USAGE Manual

uncertainty and sensitivity analysis in a GenStat environment

version 2.0

Michiel J.W. Jansen
Jac T.N.M. Thissen
Jacques C.M. Withagen

Biometris report 11.12.05

December 2005

Biometris

quantitative methods in life and earth sciences

USAGE Manual
uncertainty and sensitivity analysis in a GenStat environment

version 2.0

Michiel J.W. Jansen
Jac T.N.M. Thissen
Jacques C.M. Withagen

Biometris is the integration of the Centre for Biometry of Plant Research International and the Department of Mathematical and Statistical Methods of Wageningen University. Biometris, part of Wageningen University and Research center (Wageningen UR), was established on 20th June 2001.

For more information please visit the website www.biometris.nl or contact:

Biometris, Wageningen UR
P.O. Box 100
6700 AC Wageningen, The Netherlands
Visiting address:
Buildingnumber 116
Bornsesteeg 47
6708 PD Wageningen
Phone: +31 (0)317 476925 or +31 (0)317 484385; Fax: +31 (0)317 483554
E-mail: Biometris@wur.nl

December 2005

Contents

1. Introduction

This manual contains theory, examples and descriptions of GenStat procedures for uncertainty analysis and regression-based sensitivity analysis (GenStat Committee, 2005). For the actual use of the software presented, a moderate experience with GenStat is required. But it is hoped that a large part of the manual will also constitute interesting reading for those unfamiliar with GenStat.

USAGE contains procedures for sampling from continuous multivariate distributions of model input. Model output corresponding to the input sample is calculated outside USAGE. Various procedures are available for the subsequent analysis of uncertainty or sensitivity.

The distributions of the individual inputs are defined per input. Association between inputs is specified via rank correlation. Thus a great flexibility is achieved for defining input distributions. Restricted random samples – latin hypercube samples or samples with forced correlations – can be generated for efficiency reasons if the model runs take much computer time.

The USAGE procedures for the subsequent analysis of sensitivity focus upon regression-based methods, but an example is given of a regression-free analysis of the effect of independent groups of inputs.

In the literature diverse sensitivity measures (uncertainty contributions) have been proposed. In USAGE uncertainty and uncertainty contributions are exclusively quantified as variances and variance components.

GenStat seems to be no more and no less suitable for uncertainty and sensitivity analysis than other high-quality statistical software. But the use of advanced statistical software has a definite advantage over the use of general purpose software such as FORTRAN or C, because standard statistical routines, and routines for graphics and I/O are readily available.

Another major advantage of imbedding the routines in standard statistical software, is that the user can more easily extend the fixed menu of routines currently offered. It is shown for instance, how the sampling variability of an uncertainty measure can be assessed with a bootstrap method from the analysis of one ordinary random sample consisting of independent draws of the input vector. Other examples are the transformation of inputs or outputs before the analysis, e.g. rank-transformation, or the use of readily available model selection techniques in order to find a small number of influential inputs.

1.1. The model

We will restrict ourselves to deterministic models. From the viewpoint of the analyst of uncertainty or sensitivity, the model will be seen as follows. A scalar (one-dimensional) model output y depends on a k-vector $x = (x_1...x_k)$ of inputs:

$$y = f(x) = f(x_1...x_k).$$

The function f is deterministic; usually it is evaluated by simulation; f represents a single output. Different outputs are analyzed separately, although they are usually calculated simultaneously. The input vector x may comprise initial values, parameters, exogenous variables, etcetera.

1.2. Uncertainty and sensitivity analysis

In the analyses discussed in this manual, the uncertainty about the value of the input vector x is modelled by *randomness* of x. Usually, the study of the combined effect of all inputs on the output is called *uncertainty analysis* (UA) while the study of the contributions of components of vector x to the uncertainty of f(x) is called *sensitivity analysis* (SA) (e.g. Saltelli et al., 2000). Jansen (2005) introduces the term *stochastic sensitivity analysis* in order to distinguish the above sketched form of sensitivity analysis from all kinds of deterministic sensitivity analyses where the input is not random. Uncertainty analysis concerns the accuracy of prediction with the current knowledge, whereas sensitivity analysis pertains to the prospects to improve the accuracy by additional knowledge. Large uncertainty contributions – or large sensitivity – of individual inputs, or groups of inputs, indicate that it would be worthwhile to get to know more about these inputs, whereas it would be pointless to spend much effort gaining new information about the other inputs. Thus, the analysis provides information for decisions on research priorities. Obviously, it may explain poor validation results. And it may be of help in the selection of parameters that need to be calibrated: in particular not to calibrate parameters that cause little uncertainty.

The structure of the model is assumed to be given. But usually more information is required for model predictions: initial values have to be measured, parameters have to be estimated, exogenous variables may not be known at the time when predictions are made. The current paradigm for the study of input uncertainty propagation is to represent input uncertainty by *randomness* of the inputs. (Alternatively, input uncertainty may be represented by a *set* of plausible inputs; but that approach may be largely treated as a special case of randomness, namely uniform distributions over the set.) Uncertainty analysis studies the ensuing uncertainty in the model output. The analysis can only give an optimistic preview of prediction error, since structural errors in the model will not show up; these can only become apparent in a true validation where model predictions are compared with new observations.

Input uncertainty is represented by a multivariate probability distribution, say D, of the vector $x = (x_1 ... x_k)$:

$$x = (x_1 ... x_k) \sim D.$$

The multivariate distribution D describes the marginal distributions, i.e. the distributions of the individual inputs x_i, and their dependencies.

UA and SA start with a characterization of the output distribution, given the model and the distribution of the inputs. In this manual the variability of the distribution will be characterized by its *variance,* which is assumed to be finite. The *total uncertainty*, VTOT, is the variance of f(x) induced by the randomness of all sources x_i as described by the distribution D:

$$\mathrm{VTOT} = \mathrm{Var}[f(x)] \qquad x \sim D.$$

Increasingly often, the effect of input uncertainties is studied by computer experiments rather than analytically. The same trend occurs in general statistics, where many novel techniques rely on Monte Carlo simulation. The analytic approach requires simple, usually linear, model approximations, and it fails if no satisfactory approximation can be found. The computer-experimental approach has the advantage of conceptual simplicity, but the draw-back that it may require many model runs. Moreover, analytic results tend to be general, whereas computer-experimental results are often somewhat anecdotical by their dependence on experimental details.

Regression-based and regression-free sensitivity analyses

The procedures of USAGE perform regression-based SA, which means that the relation between the studied model output y, and the model inputs $x_1 ... x_k$, is approximated by a regression relation. Apart from the commonly used linear analysis, an analysis based on spline regression can be chosen, which often provides a more adequate description of the input/output relation. The analysis of the contribution of (groups of) inputs to prediction uncertainty is based on the regression approximation. For a satisfying analysis, the percentage of variance accounted for by the regression should be close to 100, since a regression-based analysis is blind for the variation in the model output that is not accounted for by the regression.

Individual inputs or groups of inputs?

Many common algorithms for SA focus upon uncertainty and sensitivity measures of *individual* inputs. Nevertheless, one would often prefer to study uncertainty from coherent groups of inputs, for instance all parameters associated with a subprocess, or all inputs stemming from some exogenous process. After a single-input SA, one often tries to interpret the results in terms of groups of inputs associated to specific subprocesses.

Crop growth models may provide an example of the relevance of group uncertainty contributions. These models have weather data as input, which may comprise hundreds of uncertain numbers. It is futile to study the uncertainty contribution of one weather item like the mean temperature at the tenth of June, but it makes sense to ask how much uncertainty is caused by the weather.

If the inputs consist of *stochastically independent groups*, various regression-free group-oriented sensitivity analyses are feasible. One such analysis will be discussed. A group-oriented sensitivity analysis is also possible for *dependent* input groups, but that requires quite a bit of tailoring work, and will not be discussed in this report.

Deterministic sensitivity analysis

Deterministic sensitivity analysis may be useful for inspection of the model and its software implementation. The analysis may suggest model simplifications, such as deletion of insensitive subprocesses. The questions addressed are for instance: whether some response is affected at all by some input; whether one can find a small subset of inputs dominating the response; whether the response increases or decreases according to expectation; whether the response is continuous, differentiable, etcetera.

We will briefly mention the most common types of deterministic sensitivity analysis. In *local* sensitivity analysis one studies output changes under very small input changes around some given vector value, for instance a nominal value or a calibrated value. *One-at-a-time* sensitivity analysis studies the model's response to change of one input, at fixed values of the other inputs. In particular, one may study the response to nearly continuous change over some range, for instance in order to inspect whether the response is continuous, monotonically increasing, or one-topped. For completeness we also mention *factorial sensitivity analysis*, although it will not be treated in this report (see for instance Kleijnen, 1987). In this analysis, inputs are varied according to a so-called factorial design. In the most common factorial design, the two-level design, each input has a low and a high level. Such an analysis may be used for instance to study interaction between inputs: the phenomenon that the response to one input depends on the setting of the other inputs.

Factorial designs might also be used to search for a small number of sensitive inputs between very large numbers of spurious inputs.

1.3. Communication between USAGE and model software

Typically, the software presented in this manual is used as follows. A GenStat program generates an ASCII file with a sample of model inputs. Subsequently, the user produces an ASCII file with corresponding model output, using his own modelling software. After that, another GenStat program performs an uncertainty or sensitivity analysis. The communication between the software components is left to the user, because it depends strongly on the specific modelling environment. Thus, the user himself should take care that the numbers in the intermediate files have enough decimals, for instance by using E-notation.

1.4. Outline of the manual

The estimation of the probability distribution of the inputs forms the major problem of uncertainty and sensitivity analysis, but since the subject is virtually unbounded, it falls outside the scope of USAGE in its present form. Section 2 contains a brief sketch of some subjects that often play a role in the assessment of input uncertainty of system models in agricultural and environmental research. Section 3 gives definitions of uncertainty contributions (sensitivity measures). The construction of samples from the input distribution is discussed in Section 4. These samples are used in Sections 5 and 6 for the estimation of uncertainty and sensitivity. Section 5 is devoted to regression-based sensitivity analysis for individual or grouped inputs, whereas Section 6 treats regression-free sensitivity analysis of independent input groups. In Section 7, we give some examples. Section 8 contains references.

Section 9 contains Appendix I, which discusses some mathematical details. A formal description of the USAGE procedures is given in Appendix II (Section 10).

2. Assessing input uncertainty

The estimation of the probability distribution D of the inputs $x_1 ... x_k$ constitutes the major problem of uncertainty and sensitivity analysis. Virtually any part of statistics may play a role in this estimation problem. We will only give a brief sketch of some subjects that often play a role in the assessment of input uncertainty of system models.

Various types of data may be available for the quantification of parameter uncertainty, for instance analyses from the literature, data sets that happen to be available, or experiments performed for the purpose. Presently, an increasing number of databases, including geographical databases, is becoming available via Internet. The experiments providing information on parameter uncertainty should cover a range of situations relevant for the intended model application, in particular a sufficiently large area and a sufficiently long time-span.

Sometimes a database contains a sample of model inputs that can serve directly as description of input uncertainty. For instance historical weather data, or an accurate and representative sample of soil measurements.

Parameter uncertainty is caused by natural variation between the systems modelled and by estimation error. Both may cause correlation in the simultaneous distribution describing parameter uncertainty. Natural covariation between parameters that are evaluated in separate experiments cannot be evaluated. The best solution would seem to be to assume independence unless there is counterevidence, since introducing unwarranted correlation would amount to saying that one knows more than one actually does.

Any kind of statistical technique may be required to assess parameter uncertainty, but meta-analysis, the overall analysis of analyses of separate experiments, deserves special mention (e.g. Hedges & Olkin, 1985). Meta-analysis can be applied to integrate analyses from literature. Typically, separate literature sources pertain to subsystems, so one has to perform various meta-analyses, each pertaining to a small number of parameters of a particular subprocess. Another approach that deserves to be mentioned, is to estimate parameter uncertainty from calibration on whole-system observations, i.e. the kind of observations that the model predicts (Keesman & Van Straten, 1990; Janssen & Heuberger, 1995). This approach is not without problems. A major problem arises if one has to fix some uncertain parameters, so that the other, calibrated, parameters will tend to compensate errors in the fixed ones, and thereby lose their physical meaning. Moreover, one needs a realistic measurement-error model for a realistic post-calibration uncertainty assessment. Information to formulate such an error model, for instance from duplicate measurements, is often lacking. Nevertheless, the subject of post-calibration uncertainty seems to hold great promises.

A database with soil or weather data should contain information about error in its data. And, for upscaling, also about spatial covariation of the error. Soil maps are often constructed by kriging. In such a case, the kriging interpolation error provides an estimate of map uncertainty (including covariation).

A thorough quantification of input uncertainty can be very difficult and time consuming (see for instance Metselaar & Jansen, 1995-a). In most projects, an exhaustive data-based analysis of input uncertainty will not be possible, and one will have to limit the analysis to some subset of inputs. Objective data may sometimes need to be supplemented by expert judgement. Special purpose software may be of help to translate expert opinion into a probability distribution (e.g. Van Lenthe & Molenaar, 1993). And finally, one may decide that a, less ambitious, deterministic sensitivity analysis forms a more realistic alternative.

3. Definition of uncertainty contributions (sensitivity measures)

The total uncertainty is expressed completely by the distribution of f(x) that is induced by the multivariate distribution D of input vector x. The variance of f(x), or a few selected quantiles, may serve as summary measures of uncertainty. We will use the variance as summary measure of uncertainty.

