

Bayesian Sample Size Determination for Linear and Logistic Regression in the Presence of Confounding and Measurement Error

Lawrence Joseph

Patrick Bélisle

July 30, 2019

Summary

Linear regression and logistic regression are the two most common methods employed in medical research to provide estimates of the effects of treatments or exposure variables on continuous or dichotomous outcomes, respectively, while adjusting for potential confounding variables. With small modifications to the basic models, these methods are also able to adjust for possible measurement error in covariates, either on the main exposure or a potential confounding variable. Despite their widespread use in the analysis of data, methods for designing studies that will eventually provide data for regression modelling are more rarely used. Most researchers either ignore sample size considerations entirely when planning such studies, or use simple methods that do not account for potential confounding and measurement error, both of which can have very large effects on sample size requirements. In this paper we present a method that calculates the sample size required to estimate the effect of any regression parameter to a desired degree of accuracy while adjusting for any expected confounding variables and/or measurement error. We employ both Bayesian and mixed Bayesian/likelihood sample size criteria so that our methods are applicable to studies whose data will eventually be analysed by either Bayesian or frequentist modelling. We present examples that both do and do not account for the presence of confounding and measurement error, illustrating the effect that these potential complications can have on the required sample sizes. A free software package that facilitates the use of our methods is available.

Key words: Bayesian methods; linear regression; logistic regression; measurement error; sample size determination; study design.

1 Introduction

Linear regression and logistic regression models are widely used in the analysis of data in order to estimate the effects of various independent variables on continuous or dichotomous outcomes, respectively. Although less often used in practice, with small adjustments to these models researchers can also adjust for any biases estimates in their estimates brought on by measurement error.

Both linear and logistic regression can account for confounding, providing estimates of main effects that are adjusted for the effects of confounding induced by any covariates included in the model. Since both confounding and measurement error are very commonly occurring in practice, it would seem important to account for both effects when planning any study, since larger sample sizes may be required in the presence of confounding or measurement error compared to if these problems are ignored in the modelling

A wide variety of sample size criteria have been proposed. These range from power calculations and methods based on confidence interval widths from a frequentist viewpoint (reviewed by Lemeshow et al 1990 and Desu and Raghavarao 1990) to Bayesian versions of these same criteria (Spiegelhalter and Freedman 1986, Joseph, du Berger, and Bélisle 1997), reviewed by Adcock (1997) and Wang and Gelfand (2002). Bristol (1989) showed that sample sizes based on interval widths are not directly related to those based on power, so that sample sizes guaranteeing high power may not be sufficient for accurate estimation. It is therefore important that the sample size methods match the eventual analysis. Since reporting interval estimates is preferable to hypothesis testing in most practical circumstances including regression (Gardner and Altman 1988), sample sizes should be based upon interval widths rather than power of hypothesis tests. Decision theoretic criteria have also been proposed (for example, Pallay 2000), but while interesting in theory these methods are difficult

to implement in practice, in large part because realistic loss functions are difficult to derive and are highly specific to a given application. These criteria will therefore not be discussed further here.

Frequentist sample size methods depend on accurate point estimates of the required inputs such as effect sizes and standard deviations, but these are typically not accurately known at the design stage of any study. It is therefore advantageous to consider Bayesian methods, where prior densities not only allow for uncertainty in the inputs, but incorporate this uncertainty into the sample size requirements. It is also easy to extend Bayesian regression models to include measurement error.

In this paper we develop Bayesian methods to determine sample size requirements to accurately estimate any linear or logistic regression parameter including the possibility of confounding variables and measurement error. In particular, we develop methods to derive sample sizes using the Average Coverage Criterion, (ACC), the Average Length Criterion (ALC) as well as the Modified Worst Outcome Criterion (MWOC) both accounting and not for confounding and measurement error.

The outline of this paper is as follows. Section 2 reviews various Bayesian sample size criteria based on highest posterior density (HPD) credible interval lengths. HPD intervals are optimal in that they will lead to the smallest possible sample sizes for a given desired length and coverage probability. Specific methods for calculating the sample sizes defined by the criteria in Section 2 in the case of linear and logistic regression are given in Section 3. Sample sizes from a series of prototypic examples are given in Section 4, comparing the change in sample size with and without consideration of confounding and measurement error. We end with a discussion in Section 5.

2 Bayesian sample size criteria

Let $\theta \in \Theta$ be the parameter of interest, $f(\theta)$ the prior distribution, $x = (x_1, \dots, x_n)$ the data of sample size n , \mathcal{X} the data space, $f(x)$ the predictive marginal distribution of the data, and $f(\theta|x)$ the posterior distribution for θ given data x . Then

$$\begin{aligned} f(x) &= \int_{\Theta} f(x|\theta)f(\theta)d\theta, \quad \text{and} \\ f(\theta|x) &= \frac{f(x|\theta)f(\theta)}{f(x)}. \end{aligned} \tag{1}$$

In a Bayesian approach to sample size determination inputs include the desired length of the HPD interval l , the chosen credible interval level $1 - \alpha$, and a prior distribution for θ . Since the data are not known at the design stage, we need to integrate or maximize over the predictive distribution of the data given by (1), which leads to the following three criteria.

2.1 Average coverage criterion

For a fixed posterior interval length l , one can determine the sample size by finding the smallest n such that the equation

$$\int_{\mathcal{X}} \left\{ \int_a^{a+l} f(\theta|x, n)d\theta \right\} f(x)dx \geq 1 - \alpha \tag{2}$$

is satisfied. This average coverage criterion (ACC) ensures that the mean coverage of posterior credible intervals of length l , weighted by $f(x)$, is at least $1 - \alpha$.

2.2 Average length criterion

For a fixed posterior credible interval coverage of $1 - \alpha$, one can also determine the sample size by finding the smallest n such that

$$\int_{\mathcal{X}} l'(x, n)f(x)dx \leq l, \tag{3}$$

where $l'(x, n)$ is the length of the $100(1 - \alpha)\%$ posterior credible interval for data x , determined by solving

$$\int_a^{a+l'(x,n)} f(\theta|x, n)d\theta = 1 - \alpha$$

for $l'(x, n)$ for each value of $x \in \mathcal{X}$. As above, a can be chosen to give highest posterior density intervals or symmetric intervals, which coincide for symmetric unimodal densities.

This average length criterion (ALC) ensures that the mean length of $100(1 - \alpha)\%$ posterior credible intervals weighted by $f(x)$ is at most l .

2.3 Modified worst outcome criterion

Cautious investigators may not be satisfied with the “average” assurances provided by the ACC and the ALC criteria. Therefore, a conservative sample size can also be determined by the smallest n satisfying the equation

$$\inf_{x \in \mathcal{S}} \left\{ \int_a^{a+l(x,n)} f(\theta|x, n)d\theta \right\} \geq 1 - \alpha, \quad (4)$$

where \mathcal{S} is a suitably chosen subset of the data space \mathcal{X} . For example, this modified worst outcome criterion (MWOC) ensures that if \mathcal{S} consists of the most likely 95% of the possible $x \in \mathcal{X}$, then there is 95% assurance that the length of the $100(1 - \alpha)\%$ posterior credible interval will be at most l .

One must also consider a “design prior,” that is, the prior density to plug into (1) to generate the set of all possible data x , which is not necessarily the same as the “analysis priors” used in the calculation of the posterior density. For example, one can use informative design priors to predict future data, but non-informative analysis priors to let the data alone determine the posterior density. In the next section, we will present methods for applying the above criteria to selecting sample sizes for studies involving linear and logistic regression parameters.

3 Sample size methods for linear and logistic regression parameters with and without confounding and measurement error

In this section we consider how to apply the above sample size criteria in the context of both linear and logistic regression, both with and without adjustment for confounding and/or measurement error.

3.1 Sample size methods for linear regression

Let θ be the vector of linear regression parameters including the intercept and the residual variance. We assume that all parameters are *a priori* independent, with normal densities for each intercept and regression parameter priors, and a uniform prior for the residual standard deviation. In the absence of measurement error, this completely defines the posterior density when it is combined with the standard linear regression likelihood function (please see the detailed equations in the appendix). In the case of measurement error, we assume normal errors with mean zero around each independent variable measured with error, adding only one new parameter, the standard deviation of the measurement error. We again use a uniform prior for this standard deviation term.

