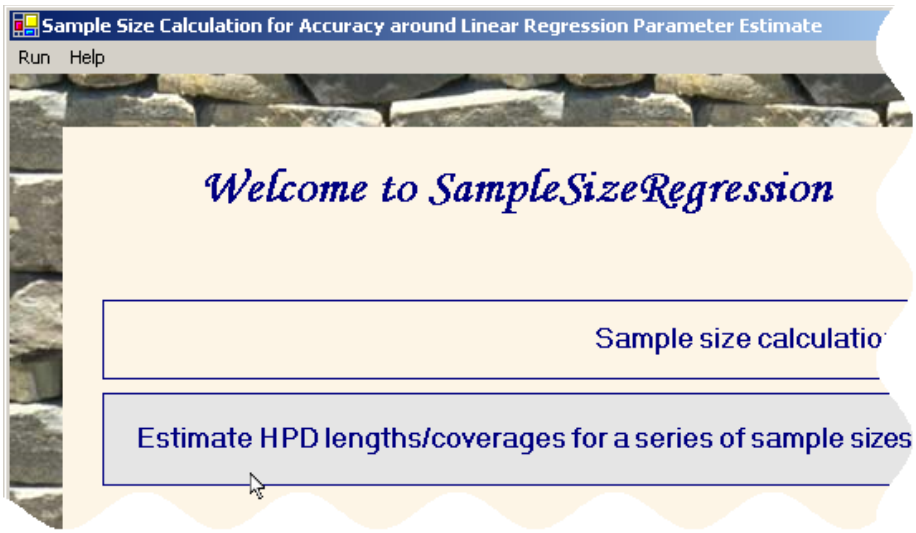


ICI peut-etre presenter des traces et les commenter sommairement??

4.4 Running SampleSizeRegression with pre-determined sample sizes

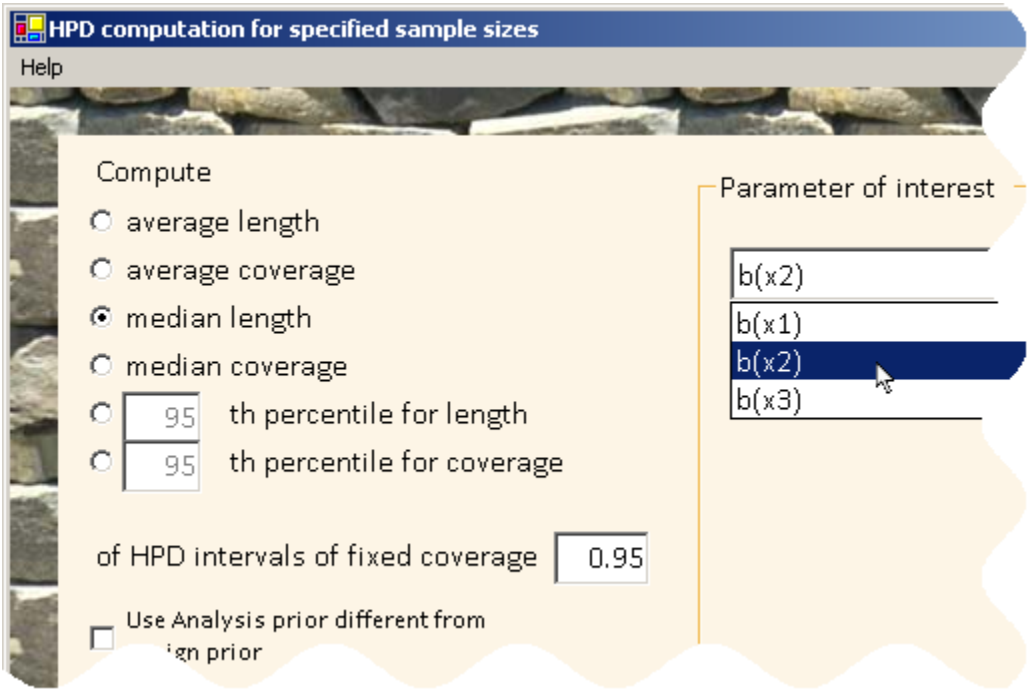
SampleSizeRegression can also be used to estimate a given outcome (e.g. the average coverage of HPD regions of fixed length) for fixed sample sizes. Indeed, suppose one knows it will not be possible to recruit more than XYZ subjects in a study, but would still like to obtain an idea, beforehand, of the coverage of HPD regions of fixed length XYZ in the same context as that illustrated in section X.Y.