Problems arise, however, with the concept of *uncertainty contributions* or *sensitivity measures*. In the literature, many types of uncertainty contributions occur; see for instance Janssen (1994) for a fairly complete overview. The many possibilities may well cause some embarrassment of choice. Our approach will be to define various kinds of uncertainty contribution as the answers to various specific questions of the type: *how much would the output variance decrease if specific* *information about the input would become available, in addition to the information contained in input distribution D.*

The specification of uncertainty as *variance* provides a practicable restriction of the abundant possibilities. It is implicitly assumed that the variance is finite. The variance is a convenient measure of prediction uncertainty, because the variance can be more easily decomposed into meaningful parts than other conceivable measures of uncertainty. In fact this has been the reason to introduce the concept of variance. With this measure, the analysis of uncertainty contributions becomes essentially a form of analysis of variance components. Application of analysis of variance components to model output, stems from the 1990s (Sobol, 1990; Jansen, Rossing and Daamen, 1994; Sobol, 1995; McKay, 1996; Saltelli, Tarantola and Chan, 1999; Jansen, 1999; Saltelli, Chan and Scott, 2000, Ch 8).

The variance of f(x), induced by the distribution D of x = (x_1...x_k) will be called VTOT

$$\text{VTOT} = \text{Var}[f(x)] \qquad x \sim D$$

Let S denote a subset of the x's, for instance one particular x_i, a group of parameters corresponding to a particular submodel, or some aggregate of exogenous variables. The uncertainty contribution of subset S will be expressed in two ways. Firstly, the *top marginal variance*, TMV_S, is the variance reduction that would occur in case one would get perfect new information about the inputs S. Secondly, the *bottom marginal variance,* BMV_S is the variance that will remain as long as one gets no new information about S. In both cases the new information is added to the information already present in input distribution D.

Stated differently, TMV_S is the *variance accounted for by S*, whereas BMV_S is the *variance not accounted for without S*.

Usually, TMV and BMV are expressed as fraction or percentage of VTOT. The concepts of top and bottom marginal variances have been introduced in UA/SA by various authors, under various names, (Krzykacz, 1990; Sobol, 1990, 1995; Jansen, 1994; McKay, 1996). The next table mentions various names used in the literature for TMV_S / VTOT and BMV_S / VTOT which have the same meaning for an *independent* group S of inputs (of course the group S may also consist of a single input).

TMV_S / VTOT	BMV_S / VTOT
relative top marginal variance correlation ratio first order sensitivity index	relative bottom marginal variance complementary correlation ratio total effect sensitivity index

Most authors concentrate on what we call top marginal variances, often for single inputs. But we will argue that bottom marginal variances should not be overlooked (see also Saltelli et al., 1999); and we have already argued that it is often better to consider the effect of groups of inputs instead of individual inputs. The top and bottom marginal variances of individual inputs are closely related to some well-known uncertainty measures e.g. the linear correlation coefficient. For instance, if x is multinormal and the response f(x) is linear in x, the top marginal variance is equal to the squared linear correlation coefficient.

Unfortunately, BMV_S and TMV_S need not be equal. When the input group S and the complementary input group, say T, are independent, one may show that $\text{BMV}_S \geq \text{TMV}_S$, with equality if f is additive in S and T, that is if f(x) is the *sum* of a function of S and a function of T (which implies that the response of the model to a change in S is the same for different values of T). Thus, in case of independence, a difference between BMV_S and TMV_S signals non-additivity of f, also called *interaction* between S and T. Differences between TMV_S and BMV_S may also be caused by *dependence* between S and T. For instance, the distribution D of x may be such that the value of S can almost be derived from the value of T, and conversely. In that extreme case, where S and T are nearly *exchangeable*, the bottom marginal variances of S and T would be small, but their top marginal variances might still be considerable. Alternatively, the dependence between S and T, and the nature of f(x), may happen to be such that S and T are *complementary* in predicting f(x), which would result in small values of the top marginal variances of S and T, but considerable values of the bottom marginal variances.

In summary, differences between the two types of variance indicate interaction and/or dependence. The situation that the TMV of a group is greater than its BMV can only be caused by dependence.

In general, TMV is a much more useful concept than BMV, and we advise to use BMV only in exceptional cases. TMV_S assesses the maximal improvement of prediction precision that might be attained by better knowledge about group S, or by better control of that group. If TMV_S is large, additional research about S might prove fruitful. If it would be utterly unrealistic, however, to expect to gain better knowledge about some input group S, you might use BMV_S to assess the uncertainty that would always remain even if you succeeded in eliminating all uncertainties about the other inputs. If, in such a case, BMV_S would be very large, research about the other inputs would seem rather futile.

Mathematical details about marginal variances are treated in Appendix I; their estimation with USAGE will be treated in Sections 5-7.

4. Generation of random samples

Monte Carlo sampling from uncertain inputs comprises firstly the sampling from univariate distributions, possibly supplemented by methods to introduce dependencies (Iman & Conover, 1982).

Additionally, spatial or temporal stochastic simulation may be required. Weather databases or weather generators may be used to account for weather uncertainty. Generators of spatio-temporal or weather processes are not implemented in USAGE.

It is desirable that the samples generated will observe the natural limitations imposed by the model. One should, for instance, take care that a positive input cannot acquire negative values. USAGE has simple facilities to impose bounds on samples.

Computer random generators are commonly initialized with a user-supplied seed. Most often the seed is an integer number; and the random generator is initialized just once, at the first call (GenStat Committee, 2005). When the same seed is used, the same random sequence will be generated. We will not philosophize about the implication that such random generators are not really random: enough has already been said elsewhere about this subject. Results of seeded Monte Carlo analyses can be reproduced, which is an asset. But anyhow, samples should be so large that the result is quite insensitive to the seed used.

4.1. Univariate samples

In USAGE a random scalar, say y, from a continuous cumulative distribution say F is drawn indirectly. First a standard homogeneous scalar u is drawn (i.e. homogeneous on the interval from 0 to 1.) Then one calculates the target scalar y such that F(y)=u. The scalar y thus obtained is random because u is random, and it has the desired distribution F. We will apply this indirect method most of the time because it allows efficient drawing of efficient samples, like latin hypercube samples that will be introduced later; moreover, the indirect method also facilitates the introduction of correlations in multivariate samples (see Sections 4.4 and 4.5).

Standard homogeneous variates, on interval (0,1) can be drawn with the GenStat directive `URAND`. Homogeneous variates on arbitrary intervals are easily derived. The directive `HISTOGRAM` can be used to obtain a summary view of the resulting sample. The USAGE procedure `SUMMARIZE` produces summary statistics such as mean, standard deviation, coefficient of variation, percentiles etcetera. Appendix II contains a formal description of the procedure.

```
\\Define the sample size and the seed for the random generator
SCALAR    n    ; 1000
SCALAR    seed ; 171096
VARIATE   [NVALUES=n] uni
\\Draw random sample uni from a uniform distribution on (0,1)
CALCULATE uni = URAND(seed)
SUMMARIZE [PRINT=#,quantiles] uni
DHISTOGRA [KEY=0] uni
```

The USAGE procedure `EDCONTINUOUS` can be used to transform homogeneous (0,1) variates into various types of continuous scalar random variables. (ED stands for 'equivalent deviate'.) The currently available distributions are: beta, gamma, lognormal, normal and uniform (in alphabetical order). In USAGE, the first four distributions can be specified either by the first two moments (mean and variance), or by a pair of quantiles. The uniform distribution can only be specified by its bounds. A lower bound can be given for the gamma and lognormal distributions (default 0); whereas a lower and an upper bound can be specified for the beta and uniform distributions (default 0 and 1). Appendix II contains a formal description of the procedure `EDCONTINUOUS`.

When producing a sample to be used as input to a model, one has to be sure that the input values lie in the ranges allowed by the model. Thus one will often need to draw from the gamma, lognormal and beta distributions rather than the unbounded normal distribution.

The lognormal and gamma distributions are useful for variates, such as masses or concentrations, that have a natural lower bound – most often 0. The two distributions are much alike, and sometimes the choice between them will be a matter of taste and tradition. Both can be bell-shaped and mirrored-j-shaped. The gamma distribution, however, tends to be less “tail heavy” than the lognormal and thus assigns smaller probability to extremes. The exponential distribution is a special case of the gamma distribution. The beta distribution is especially useful for variates that have natural upper and lower limits, for instance partitions, percentages or probabilities. The distribution can be bell-shaped, j-shaped and mirrored-j-shaped; the uniform distribution is a special case of the beta distribution.

USAGE can be used to get a feeling for the distributions mentioned. The program below shows how USAGE procedure `EDCONTINUOUS` can be used to transform homogeneous (0,1) variates into various types of continuous scalar random variables. Procedure `SUMMARIZE` is used to inspect the results.

```
\\Draw uni_ab uniform(a,b)
SCALAR    a,b; value=40,60
EDCONTINUOUS [DISTRIBUTION=uniform; LOWER=a; UPPER=b] CUMPR=uni; DEVIATE=uni_ab
SUMMARIZE [PRINT=#,quantiles] uni_ab

\\Draw nor1 from normal(mu, sigmasquare)
VARIATE   [NVALUES=n] nor
SCALAR    mu, sigmasquare; 65, 1
CALCULATE uni = URAND(0)
EDCONTINUOUS [DISTRIBUTION=normal ; MEAN=mu; VARIANCE=sigmasquare] \
          CUMPR=uni ; DEVIATE=nor1

\\Draw nor2 from standard normal with 10%-point -1 and 90%-point 1
VARIATE   vals ; !(-1, 1)
VARIATE   probs ; !(0.10, 0.90)
CALCULATE uni = URAND(0)
EDCONTINUOUS [DISTRIBUTION=normal ; METHOD=quantilES ; PROPORTIONS=probs ; \
          QUANTILES=vals] CUMPR=uni ; DEVIATE=nor2
SUMMARIZE [PRINT=#,quantiles] nor1, nor2

\\Draw lnor1 from lognormal with mean mu and variance sigmasquare
VARIATE   [NVALUES=n] lnor
CALCULATE uni = URAND(0)
EDCONTINUOUS [DISTRIBUTION=lognormal ; MEAN=65 ; VARIANCE=1] uni ; lnor1

\\Draw lnor2 from lognormal on (0, infinity), with 5%-point 5 and 95%-point 10
VARIATE   vals ; !(5, 10)
VARIATE   probs ; !(0.05, 0.95)
CALCULATE uni = URAND(0)
EDCONTINUOUS [DISTRIBUTION=lognormal ; METHOD=quantil ; PROPORTIONS=probs ; \
          QUANTILES=vals] CUMPR=uni ; DEVIATE=lnor2
SUMMARIZE [PRINT=#,quantiles ; PROPORTIONS=0.05, 0.95] lnor1, lnor2

\\Draw gam1 from gamma distribution with given mean and variance
VARIATE   [NVALUES=n] gam1
CALCULATE uni = URAND(0)
EDCONTINUOUS [DISTRIBUTION=gamma ; MEAN=65 ; VARIANCE=1] uni ; gam1

\\Draw gam2 from gamma(0, infinity), with 5%-point 5 and 95%-point 10
VARIATE   vals ; !(5, 10)
VARIATE   probs ; !(0.05, 0.95)
CALCULATE uni = urand(0)
EDCONTINUOUS [DISTRIBUTION=gamma ; METHOD=quantiles ;  ; PROPORTIONS=probs ; \
          QUANTILES=vals] CUMPR=uni ; DEVIATE=gam2
SUMMARIZE [PRINT=#,quantiles ; PROPORTIONS=0.05,0.95] gam1, gam2

\\Draw bet1 from beta distribution on (3,9) with given mean and variance
CALCULATE uni = URAND(0)
EDCONTINUOUS [DISTRIBUTION=beta ; MEAN=5 ; VARIANCE=1 ; LOWER=3 ; UPPER=9] \
          uni ; bet1

\\Draw bet2 from beta distribution on (0, 10), with 25%-point 5 and 75%-point 8
VARIATE   vals ; !(5, 8)
VARIATE   probs ; !(0.25, 0.75)
CALCULATE uni = URAND(0)
EDCONTINUOUS [DISTRIBUTION=beta ; LOWER=0 ; UPPER=10 ; METHOD=quantiles ; \
          PROPORTIONS=probs ; QUANTILES=vals] CUMPR=uni ; DEVIATE=bet2
SUMMARIZE [PRINT=#,quantiles ; PROPORTIONS=0.25,0.75] bet1, bet2
```

Warning: When a beta or gamma distribution is specified by a pair of quantiles, EDCONTINUOUS tries to find the parameters of the distribution by means of a non-linear optimization (via FITNONLINEAR) which may occasionally fail without any warning. So please check the results in this case.

