CentER

Discussion Paper

No. 2006–50

WHITE NOISE ASSUMPTIONS REVISITED: REGRESSION MODELS AND STATISTICAL DESIGNS FOR SIMULATION PRACTICE

By Jack P.C. Kleijnen

May 2006

ISSN 0924-7815

TILBURG UNIVERSITY

White noise assumptions revisited: Regression models and statistical designs for simulation practice

Jack P.C. Kleijnen

Tilburg University, Faculty of Economics and Business Administration,Tilburg, the Netherlands

Abstract

Classic linear regression models and their concomitant statistical designs assume a univariate response and white noise. By definition, white noise is normally, independently, and identically distributed with zero mean. This survey tries to answer the following questions: (i) How realistic are these classic assumptions in simulation practice? (ii) How can these assumptions be tested? (iii) If assumptions are violated, can the simulation's I/O data be transformed such that the assumptions hold? (iv) If not, which alternative statistical methods can then be applied?

Key words: metamodels, experimental designs, generalized least squares, multivariate analysis, normality, jackknife, bootstrap, heteroscedasticity, common random numbers, validation
JEL: C0, C1, C9, C15, C44

0.1 Introduction

Simulation models may be either deterministic or random (stochastic). To analyze the Input/Output (I/O) behavior of simulation models, the analysts often use *linear regression* metamodels; for example, first-order or second-order polynomial approximations of the underlying simulation model. A good analysis (e.g., such as regression analysis) requires a good *statistical design*; for example, a fractional factorial such as a 2^{k-p} design. For more mathematical details and background information see [16] and [20].

In this article, I revisit the *classic assumptions* for linear regression analysis and their concomitant designs. These classic assumptions stipulate *univariate*

output and *white noise*. In practice, however, these assumptions usually do not hold.

In general, the simulation output (say) $\widehat{\mathbf{\Theta}}$ is a *multivariate* random variable. For example, the simulation output (response) $\widehat{\Theta}_1$ may estimate the mean throughput time, and $\widehat{\Theta}_2$ may estimate the 90% quantile of the waiting time distribution. More examples will follow in Section 1.

White noise (say) u is Normally, Independently, and Identically Distributed (NIID) with zero mean: $u \sim NIID(0, \sigma_u^2)$. This definition implies the following assumptions:

(1) *normally* (Gaussian) distributed simulation responses
(2) *no Common Random Numbers* (CRN) across the (say) n factor (input) combinations simulated
(3) a *common variance* (or homoscedasticity) of the simulation responses across these n combinations
(4) a *valid* regression (meta)model; i.e., *zero* expected values for the residuals of the fitted metamodel.

In this article, I raise the following *questions*:

(1) How *realistic* are these classic assumptions?
(2) How can these assumptions be *tested* if it is not obvious that the assumption is violated (e.g., if CRN are used, then the independence assumption is obviously violated)?
(3) If an assumption is violated, can the simulation's I/O data be *transformed* such that the assumption holds?
(4) If not, which *alternative statistical methods* can then be applied?

The answers to these questions are scattered throughout the literature on statistics and simulation. In this article, I therefore try to answer these questions in a coherent way. For more details (including additional references and examples) I refer to my forthcoming book [20].

The remainder of this article is organized as follows. Section 1 discusses multivariate simulation output. Section 2 addresses possible nonnormality of the simulation output, including tests of normality, transformations of simulation I/O data, jackknifing, and bootstrapping. Section 3 covers variance heterogeneity (or heteroscedasticity) of the simulation output. Section 4 discusses cross-correlated simulation outputs, created through CRN. Section 5 discusses nonvalid low-order polynomial metamodels. Section 6 summarizes the major conclusions. An extensive list of references concludes this article.

1 Multivariate simulation output

In practice, the simulation model usually gives multivariate output. A class of practical examples concerns *inventory* simulation models with two outputs: (i) the sum of the holding and the ordering costs, averaged over the simulated periods; (ii) the service (or fill) rate, averaged over the same simulation periods. The precise definitions of these costs and the service rate vary with the applications; see [31] and also [1] and [15].

The *case study* in [18] concerns a Decision Support System (DSS) for production planning based on a simulation model. Originally, this simulation model had a multitude of outputs. However, to support decision making, it turned out that it sufficed to consider only the following two outputs (DSS criteria, bivariate response): (i) the total production of steel tubes manufactured (which was of major interest to the production manager); (ii) the 90% quantile of delivery times (which was the sales manager's concern).

A *general notation* is

$$\mathbf{w} = s(d_1, \ldots, d_k, \mathbf{r}_0) \tag{1}$$

with

$\mathbf{w}$: vector of r simulation outputs, so $\mathbf{w} = (w_0, \ldots, w_{r-1})'$ (in simulation optimization it is traditional to label the r outputs starting with zero instead of one);

$s(.)$: mathematical function implicitly defined by the computer code implementing the given simulation model;

d_j: factor (input variable) j of the simulation model, so $\mathbf{D} = (d_{ij})$ is the design matrix for the simulation experiment, with $j = 1, \ldots k$ and $i = 1, \ldots, n$ where n denotes the fixed number of combinations of the k factor levels (or values) in that experiment;

$\mathbf{r}_0$: vector of PseudoRandom Number (PRN) seeds.