From the initial form of **SampleSizeRegression**, click the second box.

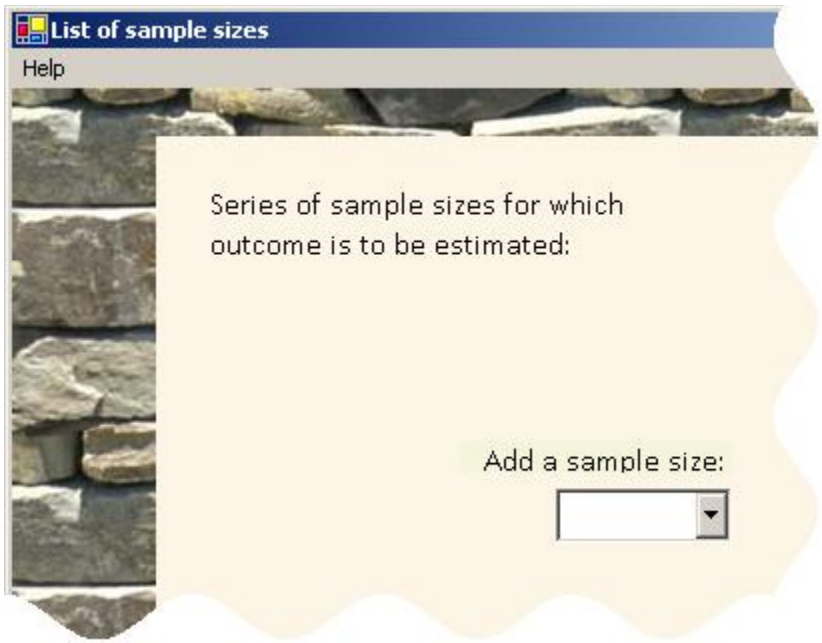


The next forms are identical to those presented in section 3 and are used to enter the number of independent variables, the prior distribution for $\boldsymbol{\beta}$, etc. We therefore omit discussion of these steps here, proceeding to the next form where there is a difference from the procedure for sample size calculations.

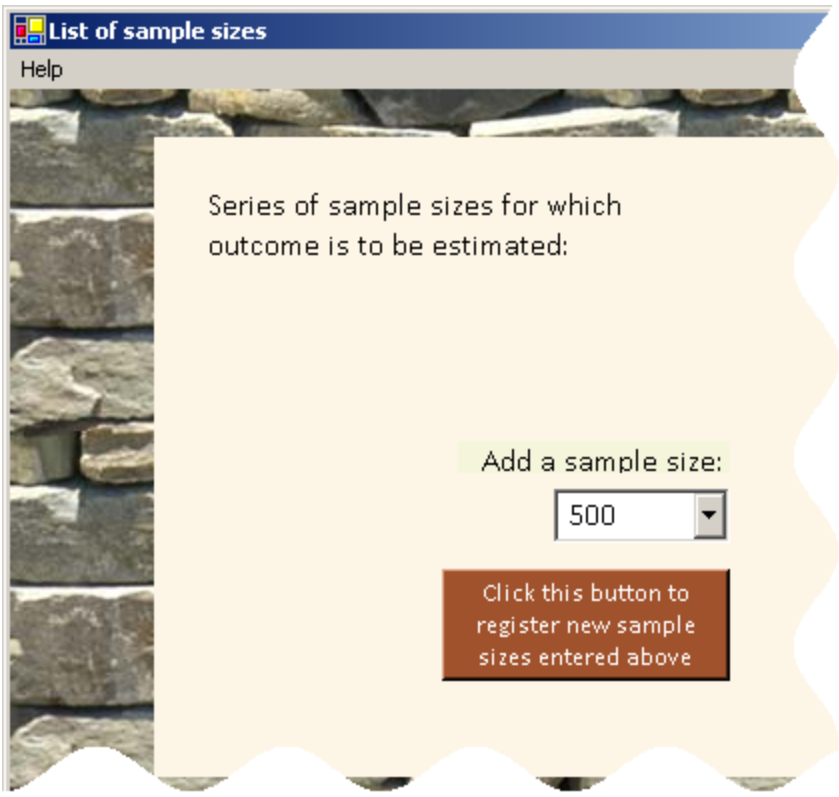
The next form allows the user to pick the outcome of interest and specify either of the fixed HPD length or coverage, as well as to pick the regression parameter of interest.



The next form is used to specify the sample sizes for which the above-specified outcome is to be estimated.