4.2. Multivariate samples

Direct sampling from multivariate normal and student distributions

The multinormal distribution is specified by a vector of means and a covariance matrix. The following example yields 1000 draws of 3 variates from a 3-dimensional normal distribution with mean zero and a given (valid) variance-covariance matrix. For details, see the formal description of USAGE procedure GMULTIVARIATE in Appendix II.

```
SCALAR    k ; 3
VARIATE   [NVALUES=1000] x[1...k]
SYMMETRIC [ROWS=k] vcov ; !( 0.0000430, \
                            -0.0015080, 1.11300, \
                            -0.0007942, 0.04264,  0.1389)
VARIATE   mu ; !(0.022, 0.796, 2.186)
GMULTIVARIATE [DISTRIBUTION=normal ; NVALUES=1000 ; PRINT=summary ; \
          SEED=768241 ; MEAN=mu ; VCOVARIANCE=vcov] NUMBERS=x
CORRELATE [PRINT=correlations] x[]
```

Similarly, the multi-student distribution is specified by a vector of means, a covariance matrix, and a number of degrees of freedom. It should be noted that the means and the covariance matrix are not equal to mean and the covariance matrix of the student distribution itself, but of a normal distribution that plays a role in the definition of the student distribution (see Section 9.3). Adding the next GMULTIVARIATE-statement to the example above, yields 1000 draws of 3 variates from a 3-dimensional student distribution with 18 degrees of freedom, with defining mean zero and some given (admissible) variance-covariance matrix. For details, see the formal description of USAGE procedure GMULTIVARIATE in this manual.

```
GMULTIVARIATE [DISTRIBUTION=student ; NVALUES=1000 ; PRINT=summary ; \
          SEED=768241 ; MEAN=mu ; VCOVARIANCE=vcov ; DF=18] NUMBERS=x
```

Indirect sampling from more general multivariate distributions

Suppose one wants to draw from a multivariate distribution with given marginal distributions of any type. If the component variates are independent, one may just draw them independently with the EDCONTINUOUS procedure, as described in Section 4.1. We will now show how to draw from a multivariate distribution with prescribed rank correlation matrix. The method stems from Iman and Conover (1982). Two remarks should be made before we proceed. Firstly, the method is only approximate: the rank correlation matrix of the distribution from which the draws are made is merely close to the prescribed rank correlation. But very close: the rank correlations of the sampled distribution differ at most 0.02 from the desired values; in virtually all cases this error will be negligible compared with the estimation error of the rank correlation. Secondly, a multivariate distribution is not uniquely defined by its marginals and its rank correlation.

The next example shows how to draw from a multivariate distribution, with given marginals that are normal, gamma and beta, each defined by its mean and variance. The rank correlation is given by a symmetric matrix, named rankcor below, which should be a valid correlation matrix.

```
SCALAR    n ; 1000
SCALAR    seed ; 161096
VARIATE   [NVALUES=n] x[1...3], uni[1...3]
SCALAR    mean[1...3] ; 0.125, 0.125, 0.0050
SCALAR    var[1...3]  ; 0.004, 0.001, 0.00001
SYMMETRIC [ROWS=3] rankcor ; !(1, 0, 1, 0.5, 0, 1)
GUNITCUBE [NVALUES=n ; RCORRELATION=rankcor ; SEED=seed] NUMBERS=uni
EDCONTINUOUS [DISTRIBUTION=normal ; MEAN=mean[1] ; VAR=var[1]] uni[1] ; x[1]
EDCONTINUOUS [DISTRIBUTION=gamma  ; MEAN=mean[2] ; VAR=var[2]] uni[2] ; x[2]
EDCONTINUOUS [DISTRIBUTION=beta   ; MEAN=mean[3] ; VAR=var[3]] uni[3] ; x[3]
PEN       NUMBER=1 ; SYMBOL=2 ; SIZE=0.2
DSCATTER  x[]
```

USAGE procedure `GUNITCUBE` produces uniform variates from a multivariate distribution with the required rank correlations. From these, the variates $x_1 ... x_3$ with the required (marginal) distributions are obtained by the procedure `EDCONTINUOUS`, which was introduced in the previous subsection. Ranks and rank correlations are unaltered by `EDCONTINUOUS`. The standard GenStat procedure `DSCATTER` visualizes the resulting multivariate distribution by means of the scatterplots of the different pairs of variates.

4.3. Restricted random sampling

Up to this point, we only discussed ordinary random sampling, where each sampled vector is drawn independently of the other ones. This is the best-understood sampling technique. There exist, however, many alternative sampling techniques ranging from slightly less random to entirely deterministic. Their reason of existence is that the estimation results of subsequent analyses are hoped to be more accurate at the same sample size. The restricted random sampling methods that will be discussed are also called estimation-variance reduction techniques.

In uncertainty and sensitivity analysis, latin hypercube samples are often used to achieve improved estimation accuracy (McKay et al., 1979; Iman & Conover, 1980; Stein, 1987; Owen, 1992). The method enforces close resemblance of the sample marginals to the marginals of the target distribution. Latin hypercube samples will be discussed in Section 4.4.

There also exist various techniques aimed at controlling sample correlations. The method of Iman & Conover (1982), which enforces rank correlations, is implemented in USAGE (see Section 4.5).

Being very simple, the theory of ordinary random sampling is well-developed. Ordinary random sampling is not maximally efficient, but it has the great advantage that one can more easily assess the accuracy of the results. For instance, the accuracy of a mean of an ordinary random sample of size n is calculated in the usual way as 1/n times the sample variance. However, the uncertainty contributions in which we are interested are not estimated by sample means, so the 1/n rule cannot be applied. Instead, with an ordinary random sample, the bias and the sampling variability of an estimate of an uncertainty contribution can be assessed via bootstrap techniques. Bootstrap techniques are not implemented in the present version of USAGE, but an example of application of the bootstrap will be given in Section 7.3.

The theory of the various restricted random samples is less developed. Some estimates of uncertainty contributions are known to be slightly biased. Most often, the variability of the estimates is inferred from the results of the analyses of a number of independent restricted random samples.

In summary, restricted random samples may be used to improve estimation accuracy, but because of possible bias it is harder to assess that accuracy. Large ordinary random samples seem to be the best choice unless the computer time required to run the model many times becomes prohibitive. The sampling variability of estimation results from restricted samples can only be assessed by repeating the whole procedure. On the other hand, bootstrap methods can be used to estimate bias and sampling variability of estimates from ordinary random samples.

4.4. Controlling sample marginals: latin hypercube sampling

Latin hypercube sampling is a much-used variance reduction technique. One may force close adherence to the required marginal distributions by means of the GUNITCUBE option setting STRATIFICATION=latin.

```
SCALAR    k ; 3
SCALAR    n ; 10
SCALAR    seed ; 291096
VARIATE   [NVALUES=n] uni[1...k]
GUNITCUBE [NVALUES=n ; STRATIFICATION=latin ; SEED=seed] NUMBERS=uni
SCALAR    lower, upper, marks ; -0.0001, 1.0001, 0.1
AXES      [EQUAL=scale] WINDOW=1 ; STYLE=grid ; XLOWER=lower ; XUPPER=upper ; \
          XMARKS=marks ; YLOWER=lower ; YUPPER=upper ; YMARKS=marks
PEN       NUMBER=1 ; SYMBOL=2
DGRAPH    [KEYWINDOW=0] uni[2] ; uni[1]
```

Figure 1 shows the graph of uni[2] versus uni[1]. The variates uni[] are stratified so that each has exactly one value in each of the intervals (0, 1/10), (1/10, 2/10) ... (9/10, 1). In a sample of size n, one has consecutive intervals of size 1/n. This is the defining property of a uniform latin hypercube sample. For the rest everything is random: the location within the intervals, and the association between the variates.

A latin hypercube sample with arbitrary marginals can be constructed by means of USAGE procedure EDCONTINUOUS. For example, one may continue the program with:

```
VARIATE   [NVALUES=n] x[1...3]
SCALAR    mean[1...3] ; 0.125, 0.125, 0.00500
SCALAR    var[1...3]  ; 0.004, 0.001, 0.00001
EDCONTINUOUS [DISTRIBUTION=normal ; MEAN=mean[1] ; VAR=var[1]] uni[1] ; x[1]
EDCONTINUOUS [DISTRIBUTION=gamma  ; MEAN=mean[2] ; VAR=var[2]] uni[2] ; x[2]
EDCONTINUOUS [DISTRIBUTION=beta   ; MEAN=mean[3] ; VAR=var[3]] uni[3] ; x[3]
```

Latin hypercube sampling can be combined with prescription of a rank correlation matrix.

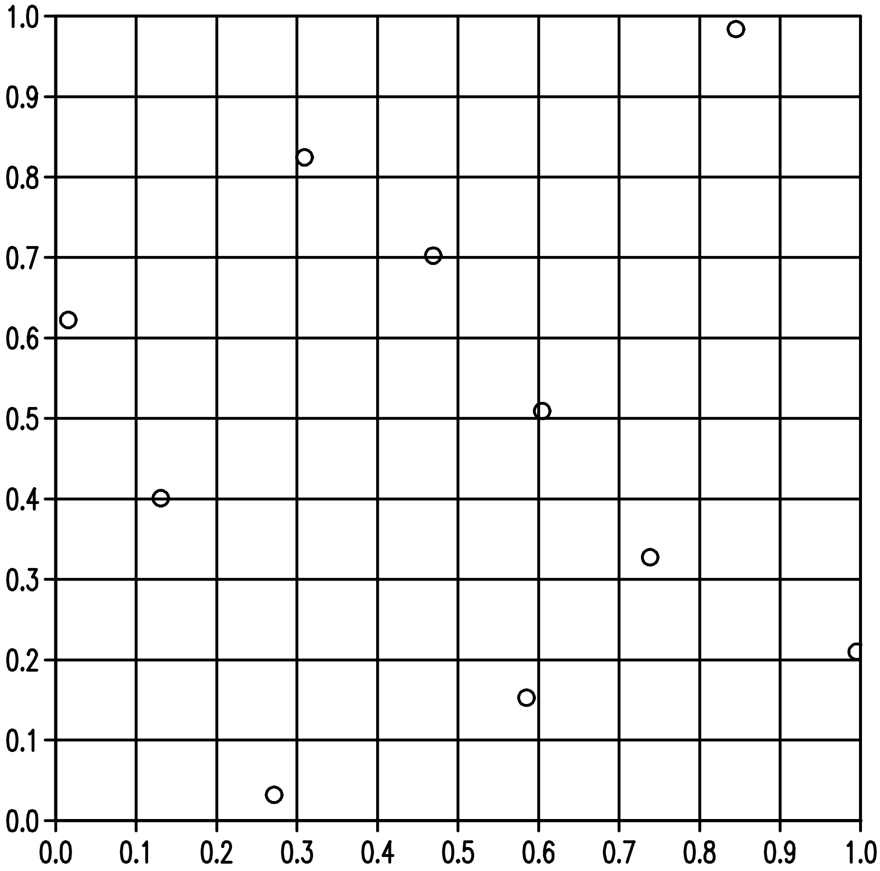

Figure 1. Two components of a uniform latin hypercube sample of size 10.

4.5. Controlling sample correlations

One may force close adherence of the sample rank correlation to the required rank correlation by means of the `GUNITCUBE` option `METHOD=iman`:

```
GUNITCUBE [NVALUES=n ; RCORRELATION=rankcor ; METHOD=iman] NUMBERS=uni
```

The method was introduced by Iman & Conover (1982). The marginal samples will remain as random as with ordinary random sampling, but the association between the vector components is much less random. The rank correlation of samples so constructed is nearly equal to the population rank correlation, especially with large samples. Iman's method for controlling correlations may be applied in combination with latin hypercube sampling:

```
GUNITCUBE [NVALUES=n ; RCORRELATION=rankcor ; METHOD=iman ; \
          STRATIFICATION=latin] NUMBERS=uni
```

5. Regression-based sensitivity analysis

This section deals with regression-based sensitivity analysis given (1) a sample of model inputs from a joint distribution representing the uncertainty about these inputs and (2) a corresponding sample of the model output studied. The model output, given its inputs, may have been produced by specialised modelling software. The procedure calculates the contributions to the variance of the model output from individual or pooled model inputs by means of regression. These contributions are expressed as percentages of the variance of the model output. The top marginal variance of a set of model inputs is calculated as the percentage of variance accounted for when that set of inputs is the only one to be fitted; it is an approximation of the correlation ratio. The calculation is successful if the percentage of variance accounted for by all inputs considered is close to 100, since the analysis only accounts for that part of the variance of the output that is explained by the regression (thus interactions between inputs are not considered). See for instance Saltelli et al (2000) for a detailed account of sensitivity analysis. The bottom marginal variance of a set of inputs is calculated as the increase of variance accounted for when that set is the last to be added to all other inputs. The calculation of single-input uncertainty contributions is only sensible when the number of inputs is moderate: it is pointless for instance when considering uncertainty due to numerous weather inputs or abundant inputs from a spatial stochastic process.

Linear analysis

Linear sensitivity analysis on variate y with model inputs $x_1...x_k$ is based on approximations of model output f by linear functions of x. The top marginal variance of an individual term x_i, TMV_i, is estimated by the variance accounted for by the least squares approximation of the form $a + b_i x_i$. The top marginal variance of a group G of terms $\{x_i \mid i \in G\}$, TMV_G, is estimated by the variance accounted for by the least squares approximation of the form $a + \Sigma_{i\in G} b_i x_i$.

The variance accounted for is given by $MS_{tot} - MS_{res}$, the total mean square minus the residual mean square. Usually it is expressed relative to MS_{tot}, by the so-called fraction of variance accounted for, also named adjusted R^2 (GenStat Committee, 2005).

Similarly, the bottom marginal variance of an individual term x_i, BMV_i, is estimated by the increase in variance accounted for, when the least squares approximation of f(x) of the type $a + \Sigma_{j\neq I} b_j x_j$ is replaced by that of the form $a + \Sigma_j b_j x_j$. The bottom marginal variance of a group G of terms $\{x_i \mid i \in G\}$, BMV_G, is estimated by the increase in variance accounted for, when the least squares approximation of f(x) of the type $a + \Sigma_{j\notin G} b_j x_j$ is replaced by that of the form $a + \Sigma_j b_j x_j$.