I assume that the multivariate I/O function $s(.)$ in (1) is approximated by r univariate low-order polynomials:

$$\mathbf{y}_h = \mathbf{X}\boldsymbol{\beta}_h + \mathbf{e}_h \text{ with } h = 0, \ldots r-1 \tag{2}$$

with

$\mathbf{y}_h$: n-dimensional vector $(y_{1;h}, \ldots, y_{n;h})'$ with the regression predictor y_h for simulation output w_h;

$\mathbf{X}$: common $n \times q$ matrix of explanatory variables $(\mathbf{x}_{ij})$ with $\mathbf{x}_{ij}$ the value

of explanatory variable j in combination i ($i = 1, \ldots, n; j = 1, \ldots, q$); for simplicity, I assume that all fitted regression metamodels are polynomials of the same order (for example, either first order or second order) (if $q > 2$ including an intercept, then the metamodel is called a *multiple* regression model);

$\boldsymbol{\beta}_h$: q-dimensional vector $(\beta_{1;h}, \ldots, \beta_{q;h})'$ with the q regression parameters for the h^{th} metamodel;

$\mathbf{e}_h$: n-dimensional vector $(e_{1;h}, \ldots, e_{n;h})'$ with the residuals for the h^{th} metamodel, in the n combinations.

The multivariate regression model in (2) violates the classic assumptions; i.e., the multivariate residuals $\mathbf{e}$ have the following two properties:

(1) The univariate residuals e_h have variances that vary with the output variable w_h ($h = 1, \ldots r$): $\sigma_h^2 \neq \sigma^2$ (e.g., simulated inventory costs and service percentages have different variances, $\sigma_1^2 \neq \sigma_2^2$).
(2) The univariate residuals e_h and $e_{h'}$ are not independent for a given input combination i: $\sigma_{h;h';i} \neq 0$ for $h \neq h'$. Obviously, if these covariances (like the variances) would not vary with the combination i, then this property could be written as $\sigma_{h;h';i} = \sigma_{h;h'} \neq 0$ for $h \neq h'$ (e.g., 'unusual' PRN streams in a given combination i may result in inventory costs that are 'relatively high'—that is, higher than expected—and a relatively high service percentage, so these two outputs are positively correlated: $\sigma_{1;2} > 0$).

These two properties violate the classic assumptions. Consequently, it seems that the univariate Ordinary Least Squares (OLS) estimators should be replaced by the *Generalized Least Squares* (GLS) estimator of the parameter vector in the corresponding *multivariate regression* model. Fortunately, [35] (a more recent reference is [36], p. 703) proves that GLS reduces to OLS computed per output if the *same design matrix* is used (as is the case in (2)); i.e., the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}_h$ in (2) is

$$\hat{\boldsymbol{\beta}}_h = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}_h \ (h = 0, \ldots, r-1) \tag{3}$$

where $\mathbf{w}_h$ was defined below (1), and $\mathbf{D} = (d_{ij})$ defined below (1) determines $\mathbf{X}$ in (2) and (3). Given this result, the simulation analysts can easily obtain confidence intervals and statistical tests for the regression parameters per type of output variable; i.e., the analysts may indeed continue to use the classic formulas.

2 Nonnormal simulation output

Least Squares (LS) is a mathematical criterion, so LS does not assume a normal distribution. Only if the simulation analysts require statistical properties—such as BLUE, confidence intervals, and tests—then they usually assume a normal distribution. In this section, I try to answer the following questions (already formulated more generally in Section 0.1): (i) How realistic is the normality assumption? (ii) How can this assumption be tested? (iii) How can the simulation's I/O data be transformed such that the normality assumption holds? (iv) Which statistical methods can be applied that do not assume normality?

2.1 Asymptotic normality

By definition, *deterministic* simulation models do not have a normally distributed output for a given factor combination; this output is a single fixed value. In practice, simulation analysts often assume a normal distribution for the *residuals* of the fitted metamodel. An example is the case study in [19] on coal mining using deterministic System Dynamics simulation; another example is the case study in [30] on global heating caused by the CO_2 greenhouse effect. Indeed, the simulation analysts might argue that so many things affect the residuals that the classic Central Limit Theorem (CLT) applies; i.e., a normal distribution is a good assumption for the residuals of a metamodel fitted to a deterministic simulation's I/O data.

In the remainder of this subsection, I focus on *random* simulation models. Simulation responses within a run are *autocorrelated* (serially correlated). By definition, a *stationary covariance process* has a constant mean (say) $E(w_t) = \mu$ and a constant variance $var(w_t) = \sigma^2$; its covariances depend only on the *lag* $|t - t'|$ between the variables w_t and $w_{t'}$; that is, $cov(w_t, w_{t'}) = \sigma_{|t-t'|}$. The *average* of a stationary covariance process is *asymptotically* normally distributed if the covariances tend to zero sufficiently fast for large lags; see [32], Chapter 2.8. For example, in inventory simulations the output is often the costs averaged over the simulated periods; this average is probably normally distributed. Another output of an inventory simulation may be the service percentage calculated as the fraction of demand delivered from on-hand stock per (say) week, so 'the' output is the average per year computed from these 52 weekly averages. This yearly average may be normally distributed—unless the service goal is 'close' to 100%, so the average service rate is cut off at this threshold and the normal distribution is a bad approximation.

Note that confidence intervals based on Student's t statistic are quite insen-

sitive to nonnormality, whereas the lack-of-fit F-statistic is more sensitive to nonnormality; see [16] for details including references.

In summary, a limit theorem may explain why random simulation outputs are asymptotically normally distributed. Whether the actual simulation run is long enough, is always hard to know. Therefore it seems good practice to check whether the normality assumption holds (see the next subsection).

2.2 Testing the normality assumption

Basic statistics textbooks (also see the recent, 2006, article [2]) and simulation textbooks (see [16] and [31]) propose several *visual plots* and *goodness-of-fit statistics* to test whether a set of observations come from a specific distribution type such as a normal distribution. A basic assumption is that these observations are IID. Simulation analysts may therefore obtain 'many' (say, $m = 100$) replicates for a specific factor combination (e.g., the base scenario) if such an approach is computationally feasible. However, if a single simulation run takes relatively much computer time, then only 'a few' (say, $2 \leq m \leq 10$) replicates are feasible, so the plots are too rough and the goodness-of-fit tests lack power.