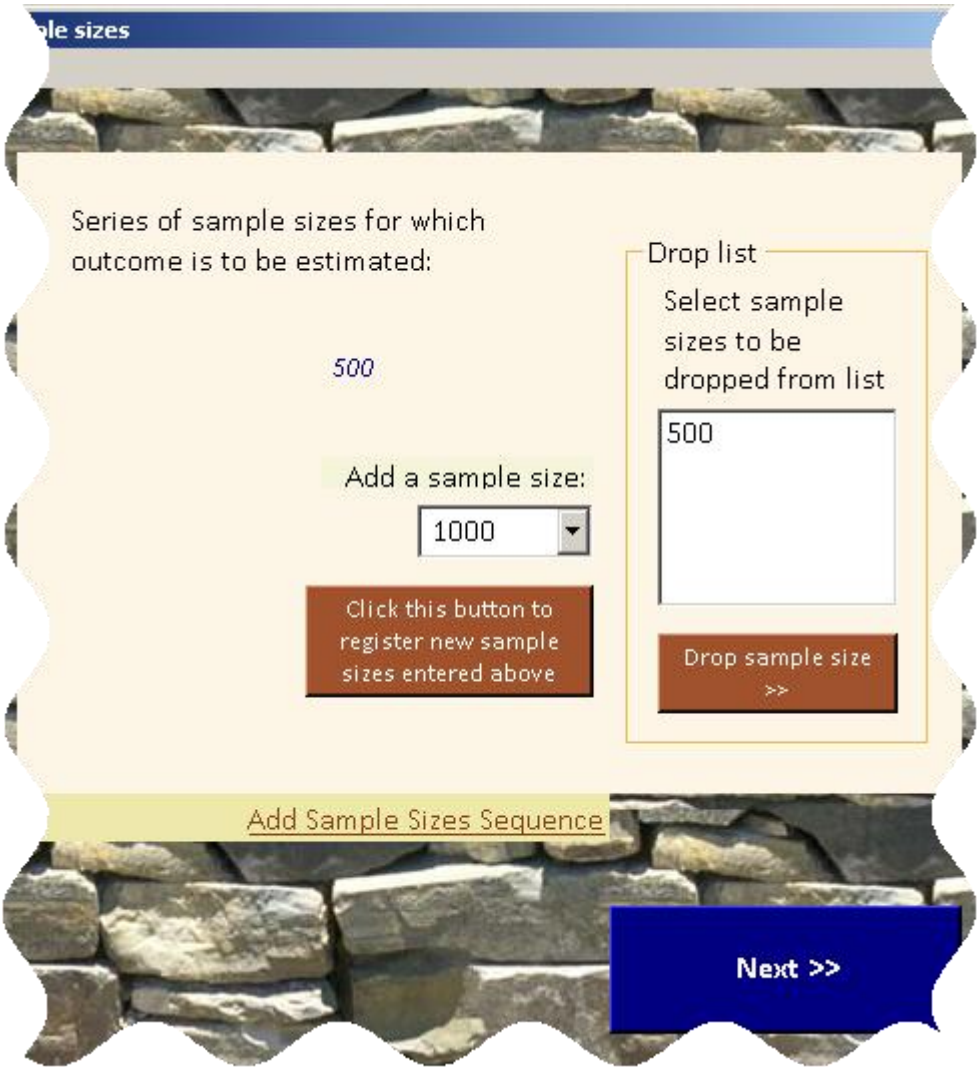


Enter 500 in the *Add a sample size* text box and click the button underneath to register this sample size.