In USAGE, the linear analysis for all individual x's can be performed as follows with procedure `RUNCERTAINTY`:

```
VARIATE   [NVALUES=n] x[1...k], y
...
MODEL     y
RUNCERTAINTY X=x[1...k]
```

The linear analysis for pooled x's, e.g. x[1,2], x[3] and x[4…k], can be performed by using pointers as follows:

```
VARIATE   [NVALUES=n] x[1...k], y
...
POINTER   pool1 ; !p(x[1], x[2])
POINTER   pool2 ; !p(x[3])
POINTER   pool3 ; !p(x[4...k])
MODEL     y
RUNCERTAINTY X=pool1, pool2, pool3
```

Obviously, such a linear sensitivity analysis will only work if f(x) can be well approximated by a linear function, as evidenced by a close-to-one value of the adjusted-R^2 of the full approximation $a + \Sigma_j b_j x_j$.

Additive analysis

If model output f cannot be well approximated by a linear function of the x's, one may try to approximate f(x) by a more general additive function, using splines for instance. Analogous to linear analysis, spline sensitivity analysis is based on comparison of the variances accounted for by different least squares approximations of model output f(x).

The top marginal variance of an individual x_i, TMV_i, is estimated by the increase in variance accounted for by an approximation of the form $a + s_i(x_i)$, where $s_i(x_i)$ denotes a smoothing spline in x_i. The top marginal variance of a group G of terms $\{x_i \mid i \in G\}$, TMV_G, is estimated by the variance accounted for by the least squares approximation of the form $a + \Sigma_{i \in G}\ s_i(x_i)$.

Similarly, the bottom marginal variance of individual x_i, BMV_i, is estimated by the increase in variance accounted for, when an approximation of f(x) of the type $a + \Sigma_{j \neq i}\ s_j(x_j)$ is replaced by one of the form $a + \Sigma_j\ s_j(x_j)$. The bottom marginal variance of a group G of terms $\{x_i \mid i \in G\}$, BMV_G, is estimated by the increase in variance accounted for, when the least squares approximation of f(x) of the type $a + \Sigma_{j \notin G}\ s_j(x_j)$ is replaced by that of the form $a + \Sigma_j\ s_j(x_j)$.

The smoothness of the splines can be controlled by means of the so-called *effective number of degrees of freedom* (DF; see GenStat Committee, 2005). DF acts much the same as the number of degrees of freedom of a polynomial: a larger DF results in closer adherence to the data, at the price of less smoothness. When DF is not set, the default value 2 is used.

The analysis is performed by USAGE procedure `RUNCERTAINTY`, using the option `CURVE=SPLINE`:

```
  MODEL     y
  RUNCERTAINTY [CURVE=spline ; DF=2] X=x[]
```

The default setting of the `CURVE` option is `linear`. The effective degrees of freedom of the, individual or pooled, x's is defined by parameter `DF` with default value 2.

Approximation of f() by a sum of splines in the individual inputs may constitute an improvement upon linear approximation, since nonlinearities in the response of f() to individual inputs are allowed for. But if the model is strongly non-additive in the x's, the method fails. Success is evidenced by a close-to-one value of the adjusted-R^2 of the full approximation $a + \Sigma_j\ s_j(x_j)$.

With a sufficiently large DF, the difference between 1 and adjusted-R^2 is the fraction of variance due to interactions.

Since the `RUNCERTAINTY` approximation, with linear functions or splines, can not incorporate interactions, differences between estimates of the TMV and the BMV of an input x_i can only be caused by correlations between the inputs and by sampling variability (see Section 3).

6. Regression-free sensitivity analysis of independent input groups

Suppose that the inputs can be divided into two independent groups S and T. Write the model output studied as f(S, T). The bottom marginal variance of S and the top marginal variance of T can be estimated from a sample of the following structure :

$$\begin{array}{ll} f(S_{11}, T_1) & f(S_{12}, T_1) \\ f(S_{21}, T_2) & f(S_{22}, T_2) \\ \dots & \dots \\ f(S_{N1}, T_N) & f(S_{N2}, T_N) \end{array}$$

where all S_{ij} and T_i are independent realizations of S and T. Denote the above two columns by y_1 and y_2. The bottom marginal variance of S and the top marginal variance of T may be estimated by:

$$\mathrm{B\hat{M}V(S)} = \tfrac{1}{2}\mathrm{Var}(y_1 - y_2)$$

$$\mathrm{T\hat{M}V(T)} = \mathrm{Cov}(y_1, y_2)$$

The two are complementary and add up to the total variance. The example is a very simple case of a class of ANOVA designs and analyses for sensitivity analysis. For methods that enable the estimation of top and bottom marginal variances of more than two independent groups of inputs see Jansen (1994, 1996, 1999) and Sobol (1990, 1995).

7. Examples

The examples in the next subsections are simpler than is typical for uncertainty and sensitivity analysis, except for the last example. Often, principles can be better explained by simple examples. The first subsection shows the similarities between sensitivity analysis and standard regression.

7.1. Parameter uncertainty after regression

Regression analysis leads to an estimate b of the parameter vector β. The estimation quality is evidenced by an estimate V of the covariance matrix Σ of b, and by the number v of residual degrees of freedom. Under favourable circumstances, in particular a large and informative data-set, the estimate b of parameter vector β is approximately multinormally distributed with mean β and covariance matrix Σ.

The simplest way to describe uncertainty about β is to express it by a multinormal distribution with b as vector of means, and V as covariance matrix:

$$\beta \sim N(b, V) .$$

This approach, however, neglects that V is only an approximation of the true covariance matrix Σ. The inaccuracy in V is accounted for by the so-called multivariate student distribution characterized by b, V and ν:

$\beta \sim t_\nu$ (b, V).

Provided that ν > 2, the multivariate student distribution has mean b, and variance-covariance matrix [ν /(ν–2)] V. So the student uncertainty distribution conveys more uncertainty than the normal one, especially for small degrees of freedom. (For more details, see Appendix I.)

NOTE. The multinormal and the multistudent distribution have unrestricted ranges. But often parameters are known to be restricted, for instance to positive values, or to fractions between 0 and 1. Sampling from an unrestricted uncertainty distribution may then give rise to problems. The following stratagem may be used to circumvent these problems. Define the uncertainty distribution so that its marginals satisfy the restrictions and have the required means and variances, for instance by means of beta and gamma distributions. Define the correlation structure by means of rank correlation matrix corrmat(V), or better still by corrmat{[ν /(ν–2)] V}. Procedure `GUNITCUBE` has an option to specify the rank correlation matrix. The method just sketched may be justified by the fact that for a multinormal distribution, the correlation matrix and the rank correlation matrix are very nearly equal (see Appendix I).

Linear regression

As an example we re-analyse the 'Tribolium beetles weight loss' data, which are discussed in Sokal & Rolf (1981; chapter 14).

```
VARIATE   loss; !(8.98, 8.14, 6.67, 6.08, 5.90, 5.83, 4.68, 4.20, 3.72)
VARIATE   humidity; !(0, 12.0, 29.5, 43.0, 53.0, 62.5, 75.5, 85, 93.0)
MODEL     loss
FIT       humidity
RKEEP     ESTIMATES=mean; VCOV=vcov ; DF=df
```

The sensitivity analysis can easily be done analytically (loc. cit.). But we take the simulatory road for illustration.

```
\\Draw a multivariate student parameter sample
SCALAR    n ; 1000
VARIATE   [NVALUES=n] a, b
GMULTIVARIATE [DISTRIBUTION=student ; NVALUES=n ; MEAN=mean ; \
          VCOVARIANCE=vcov ; DF=df ; SEED=111296] NUMBERS=!p(a, b)
```

We start with a graphical analysis with humidity in the range 0...100, see Figure 2; and compare it with loc. cit., Figure 14.11.

```
DELETE    [REDEFINE=yes] loss, humidity
CALCULATE humidity = 100 * !(1...n) / n
CALCULATE loss = a + b*humidity                  "Corresponding loss values"
CALCULATE lm = mean$[1] + mean$[2] * humidity    "loss at mean parameter values"
PEN       1,2 ; METHOD=point,line ; SYMBOL=2,0 ; LINESTYLE=0,1 ; COLOUR=1 ; \
          SIZE=0.3,* ; THICKNESS=*,2
DGRAPH    [KEYWIN=0; TITLE='loss vs humidity'] loss,lm ; humidity ; PEN=1,2
```

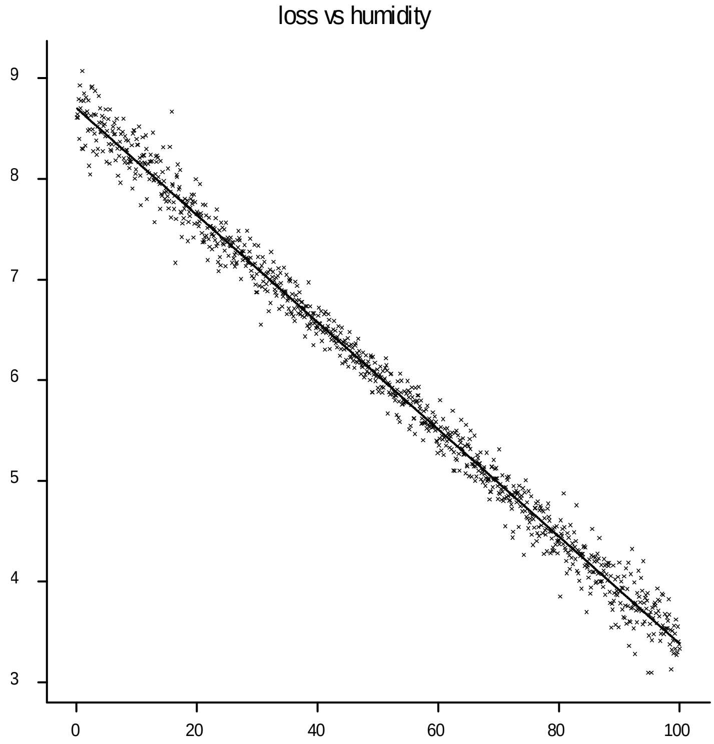

Figure 2: regression line and simulated predicted values at mean parameter values

Next we can calculate the median and 95% two-sided confidence limits at humidity say 100% (loc. cit., Box 14.3, Item 7).

```
CALCULATE humidity = 100
CALCULATE loss = a + b*humidity
SUMMARIZE [PRINT=#,quantiles; PROPORTIONS=0.025, 0.5, 0.975] loss
```

The above lines yield the 95% interval (3.0, 3.8), and the median 3.4. The interval approximates the exact interval, which can also be calculated analytically. A larger sample should improve the approximation.

Photosynthesis

In the previous linear regression, the sensitivity analysis could be most efficiently performed by hand. The only possible advantage of the Monte Carlo analysis lies in its conceptual transparency. With nonlinear models, however, Monte Carlo simulation may be preferable in every aspect to hand calculation, since hand calculation will often involve an approximation error that is worse than the Monte Carlo sampling error. The sampling error can be made as small as desired, given enough computer time. The nonlinear saturation curve for photosynthesis

$$\text{photos} = \text{amax} + (\text{rd} - \text{amax}) \times \exp[\text{rad} \times \text{eff} / (\text{rd} - \text{amax})]$$

is fitted to a small but very accurate dataset.

```
VARIATE   rad ; !(0.00, 103.5, 327.02, 484.78, 736.96, 996.12)
VARIATE   photos ; !(-0.81, 3.61, 10.48, 12.99, 14.48, 14.91)
EXPRESSION e; !e(fit = amax + (rd-amax)*EXP(rad*eff/(rd-amax)))
MODEL     photos ; FIT=fit
\\Initial values stem from analysis of similar data
RCYCLE    amax, rd, eff; ini=15,  1, .05
```

```
FITNONLINEAR [CALCULATION=e; PRINT=model,summ,esti,corr]
RKEEP     ESTIMATES=m ; VCOVARIANCE=vcov ; DF=df
```

The sensitivity analysis starts with random draws from the parameter distribution.

```
DELETE    [REDEFINE=yes] photos, rad, amax, rd, eff
SCALAR    n ; 1000
VARIATE   [NVALUES=n] amax, rd, eff
\\Draw multivariate student parameter sample
GMULTIVARIATE [DISTRIBUTION=student ; NVALUES=n ; MEAN=mean ; \
          VCOVARIANCE=vcov ; DF=df ; SEED=111296] !p(amax, rd, eff)
```

We start with a graphical analysis with radiation in the range 0...1000 (see Figure 3).

```
CALCULATE rad = 1000 * !(1...n) / n
\\Calculate corresponding photosynthesis values
CALCULATE photos= amax + (rd-amax)*EXP(rad*eff/(rd-amax))
\\pm is photosynthesis curve at mean parameter values
CALCULATE pm = mean$[1] + (mean$[2]-mean$[1]) * \
          EXP(rad*mean$[3]/(mean$[2]-mean$[1]))
PEN       1,2 ; METHOD=point,line ; SYMBOL=1,0 ; LINESTYLE=0,1 ; COLOUR=1 ; \
          SIZE=0.3,* ; THICKNESS=*,2
DGRAPH    [KEYWIN=0 ; TITLE='photos vs rad'] photos,pm ; rad ; PEN=1,2
```

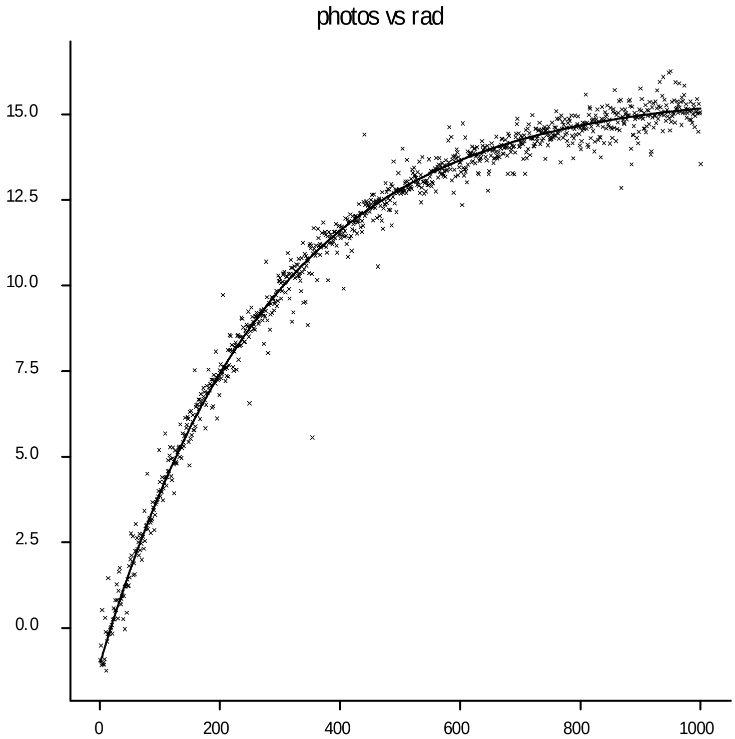

Figure 3: fitted curve and simulated predicted values at mean parameter values

By way of example, we calculate median and 95% two-sided confidence limits at rad say 500

```
CALCULATE rad = 500
CALCULATE photos= amax + (rd-amax)*EXP(rad*eff/(rd-amax))
SUMMARIZE [PRINT=#,quantiles ; PROPORTIONS=0.025, 0.5, 0.975] photos
```

These lines yield the 95% interval (11.8, 13.4), with 12.8 as median.