The above information completely defines the likelihood function and priors for our model, both with and without measurement error. Therefore all we need to complete our analysis are the design priors. We assume these will, take the same form as the analysis priors, although each parameter may or may not take on different numerical values compared to their analysis prior counterparts. With all necessary functions now defined, applying the Bayesian sample size criteria of Section 2 is straightforward, and is implemented by MCMC.

Given a bounded range that contains the correct sample size, it can be found through

a bisectional search algorithm, for example as described by Thisted (1988). In practice, a starting sample size is selected, and the above criterion is checked. The sample size is then increased or decreased according to whether the criterion is above or below the required threshold, continuing until the optimal sample size is found.

3.2 Sample size methods for logistic regression

Let θ be the vector of logistic regression parameters. We again assume that all parameters are *a priori* independent, with normal densities for each intercept and regression parameter priors. In the absence of measurement error, this completely defines the posterior density when it is combined with the standard logistic regression likelihood function (please see details in the appendix). In the case of measurement error, we again assume normal errors with mean zero around each independent variable measured with error, adding only one new parameter, the standard deviation of the measurement error. We again use a uniform prior for this standard deviation term.

The above information completely defines the likelihood function and priors for our model, both with and without measurement error. Therefore all we need to complete our analysis is the design priors. We assume these will take the same form as the analysis priors, although each parameter may or may not take on different numerical values compared to their analysis prior counterparts. With all necessary functions now defined, applying the Bayesian sample size criteria of Section 2 is straightforward, and is implemented by MCMC. As above we find the final sample size through a bisectional search algorithm.

A user-friendly R package called “RegressionSampleSize” that implements all of the above methods are available from the first author’s website, at

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

We next use this software to determine sample sizes for various scenarios that may occur in the planning of studies where parameters of interest will be estimated with linear or logistic regression.

4 Sample sizes for prototypic scenarios

We first look at a typical example when estimating a linear regression parameter, followed by an application of our methods to logistic regression parameters. In each case we calculate the required sample size to ensure the desired degree accuracy using the ACC criterion. Of course, similar results will be produced when using any of the other Bayesian sample size criteria. In each case we compare results with no confounding and no measurement error to results with one or both of these problems.

4.1 Example for linear regression

Suppose one is designing a study that will be analysed by linear regression, and where the outcome will be predicted by two independent variables, a main variable of interest, say x_1 which may possibly be measured with error, and a second possibly correlated variable which also may be measured with error, x_2 . We will assume that the true population distribution of these two variables are both $N(0,1)$. We would like to accurately estimate the beta coefficient of the first variable, β_1 while adjusting for the possible confounding effects from the second variable, also accounting for any measurement error. We will use a 95% HPD credible interval to estimate β_1 , whose length we would like to be at most 0.8. The true value of all beta parameters is assumed to be 1, and the true residual variance was taken to be 1, with priors uniformly distributed on the interval $[0.1, 10]$. What sample size is required using the ACC criterion?

If the two independent variables are uncorrelated and there is no measurement error, then the sample size we will need is 990. If there is a weak correlation of 0.1 between x_1 and x_2 then the sample size is increased to 1111, and larger correlations of 0.5 and 0.9 increase the required sample sizes to 1314 and 5183, respectively. We can see that correlation between the variables can have a very large effect on the sample size.

The above sample sizes are what is required if there is no measurement error. Suppose that both x_1 and x_2 are both measured with error, with mean error of zero, but with standard deviation uniformly distributed within the interval $[0.1, 0.3]$. The sample size when there is no correlation between the two variables now increases to 1090, and the sample size changes to 1127, 1501, and 4312 when there is correlation of 0.1, 0.5, and 0.9, respectively. Therefore we see that correlation has a larger effect on the required sample size compared to measurement error in this example.

4.2 Example for logistic regression

We will now turn to the design of study whose data will be analysed by logistic regression.

Suppose one is designing a study that will be analysed by logistic regression, and where the outcome will be predicted by two independent variables, a main variable of interest, say x_1 which may possibly be measured with error, and a second possibly correlated variable which also may be measured with error, x_2 . We will assume that the true population distribution of these two variables are both $N(0,1)$. We would like to accurately estimate the beta coefficient of the first variable, β_1 while adjusting for the possible confounding effects from the second variable, also accounting for any measurement error. We will use a 95% HPD credible interval to estimate β_1 , whose length we would like to be at most 1 on an odds ratio scale. The true value of all beta parameters is assumed to be 0.1, with priors selected such that 95% of their prior density is between 0.2 and 2.5, again on the odds ratio scale.

What sample size is required using the ACC criterion?

If the two independent variables are uncorrelated and there is no measurement error, then the sample size we will need is 1269. If there is a correlation of 0.5 between x_1 and x_2 then the sample size is increased to 1680. Similar to linear regression, we can see that correlation between the variables can have a large effect on the sample size.

The above sample sizes are what is required if there is no measurement error. Suppose that both x_1 and x_2 are both measured with error, with mean error of zero, but with standard deviation uniformly distributed within the interval $[0.2, 0.3]$. The sample size when there is no correlation between the two variables now changes to 1156, and the sample size changes to 1482 when there is correlation of 0.5. Therefore we once again see that correlation has a larger effect on the required sample size compared to measurement error in this example.

5 Discussion

Many sample size criteria have been proposed from both hypothesis testing and interval estimation viewpoints. Since the advent of fast desktop computers and the development of MCMC algorithms for Bayesian analysis, increasing numbers of statisticians are using Bayesian methods. Accompanying these advances have been important modeling advances in bias correction, such as adjusting analyses for measurement error. Although these methods have increased flexibility in modelling, there are fewer design methods to plan studies where these new analytic techniques will be used. This paper addresses this gap in providing Bayesian methods to determine sample sizes for linear or logistic regression modelling, including adjusting sample size requirements for possible confounding variables and measurement error in covariates.

As both of the prototypic examples we presented in Section 4 show, sample size require-

ments can considerably change if confounding and/or measurement error is accounted for at the design stage. The accompanying software, available from the first author's webpage at <http://www.medicine.mcgill.ca/epidemiology/Joseph/>, allows for easy implementation of all methods discussed in this paper.

Appendix

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 program¹ 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 search 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 $X_i = (X_{i1}, X_{i2}, \dots, X_{iP})$ are to be collected, with or without **measurement error**, on N subjects as well as the outcome (or response) variable Y_i , $i = 1, 2, \dots, N$, and that the relationship between the outcome and the P independent variables will be modeled through either

- a) a linear model, that is,

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

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

when the outcome is binary.

In both linear and logistic model, we assume a multivariate normal prior distribution on the regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)'$. 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

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

where

\mathbf{Y}	is the response vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_N)'$
\mathbf{X}	is the design matrix
$f(\mathbf{Y} \boldsymbol{\beta}, \mathbf{X}, \sigma)$	is the likelihood of the data \mathbf{Y} – that is, a multivariate normal distribution with mean $\mathbf{X}\boldsymbol{\beta}$ 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 $\prod_{k \in S} \prod_{i=1}^N \int_{\sigma_k} f(X_{ik} X_{ik}^*, \sigma_k) f(\sigma_k) d\sigma_k = 1$
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 β 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 β analysis prior distribution). The WinBUGS Markov Chain Monte Carlo process leads to the approximation of the posterior distribution of the regression parameters β .

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.

Sample Size Calculations for Linear and Logistic Regression

Select the regression model of your choice by clicking the appropriate button below

Linear Regression	Logistic Regression
<p>The model is $Y \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_q)$</p> <p>where β is the vector of regression coefficients, \mathbf{X} is the design matrix, \mathbf{I}_q is the $q \times q$ identity matrix and σ is the residual standard error</p>	<p>The model is $Y_i \sim \text{Bernoulli}(\pi_i)$</p> <p>where $\text{logit}(\pi_i) = \beta_0 + \sum_{k=1}^p \beta_k x_{ik}$</p> <p>$\beta_0$ is the intercept β_1, \dots, β_p are the regression coefficients and $x_{ik} \in \{-1, \dots, N\}; k=1, \dots, p$</p>

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.