Actually, the white noise assumption concerns the metamodel's *residuals* e—not the simulation model's outputs w. The estimated residuals are $\widehat{e}_i = \widehat{y}_i - w_i$ with $i = 1, \ldots n$ and $\widehat{y}_i = \mathbf{x}_i \widehat{\boldsymbol{\beta}}$; an alternative definition is $\widehat{\overline{e}}_i = \widehat{y}_i - \overline{w}_i$ where $\overline{w}_i = \sum_{r=1}^{m_i} w_{i;r}/m_i$ is the simulation output averaged over the m_i replicates. I assume that the simulation analysts obtain at least a few replicates, $m_i > 1$. For simplicity of presentation, I further assume that the number of replicates is constant: $m_i = m$ (> 1). If the simulation outputs w have a constant variance (σ_w^2), then $\sigma_{\overline{w}}^2$ $(= \sigma_w^2/m)$ is also constant. Unfortunately, the *estimated* residuals do not have constant variances and are not independent; it can be proven that

$$\mathbf{cov}(\widehat{\overline{\mathbf{e}}}) = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma_{\overline{w}}^2 \tag{4}$$

where $\mathbf{X}$ is the $n \times q$ matrix of explanatory regression variables defined below (3). Nevertheless, analysts (e.g., [4]) apply visual inspection of residual plots, which are standard output of many statistical packages. Note that (4) uses the well-known *hat matrix* $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Also see [3].

2.3 Transformations of simulation I/O data, jackknifing, and bootstrapping

The simulation output w may be transformed to obtain better normality. Well-known is the *Box-Cox power transformation*:

$$v = \frac{w^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0; \text{ else } v = \ln(w). \tag{5}$$

A complication is that the metamodel now explains not the behavior of the original output, but the behavior of the transformed output! See [3], p. 82 and [11].

In case of nonnormal output, *outliers* occur more frequently when the actual distribution has 'fatter' tails. *Robust regression analysis* might then be applied; see [3] and [37]. However, I have not seen any applications of this approach in simulation.

Normality is not assumed by the following two general computer-intensive statistical procedures that use the original simulation I/O data $(\mathbf{D}, \mathbf{w})$: jackknifing and bootstrapping (actually, the jackknife is a linear approximation of the bootstrap; see [10]). Both procedures have become popular since powerful and cheap computers have become available to the analysts.

2.3.1 Jackknifing

In general, *jackknifing* solves the following two types of problems:

(1) How to compute *confidence intervals* in case of nonnormal observations?
(2) How to reduce possible *bias* of estimators?

Examples of nonnormal observations are the estimated service rate close to one in inventory simulations, and extreme quantiles such as the 99.99% point in risk simulations (see the nuclear waste simulations in [24]). Examples of biased estimators will follow in Section 3.

Suppose the analysts want a confidence interval for the regression coefficients $\boldsymbol{\beta}$ in case the simulation output has a very nonnormal distribution. So the linear regression metamodel is still (2) with $r = 1$. Assume that each factor combination i is replicated an equal number of times, $m_i = m > 1$. The original OLS estimator (also see (3)) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\overline{\mathbf{w}}. \tag{6}$$

Jackknifing deletes the r^{th} replicate among the m IID replicates, and recomputes the estimator for which a confidence interval is wanted:

$$\hat{\boldsymbol{\beta}}_{-r} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\overline{\mathbf{w}}_{-r} \ (r = 1, \ldots, m) \tag{7}$$

where $\overline{\mathbf{w}}_{-r}$ is the n-dimensional vector with components that are the averages of the $m-1$ replicates after deleting replicate r:

$$\overline{w}_{i;-r} = \frac{\sum_{r' \neq r}^{m} w_{i;r'}}{m-1} \tag{8}$$

where for the case $r = m$ the summation runs from 1 to $m-1$ (not m).

Obviously, (7) gives the m correlated estimators $\hat{\boldsymbol{\beta}}_{-1}, \ldots, \hat{\boldsymbol{\beta}}_{-m}$. For ease of presentation, I focus on β_q (the last of the q regression parameters in the vector $\boldsymbol{\beta}$). Jackknifing uses the *pseudovalue* (say) J, which is the following weighted average of $\widehat{\beta}_q$ (the original estimator) and $\widehat{\beta}_{q;-r}$ (the q^{th} component of the jackknifed estimator $\hat{\boldsymbol{\beta}}_{-r}$ defined in (7)) with the number of observations as weights:

$$J_r = m\widehat{\beta}_q - (m-1)\widehat{\beta}_{q;-r}. \tag{9}$$

In this example both the original and the jackknifed estimators are unbiased, so the pseudovalues also remain unbiased estimators. Otherwise it can be proven that the bias is reduced by the *jackknife point estimator*

$$\overline{J} = \frac{\sum_{r=1}^{m} J_r}{m}, \tag{10}$$

which is simply the average of the m pseudovalues defined in (9).

To compute a *confidence interval*, jackknifing treats the pseudovalues as if they were NIID:

$$P(\overline{J} - t_{m-1;1-\alpha/2}\widehat{\sigma}_{\overline{J}} < \beta_q < \overline{J} + t_{m-1;1-\alpha/2}\widehat{\sigma}_{\overline{J}}) = 1 - \alpha \tag{11}$$

where $t_{m-1;1-\alpha/2}$ denotes the $1-\alpha/2$ quantile (upper $\alpha/2$ point) of the distribution of Student's t statistic with $m-1$ degrees of freedom, and

$$\widehat{\sigma}_{\overline{J}} = \sqrt{\frac{\sum_{r=1}^{m}(J_r - \overline{J})^2}{m(m-1)}}.$$

The interval in (11) may be used to test the null-hypothesis that the true regression parameter has a specific value, such as zero.