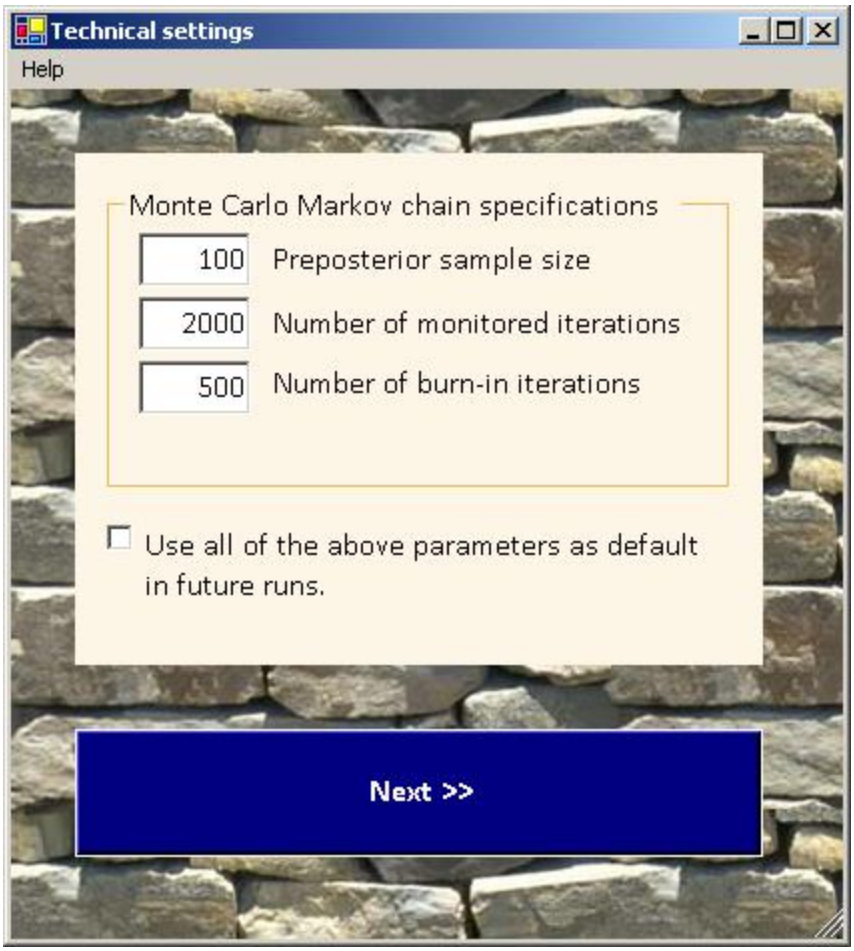


Proceed the same way for each sample size of interest.

Click *Next>>* when done.



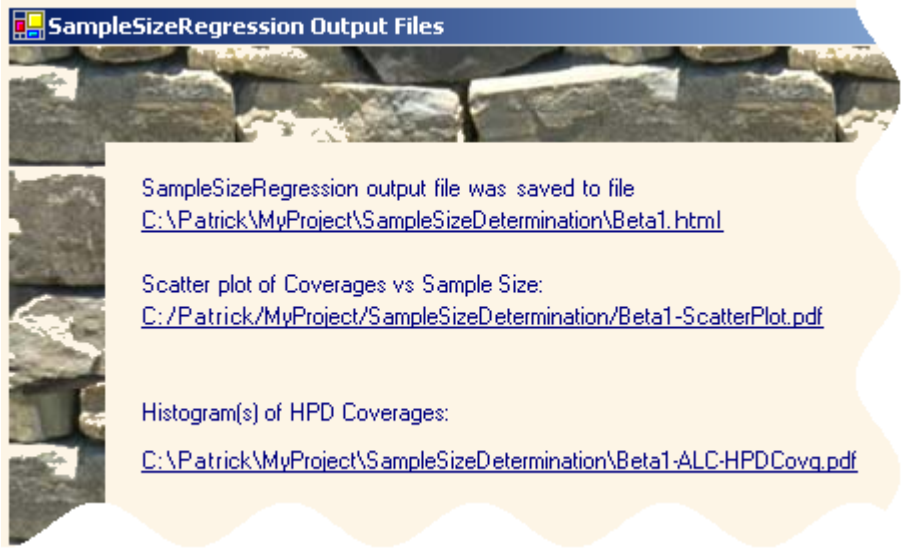
The next form – Technical settings – is simplified compared to that presented in section 3, as the sample size search algorithm settings are irrelevant here.



The last form is the Problem Reviewal form, similar to that presented in section X.Y.

When done, **SampleSizeRegression** will pop-up a form with links to the main html output file and to secondary graphical pdf output files.

Whether this form will be opened full size or minimized (then only noticeable in the taskbar) depends on the corresponding option in the Problem Reviewal form.