Bayesian Sample Size Calculations for Linear and Logistic Regression

Help

Linear Regression

The model is $Y \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_q)$

where $\boldsymbol{\beta}$ is the vector of regression coefficients,
 \mathbf{X} is the design matrix,
 \mathbf{I}_q is the $q \times q$ identity matrix and
 σ is the residual standard error

with $\sigma \sim \text{Uniform}(a, b)$
where a = and b =

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 $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ 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

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

$x_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

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

Variable names

Please fill in alternative variable names or labels (optional)

	Variable #	Variable name/label
▶	1	age
	2	x.2
	3	x.3

Ok>>

Variable Mean

	Variable name	Mean
	x.1	(null)
▶	x.2	(null)
	x.3	(null)

Covariance matrix

		x.1	x.2	x.3
▶	x.1	(null)	(null)	(null)
	x.2	(null)	(null)	(null)
	x.3	(null)	(null)	(null)

Variance of
x.1

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*.

Precision/Covariance Matrix

Edit Precision/Covariance matrix Help

- Round-off Covariance Matrix entries ▶
- Set Covariance Off-Diagonal elements to 0
- Clear Covariance Matrix entries
- Upper triangular matrix elements can be visited ▶
- Invert matrix and display Precision matrix
- Switch to Precision matrix entry

Variable Mean

	Variable name	Mean
	x.1	(null)
▶	x.2	(null)
	x.3	(null)

Covariance matrix

		x.1	x.2
▶	x.1	(null)	(null)
	x.2	(null)	(null)
	x.3	(null)	(null)

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

Precision matrix

	x.1	x.2	x.3
x.1	(null)	(null)	(null)
x.2	(null)	(null)	(null)
x.3	(null)	(null)	(null)

Precision of x.2

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.

By default, each variable is considered to be measured exactly.
 If one or more variables is measured with error, select it from the list below and enter the prior distribution for the standard deviation of the measurement error.

Variable

x.1	<input type="radio"/> is known exactly	modeled as measured value $\sim N(\text{true value, s.d.} = \sigma)$ with $\sigma \sim \text{Uniform}(a, b)$ where a = <input type="text"/> and b = <input type="text"/>
x.2	<input checked="" type="radio"/> is measured with error	
x.3	<input type="radio"/> is known exactly	

The next form (below) is for entry of the parameters of the multivariate normal prior distribution for the β 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.

Parameter Prior Means

	Parameter	Mean
▶	Intercept	(null)
	beta(x.1)	(null)
	beta(x.2)	(null)
	beta(x.3)	(null)

Covariance matrix

	Row name	Intercept	beta(x.1)	beta(x.2)	beta(x.3)
▶	Intercept	(null)	(null)	(null)	(null)
	beta(x.1)	(null)	(null)	(null)	(null)
	beta(x.2)	(null)	(null)	(null)	(null)
	beta(x.3)	(null)	(null)	(null)	(null)

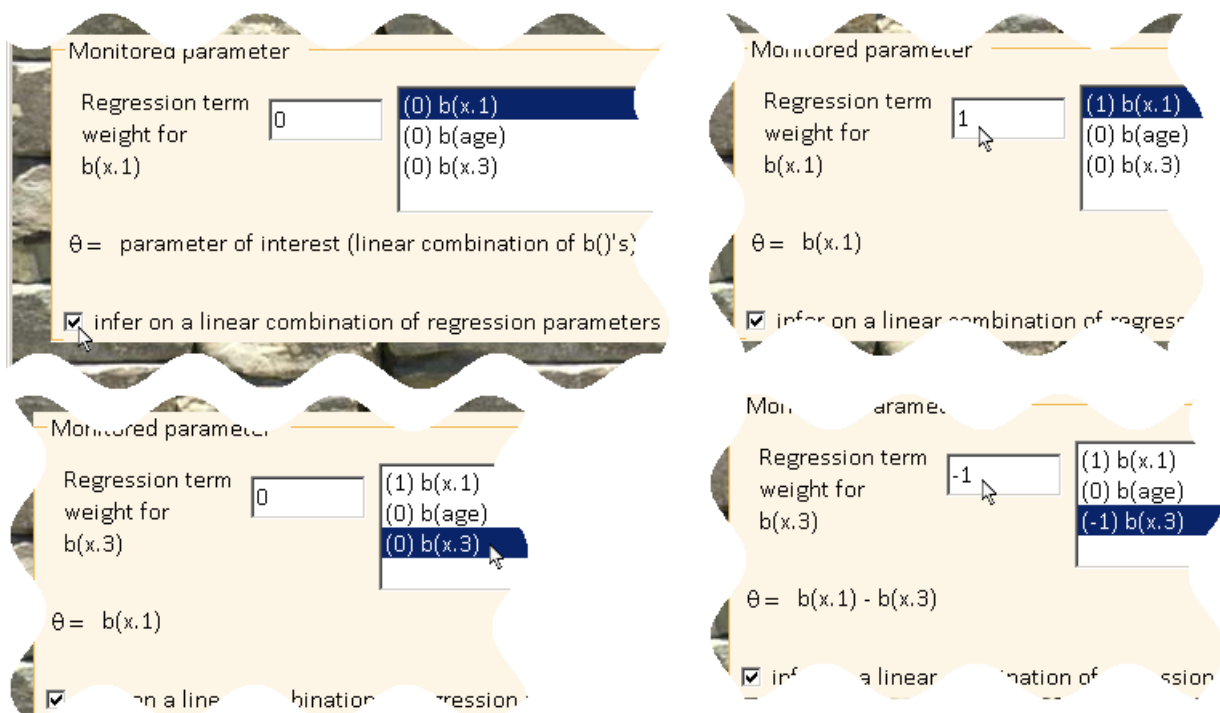
If sample size determination was selected in the initial window, the next form (below) allows selecting one of the ten sample size criteria available. This window is also used to specify the fixed or target (depending on criterion selected) HPD length and coverage for the parameter of interest, also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

The screenshot shows a software window titled "Criterion" with a "Help" button. The window is divided into three main sections:

- Criterion:** A list of five criteria with radio buttons:
 - ALC** Average length criterion
 - ACC** Average coverage criterion
 - MLC** Median length criterion
 - MCC** Median coverage criterion
 - MWOC** Modified worst outcome criterionA checkbox labeled "Use Mixed Bayesian/Likelihood approach" is present and unchecked.
- Targets:** Two input fields:
 - HPD length: [Empty text box]
 - HPD coverage: [0.95]
- Monitored parameter:** A dropdown menu with a list of parameters:
 - b(x.2)
 - b(x.1)
 - b(x.2)** (highlighted)
 - b(x.3)

When the Mixed Bayesian/Likelihood approach is chosen in form above, the prior distributions for β 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 β): 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 and 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.

Monitoring parameter

Regression term weight for b(age)

$\theta = 10 * b(\text{age})$

infer on a linear combination of regression r

(0) b(x.1)
(10) b(age)
(0) b(x.3)

The image shows a software interface with a yellow background and a scalloped border. At the top, it says "Monitoring parameter". Below that, there is a label "Regression term weight for" followed by a text input field containing the number "10" and a mouse cursor. To the right of the input field is a dropdown menu with three options: "(0) b(x.1)", "(10) b(age)", and "(0) b(x.3)". The second option, "(10) b(age)", is highlighted in blue. Below the input field and dropdown, the equation $\theta = 10 * b(\text{age})$ is displayed. At the bottom, there is a checkbox that is checked, followed by the text "infer on a linear combination of regression r".

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.