Applications of jackknifing in simulation are numerous. For example, jackknifing gave confidence intervals for Weighted LS (WLS) with estimated covariance matrix $\widehat{\mathbf{cov}}(\mathbf{w})$; see [25]. Jackknifing reduced the bias and gave confidence

intervals for a Variance Reduction Technique (VRT) called control variates or regression sampling; see [26]. Jackknifing may also be applied in the renewal analysis of steady-state simulation; see [28], pp. 202-203.

2.3.2 Bootstrapping

Textbooks on bootstrapping are [8], [10], [13], and [33]; a recent article is [7] (more references will follow below). Bootstrapping may be used for two types of situations:

(1) The relevant distribution is not Gaussian.
(2) The statistic is not standard.

Sub 1: Reconsider the example used for jackknifing; i.e., the analysts want a. confidence interval for the regression coefficients $\boldsymbol{\beta}$ in case of nonnormal simulation output. Again assume that each of the n factor combinations is replicated an equal number of times, $m_i = m > 1$ $(i = 1, \ldots, n)$. The original LS estimator was given in (6).

The bootstrap distinguishes between the *original observations* w and the *bootstrapped observations* (say) w^* (note the superscript). Standard bootstrapping assumes that the original observations are IID. In the example, there are $m_i = m$ IID original simulated observations per factor combination i, namely $w_{i;1}, \ldots, w_{i;m}$ (these observations give $\overline{w}_i$, which give the vector $\overline{\mathbf{w}}$, which occurs in (6)).

The bootstrap observations are obtained by *resampling with replacement* from the original observations, while the sample size is kept constant, at m. In the example, the bootstrapped observations $w^*_{i;1}, \ldots, w^*_{i;m}$ occur with frequencies $f_1, \ldots, f_m$ such that $f_1 + \ldots + f_m = m$; i.e., these frequencies follow the multinomial (or polynomial) distribution with parameters m and $p_1 = \ldots = p_m = 1/m$. This resampling is executed for each combination i $(i = 1, \ldots n)$. These bootstrapped outputs $w^*_{i;1}, \ldots, w^*_{i;m}$ give the bootstrapped average simulation output $\overline{\mathbf{w}}^*$. Substitution into (6) gives the bootstrapped LS estimator

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\overline{\mathbf{w}}^*. \tag{12}$$

To reduce sampling variation, this resampling is repeated (say) B times; B is known as the *bootstrap sample size* (typical values for B are 100 and 1,000). This gives $\hat{\boldsymbol{\beta}}^*_1, \ldots, \hat{\boldsymbol{\beta}}^*_B$ (or $\hat{\boldsymbol{\beta}}^*_b$ with $b = 1, \ldots, B$).

Let's again focus on the single regression parameter, β_q. The bootstrap literature gives several confidence intervals, but most popular is

$$P(\widehat{\beta}^*_{q;(\lfloor B\alpha/2 \rfloor)} < \beta_q < \widehat{\beta}^*_{q;(\lfloor B(1-\alpha/2) \rfloor)}) = 1 - \alpha$$

where $\widehat{\beta}^*_{q;(\lfloor B\alpha/2 \rfloor)}$ is the $\alpha/2$ quantile of the Empirical Density Function (EDF) of the bootstrap estimate $\widehat{\beta}^*_q$, and $\widehat{\beta}^*_{q;(\lfloor B(1-\alpha/2) \rfloor)}$ is its $1-\alpha/2$ quantile.

Applications of bootstrapping include [21], which bootstraps to validate trace-driven simulation models in case of serious nonnormal outputs.

Sub 2: Besides classic statistics such as the t and F statistics, the simulation analysts may be interested in statistics that have no tables with critical values, which provide confidence intervals—assuming normality. For example, [23] bootstrapped R^2 to test the validity of regression metamodels in simulation.

3 Heterogeneous simulation output variances

By definition, *deterministic* simulation models give a single fixed value for a given factor combination, so the conditional variance is zero: $var(w|\mathbf{x}) = 0$. Simulation analysts often assume a normal distribution for the residuals of the metamodel fitted to the I/O data of the deterministic simulation model (see Section 2.1). Usually, the analysts then assume a normal distribution with a *constant* variance (Kriging models also assume a constant variance). I do not know a better assumption that works in practice for deterministic simulation models.

I further focus on *random* simulation models, and try to answer the following questions:

(1) How realistic is the common variance assumption?
(2) How can this assumption be tested?
(3) How can the simulation's I/O data be transformed such that the common variance assumption holds?
(4) Which statistical analysis methods can be applied that allow nonconstant variances?
(5) Which statistical design methods can be applied to account for variance heterogeneity?

Sub 1: In practice, the variances of random simulation outputs change when factor combinations change. For example, in the M/M/1 queueing simulation not only the mean of the steady-state waiting time changes as the traffic rate changes—the variance of this output changes even more!

Sub 2: Though it may be *a priori* certain that the variances of the simulation outputs are not constant, the analysts may hope that the variances are (nearly) constant in their particular application. Unfortunately, the variances are unknown so they must be estimated. This estimator itself has high vari-

ance; in case of normally distributed output,$var(\widehat{\sigma}^2) = 2\sigma^4/m$. Actually, there are n combinations of the k factors in the simulation experiment, so n variance estimators $\widehat{\sigma}_i^2$ need to be compared. This problem may be solved in many different ways, but I recommend the distribution-free test in [5], p. 241.

Sub 3: The logarithmic transformation in (5) may be used not only to obtain normal output but also to obtain outputs with constant variances. A problem may again be that the metamodel now explains the transformed output instead of the original output.