7.2. A historical example: smallpox inoculation

Smallpox was a dangerous infectious disease, which has been compared in virulence to the plague (Carey, 1995). It killed, for instance, five reigning monarchs during the eighteenth century. Early in the eighteenth century, in some upper circles in Europe, an old Turkish medical technique came into vogue, called inoculation or variolation. It was a method of immunization against smallpox by means of a slight artificial infection with this disease (i.e. human smallpox). Inoculation, however, was not altogether harmless: one might die from the artificial infection. This risk had to be compared with the permanent risk of natural infection. The physician and mathematician Daniel Bernoulli constructed a model that should enable comparison of the two risks. The following analysis proceeds from a very individualistic point of view. For instance, the risk of contaminating others under either action is neglected. Bernoulli, however, also took the risks and benefits for society into account.

The simple model used by Bernoulli (1760) goes as follows. A susceptible has a constant probability density α in time to get infected. If infected, he has probability β to die from the infection, and probability $1\text{-}\beta$ to recover and stay immune for the rest of his life. Probability density α was estimated at 1/8 per year, and probability β at 1/8; independent of the number of infected in the neighbourhood, of one's age, of place, time etcetera. The equality of the two numbers is a coincidence. Bernoulli was well aware that his model, and its parameters, were only approximate, but he stated that it conformed reasonably to the facts known. To test the validity of the model, Bernoulli calculated the fraction, π say, of every generation that would be killed by smallpox, taking into account all other competing causes of death. This fraction π would amount to about 1/14, which was accepted as a realistic figure. In the further analysis, we will assume that the 95% interval of π ranges from about 1/28 to about 1/7.

On the other hand, there is the risk, say γ, to die through the inoculation. The victim of the artificial infection would die shortly afterwards, but for simplicity the model assumes immediate death. The estimates for γ ranged from a very optimistic 1/1200 to a very pessimistic 1/60. Bernoulli worked with the estimate 1/200.

Under this simple model, Bernoulli calculated analytically the relative gain from inoculation at time t reckoned from the moment t_0 of inoculation. The relative gain R(t) is defined as the difference of the probabilities to live for at least t years after t_0, when inoculated and when not inoculated, relative to the latter one.

$$R(t) = (1 - \gamma) / (1 - \beta + \beta e^{-\alpha t}) - 1 \qquad (t > 0).$$

The relative gain equals $-\gamma$ just after inoculation; for large t, it will tend to the limit $(1-\gamma)/(1-\beta) - 1$. If $\beta > \gamma$, the gain will eventually become positive: the first positive value occurs after time lag

$$\tau = (1 / \alpha) \ln [\beta / (\beta - \gamma)].$$

Since the model is a simplification, and since parameters α, β and γ were only known inaccurately, the decision to inoculate was no easy one. The mathematician d'Alembert (1761) fiercely criticized the model of Bernoulli because of its simplifications and uncertainties. Just like Bernoulli, he was an ardent advocate of inoculation, but he found that more convincing arguments were required. D'Alembert hoped that by a refinement of the inoculation technique, its risk would decrease to the level of the risk of deadly contagion, within a month say, by natural smallpox. In 1798, these hopes were realized by Jenner, a country doctor, who developed *vaccination*, a novel immunization technique based on cowpox instead of human pox.

One may ask whether the fierce critique of d'Alembert was truly rational. In order to shred some light on this question, we shall perform a sensitivity analysis. The analysis is intended as an amusing illustration, and should not be taken very seriously, because the parameter uncertainties are defined rather loosely. Moreover, one should be aware that structural model errors are not addressed by the analysis (as usual).

In the sensitivity analysis we assume that α, β and γ are independent, with the distributions given below; the 95%-intervals of the distributions are given in the columns low and high:

Uncertainty distributions of smallpox model parameters. The parameters are assumed to be independent. The means are Bernoulli's estimates.

parameter	Type	mean	variance	minimum	maximum	low	high
α	gamma	0.125	0.004	0	∞	0.033	0.28
β	beta	0.125	0.001	0	1	0.070	0.19
γ	beta	0.005	0.00001	0	1	0.00082	0.013

The means of these distributions are equal to the figures used by Bernoulli. The variances have been chosen after some computer experimentation so as to conform reasonably well to the uncertainties mentioned above. The 95%-interval of γ conforms to the optimistic and pessimistic estimates from d'Alembert. The intervals for α and γ have been taken quite large, but such that the interval for π would not become too wide. It appeared that β contributed most to the uncertainty in π: for that reason the variance of β was taken smaller than that of α. With the above parameter uncertainties, the ensuing 95% interval for π ranges approximately from 0.040 to 0.13, as intended.

A sample of size 1000 is constructed as follows with GenStat.

```
SCALAR    k ; 3
SCALAR    n ; 1000
SCALAR    seed ; 161096
VARIATE   [NVALUES=n] alpha, beta, gamma, pi, tau, rgain1, rgain18, uni[1...k]
SCALAR    mean[1...k] ; 0.125, 0.125, 0.00500
SCALAR    var[1...k]  ; 0.004, 0.001, 0.00001
GUNITCUBE [NVALUES=n ; STRATIFICATION=latin ; SEED=seed] NUMBERS=uni
EDCONTINUOUS [DISTRIBUTION=gamma ; MEAN=mean[1] ; VAR=var[1]] uni[1] ; alpha
EDCONTINUOUS [DISTRIBUTION=beta  ; MEAN=mean[2] ; VAR=var[2]] uni[2] ; beta
EDCONTINUOUS [DISTRIBUTION=beta  ; MEAN=mean[3] ; VAR=var[3]] uni[3] ; gamma
SUMMARIZE [PRINT=#,quantiles ; PROPORTIONS=0.025, 0.975] alpha, beta, gamma
```

Next, outside of GenStat the model software calculates the corresponding model outputs, namely the probability π, the time lag τ (if $\gamma \geq \beta$ a missing value is produced), and the relative gains, R(1) and R(18). The subsequent uncertainty and sensitivity analysis is done as follows in GenStat.

```
SUMMARIZE [PRINT=#,quantiles ; REPRESENTATION=stand ; \
          PROPORTIONS=0.05,0.25,0.5,0.75,0.95] pi, tau, rgain1, rgain18
FOR yy=pi, tau, rgain1, rgain18
  MODEL     yy
  RUNCERTAINTY x=alpha, beta, gamma
  RUNCERTAINTY [CURVE=spline] X=alpha, beta, gamma ; DF=4
ENDFOR
```

The main results of the analysis are the estimates of the characteristics (mean, quantiles etcetera) of the distributions of major model outputs. The table below contains some quantiles.

quantile	π	τ	R(1)	R(18)
0.050	0.044	0.065	-0.0019	0.052
0.250	0.061	0.170	0.0037	0.083
0.500	0.074	0.318	0.0083	0.107
0.750	0.090	0.599	0.0146	0.136
0.950	0.116	1.332	0.0254	0.183

In the present case, the estimates of uncertainty contributions are not very useful, since the main question is whether to be inoculated, and not what research might be most effective in order to reduce the uncertainty. But it does no harm to estimate uncertainty contributions just for the exercise.

Uncertainty about time lag τ

The probability that $\gamma < \beta$ is estimated to be 1 (no missing values occurred in τ). The length τ of the period in which inoculation has a negative impact, is less than 1.3 year with probability 0.95; the median value equals 0.3 year. The spline sensitivity analysis has adjusted R^2=71%, considerably more than the 58% of the linear analysis; the difference is largely caused by nonlinearity of the response to α (see Figure 4). It is seen that the infection pressure α and the inoculation risk γ contribute much to the uncertainty about τ.

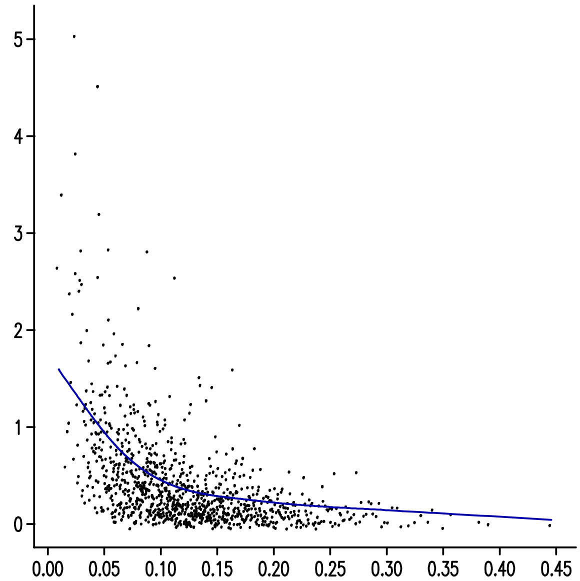

Figure 4 Sampled values of τ versus those of α; a spline fit has been added.

Uncertainty about relative gain one year after inoculation

It is seen that the mean and median relative gain one year after inoculation are merely 1%. The probability of negative gain after one year is still over 5% (which corresponds with the above analysis of τ). A linear analysis has a nice adjusted R^2=94.8%, which is hardly improved by a spline analysis. Infection pressure α contributes most to the uncertainty.

Uncertainty about relative gain 18 years after inoculation

The relative gain 18 years after inoculation is much larger; the median gain is estimated at 11%. A spline analysis accounts for 97% of the uncertainty: quite satisfying, and appreciably better than the 88% of the linear analysis. Smallpox death risk parameter β causes most of the uncertainty, whereas the contribution of inoculation risk γ is negligible: that minor risk is almost forgotten after 18 years.

Indeed, it seems that d'Alembert was right that a diminishment of the risk of inoculation would be of great help in convincing people of the advantages of inoculation. But even with the state of the art and the uncertainties of that time, the advantages of inoculation seem to outnumber the disadvantages.

7.3. Bootstrap percentile confidence interval for analysis results

In this subsection we perform sensitivity analysis on a test function rather than on a model. The test function is given by

$$f(x) = x_1^2/ \sqrt{2} + (x_2+x_3) / \sqrt{(7/4)} + 2\ (x_4-x_5) + x_6x_7 \sqrt{2} + x_8.$$

The arguments are assumed to have the following distribution: the marginal distribution of each x_i is normal with mean 0 and variance 1; all correlations are equal 0, except $\rho(x_2 , x_3)$ and $\rho(x_4 , x_5)$, which equal 3/4.

The function f(x) may be written as a sum of functions of independent groups, namely the groups $\{x_1\}$, $\{x_2 , x_3\}$, $\{x_4 , x_5\}$, $\{x_6 , x_7\}$, $\{x_8\}$. They may be described as independent groups with additive effects. For such groups, the top and bottom marginal uncertainty contributions are equal. Thus one might speak unequivocally about the uncertainty contributions of these groups. They are listed in the following table.

Uncertainty contributions of independent groups with additive effects.

Group	absolute	relative (%)
x_1	1	12.5
x_2, x_3	2	25
x_4, x_5	2	25
x_6, x_7	2	25
x_8	1	12.5

The sensitivity analysis in the next example is based on spline regression. It estimates the uncertainty contributions of the groups mentioned. Since all terms in f(x) can be ‘seen’ by a spline of this type, we should expect some 25% of the variance to remain out of sight of the sensitivity analysis.