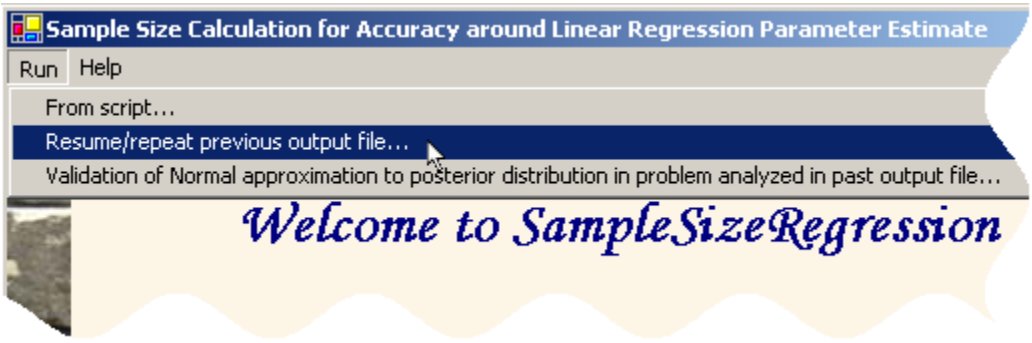


When running multiple scripts through the *Run from script*… top-left menu from the initial form, the above final form with links to output files will NOT appear.

4.5 Resuming/repeating a previous analysis

It is easy to resume or repeat a problem that was already run with **SampleSizeRegression**. This can be useful if one wants to modify the prior distribution previously used for one or more parameters, or to check how the sample size may change under a different sample size criterion, for example.

The top-left menu from the initial form offers the option to resume or repeat a problem that was previously run.



4.5.1 Resuming a previous analysis

Some errors, such as due to a computer crash, to the need to reboot your system while **SampleSizeRegression** was still running, or if you inadvertently stopped a WinBUGS script that was launched by **SampleSizeRegression,** might lead you to want to continue running a previous instance of the program. Regardless of the reason for an interruption of the program, an error message such as the one reproduced below will be printed at the bottom of the html output file.

Error message

Program started on Wed Jul 31 16:22:10 2013.

Program aborted on Wed Jul 31 16:22:22 2013.
Cannot read WinBUGS output file (C:\Users\Patrick.Belisle\AppData\Local\Temp\SSReg\20130731-162210-wbstats.txt)

To resume calculations started above, open SampleSizeRegression and browse to this file through the top-left menu item **Complete/repeat past output file…** from the initial form.

To resume the problem, select the *Resume/repeat previous output file…* top-left menu item from the initial **SampleSizeRegression** form and load the incomplete html output file: **SampleSizeRegression** will then resume from where it stopped (or was stopped!).

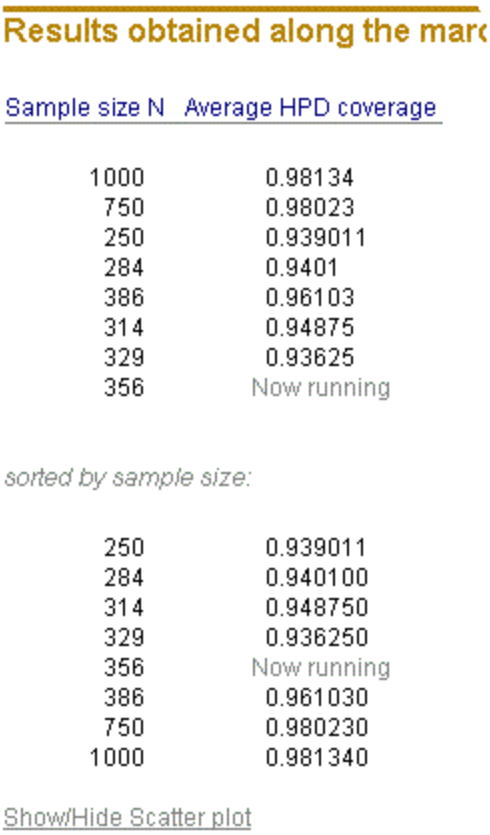
If the same error occurs again, it might be a sign of a problem inherent to **SampleSizeRegression**. Please do not hesitate to contact us if you cannot think of a solution to the problem or if the error message is unclear. See our contact email address at the end of this document.

4.5.2 Repeating a previous analysis

The *Resume/repeat previous output file…,* as its name indicates, can also be used to repeat an analysis previously done with **SampleSizeRegression**. If the original analysis was done with fair precision (i.e., with a large number of samples from preposterior and a decent number of monitored iterations), there is not much interest in actually repeating the same analysis (even though you would almost surely obtain at least slightly different results by doing so, as the whole process is subject to Monte Carlo error). This option, however, can advantageously be used to rerun an analysis with slight modifications, such as different hyperparameters for the prior distribution of one or more parameter, a different sample size criterion, or even for focusing the inference on a different regression parameter. When loading a **SampleSizeRegression** html output file that completed with success, this option will bring you directly to the final Problem Reviewal form which, as already seen earlier in this document, allows the user to modify almost every single problem description parameter.