Technical settings

Help

Search algorithm

- model-based
- bisectional

Monte Carlo Markov chain specifications

1000 Preposterior sample size

2000 Number of monitored iterations

4000 Number of burn-in iterations

[More](#)

Sample size

200 Starting value

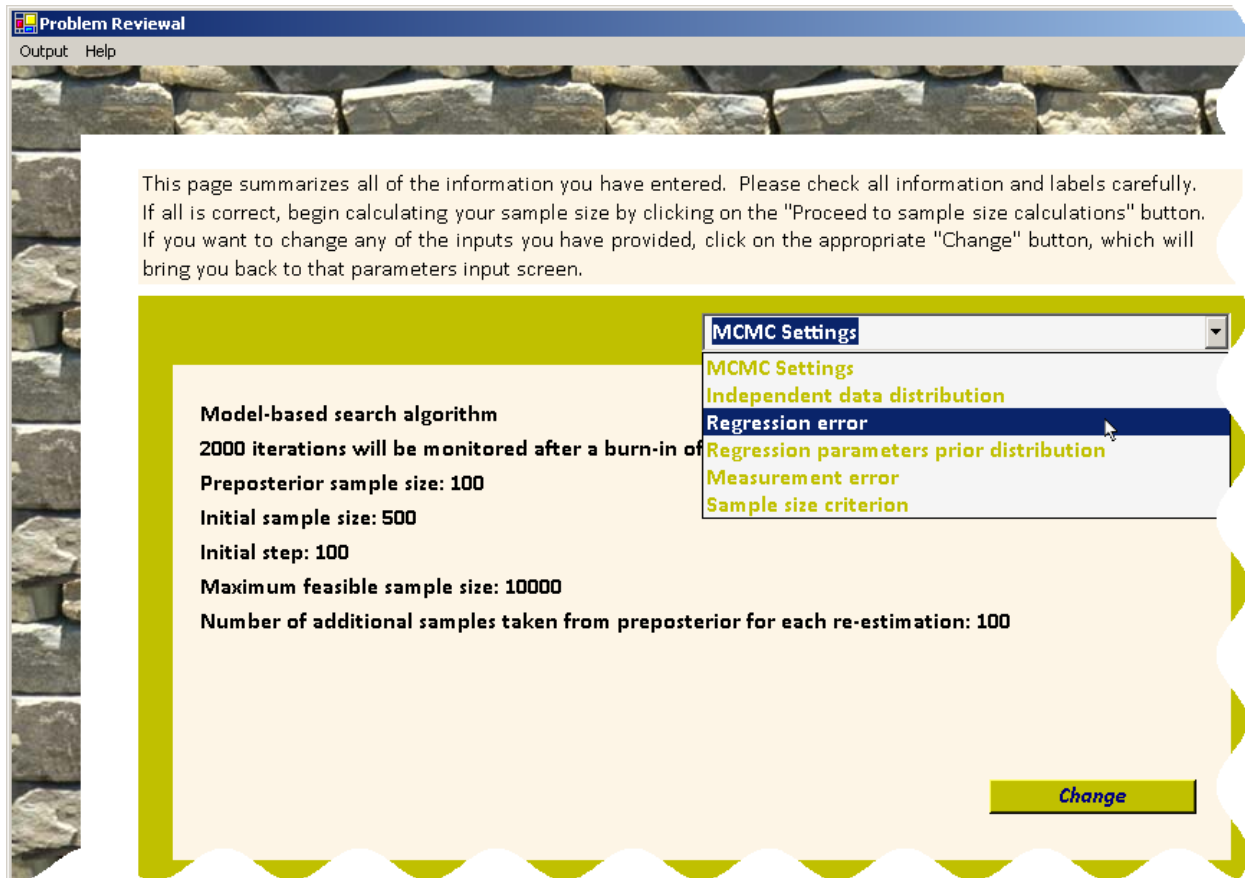
100 Initial step

15000 Maximum feasible size

20 Lower sample size to estimate
(protection against WinBUGS crashes)

Use all of the above parameters as default in future runs.

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.

parameter(s) prior distribution
Covariance matrix Help

95% prior interval for
Pr{Outcome = 1 | x.cont = avg, x.dichotomous = 0}

from: 0.25 to: 0.6

95% prior interval for Odds Ratio
exp(b(x.2))

from: 0.8 to:

Odds Ratio
Summary / Pick list
exp(b(x.1)): 0.6 to 3
exp(b(x.2)): 0.8 to ?
exp(b(x.3)): ? to ?

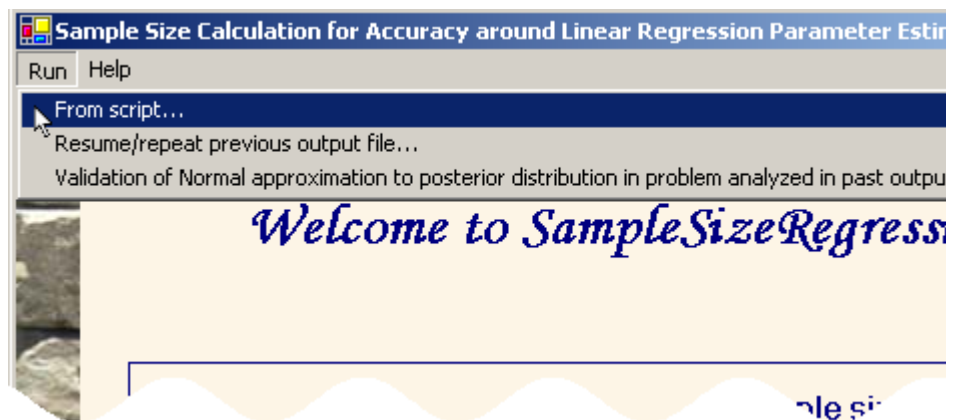
The top 2 boxes are used to enter the 95% interval for the prior probability of a positive outcome ($Y_i = 1$) when the continuous independent variables are equal to their average and the dichotomous independent variables are 0's. The other values to enter are the 95% prior intervals for each of the 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 , $i=1, 2, \dots, P$, and (μ_0, σ_0) such that $\exp(\mu_0 + z_{0.025} \sigma_0)$ are equal to the logit() value of the two endpoints of the 95% prior interval for $\Pr\{Y = 1 \mid \text{continuous X variables} = \text{their average and dichotomous X variables} = 0\}$.

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.

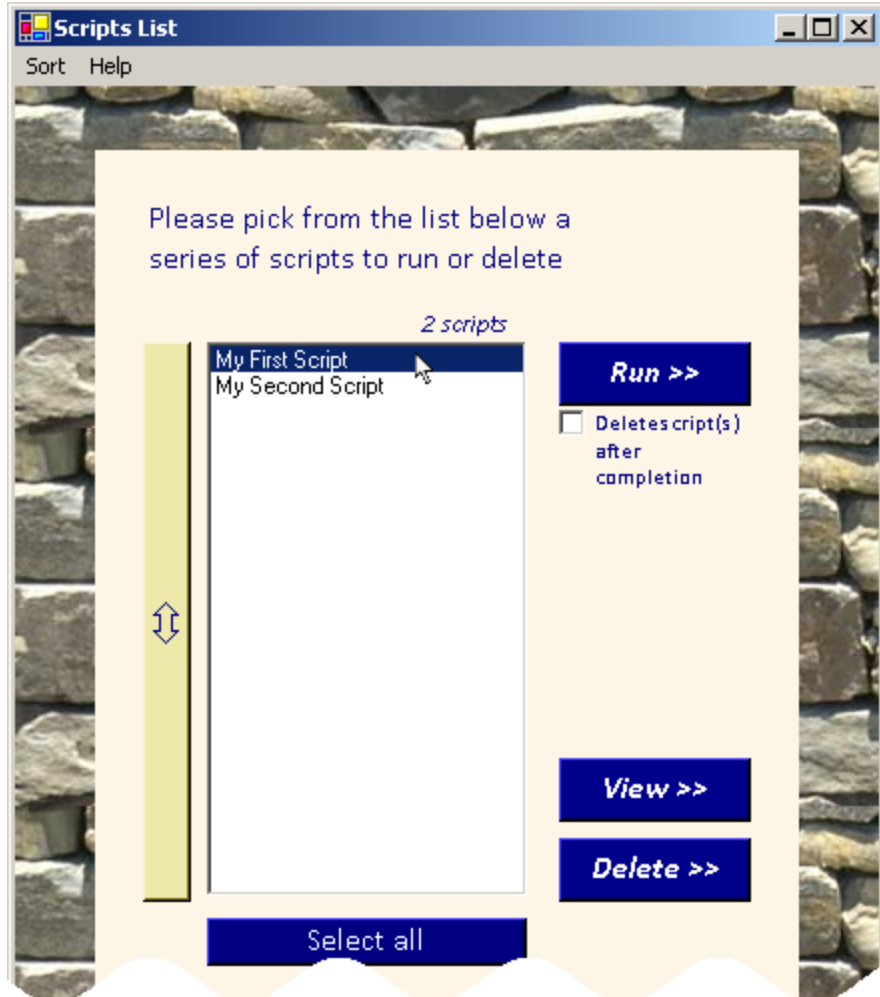


Running and submitting a script is useful when computing sample sizes for a number of variants (e.g., with different criteria or with different prior distributions), in that it eliminates delays between each run.