The analysis for a random sample with 1000 draws is done with the next program fragment.

```
SCALAR    seed ; 231205
SCALAR    n ; 1000
VARIATE   [NVALUES=n] x[1...8], y
SYMMETRIC [ROWS=8] vcov
DIAGONAL  [ROWS=8] identity ; !(8(1))
CALCULATE vcov = identity
CALCULATE vcov$[3][2] = 0.75
CALCULATE vcov$[5][4] = 0.75
\\Draw input sample
GMULTIVARIATE [DISTRIBUTION=normal ; NVALUES=1000 ; VCOVARIANCE=vcov ; \
          SEED=seed] NUMBERS=x
\\Calculate test function for sampled values
CALCULATE y = x[1]*x[1]/SQRT(2) + (x[2]+x[3])/SQRT(1.75) + \
          (x[4]-x[5])*2 + x[6]*x[7]*SQRT(2) + x[8]
\\Sensitivity analysis for partially grouped model inputs
POINTER   p1, p23, p45 ; !p(x[1]), !p(x[2,3]), !p(x[4,5])
POINTER   p67, p8 ; !p(x[6,7]), !p(x[8])
MODEL     y
RUNCERTAINTY [CURVE=spline ; TOP%=top] X=p1, p23, p45, p67, p8
```

A fragment from the resulting output is given below

```
Uncertainty analysis
====================

  Response variate: y
   Number of units: 1000
              Mean: 0.821
          Variance: 8.330
       R2-adjusted: 75.5

Bottom and top uncertainty contributions based on smoothing spline fit
----------------------------------------------------------------------

         input       bottom%      top%   Sumdf
            p1          11.3      11.1       2
           p23          25.0      28.0       4
           p45          23.9      25.6       4
           p67           0.1       0.0       4
            p8          12.4      13.1       2
```

Note that the estimates of the top and bottom marginal variances are close to each other, which is an expression of the fact that in this example the theoretical values of both types of variance components are equal, while the estimates are calculated differently. The blindness of the analysis for the effect of x_6 and x_7 is precisely as expected. The low value (75%) of adjusted R^2 might form a reason to turn to a regression-free alternative.

We will nevertheless continue this example to demonstrate how one can calculate confidence limits for an analysis based on a sample of independent consecutive draws, instead of say a latin hypercube sample. The calculation is quite elementary. You just draw nboot samples of size N, from the original sample of size N. In the new sample, some elements of the original may occur more than once, while some other elements may be absent: one draws ‘with replacement’. Repeat the original sensitivity analysis for each new sample and store the sensitivity coefficients. The α and $(1-\alpha)$ percentiles of the nboot values thus obtained for each result, constitute an $(1-2\alpha)$ bootstrap percentile confidence interval (Efron & Tibshirani, 1993). The calculation is straightforward:

```
\\90% bootstrap confidence interval for top marginal variances
SCALAR    nboot, nsample ; 100, 1000
VARIATE   [NVALUES=nsample] xsample[1...8], ysample, index
VARIATE   [NVALUES=nboot] t1, t23, t45, t67, t8
POINTER   psample1, psample23, psample45, psample67, psample8 ; \
          !p(xsample[1]), !p(xsample[2,3]), !p(xsample[4,5]), \
          !p(xsample[6,7]), !p(xsample[8])
FOR [NTIMES=nboot ; INDEX=ii]
  CALCULATE index = 1 + INTEGER(nsample * URAND(0 ; nsample))
  CALCULATE ysample,xsample[] = (y,x[])$[index]
  MODEL     ysample
  RUNCERTAINTY [PRINT=* ; CURVE=spline ; TOP%=top] X=psample1, psample23, \
            psample45, psample67, psample8
  CALCULATE (t1,t23,t45,t67,t8)$[ii] = top$[1...5]
ENDFOR
SUMMARIZE [PRINT=#,quantiles ; PROPORTIONS=0.05, 0.95] t1, t23, t45, t67, t8
```

A fragment from the result is given below:

```
Summary description
-------------------

   Variate        Mean          Sd      Median   Nmv    Nval
        t1   1.150E+01   2.472E+00   1.128E+01     0     100
       t23   2.823E+01   2.296E+00   2.817E+01     0     100
       t45   2.595E+01   2.235E+00   2.582E+01     0     100
       t67   8.672E-01   5.855E-01   8.785E-01     0     100
        t8   1.336E+01   1.991E+00   1.323E+01     0     100

* Quantiles      0.050       0.950
   Variate
        t1   8.067E+00   1.584E+01
       t23   2.453E+01   3.188E+01
       t45   2.184E+01   2.968E+01
       t67   8.303E-03   1.833E+00
        t8   9.954E+00   1.683E+01
```

8. References

Abramowitz, M. & Stegun, I.A., 1965, *Handbook of mathematical functions*, Dover, New York.

Box, G.E.P. & Tiao, G.C., 1973, *Bayesian inference in statistical analysis*, Addison-Wesley, 1973 (reprint: Wiley 1992).

Efron, B. & Tibshirani, R.J., 1993, *An introduction to the bootstrap*, Chapman & Hall, London.

Carey, J. (ed.), 1995, *The Faber book of science*, Faber, London.

GenStat Committee, 2005, *GenStat Eighth Edition*, VSN International, Hemel Hempstead.

Hedges, L.V. & Olkin, I., 1985, *Statistical methods for meta-analysis*, Academic Press.

Iman, R.L. & Conover, W.J., 1980, *Small sample sensitivity analysis techniques for computer models with an application to risk assessment*, Communications in Statistics A: theory and methods, 9, 1749-1842.

Iman, R.L. & Conover, W.J., 1982, *A distribution-free approach to inducing rank correlation among input variables*, Commun. statist.-simula. computa.,11(3), 311-334.

Jansen, M.J.W. & Rossing, W.A.H. & Daamen, R.A., 1994, *Monte Carlo estimation of uncertainty contributions from several independent multivariate sources*, In: Grasman, J. & Van Straten, G. (eds.), Predictability and Nonlinear Modelling in Natural Sciences and Economics, p334-343, Kluwer, Dordrecht.

Jansen, M.J.W., 1996, *Winding stairs sample analysis program WINDINGS 2.0*, GLW-note MJA-96-2, GLW-DLO, Wageningen.

Jansen, M.J.W., 1997, *Maximum entropy distributions with prescribed marginals and normal score correlations*, Pages 87-92 in: Beneš V. & Štěpán (Eds.), Distributions with given marginals and moment problems, Kluwer, Dordrecht.

Jansen, M.J.W., 1999, *Analysis of variance designs for model output*, Computer Physics Communications 117, 35-43.

Jansen, M.J.W., 2005, *ASSA: Algorithms for stochastic sensitivity analysis*. Version 1.0. Wettelijke Onderzoekstaken Natuur & Milieu, Werkdocument 4. Website: www.wotnatuurenmilieu.wur.nl > publicaties > werkdocumenten > 2005.

Janssen, P.H.M. & Heuberger, P.S.C. & Sanders, R., 1992, *UNCSAM 1.1: a software package for sensitivity and uncertainty analysis. Manual*, report 959101004, National Institute of Public Health and Environmental Protection, Bilthoven.

Janssen, P.H.M., 1994, *Assessing sensitivities and uncertainties in models: a critical evaluation*, In: Grasman, J. & Van Straten, G. (eds.), Predictability and Nonlinear Modelling in Natural Sciences and Economics, p344-361, Kluwer, Dordrecht.

Janssen, P.H.M. & Heuberger, P.S.C., 1995, *Rotated-Random-Scanning: a simple method for set-valued model calibration*. Report 733001004, National Institute for Public Health and Environmental Protection, Bilthoven.

Keesman, K. & Van Straten, G., 1990, *Set membership approach to identification and prediction of lake eutrophication*. Water Resources research 26: 2643-2652

Kleijnen, J.P.C., 1987, *Statistical tools for simulation practitioners*, Dekker, New York.

Krzykacz, B., 1990, *SAMOS: a computer program for the derivation of empirical sensitivity measures of results from large computer models*, Gesellschaft für Reaktorsicherheit, GRS-A-1700, 1990.

McKay, M.D. & Beckman, R.J. & Conover, W.J., 1979, *A comparison of three methods for selecting values of input variables in the analysis of output from a computer code*, Technometrics, 21, 239-245.

McKay, M.D., 1996, *Variance-based methods for assessing uncertainty importance in NUREG-1150 analyses*, Los Alamos National Laboratory, LA-UR-96-2695. Also available as http://www.jrc.it/uasa/services/samo_group/download/paper.ps

Owen, A.B., 1992, *A central limit theorem for latin hypercube sampling*, J.R.Statist.Soc.B 54, 541-551.

Press, W.H. & Flannery, B.P. & Teukolsky, S.A. & Vettering, W.T., 1992, *Numerical recipes in C: the art of scientific computing*, second edition, Cambridge University Press, Cambridge.

Saltelli, A. & Tarantola, S. & Chan, K.P.S., 1999, *A quantitative model-independent method for global sensitivity analysis of model output*, Technometrics 41, 39-56.

Saltelli, A. & Chan, K. & Scott, E.M., 2000, *Sensitivity analysis*, Wiley, Chichester.

Sobol, I.M., 1990, *Sensitivity estimates for nonlinear mathematical models*, Matematicheskoe Modelirovanie 2, 112-118 (in Russian), translated in Mathematical Modelling and Computational Experiments, vol 1, 407-414.

Sobol, I.M., 1995, *Sensitivity analysis of nonlinear models using sensitivity indices*, International symposium SAMO95: theory and applications of sensitivity analysis of model output in computer simulation, September 1995, Belgirate, Italy.

Sokal, R.R & Rolf, F.J., 1981, *Biometry: the principles and practice of statistics in biological research*, Freeman, New York.

Stein, M., 1987, *Large sample properties of simulations using latin hypercube sampling*, Technometrics, 29, 143-151.

Van Lenthe, J. & Molenaar, W., 1993, *ELI: ELIciatation of uncertain knowledge. Preliminary manual.*, iecProGAMMA, Groningen.

9. Appendix I: Some mathematical details

9.1. GUNITCUBE

Ordinary random samples

In this subsection it will be shown how procedure GUNITCUBE with options RCORRELATION=rc, METHOD=simple & STRATIFICATION=none draws a sample from a continuous multivariate distribution with standard homogeneous marginals and rank correlations very close to the desired rank correlation.,

The procedure is based on the property that the Pearson and rank correlations of a multinormal distribution are very nearly equal (see Figure 5). Applying this property, GUNITCUBE works as follows. Firstly, a multinormal sample of k variates, say z_1 ... z_k, is drawn with mean 0, and covariance matrix C. The standard normal marginals z_i are transformed into standard homogeneous x_i by means of the mapping $x_i = \Phi(z_i)$, where Φ denotes the standard normal distribution function.

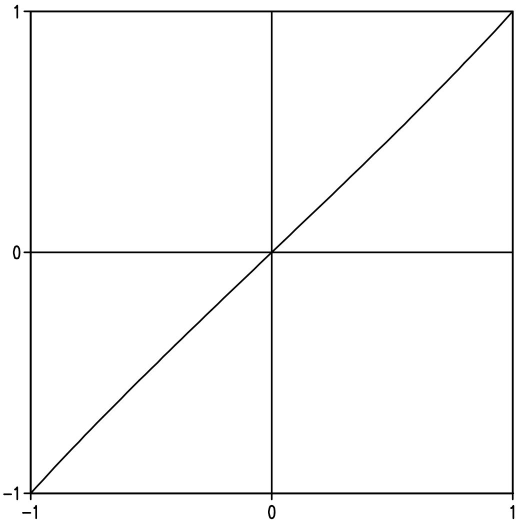

Figure 5 Bivariate normal distribution: rank correlation versus ordinary correlation.

Pearson and rank correlation of the normal distribution

Before we can demonstrate the property mentioned, we have to introduce the concept of rank correlation for random variates, since, originally, rank correlation is only defined for samples. The distributional rank correlation between two continuous random variates x_1 and x_2, with marginals F_1 and F_2 is defined as the correlation between the corresponding standard homogeneous variates $F_1(x_1)$ and $F_2(x_2)$. Some authors use the term grade correlation instead.

Obviously, the distributional rank correlation between two standard homogeneous variates is equal to their ordinary Pearson correlation. Moreover, the distributional rank correlation between two variates is invariant under monotonically increasing transformations per variate. There exists a close connection between distributional and sample rank correlation: the sample rank correlation of a large ordinary random sample from a pair of variates with distributional rank correlation ρ^*, will tend to ρ.

It will be shown that the distributional rank correlation between any two standard homogeneous variates x_i and x_j from which GUNITCUBE draws a sample is close to the desired value c_{ij}:

$$\begin{aligned} \mathrm{rcorr}(x_i, x_j) &= \mathrm{corr}(x_i, x_j) \\ &= (6/\pi) \arcsin(c_{ij}/2) \\ &= c_{ij} + \eta_{ij} \end{aligned}$$

in which the approximation error η_{ij} satisfies $|\eta_{ij}| \leq 0.018$.

We will give a proof for the first two variates x_1 and x_2. Amazingly, no such proof was given by Iman and Conover (1982), who proposed the method. Denote their desired rank correlation c_{12} by ρ, and denote their actual rank correlation by ρ^*. The variates z_1 and z_2 are bivariate normal with standard normal marginals and correlation ρ. Thus, $x_1=\Phi(z_1)$ and $x_2=\Phi(z_2)$ are standard homogeneous; so both have mean 1/2 and variance 1/12. Their correlation may be calculated via the introduction of two auxiliary standard normal variates, ε_1 and ε_2 that are independent of each other and of z_1 and z_2. By the definition of Φ, one has

$$x_i = \Phi(z_i) = P(\varepsilon_i < z_i) = P(\varepsilon_i - z_i < 0) ,$$

so that the expectation $E[x_1x_2]$ satisfies

$$E[\Phi(z_1)\ \Phi(z_2)] = E[P(\varepsilon_1 - z_1 < 0 \mid z_1)\ P(\varepsilon_2 - z_2 < 0 \mid z_2)] = P(\varepsilon_1 - z_1 < 0\ \cap \varepsilon_2 - z_2 < 0).$$

Now $\varepsilon_1 - z_1$ and $\varepsilon_2 - z_2$ have normal distributions with mean 0, variance 2, and correlation $\rho/2$. The probability that both are negative is given by

$$P(\varepsilon_1 - z_1 < 0 \cap \varepsilon_2 - z_2 < 0) = 1/4\ + \arcsin(\rho/2) / (2\pi)$$

(see for instance Abramowitz and Stegun, 1964; formula 26.3.19). So that

$$E[x_1\ x_2] = E[\Phi(z_1)\ \Phi(z_2)] = 1/4\ + \arcsin(\rho/2) / (2\pi).$$

The correlation between x_1 and x_2 follows as $\rho^* = (6/\pi) \arcsin(\rho/2)$; which concludes the first part of the proof. The closeness of ρ^* to ρ is ascertained numerically: $\max(|\rho - \rho^*|)$ appears to have the value 0.018 (see Figure 5).