Sub 4: In case of heterogeneous variances, the LS criterion still gives an *unbiased* estimator (it suffices that the residuals have zero mean, $E(\mathbf{e}) = 0$). The variance of the LS (or OLS) estimator, however, now is

$$\mathbf{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}_N'\mathbf{X}_N)^{-1}\mathbf{X}_N'\mathbf{cov}(\mathbf{w})\mathbf{X}_N(\mathbf{X}_N'\mathbf{X}_N)^{-1} \tag{13}$$

where $\mathbf{X}_N$ is $N \times q$ with $N = \sum_{i=1}^{n} m_i$, and $\mathbf{cov}(\mathbf{w})$ has the same dimensions as this $\mathbf{X}$ has, and the first m_1 elements on its main diagonal are $var(w_1)$, ..., the last m_n elements on this main diagonal are $var(w_n)$. In Section 4, I shall present a simple method to derive confidence intervals for the q individual OLS estimators $\hat{\beta}_j$ (see equation 24). If the number of replicates is constant ($m_i = m$), then the LS estimator may be written as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\overline{\mathbf{w}} \tag{14}$$

where $\mathbf{X}$ is $n \times q$ and $\overline{\mathbf{w}} = (\overline{w}_i)'$ denotes the vector with the n simulation outputs averaged over the m replicates; also see (2).

Though the OLS estimator remains unbiased, it is no longer the BLUE. It can be proven that the BLUE is now the Weighted LS (or WLS) estimator

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}_N'[\mathbf{cov}(\mathbf{w})]^{-1}\mathbf{X}_N)^{-1}\mathbf{X}_N'[\mathbf{cov}(\mathbf{w})]^{-1}\mathbf{w}. \tag{15}$$

where I explicitly denote the number of rows $N = \sum_{i=1}^{n}$ of $\mathbf{X}_N$, which is an $N \times q$ matrix. The reason is that—analogously to (14)—for a constant number of replicates ($m_i = m$) the WLS estimator may be written as

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'[\mathbf{cov}(\overline{\mathbf{w}})]^{-1}\mathbf{X})^{-1}\mathbf{X}'[\mathbf{cov}(\overline{\mathbf{w}})]^{-1}\overline{\mathbf{w}} \tag{16}$$

where $\mathbf{X}$ is $n \times q$ (also see (2)) and $\mathbf{cov}(\overline{\mathbf{w}}) = \mathbf{cov}(\mathbf{w})/m$. The covariance matrix of this WLS estimator is

$$\mathbf{cov}(\widetilde{\boldsymbol{\beta}}) = (\mathbf{X}'[\mathbf{cov}(\overline{\mathbf{w}})]^{-1}\mathbf{X})^{-1}. \tag{17}$$

In practice, $\mathbf{cov}(\mathbf{w})$ is unknown so this covariance matrix must be estimated. The elements on this diagonal matrix are estimated through the classic unbi-

ased variance estimator

$$\widehat{var}(w_i) = \widehat{\sigma}^2(w_i) = s_i^2(w) = \frac{\sum_{r=1}^{m}(w_{ir} - \overline{w}_i)^2}{m-1} (i = 1, \ldots n), \tag{18}$$

which gives $\widehat{\mathbf{cov}}(\mathbf{w})$. Substituting this estimated matrix into the classic WLS formula (15) gives the *Estimated WLS* (EWLS) or Aitken estimator:

$$\widehat{\widehat{\boldsymbol{\beta}}} = (\mathbf{X}'[\widehat{\mathbf{cov}}(\mathbf{w})]^{-1}\mathbf{X})^{-1}\mathbf{X}'[\widehat{\mathbf{cov}}(\mathbf{w})]^{-1}\mathbf{w} \tag{19}$$

where $\mathbf{X}$ is again $N \times q$. *This EWLS is not a linear estimator.* Consequently, the statistical analysis becomes more complicated. For example, the analogue of (17) holds only asymptotically (under certain conditions); see, for example, [12] and [22]:

$$\mathbf{cov}(\widehat{\widehat{\boldsymbol{\beta}}}) \approx (\mathbf{X}'[\widehat{\mathbf{cov}}(\overline{\mathbf{w}})]^{-1}\mathbf{X})^{-1}. \tag{20}$$

Classic confidence intervals do no longer hold.

Relatively simple solutions for this type of problem were already presented in Sections 2.3.1 and 2.3.2, namely jackknifing and bootstrapping. *Jackknifing* the EWLS estimator was done in [25], as follows. Delete the r^{th} replicate among the m IID replicates, and recompute the EWLS estimator (analogous to (7)):

$$\widehat{\widehat{\boldsymbol{\beta}}}_{-r} = (\mathbf{X}'[\widehat{\mathbf{cov}}(\mathbf{w})_{-r}]^{-1}\mathbf{X})^{-1}\mathbf{X}'[\widehat{\mathbf{cov}}(\mathbf{w})_{-r}]^{-1}\mathbf{w}_{-r} \ (r = 1, \ldots, m)$$

where $\overline{\mathbf{w}}_{-r}$ consists of n averages computed from $m-1$ replicates after deleting replicate r, and $\widehat{\mathbf{cov}}(\mathbf{w})_{-r}$ is computed from the same replicates. Use these $\widehat{\widehat{\boldsymbol{\beta}}}_{-r}$ and the original $\widehat{\widehat{\boldsymbol{\beta}}}$ computed through (19) to compute the pseudovalues, which give the desired confidence interval. *Bootstrapping* the EWLS estimator is discussed in [23].