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

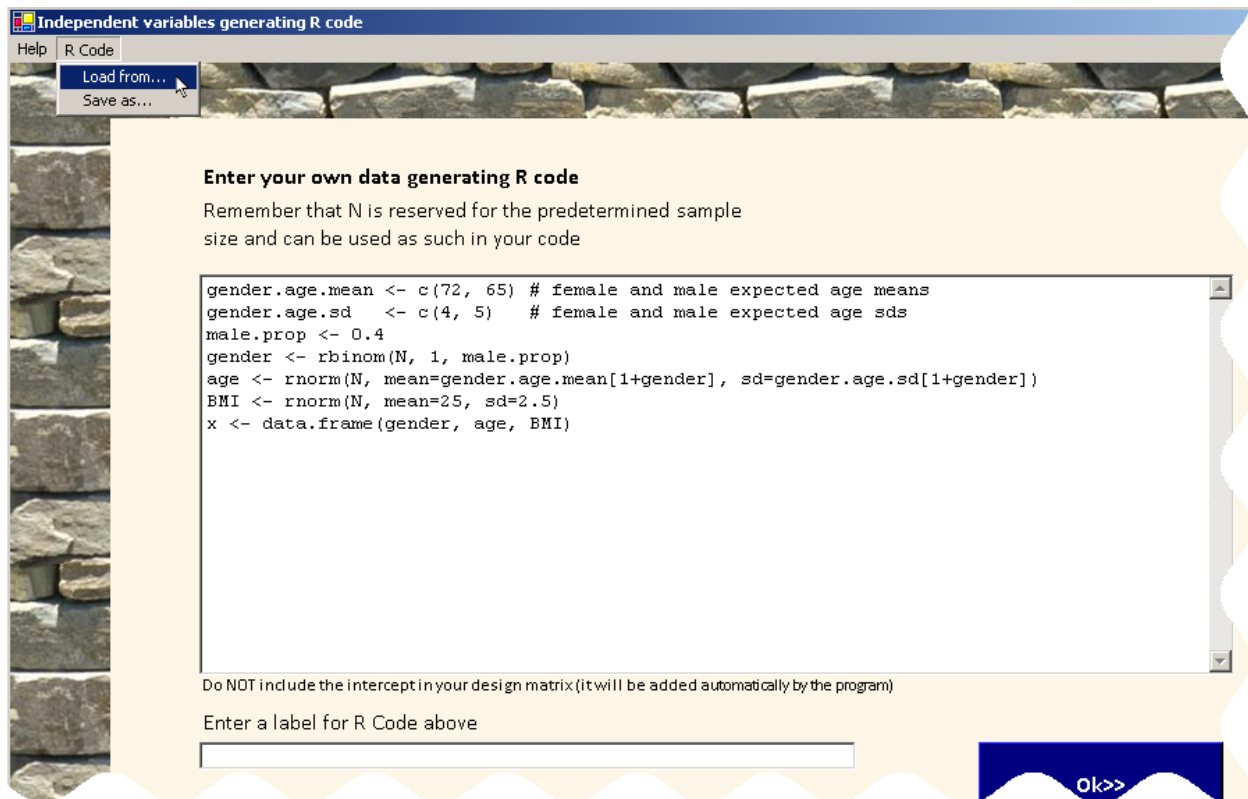
4.1 Sample size calculation

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 \mathbf{X} is not multivariate normally distributed

The independent variables $X_i = (X_{i1}, X_{i2}, \dots, X_{iP})$ 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 screenshot shows a web browser window with the title "Independent variables generating R code". The browser's address bar shows "R Code". The page has a menu with "Help", "Load from...", and "Save as...". The main content area has a yellow background with a stone wall border on the left. It contains the following text:

Enter your own data generating R code
Remember that N is reserved for the predetermined sample size and can be used as such in your code

```
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)
```

Do NOT include the intercept in your design matrix (it will be added automatically by the program)

Enter a label for R Code above

Ok>>

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.

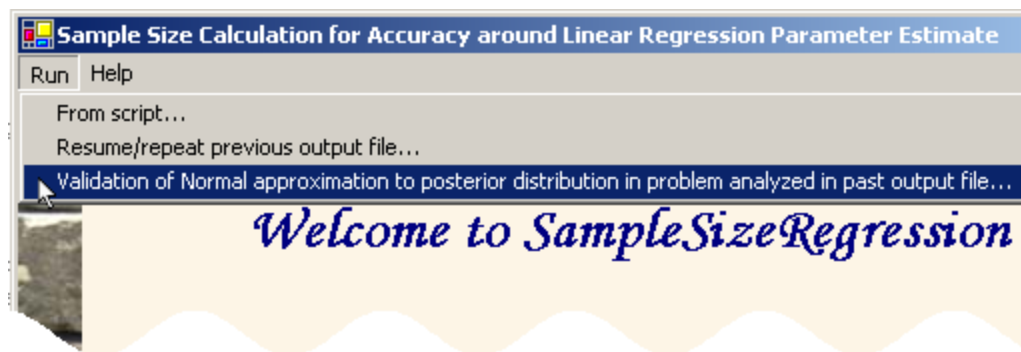
We suggest that you enter a label in the text box under the label *Enter a label for R code above*: doing so will make the use of your R code only one click away the next time you run **SampleSizeRegression**.

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.

Validating Normal approximation to posterior distribution

Help

Validate Normal approximation for problem addressed in
<C:\Patrick\SampleSize\SampleSizeRegression\programmers.cornet\log\SSizeSearch-dmnorm.html>

Validation Sampling

Number of values sampled from preposterior to assess the approximation of the posterior distribution by a Normal distribution:

Number of plots per page:

rows X columns

Use parameters above as default in future runs

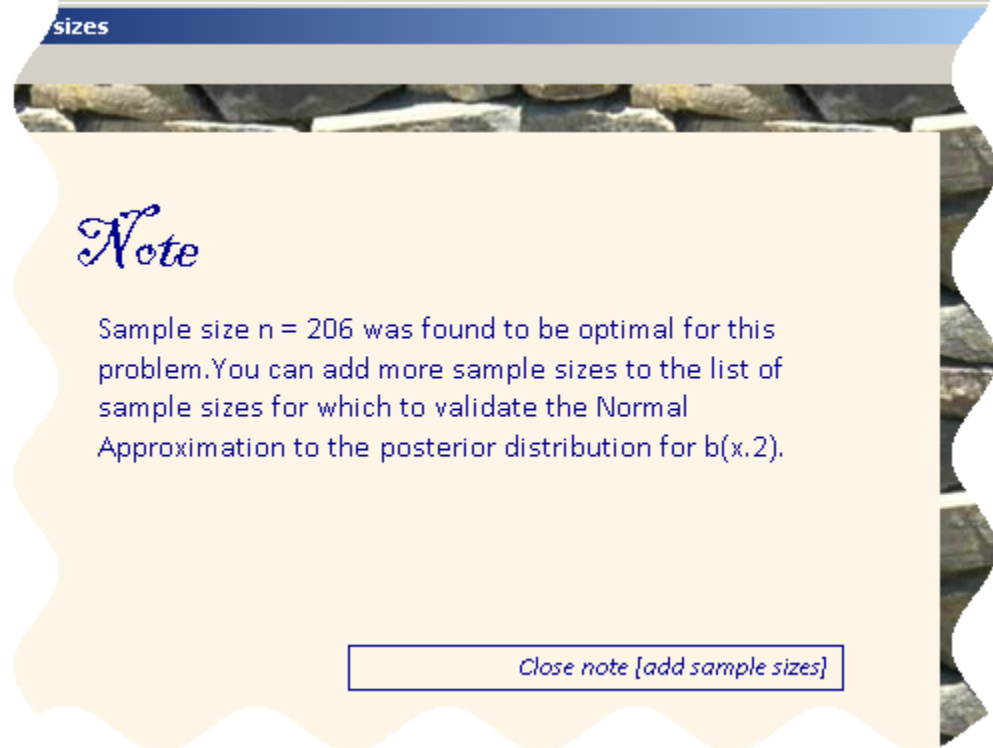
[Save output plot to file:](#)

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.