A different interpretation

In the previous section it was shown that `GUNITCUBE` with a rank correlation matrix draws a sample from a continuous multivariate distribution with standard homogeneous marginals and rank correlations very close to the desired rank correlation. Note that a distribution is not uniquely defined by its marginals and correlation matrix, so that there are more distributions satisfying the specifications.

Amazingly, the procedure can also be interpreted in a different way. `GUNITCUBE` draws from a maximum-entropy distribution with standard homogeneous marginals and normal-score correlation matrix C. This distribution is unique, and its property of maximal entropy is attractive in the context of uncertainty and sensitivity analysis: of all distributions satisfying the given constraints, the one with maximal entropy contains the least information. Adopting any other distribution would be tantamount to assuming that we know more than we actually do (Jansen, 1997).

Restricted random samples

When `GUNITCUBE` is called with the option `METHOD=iman`, the procedure takes a somewhat different road. But again, the procedure is based on the near equality of rank and Pearson correlations in the multinormal distribution. Details of the procedure are given in Iman and Conover (1982). `GUNITCUBE` implements the procedure described there, using van der Waerden scores.

When `GUNITCUBE` is called with the option `STRATIFICATION=latin`, the sample of x's produced thus far is not the final output. After the x's have been drawn as described above, a second sample is drawn: a simple uncorrelated latin hypercube sample, independent of the first sample, having the same dimensions. The final output consists of the values of the latin hypercube sample, ranked component wise according to the x-sample.

9.2. Marginal variances

The variance of y = f(x), induced by the distribution D of x = (x_1...x_k) will be called VTOT

VTOT = Var[y]   y = f(x),   x ~ D

Let S denote a subset of the x's, possibly one single x. The uncertainty contribution of subset S will be expressed in two ways. By the top marginal variance: the variance reduction that would occur in case one would get perfect new information about the inputs S. And by the bottom marginal variance: the variance that will remain as long as one gets no new information about S. In both cases the new information is added to the information already present in input distribution D.

More formally, the variance that would remain in case input group S should become perfectly known, has the expectation E[Var[f(x) | S]]. Accordingly, the top marginal variance TMV_S of S is defined as

TMV_S = VTOT - E[Var[f(x) | S]].

Let -S indicate the complementary subset of all inputs not comprised in S. The variance that would remain in case -S should become perfectly known, has the expectation E[Var[f(x) | -S]]. Thus we define the bottom marginal variance of S as

BMV_S = E[Var[f(x) | -S]].

Obviously

$BMV_S + TMV_{-S}$ = VTOT.

The following well-known variance decomposition rule for conditional distributions

Var[y] = Var[E[y | S]] + E[Var[y | S]]

leads to an equivalent expression for TMV_S:

TMV_S = Var[E[y | S]].

In USAGE, marginal variances are expressed as fractions of VTOT. When S consists of a single input x_i, $\eta^2_i \equiv TMV_i$ / VTOT is equal to the square of the so-called correlation ratio of y and x_i. Note that the correlation ratio is not the same as the correlation coefficient. When E[y | x_i] is linear in x_i, η^2_i is equal to the squared correlation coefficient between y and x_i, say ρ^2_i. But when E[y | x_i] is nonlinear in x_i, η^2_i is greater than ρ^2_i.

Analogously, when S consists of more than one component, $R^2_S \equiv TMV_S$ / VTOT is called the (theoretical) *squared multiple correlation, coefficient of determination*, or *fraction of variance accounted for*; the adjective 'theoretical' is used to convey that the concept is not based upon a specific form of E[y | S] as function of S and because it applies to a distribution rather than to a finite sample.

9.3. Multivariate Student distribution

The multivariate student distribution describes the uncertainty about the coefficients of an ordinary linear regression when the variance of the observations is unknown. The output of such a regression contains a vector of estimates m, a variance-covariance matrix V, and a number of degrees of freedom ν. These characterize the multivariate student distribution.

A multivariate student vector, with parameters b, V and ν, is generated as

$$t_\nu(b, V) \sim b + N(0, V) / \sqrt{}(\chi_\nu^2 / \nu)$$

where N(0, V) is multinormal with mean 0 and variance V; and where the scalar χ_ν^2 has a chi-square distribution with ν degrees of freedom. Note that parameter b is not always the mean of the target distribution: when ν=1, the multistudent distribution has no mean! But reassuringly, when ν > 1, the multistudent distribution has a mean that is equal to b. The parameter V is never equal to the covariance matrix of the multistudent distribution: the latter covariance matrix exists only if ν > 2, and then it equals [ν/(ν-2)] V.

10. Appendix II: Description of USAGE procedures

Contents:

`EDCONTINUOUS`	calculates equivalent deviates for continuous distributions
`GMULTIVARIATE`	generates random numbers from multivariate normal or Student’s t distribution
`GUNITCUBE`	generates random numbers from a distribution with marginal uniform distributions
`RUNCERTAINTY`	calculates contributions of model inputs to the variance of a model output
`SUMMARIZE`	prints summary statistics for variates

Procedure EDCONTINUOUS

M.J.W. Jansen, J.C.M. Withagen & J.T.N.M. Thissen

EDCONTINUOUS calculates equivalent deviates for continuous distributions

Options

DISTRIBUTION = *string* Type of distribution required (beta, gamma, lognormal, normal, uniform); default normal

METHOD = *string* Method by which the defining parameters of the distribution are specified (moments, quantiles); default moments

MEAN = *scalar* Mean of distribution; default *

VARIANCE = *scalar* Variance of distribution; default *

PROPORTIONS = *variate* Two cumulative lower probabilities of distribution; default *

QUANTILES = *variate* Two quantiles (equivalent deviates) corresponding to PROPORTIONS; default *

LOWER = *scalar* Lower bound of beta, gamma, lognormal or uniform distribution; default 0

UPPER = *scalar* Upper bound of beta or uniform distribution; default 1

Parameters

CUMPROBABILITY = *variates* or *scalars*
Cumulative lower probabilities for which equivalent deviates are required; must be set

DEVIATE = *variates* or *scalars* To save equivalent deviates corresponding to CUMPROBABILITY

Description

Procedure EDCONTINUOUS calculates equivalent deviates corresponding to given cumulative lower probabilities for five continuous distributions: beta, gamma, lognormal, normal and uniform. The CUMPROBABILITY parameter specifies the cumulative lower probabilities and the corresponding equivalent deviates are saved by means of the DEVIATE parameter. The DISTRIBUTION option specifies the type of distribution. The METHOD option specifies how the parameters of the distribution are defined. When METHOD=moments the first two moments must be set by the MEAN and VARIANCE options. Alternatively, when METHOD=quantiles the distribution is characterised by a pair of cumulative lower probabilities with corresponding quantiles, and options PROPORTIONS and QUANTILES must be set. The uniform distribution is characterised by the LOWER and UPPER option settings, and other options are ignored. Lower and upper bounds for the other distributions can be specified by options UPPER and LOWER; these must be compatible with other option settings.

Options: DISTRIBUTION, METHOD, MEAN, VARIANCE, PROPORTIONS, QUANTILES, LOWER, UPPER.
Parameters: CUMPROBABILITY, DEVIATE.

Method

Internal calls are made to GenStat's ED-functions EDNORMAL, EDBETA and EDGAMMA. In most cases, the required ED-function parameters are derived from simple, well-known relations between ED-function parameters and moments or quantiles. However, when a beta or gamma distribution is specified by two quantiles, the ED-function parameters are derived by means of the FITNONLINEAR directive, which may cause numerical problems.

Action with RESTRICT

Deviates are only calculated for the set of units to which CUMPROBABILITY is restricted. Other units will remain unaffected.

References

None.

Procedures Used

None.

Similar procedures

GRANDOM generates pseudo-random numbers from probability distributions. GMULTIVARIATE generates pseudo-random numbers from multivariate normal or Student’s t distribution. GRMULTINORMAL generates pseudo-random numbers from the multivariate normal distribution

Example

```
PRINT     !t('Examples of how to use Biometris procedure EDCONTINUOUS') ; \
             JUSTIFICATION=left
VARIATE      cum ; !(0.01, 0.02 ... 0.99)
EDCONTINUOUS [DIST=normal ; METHOD=quantiles ; PROPORTION=!(.05, .95) ; \
             QUANTILES=!(6.9, 8.2)] CUMPROBABILITY=cum ; DEVIATE=v[1]
EDCONTINUOUS [DIST=beta ; METHOD=quantiles ; PROPORTION=!(.25, .75) ; \
             QUANTILES=!(0.3, 0.5)] CUMPROBABILITY=cum ; DEVIATE=v[2]
EDCONTINUOUS [DIST=gamma ; MEAN=2 ; VARIANCE=1] CUMPROBABILITY=cum ; \
             DEVIATE=v[3]
TEXT         title ; 'Example of EDCONTINUOUS: v[1]'
DHISTOGRAM   [WINDOW=5 ; KEY=0 ; TITLE=title    ; SCREEN=keep] v[1]
DHISTOGRAM   [WINDOW=6 ; KEY=0 ; TITLE='v[2]'   ; SCREEN=keep] v[2]
DHISTOGRAM   [WINDOW=7 ; KEY=0 ; TITLE='v[3]'   ; SCREEN=keep] v[3]
DGRAPH       [WINDOW=8 ; KEY=0 ; TITLE='v[2,3]' ; SCREEN=keep] v[2] ; v[3]
```

Procedure GMULTIVARIATE *M.J.W. Jansen, J.C.M. Withagen & J.T.N.M. Thissen*

GMULTIVARIATE generates random numbers from multivariate normal or Student t distribution

Options

PRINT = *string*	Whether to print a summary (summary); default * prints no output
DISTRIBUTION = *string*	Type of distribution required (normal, student); default normal
NVALUES = *scalar*	Number of values to generate; default 1
MEANS = *variate*	The mean for the multivariate Normal or Student's t distribution; default is a variate with values all equal to 0
VCOVARIANCE = *diagonal matrix* or *symmetric matrix*	The variance-covariance matrix for the multivariate Normal or Student's t-distribution; default is to use an identity matrix
DF = *scalar*	Number of degrees of freedom for Student's t distribution; default *
SEED = *scalar*	Seed to generate the random numbers; default 0 continues an existing sequence or initialises the sequence automatically if no random numbers have been generated in this job

Parameters

NUMBERS = *pointers* or *matrices*	Saves the random numbers as either a pointer to a set of variates or a matrix

Description

Procedure GMULTIVARIATE generates pseudo-random numbers from a multivariate Normal or from a multivariate Student's t distribution. The type of distribution can be set by the DISTRIBUTION option. The mean *mu* is specified by the option MEANS as a variate of length *p*; the variance-covariance matrix *Sigma* is specified by the option VCOVARIANCE as a diagonal or symmetric matrix with *p* rows and columns; and the option NVALUES specifies the number of values to be generated. Note that VCOVARIANCE must be positive semi-definite. The DF option must be used to specify the number of degrees of freedom for the Student distribution and must be at least 3.

The SEED option can be set to initialise the random-number generator, hence giving identical results if the procedure is called again with the same options. If SEED is not set, generation will continue from the previous sequence in the program, or, if this is the first generation, the generator will be initialised by CALCULATE.

The numbers can be saved using the NUMBERS parameter, in either a pointer to a set of variates, or a matrix. If the NUMBERS structure or structures are already declared, their dimensions must be compatible with the settings of the NVALUES, MEANS and VCOVARIANCE options. The dimensions are also used, if necessary, to set defaults for the options. By default, MEANS is taken to be a variate of zero values, and VCOVARIANCE is taken to be the identity matrix. If the setting of NUMBERS is not already declared, it will be defined as a pointer to a set of variates with dimensions deduced from the option settings.

Options: PRINT, DISTRIBUTION, NVALUES, MEANS, VCOVARIANCE, DF, SEED.
Parameters: NUMBERS.

Method

Pseudo-random numbers from a multivariate Normal distribution are generated by forming a matrix Y of columns of univariate Normal random numbers, using the Box-Muller method (Box & Muller 1958), followed by a linear transformation

X = A Y + *mu*,

where A is calculated by a Choleski decomposition, AA' = *Sigma*. See, for example, Johnson (1987, pages 52-55) or Tong (1990, pages 181-186). Pseudo-random numbers from the multivariate Student distribution are generated according to the definition of the multivariate Student distribution:

t(*mu*, *Sigma*, *df*) ~ *mu* + MN(0, *Sigma*) / Sqrt(Chi-squared(*df*)/*df*) where MN(0, *Sigma*) is multivariate normal with mean 0 and variance-covariance *Sigma*; and where the scalar Chi-squared(*df*) has a chi-square distribution with *df* degrees of freedom. See, for example, Box & Tiao (1973). Note that the variance-covariance matrix of the multivariate Student distribution equals [*df* / (*df* - 2)] *Sigma*.

Action with RESTRICT

Variates that have been restricted will receive output from `GMULTIVARIATE` only in those units that are not excluded by the restriction. Values in the excluded units remain unchanged. Note that the `NVALUES` option must equal the full size of the variates. Restrictions on the `MEANS` variate are ignored.

References

Box, G.E.P. and Muller, M.E. (1958). A note on generation of normal deviates. *Annals of Mathematical Statistics*, **28**, 610-611.