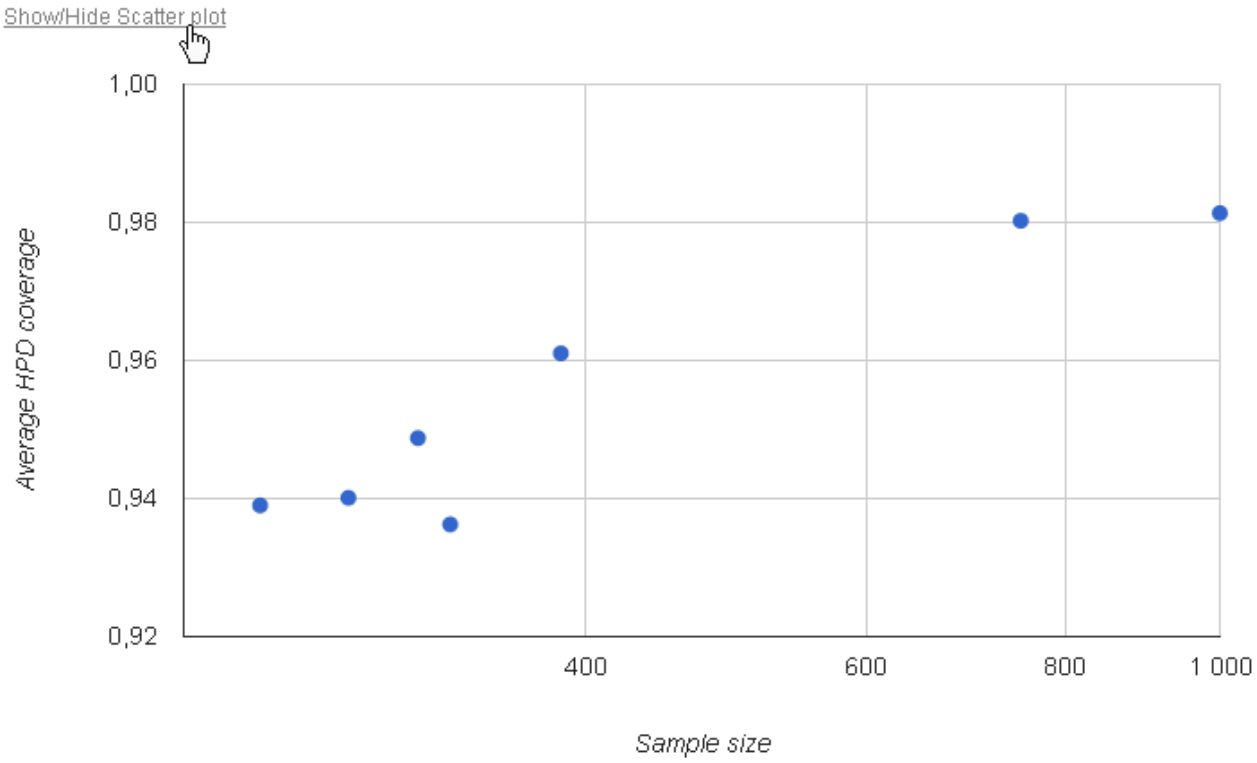
4.1 Iterative html output file updates

Whether you are running **SampleSizeRegression** for a sample size calculation or for outcome estimation for a series of predetermined sample sizes, you may be interested in having a look at intermediate results while **SampleSizeRegression** is running. The main html output file is updated after the outcome of interest is estimated for each sample size, and viewing it in your favorite browser is possible at any point in time.



The output section labeled **Results obtained along the march to optimal sample size** is divided into two subsections: the upper portion lists the series of sample sizes along with their corresponding estimated outcome, in the order in which they were run. This is followed by a subsection where the same results are listed in ascending order of sample size, which may be easier to follow. Note that if the outcome had to be re-estimated for a given sample size (along the search for optimal sample size), that sample size will appear two or more times in the upper section, once per re-estimation, while only the final estimate (which summarizes information from each intermediate outcome estimate) will appear in the lower (sorted) section.