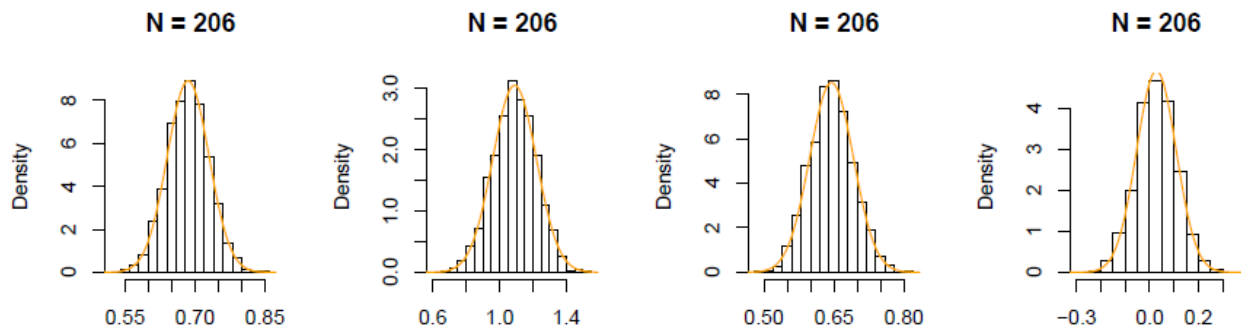
By default, the normal approximation validation check will be run for that optimal sample size only, but you can also add additional sample sizes, should you consider sampling more or less subjects than indicated and wish to validate the normal approximation for these alternative sample sizes as well. 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 β_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, $i = 1, 2, \dots, N$.

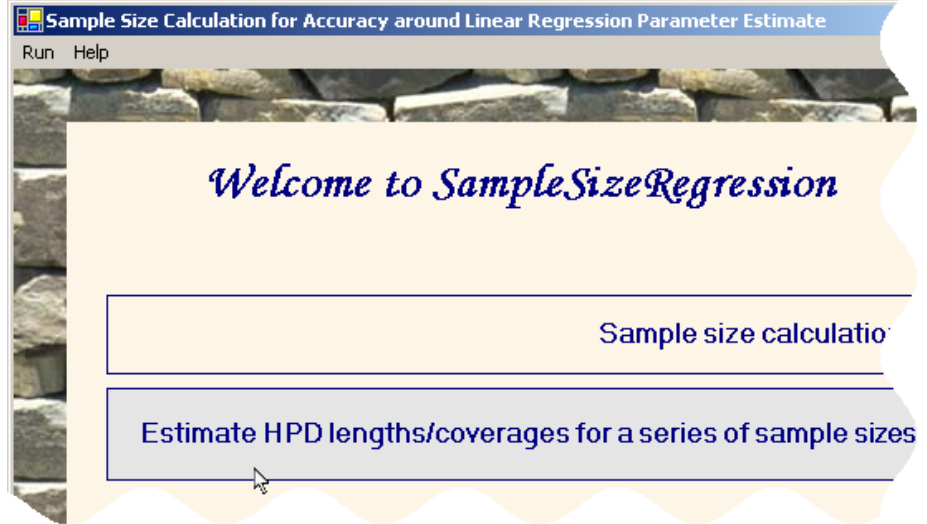


ICI peut-etre presenter des traces et les commenter sommairement??

4.4 Running SampleSizeRegression with pre-determined sample sizes

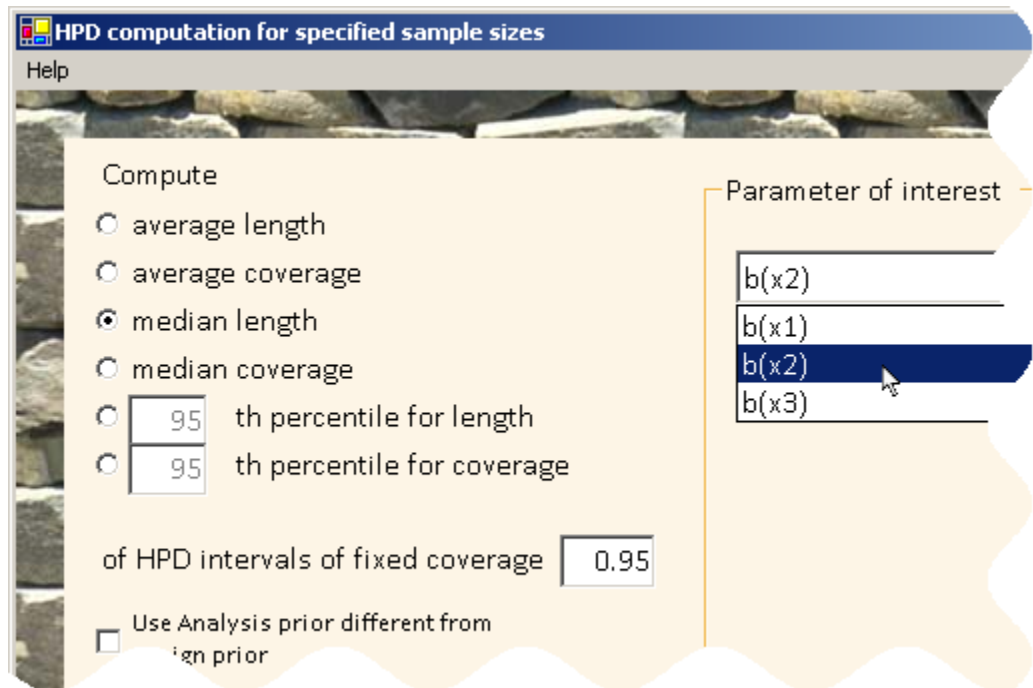
SampleSizeRegression can also be used to estimate a given outcome (e.g. the average coverage of HPD regions of fixed length) for fixed sample sizes. Indeed, suppose one knows it will not be possible to recruit more than XYZ subjects in a study, but would still like to obtain an idea, beforehand, of the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

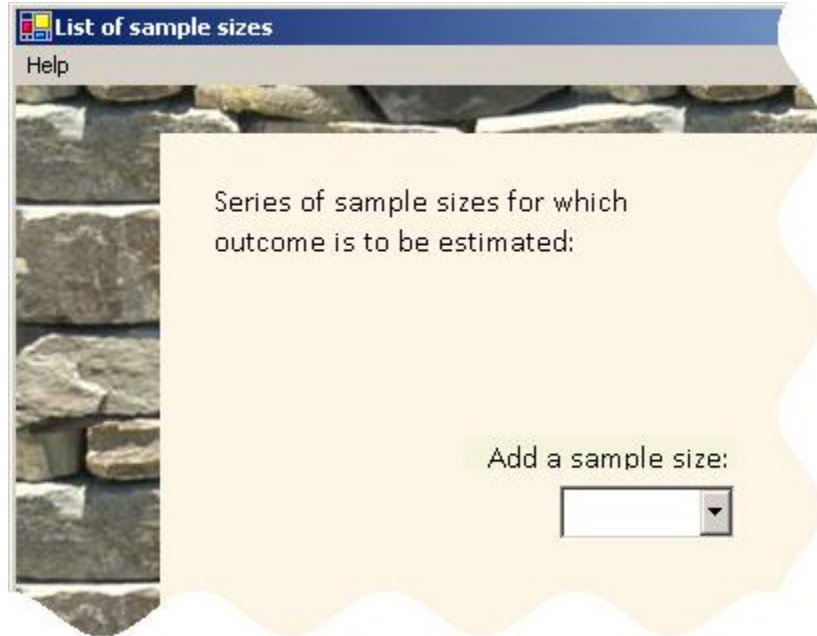


The next forms are identical to those presented in section 3 and are used to enter the number of independent variables, the prior distribution for β , etc. We therefore omit discussion of these steps here, proceeding to the next form where there is a difference from the procedure for sample size calculations.

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.