Sub 5: If the output variances are not constant, *classic designs* still give the *unbiased* OLS estimator $\widehat{\boldsymbol{\beta}}$ and WLS estimator $\widetilde{\boldsymbol{\beta}}$. The literature pays little attention to the derivation of alternative designs for heterogeneous output variances. In [29], we investigated designs in which the n factor combinations are replicated so many times that the estimated variances of the averages per combination are (approximately) constant. Because $var(\overline{w}_i) = \sigma_i^2/m_i$ $(i = 1, \ldots, n)$, the number of replicates should satisfy

$$m_i = c_0\sigma_i^2 \tag{21}$$

where c_0 is a common positive constant such that the m_i become integers. This equation implies that the higher the variability of the simulation output w_i is, the more replicates are simulated. The allocation of the total number of simulation runs ($N = \sum_{i=1}^{n} m_i$) according to (21) is not necessarily optimal, but it simplifies the regression analysis and the design of the simulation

experiment (an alternative replaces σ_i^2 by σ_i). Indeed the regression analysis can now apply OLS to the averages $\overline{w}_i$ to get BLUE.

In practice, however, the variances of the simulation outputs must be estimated. A *two-stage* procedure takes a *pilot sample* of (say) $m_0 \geq 2$ replicates for each factor combination, and estimates the variances σ_i^2 through

$$s_i^2(m_0) == \frac{\sum_{r=1}^{m_0} [w_{ir} - \overline{w}_i(m_0)]^2}{m_0 - 1} (i = 1, \ldots n) \tag{22}$$

with $\overline{w}_i(m_0) = \sum_{r=1}^{m_0} w_{ir}/m_0$. Combining (22) and (21), [29] selects additional replicates $\widehat{m}_i - m_0$ where

$$\widehat{m}_i = m_0 \left\lfloor \frac{s_i^2(m_0)}{\min_{1 \leq i \leq n} s_i^2(m_0)} \right\rfloor$$

with $\lfloor x \rfloor$ denoting the integer closest to x (so, in the second stage no additional replicates are simulated for the combination with the smallest estimated variance). After the second stage all $\widehat{m}_i$ replicates are used to estimate the average output and its variance. OLS is applied to these averages. The covariance matrix $\mathbf{cov}(\widehat{\boldsymbol{\beta}})$ is estimated through (13) with $\mathbf{cov}(\mathbf{w})$ estimated through a diagonal matrix with diagonal elements $s_i^2(\widehat{m}_i)/\widehat{m}_i$. Confidence intervals are based on the classic t statistic with degrees of freedom equal to $m_0 - 1$.

Because these $s_i^2(\widehat{m}_i)/\widehat{m}_i$ may still differ considerably, this two-stage approach may be replaced by a *sequential* approach. The latter approach adds one replicate at a time, until the estimated variances of the average simulation outputs have become constant; see [29]. The sequential procedure requires fewer simulation responses, but is harder to understand, program, and implement.

4 Cross-correlated simulation outputs: Common random numbers

Obviously, CRN implies *random* simulation. In this section, I try to answer the following questions:

(1) How realistic is the assumption of independent simulation outputs?
(2) Which statistical analysis methods can be applied that allow correlated outputs?
(3) Which statistical design methods can be applied to account for correlated outputs?

Sub 1: In practice, simulation analysts often use CRN; actually, CRN is the default of much simulation software. CRN implies that the simulation outputs of different factor combinations are correlated across these combinations. The

goal of CRN is to reduce $var(\widehat{\beta}_j)$ with $j = 1, \ldots, q$ (actually, the variance of the intercept increases when CRN are used). So CRN is useful to better explain the factor effects, and to better predict the output of combinations not yet simulated (provided the inaccuracy of the estimated intercept is outweighed by the accuracy of all other estimated effects). Because CRN violates the classic assumptions of regression analysis, the analysts have two options:

(i) Continue to use OLS

(ii) Switch to GLS.

Sub (i): The variance of the OLS estimator is given by (13), but now $\mathbf{cov}(\mathbf{w})$ is not a diagonal matrix. It is simple to derive confidence intervals and test null-hypotheses—provided there are $m \geq 2$ replicates (also see [31], p. 630, 642). From replicate r, compute

$$\hat{\boldsymbol{\beta}}_r = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}_r \ (r = 1, \ldots, m). \tag{23}$$

The n components of the vector $\mathbf{w}_r$ are correlated (because they use CRN) and may have different variances, but the m estimators $\hat{\beta}_{j;r}$ of a specific regression parameter β_j are independent (because they use non-overlapping PRN streams) and have a common standard deviation (say) $\sigma(\hat{\beta}_j)$. So

$$t_{m-1} = \frac{\overline{\widehat{\beta}}_j - \beta_j}{s(\overline{\widehat{\beta}}_j)} \text{ with } j = 1, \ldots, q \tag{24}$$

with

$$s(\overline{\widehat{\beta}}_j) = \sqrt{\frac{\sum_{r=1}^{m}(\widehat{\beta}_{j;r} - \overline{\widehat{\beta}}_j)^2}{m(m-1)}}.$$

Sub (ii): CRN implies that the BLUE is the GLS estimator; see (15) where $\mathbf{cov}(\mathbf{w})$ is now not diagonal. Obviously, $\mathbf{cov}(\widetilde{\boldsymbol{\beta}})$ is analogous to (17). Again, in practice $\mathbf{cov}(\mathbf{w})$ is estimated by $\widehat{\mathbf{cov}}(\mathbf{w})$ which has the elements

$$\widehat{cov}(w_i, w_{i'}) = \frac{\sum_{r=1}^{m}(w_{i;r} - \overline{w}_i)(w_{i';r} - \overline{w}_{i'})}{m-1} \ (i, i' = 1, \ldots, n) \tag{25}$$

where $m = \min(m_i, m_{i'})$; usually, $m_i = m_{i'} (= m)$ if CRN is applied. This $\widehat{\mathbf{cov}}(\mathbf{w})$ is *singular* if the number of replicates is 'too small'; that is, if $m \leq n$; see [9]. Substituting $\widehat{\mathbf{cov}}(\mathbf{w})$ into the classic GLS formula gives the *Estimated GLS* (EGLS), analogous to the EWLS estimator in (19). The EGLS estimator can again be analyzed through jackknifing and bootstrapping. In [17], I compared OLS and EGLS relying on the asymptotic covariance matrix (20) with nondiagonal $\widehat{\mathbf{cov}}(\mathbf{w})$; [7], however, claims that 'bootstrap tests ... yield more reliable inferences than asymptotic tests in a great many cases'. In conclusion, CRN with EGLS may give better point estimates of the factor effects

(except for the intercept), but a proper statistical analysis may require 'many' replicates, namely $m > n$.