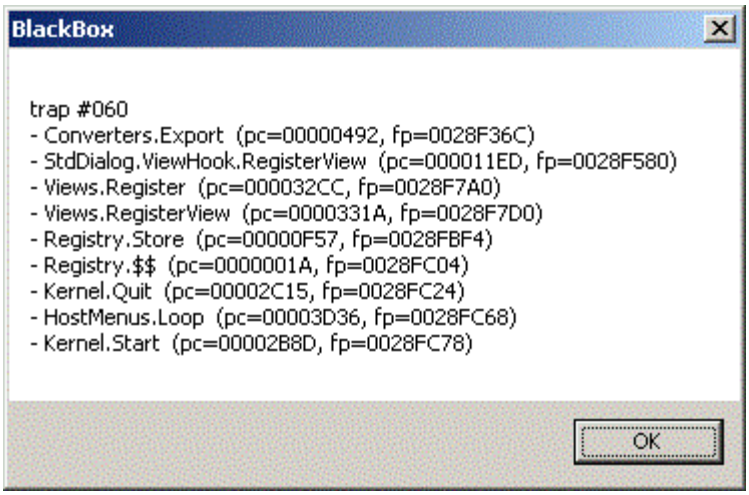
Finally, a link labeled *Show/Hide Scatter plot* opens a scatter plot (see example below) which helps visualize the relationship between outcome and sample size; it may even give enough information to the user with regards to the final sample size, even though the program has not formally reached convergence, and the user may decide to stop the program before it does converge. Note that when **SampleSizeRegression** finishes, a .pdf document presents the same scatter plot with a little bit more information.



5. Troubleshooting

5.1 Avoiding trap errors from permission settings on Windows 7 and Windows Vista platforms

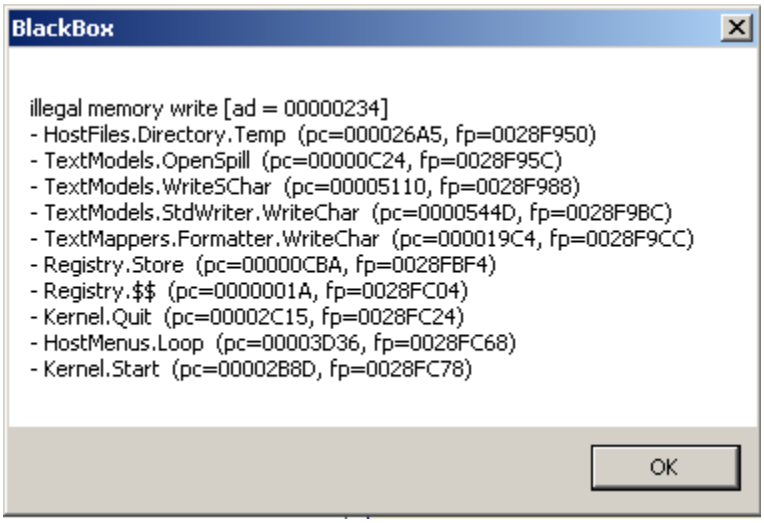
If you are working on a Windows 7 or Windows Vista platform and have run WinBUGS before, you may have already run into the cryptic **Trap #060** error message illustrated to the right. This is due to restricted write permissions in c:\Program Files, where you may have installed WinBUGS.



WinBUGS **must** be installed in a directory where you have write permissions (e.g.C:\Users\user name \Documents) for **SampleSizeRegression** to run smoothly.