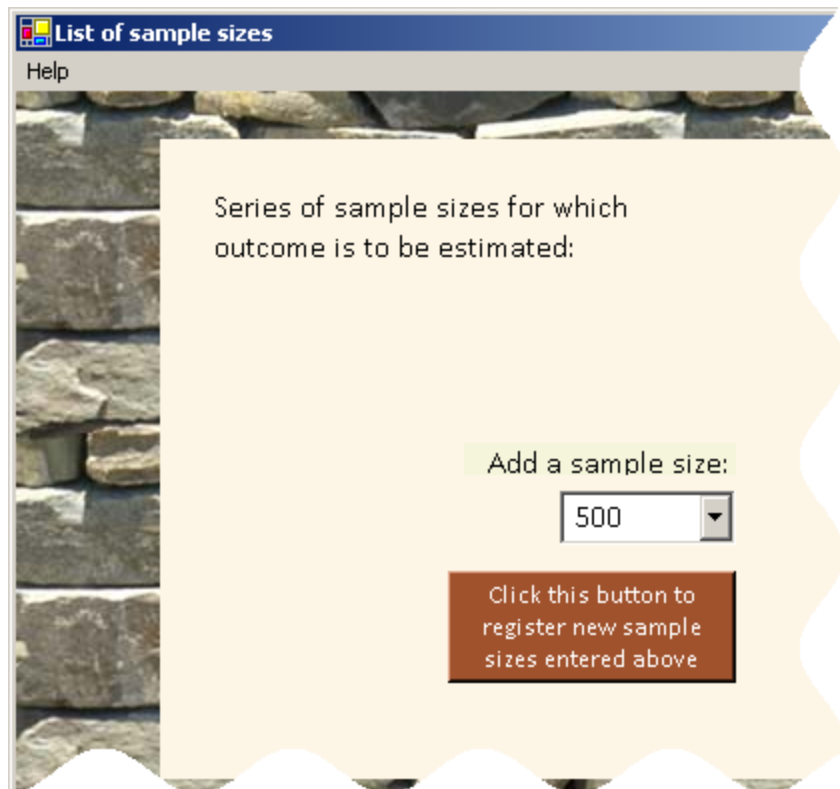


The next form is used to specify the sample sizes for which the above-specified outcome is to be estimated.



The screenshot shows a window titled "List of sample sizes" with a "Help" button. The main content area has a stone wall background and a yellow callout box. Inside the callout box, the text reads "Series of sample sizes for which outcome is to be estimated:". Below this text, there is a label "Add a sample size:" followed by a text input field that is currently empty.

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



The screenshot shows the same "List of sample sizes" window. The text input field now contains the number "500". Below the input field, a brown button with white text is visible, which reads "Click this button to register new sample sizes entered above".

Proceed the same way for each sample size of interest.

Click *Next*>> when done.

Sample sizes

Series of sample sizes for which outcome is to be estimated:

500

Add a sample size:

1000

Click this button to register new sample sizes entered above

Drop list

Select sample sizes to be dropped from list

500

Drop sample size >>

Add Sample Sizes Sequence

Next >>

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

Technical settings

Help

Monte Carlo Markov chain specifications

100 Preposterior sample size

2000 Number of monitored iterations

500 Number of burn-in iterations

Use all of the above parameters as default in future runs.

Next >>

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.

SampleSizeRegression Output Files

SampleSizeRegression output file was saved to file
<C:\Patrick\MyProject\SampleSizeDetermination\Beta1.html>

Scatter plot of Coverages vs Sample Size:
<C:/Patrick/MyProject/SampleSizeDetermination/Beta1-ScatterPlot.pdf>

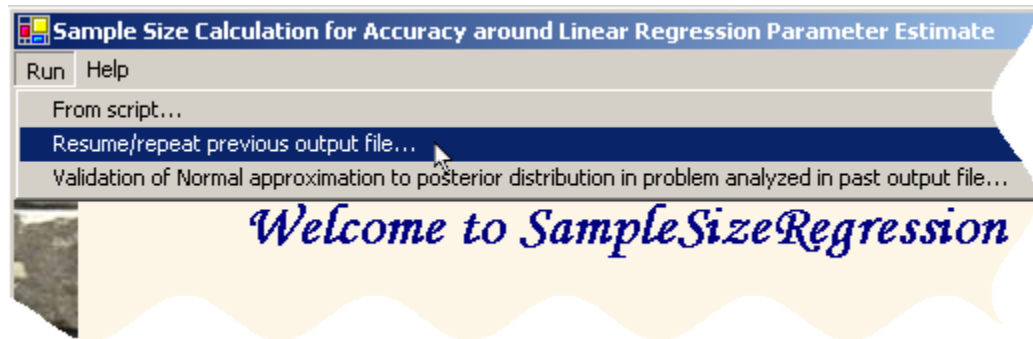
Histogram(s) of HPD Coverages:
<C:\Patrick\MyProject\SampleSizeDetermination\Beta1-ALC-HPDCovq.pdf>

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.

Results obtained along the march

Sample size N Average HPD coverage

1000	0.98134
750	0.98023
250	0.939011
284	0.9401
386	0.96103
314	0.94875
329	0.93625
356	Now running

sorted by sample size:

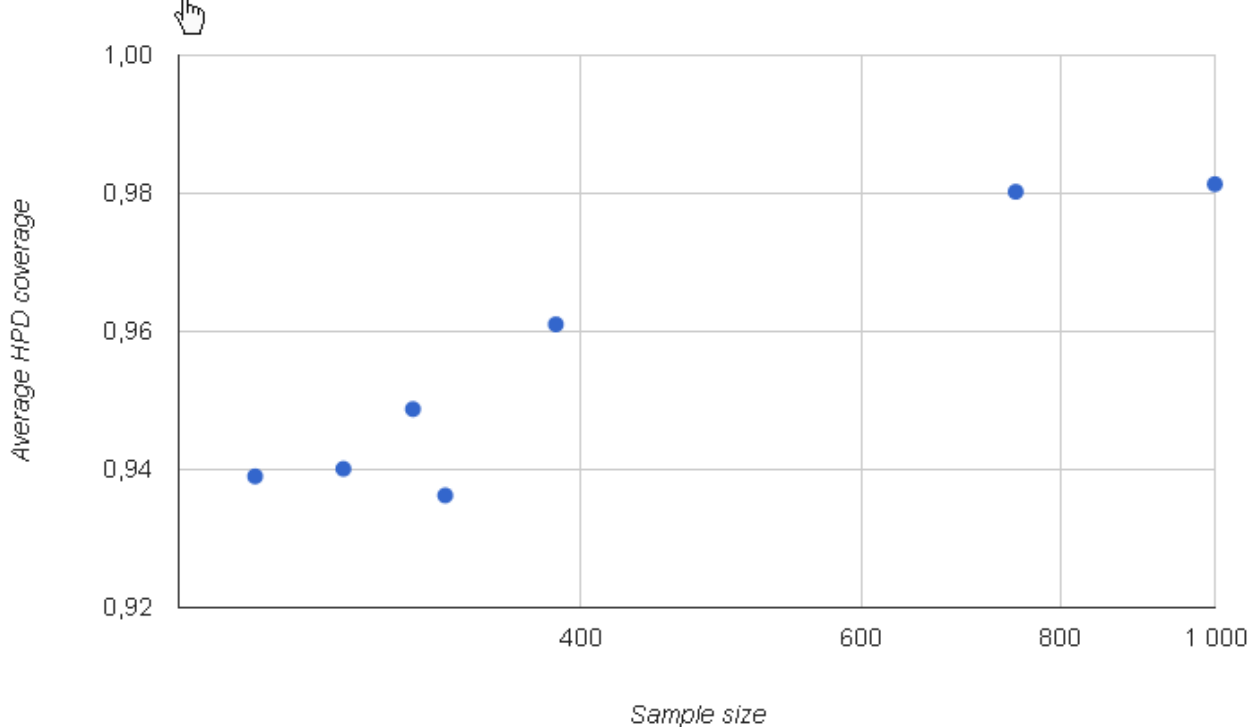
250	0.939011
284	0.940100
314	0.948750
329	0.936250
356	Now running
386	0.961030
750	0.980230
1000	0.981340

Show/Hide Scatter plot

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.