Sub 3: The literature pays no attention to the derivation of alternative designs for CRN. *Sequential* procedures are proposed in [27] and [40], to select the next factor combination to be simulated, where the simulation model may be either deterministic or random—assuming the simulation I/O data $(\mathbf{D}, \mathbf{w})$ are analyzed through Kriging (instead of linear regression), which allows the simulation outputs to be correlated.

5 Nonvalid low-order polynomial metamodel

Now, I try to answer the following questions:

(1) How can the validity of the low-order polynomial metamodel be tested?
(2) If this metamodel is not valid, how can the simulation's I/O data be transformed such that a low-order polynomial becomes valid ?
(3) Which alternative metamodels can be applied?

Sub 1: A valid metamodel has zero mean residuals, so $H_0 : E(e) = 0$. To test this null-hypothesis, the analysts may apply the classic *lack-of-fit F-statistic* assuming white noise. However, if the analysts apply CRN, then they may apply Rao's variant derived in [34] (and evaluated in [17]):

$$F_{n-q;m-n+q} = \frac{m-n+q}{(n-q)(m-1)}(\overline{\mathbf{w}} - \widehat{\overline{\mathbf{y}}})'[\widehat{\mathbf{cov}}(\overline{\mathbf{w}})]^{-1}(\overline{\mathbf{w}} - \widehat{\overline{\mathbf{y}}}) \qquad (26)$$

where $n > q$, $m > n$, and $\widehat{\overline{\mathbf{y}}}$ denotes the EGLS estimator. Obviously, this test also allows EWLS instead of EGLS. Normality of the simulation output is an important assumption for both the classic test and Rao's test. In case of nonnormality, the analysts may apply jackknifing or bootstrapping; [23] bootstraps Rao's statistic and the classic R^2 statistic.

An alternative test uses *cross-validation* and the t statistic, which is less sensitive to nonnormality than the F statistic; see [17]. Moreover, this t statistic requires fewer replications, namely $m > 1$ instead of $m > n$ if EWLS or EGLS is used.

Besides these quantitative tests, the analysts may use *graphical* methods to judge the validity of a fitted metamodel (be it a linear regression model or some other type of metamodel such as a Kriging model). Scatterplots are well known. The recent panel publication [39] also emphasizes the importance of visualization; also see [14]. If these validation tests reject H_0, then the analysts may consider the following alternatives.

Sub 2: A well-known transformation in queueing simulations combines two simulation inputs—namely, the arrival rate λ and the service rate μ— into a single independent regression variable—namely, the traffic rate $x = \lambda/\mu$. Another transformation replaces y, λ, and μ by $\log(y)$, $\log(\lambda)$, and $\log(\mu)$, to make the first-order polynomial approximate relative changes.

Another simple transformation assumes that the I/O function of the underlying simulation model is *monotonic*. Then the dependent and independent variables may be replaced by their ranks, which results in so-called *rank regression*; see [6] and [38]. Note that Spearman's correlation coefficient uses the same transformation for two correlated random variables. For example, [24] applies rank regression and Spearman's coefficient to find the most important factors in a simulation model of nuclear waste disposal.

Transformations may also be applied to make the simulation output (dependent regression variable) better satisfy the assumptions of normality (see (5)) and variance homogeneity. Unfortunately, different goals of the transformation may conflict with each other; for example, the analysts may apply the logarithmic transformation to reduce nonnormality, but this transformation may give a metamodel in variables that are not of immediate interest.

If classic designs do not give valid metamodels, then I recommend to look for transformations, as discussed above. I do not recommend routinely adding higher-order terms to the metamodel, because these terms are hard to interpret. However, if the goal is not to better *understand* the underlying simulation model but to better *predict* the output of an expensive simulation model, then high-order terms may be added. Indeed, full factorial 2^k designs enable the estimation of all interactions (for example, the interaction among all k factors). If more than two levels are simulated per factor, then the following types of metamodels may be considered.

Sub 3: There are several alternative metamodel types; for example, Kriging models. These alternatives may give better predictions than low-order polynomials do. However, these alternatives are so complicated that they do not help the analysts better understand the underlying simulation model—except for sorting the simulation inputs in order of their importance. Furthermore, these alternative metamodels require *alternative design types*. This is a completely different issue, so I refer to the extensive literature on this topic (including [20]).

6 Conclusions

In this survey, I discussed the assumptions of classic linear regression analysis and the concomitant statistical designs. In Section 1, I pointed out that multivariate simulation output can still be analyzed through OLS. In Section 2, I addressed possible nonnormality of simulation output, including normality tests, transformations of simulation I/O data, jackknifing, and bootstrapping. In Section 3, I presented analysis and design methods for heteroscedastic simulation output. In Section 4, I discussed how to analyze cross-correlated simulation outputs created by CRN. In Section 5, I discussed possible lack-of-fit of low-order polynomial metamodels, and possible remedies. I gave many references for further study of these issues.