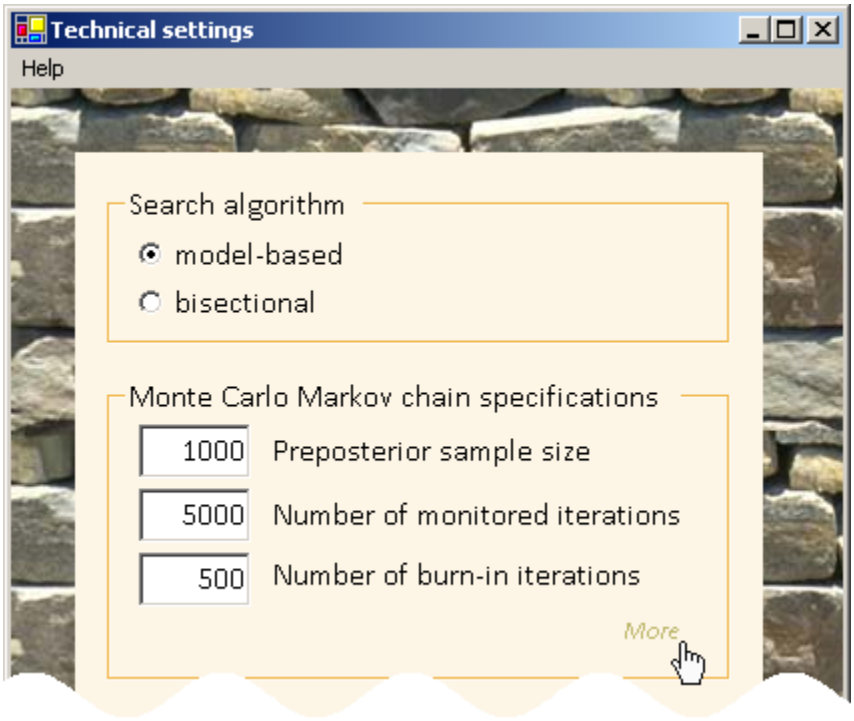
5.2 Illegal Vista platforms write

You may hit a **Illegal memory write** error message (prompted by WinBUGS) for high dimensional problems. This typically happens when WinBUGS tries to save results to disk (just before closing) when the number of monitored iterations is large; the work-around is to save intermediate results and clear WinBUGS memory after a certain number of iterations. By default, this is done after each 1000 iterations, but that may not be sufficient in complex problems.

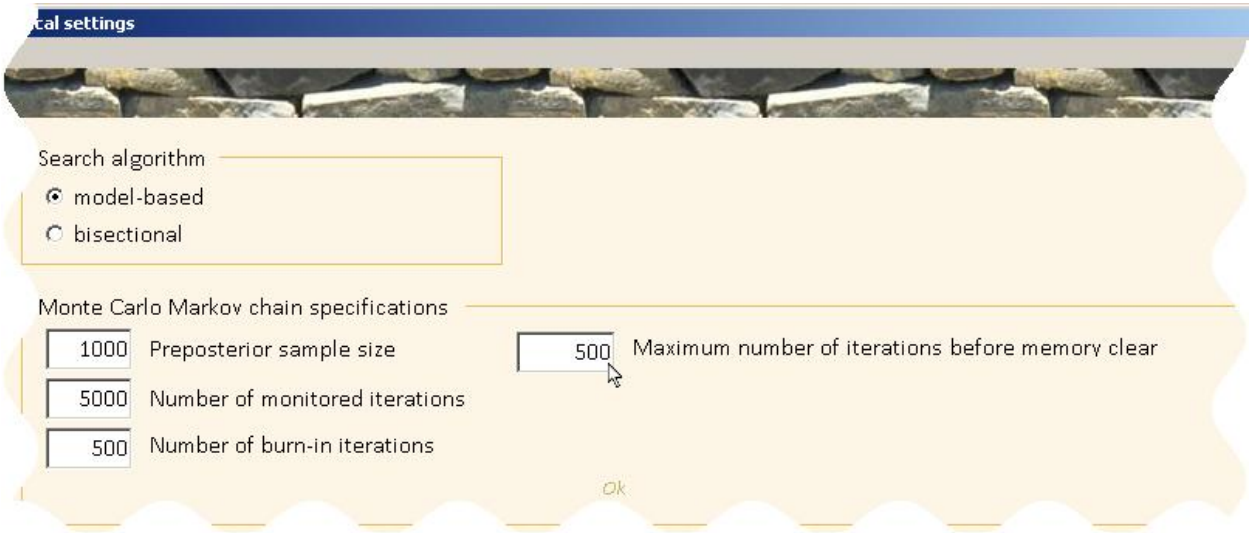


The necessary frequency of WinBUGS memory wash-outs depends on the agreement table size and your system environment. If wash-outs at every 1000 iterations are not sufficient, we suggest that you try with wash-outs at every 500 iterations, 200 iterations, and so on, until **SampleSizeRegression** runs to completion. Changing the maximum number of iterations before each memory clear is done through the *Technical settings* form, as shown below.

Click the *More* button in the Monte Carlo Markov chain specifications box.

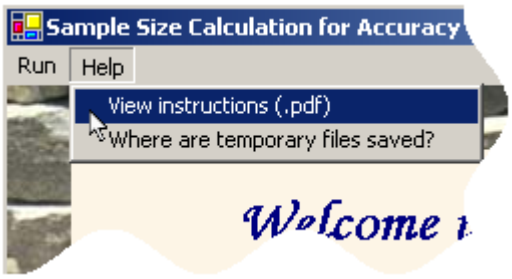


Then modify the value of *Maximum number of iterations before memory clear*.



6. Need help?

The present instruction manual is available from the top menu *Help* item on each form of **SampleSizeRegression** GUI.



Questions? Comments? Please send email to: lawrence.joseph@mcgill.ca

---

Other Bayesian software packages are available at

http://www.medicine.mcgill.ca/epidemiology/Joseph