Johnson, M.E. (1987). *Multivariate Statistical Simulation*. John Wiley & Sons, New York.

Tong, Y.L. (1990). *The Multivariate Normal Distribution*. Springer-Verlag, New York.

Box, G.E.P. & Tiao, G.C. (1973). *Bayesian inference in statistical analysis*. John Wiley & Sons, New York.

Related Procedures

None.

Similar Procedures

`GRMULTINORMAL` generates pseudo-random numbers from a multivariate normal distribution.

Example

```
PRINT      !t('Examples of how to use Biometris procedure GMULTIVARIATE') ; \
              JUSTIFICATION=left
VARIATE       [VALUES=1,2,3] mean
SYMMETRIC     [ROWS=3 ; VALUES=1, 0,4, 1,3,9] vcov
GMULTIVARIATE [NVALUES=100 ; MEANS=mean ; VCOVARIANCE=vcov ; SEED=52] norm
GMULTIVARIATE [PRINT=summary ; DISTRIBUTION=student ; NVALUES=100 ; \
              MEANS=mean ; VCOVARIANCE=vcov ; DF=10 ; SEED=52] stud
DSCATTER      norm[]
DSCATTER      stud[]
```

Procedure GUNITCUBE

M.J.W. Jansen, J.C.M. Withagen & J.T.N.M. Thissen

GUNITCUBE generates random numbers from a distribution with marginal uniform distributions

Options

NVALUES = *scalar* — Number of values to generate; default 1 or deduced from the NUMBERS parameter

RCORRELATION = *scalar* or *symmetricmatrix* — Required rank correlation matrix; default the identity matrix

SEED = *scalar* — Seed to generate the random numbers; default 0 continues an existing sequence or initializes the sequence automatically if no random numbers have been generated in this job

STRATIFICATION = *string* — Stratification (none, latin); default none

METHOD = *string* — Method to achieve rank correlation (simple, iman); default simple

Parameters

NUMBERS = *pointers* or *matrices* — Saves the random numbers as either a pointer to a set of variates or a matrix

Description

Procedure GUNITCUBE generates pseudo-random numbers from a multivariate distribution with marginal distributions that are uniform on the interval from 0 to 1, and with a given rank-correlation matrix RCORRELATION. The numbers can be saved using the NUMBERS parameter, in either a pointer to a set of variates, or a matrix. If the NUMBERS structures are already declared, their dimensions must be compatible with the settings of the NVALUES and RCORRELATION options. Otherwise the dimensions of the NUMBERS pointer are deduced from these options. The dimensions of NUMBERS are also used, if necessary, to set defaults for the options. If NUMBERS is not declared in advance, RCORRELATION must be set. By default RCORRELATION is taken to be the identity matrix. If the setting of NUMBERS is not already declared, it will be defined as a pointer to a set of variates with dimensions deduced from the option settings.

An ordinary random sample is obtained by the option settings STRATIFICATION=none and METHOD=simple. Option setting STRATIFICATION=latin can be used to obtain Latin-hypercube samples, with marginal sample distributions that are very nearly uniform, while option setting METHOD=iman imposes close resemblance between the sample correlation matrix and RCORRELATION.

If RCORRELATION is set, the required rank correlation will be introduced according to the specified METHOD option (thus, METHOD has no effect if RCORRELATION is unset). The combination of RCORRELATION set to an identity matrix and METHOD=simple is stochastically equivalent to RCORRELATION unset.

To avoid values very close to 0 and 1, NUMBERS smaller than 0.000005 and larger than 0.999995 are set to these respective limits.

Options: NVALUES, RCORRELATION, SEED, STRATIFICATION, METHOD.
Parameters: NUMBERS.

Method

The method to construct a latin hypercube sample stems from McKay e.a. (1979). The method to introduce the required rank correlation stems from Iman & Conover (1982).

Action with RESTRICT

Any restrictions on variates of the NUMBERS pointer will be cancelled and all units will be used.

References

Iman, R.L. & Conover, W.J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*, **11(3)**, 311-334.

McKay, M.D. & Beckman, R.J. & Conover, W.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239-245.

Procedures Used

None.

Similar procedures

None.

Example

```
PRINT   !t('Example of how to use Biometris procedure GUNITCUBE') ; \
           JUSTIFICATION=left
SCALAR     nvariates, nvalues, seed ; VALUE=3, 100, 937456
SYMMETRIC  [ROWS=nvariates] corr
CALCULATE  corr = DIAGONAL(!(#nvariates(1)))
CALCULATE  corr$[2,3;1] = -0.8, 0.4
GUNITCUBE  [NVALUES=nvalues ; RCORRELATION=corr ; SEED=seed ; \
           STRATIFICATION=latin ; METHOD=iman] uni
PRINT      MEAN(uni[])
PRINT      VARIANCE(uni[])
CORRELATE  [PRINT=correlations] uni[]
PRINT   !t('Marginal distributions are nearly uniform') ; JUSTIFICATION=left
GROUPS     uni[1...3] ; funi[1...3] ; LIMITS=!(0.1,0.2...0.9)
TABULATE   [CLASSIFICATION=funi[1] ; COUNT=count[1]]
TABULATE   [CLASSIFICATION=funi[2] ; COUNT=count[2]]
TABULATE   [CLASSIFICATION=funi[3] ; COUNT=count[3]]
PRINT      [SERIAL=yes] count[]
DSCATTER   uni[]
```

Procedure RUNCERTAINTY

M.J.W. Jansen, J.C.M. Withagen & J.T.N.M. Thissen

RUNCERTAINTY calculates contributions of model inputs to the variance of a model output

Options

PRINT = *strings* — What to print (fullmodel, uncertainty); default fullmodel, uncertainty

PLOT = *string* — Graphical output required (histogram); default *

CURVE = *string* — Type of curve to be fitted (linear, spline); default linear

ESTIMATES = *variate* — To save regression coefficients of all X variates (only when CURVE=linear)

BOTTOM% = *variate* — To save bottom marginal variances as percentage of the variance of the model output. Increase of percentage variance accounted for when an X structure is last to be added

TOP% = *variate* — To save top marginal variances as percentage of variance of the variance of the model output. Percentage variance accounted for when an X structure is the only one to be fitted.

ADJUSTEDR2 = *scalar* — To save adjusted percentage of variance accounted for by all X variates

Parameters

X = *pointers* or *variates* — Set of model inputs for which uncertainty contributions are to be calculated. If a pointer is specified it must only point to variates

DF = *scalars* — Effective degrees of freedom of the smoothing splines to fit for each X structure; default 2

FITTEDVALUES = *variates* — Variates to store the fitted values for each X structure when that input is the only one to be fitted

Description

Procedure RUNCERTAINTY performs uncertainty analysis given (1) a sample of model inputs from a joint distribution representing the uncertainty about these inputs and (2) a corresponding sample of the model output studied. The model output, given its inputs, may have been produced by specialised modelling software. The procedure calculates the contributions to the variance of the model output from individual or pooled model inputs by means of regression. These contributions are expressed as percentages of the variance of the model output. The top marginal variance of a model input is calculated as the percentage of variance accounted for when that input is the only one to be fitted; it is an approximation of the correlation ratio. The bottom marginal variance of an input is calculated as the increase of variance accounted for when that input is the last to be added to all other inputs. The calculation is successful if the percentage of variance accounted for by all inputs is close to 100, since the analysis only accounts for that part of the variance of the output that is explained by the regression (thus interactions between inputs are not considered). See Jansen et al (2002) and Saltelli et al (2000) for a detailed account of uncertainty analysis.

A call to RUNCERTAINTY must be preceded by a MODEL statement which defines the response variate with the model outputs. Only the first response variate is analysed and options other than WEIGHTS should not be set in the MODEL statement. Generalized models are not allowed. The model inputs are specified by the X parameter that can consist of variates or pointers to one or more variates. If a pointer is specified the total contribution of the variates of the pointer is calculated. The calculation applies multiple linear regression or spline regression of Y on the X structures plus a constant term. The choice between linear and spline regression can be made by means of the CURVE option. When using CURVE=spline, the degrees of freedom of the smoothing spline can be set separately for each X structure by means of the DF parameter. On output the full model has been fitted, and RKEEP and RDISPLAY can be used to further store and display the fit of the full model.

Cases with one or more missing values in the response variate, weight vector or any term in the full model are excluded from the analysis. This implies that, when terms have missing values for different units, FIT used on a subset of model inputs may give different results than RUNCERTAINTY.

The option setting PRINT=fullmodel prints the fit of the full model while suppressing all warning messages. Setting PRINT=uncertainty prints the top and bottom marginal %variances of the X structures and, in case CURVE=linear, the parameter estimates of the full model. The option setting PLOT=histogram option draws a histogram of the top and bottom marginal %variances side by side for each of the X structures. The results of the uncertainty analysis can be saved by means of options ESTIMATES (in case CURVE=linear), BOTTOM%, TOP% and ADJUSTEDR2. The fitted values of the models with individual X structures only (pointers and/or variates) can be saved by means of the FITTEDVALUES parameter. These fittedvalues correspond to the top marginal %variances.

Options: PRINT, PLOT, CURVE, ESTIMATES, BOTTOM%, TOP%, ADJUSTEDR2.
Parameters: X, DF, FITTEDVALUES.

Method

The procedure calculates the percentage of variance accounted for the relevant regressions. The top marginal %variance for an input X is calculated as 100(vary-rmstop)/vary, where vary is the variance of the response and rmstop is the residual mean square of the model with only input X. The bottom marginal %variance for an input X equals 100(rmsbottom-rmsall)/vary, where rmsall is the residual mean square of the full model with all inputs, and rmsbottom is the residual mean square of the full model without input X. A TERMS statement in the procedure deals with missing values in the X variates.

Action with RESTRICT

Only the response variate can be restricted. The analysis is restricted accordingly. Restrictions on the X structures are not allowed. The saved FITTEDVALUES variates will be unrestricted, but only units not excluded by the restriction will have values.

References

Jansen M.J.W. ,Withagen J.C.M. & Thissen J.T.N.M. (2005). *USAGE: uncertainty and sensitivity analysis in a GenStat environment. Manual. Version 2.0.* Wageningen: Biometris.

Saltelli, A. & Chan, K. & Scott, E.M. (2000; eds.). *Sensitivity analysis*. Chichester: Wiley.

Procedures Used

None.

Similar procedures

GMULTIVARIATE and GUNITCUBE can be used to generate random inputs. RSELECT selects best subsets of predictor variables in regression. RSCREEN performs screening tests for generalised or multivariate linear models. RSEARCH helps search through models for a regression or generalised linear model.

Example

```
PRINT     !t('Examples of how to use Biometris procedure RUNCERTAINTY') ; \
             JUSTIFICATION=left
POINTER      par  ; !p(a0, a1, a2)
POINTER      soil ; !p(ph, cd)
READ         par[1...3], esp, soil[1,2], lcdp ; DECIMALS=1
 59 42 43 69 59 66 2199    55 39 48 52 57 54 1726    60 59 50 46 58 43 1631
 53 43 49 53 50 30 1134    49 48 71 52 29 73 1292    64 52 44 55 51 43 1411
 67 67 32 64 62 53 2042    51 49 52 51 47 44 1224    47 33 54 48 30 51 1043
 44 45 48 52 34 42  870    43 44 54 59 64 66 2028    55 40 38 59 46 62 1435
 50 50 48 42 56 47 1374    64 69 54 55 61 50 2004    47 55 57 46 59 40 1405
 39 62 58 53 68 50 1894    61 39 59 47 35 47  948    73 37 52 41 45 38  992
 40 61 50 64 58 49 1616    44 55 56 51 52 50 1388    48 50 41 35 42 60 1167
 51 48 44 58 45 54 1147    76 49 48 48 50 37 1190    47 51 46 28 66 64 1973
 53 44 47 65 44 64 1354  :
MODEL        lcdp
RUNCERTAINTY [CURVE=linear] X=par,esp,soil
RUNCERTAINTY [CURVE=spline] X=par,esp,soil ; DF=1,1,2
```

Procedure SUMMARIZE

J.C.M. Withagen

SUMMARIZE prints summary statistics for variates

Options

PRINT = *strings*	What characteristics to print (mean, sd, %cv, median, min, max, nmv, nvalues, quantiles); default mean, sd, median, nmv, nvalues
PROPORTIONS = *numbers*	Proportions at which to calculate quantiles; default .10, .25, .50, .75, .90
REPRESENTATION = *string*	Representation of values of summary statistics (exponential, standard); default exponential

Parameters

DATA = *variates*	Data to summarize; must be set

Description

Procedure SUMMARIZE calculates summary statistics for values stored in a variate as specified by the DATA parameter. The statistics to be calculated are indicated by the PRINT option. The summary is printed in a table with variate identifiers as rows and names of the summary statistics as columns. If PRINT=quantiles quantiles are calculated at the proportions specified by the PROPORTIONS option and printed in a separate table. By default values are presented in E-format. They can be presented in standard output format by the setting the REPRESENTATION option to standard.

Options: PRINT, PROPORTIONS, REPRESENTATION.
Parameters: DATA.

Method

The procedure uses standard GenStat directives.

Action with RESTRICT

Any restriction on the data will be applied to all calculations.

References

None.

Procedures Used

None.

Similar procedures

DESCRIBE saves and/or prints summary statistics for variates, but in a different format.

Example

```
PRINT  !t('Example of how to use Biometris procedure SUMMARIZE') ; \
         JUSTIFICATION=left
CALCULATE data[1...5] = URAND(50697,4(0) ; 100)
SUMMARIZE [PRINT=#,quantiles ; REPRESENTATION=standard] data[]
```