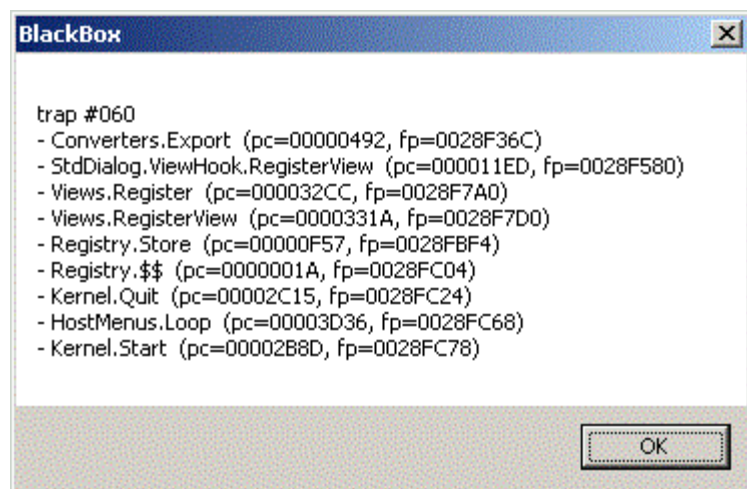
Show/Hide Scatter plot



5. Troubleshooting

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

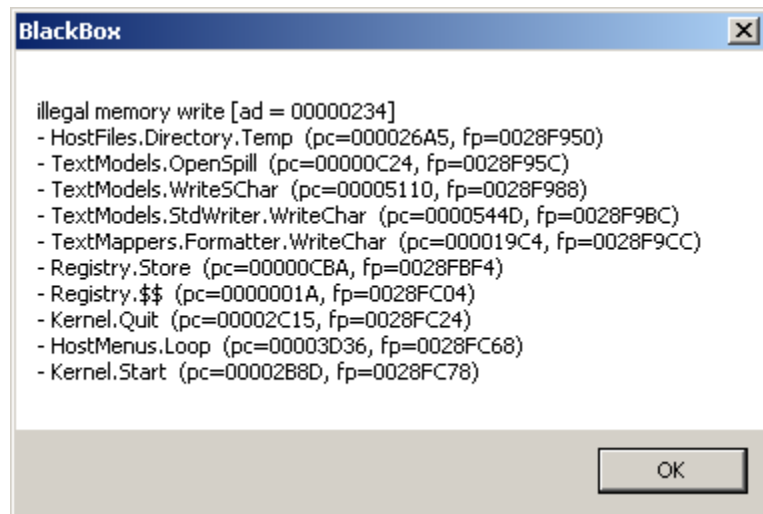
If you are working on a Windows 7 or Windows Vista platform and have run WinBUGS before, you may have already run into the cryptic **Trap #060** error message illustrated to the right. This is due to restricted write permissions in c:\Program Files, where you may have installed WinBUGS.



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

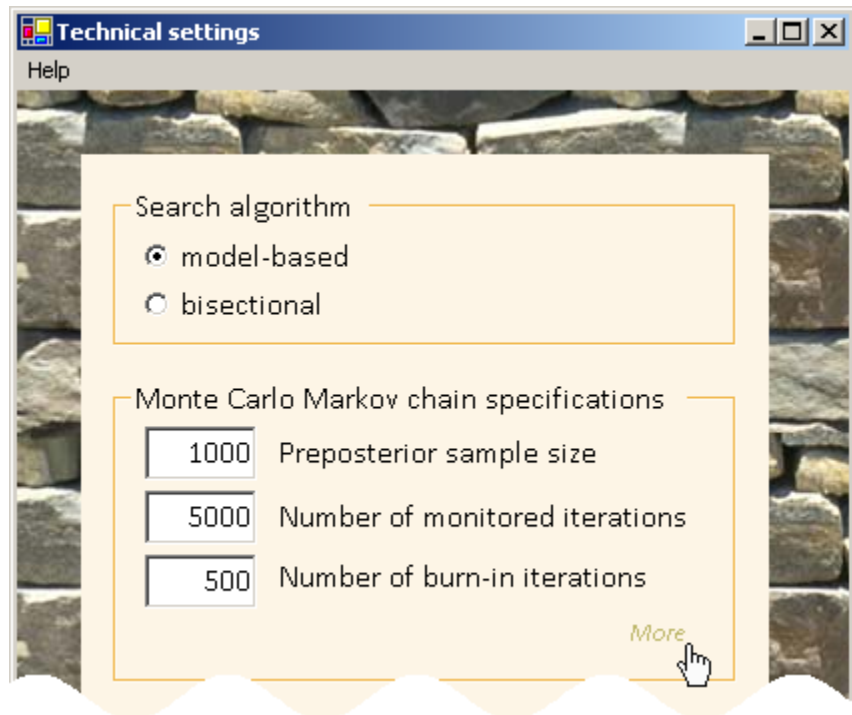
5.2 Illegal Vista platforms write

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.

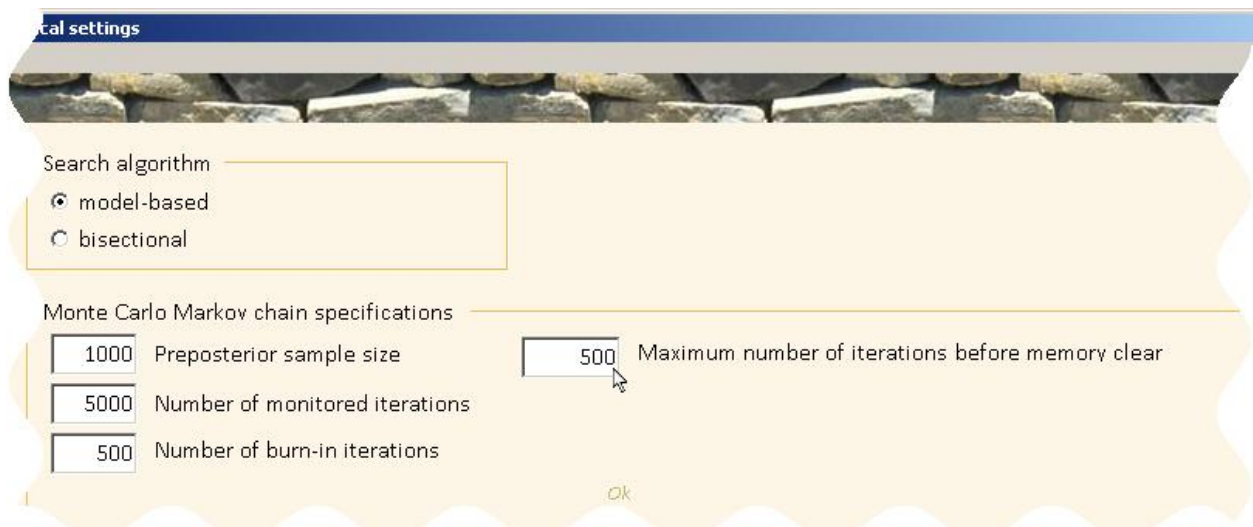


The necessary frequency of WinBUGS memory wash-outs depends on the agreement table size and your system environment. If wash-outs at every 1000 iterations are not sufficient, we suggest that you try with wash-outs at every 500 iterations, 200 iterations, and so on, until **SampleSizeRegression** runs 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*.



References

- Adcock, C. Sample size determination: A review. (1997). *Journal of the Royal Statistical Society. Series D (The Statistician)* **46(2)**, 261-283.
- Bristol, D. R. Sample sizes for constructing confidence intervals and testing hypotheses. (1989). *Statistics in Medicine* **8**, 803-811.
- Desu M., and Raghavarao D. (1990). *Sample Size Methodology*. New York, Academic Press.
- Gardner, M., and Altman, D. (1988). Estimating with confidence. *BMJ* **296(6631)**, 1210-1211.
- Lemeshow, S., Hosmer, D., Klar, J., Lwanga, S. (1990). *Adequacy of sample size in health studies*. Chichester, England: John Wiley & Sons Ltd.
- M'Lan, E., Joseph, L., and Wolfson, D. (2008). Bayesian sample size determination for binomial proportions. *Bayesian Analysis* **3(2)**, 269-296.
- Pallay, A. A decision analytic approach to determining sample sizes in a phase III program. *Drug Information Journal* 2000, **34**, 365-377.
- Thisted R. *Elements of Statistical Computing*. (1988). CRC Press, New York.