References

[1] Angün, E., D. den Hertog, G. Gürkan, and J.P.C. Kleijnen (2006), Response surface methodology with stochastic constrains for expensive simulation. Working Paper, Tilburg University, Tilburg, Netherlands

[2] Arcones, M.A. and Y. Wang (2006), Some new tests for normality based on U-processes *Statistics & Probability Letters*, 76, no. 1, pp. 69-82

[3] Atkinson, A. and M. Riani (2000), *Robust diagnostic regression analysis.* Springer, New York

[4] Ayanso, A., M. Diaby, and S.K. Nair (2006), Inventory rationing via drop-shipping in Internet retailing: a sensitivity analysis. *European Journal of Operational Research*, 171, no. 1, pp. 135-152

[5] Conover, W.J. (1980), *Practical nonparametric statistics: second edition.* Wiley, New York

[6] Conover, W.J. and R.L. Iman (1981), Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35, no. 3, pp124-133

[7] Davidson, R. and J.G. MacKinnon (2006), Improving the reliability of bootstrap tests with the fast double bootstrap *Computational Statistics & Data Analysis*, in press

[8] Davison, A.C. and D.V. Hinkley (1997), *Bootstrap methods and their application.* Cambridge University Press, Cambridge

[9] Dykstra, R.L. (1970), Establishing the positive definiteness of the sample covariance matrix. *The Annals of Mathematical Statistics*, 41, no. 6, pp. 2153-2154

[10] Efron, B. and R.J. Tibshirani (1993), *An introduction to the bootstrap*. Chapman & Hall, New York

[11] Freeman, J. and R. Modarres (2006), Inverse Box Cox: the power-normal distribution *Statistics & Probability Letters*, 76, no. 8, pp. 764-772

[12] Godfrey, L.G. (2006), Tests for regression models with heteroskedasticity of unknown form *Computational Statistics & Data Analysis,* 50, no. 10, pp. 2715-2733

[13] Good, P.I. (2005), *Resampling methods: a practical guide to data analysis; third edition*. Birkhäuser, Boston

[14] Helton, J.C , J.D. Johnson, C.J. Sallaberry, and C.B. Storlie (2006), Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering and Systems Safety*, in press

[15] Ivanescu, C., W. Bertrand, J. Fransoo, and J.P.C. Kleijnen (2006), Bootstrapping to solve the limited data problem in production control: an application in batch processing industries. *Journal of the Operational Research Society*, 57, number 1, pp. 2-9

[16] Kleijnen, J.P.C. (1987), *Statistical tools for simulation practitioners*. Marcel Dekker, New York

[17] Kleijnen, J.P.C. (1992), Regression metamodels for simulation with common random numbers: comparison of validation tests and confidence intervals. *Management Science*, 38, no. 8, pp. 1164-1185

[18] Kleijnen, J.P.C. (1993), Simulation and optimization in production planning: a case study. *Decision Support Systems*, 9, pp. 269-280

[19] Kleijnen, J.P.C. (1995), Sensitivity analysis and optimization of system dynamics models: regression analysis and statistical design of experiments. *System Dynamics Review*, 11, no. 4, pp. 275-288

[20] Kleijnen, J.P.C. (2007), *DASE: Design and analysis of simulation experiments.* Springer Science + Business Media

[21] Kleijnen, J.P.C., R.C.H. Cheng and B. Bettonvil (2001),Validation of trace-driven simulation models: bootstrapped tests. *Management Science*, 47, no. 11, pp. 1533-1538

[22] Kleijnen, J.P.C., P. Cremers and F. van Belle (1985), The power of weighted and ordinary least squares with estimated unequal variances in experimental designs. *Communications in Statistics, Simulation and Computation*, 14, no. 1, pp. 85-102

[23] Kleijnen, J.P.C. and D. Deflandre (2006), Validation of regression metamodels in simulation: Bootstrap approach.*European Journal of Operational Research*, 170, no. 1, pp. 120-131

[24] Kleijnen, J.P.C. and J. Helton (1999), Statistical analyses of scatter plots to identify important factors in large-scale simulations,1: review and comparison of techniques. *Reliability Engineering and Systems Safety*, 65, no. 2, pp. 147-185

[25] Kleijnen, J.P.C., P.C.A. Karremans, W.K. Oortwijn, and W.J.H. van Groenendaal (1987), Jackknifing estimated weighted least squares: JEWLS. *Communications in Statistics, Theory and Methods*, 16, no. 3, pp. 747-764

[26] Kleijnen, J.P.C., J. Kriens, H. Timmermans, and H. Van den Wildenberg (1989), Regression sampling in statistical auditing: a practical survey and evaluation (including Rejoinder). *Statistica Neerlandica*, 43, no. 4, pp. 193-207 (p. 225)

[27] Kleijnen, J.P.C. and W.C.M. van Beers (2004), Application-driven sequential designs for simulation experiments: Kriging metamodeling. *Journal of the Operational Research Society,* 55, no. 9, pp. 876-883

[28] Kleijnen, J.P.C. and W. van Groenendaal (1992), *Simulation: a statistical perspective*. John Wiley, Chichester (England)

[29] Kleijnen, J.P.C. and W. van Groenendaal (1995), Two-stage versus sequential sample-size determination in regression analysis of simulation experiments. *American Journal of Mathematical and Management Sciences*, 15, nos. 1&2, pp. 83-114

[30] Kleijnen, J.P.C., G. van Ham, and J. Rotmans (1992), Techniques for sensitivity analysis of simulation models: a case study of the CO2 greenhouse effect. *Simulation*, 58, no. 6, pp. 410-417

[31] Law, A.M. and W.D. Kelton (2000), *Simulation modeling and analysis; third edition*. McGraw-Hill, Boston

[32] Lehmann, E. L. (1999), *Elements of large-sample theory*, Springer, New York

[33] Lunneborg, C.E. (2000), *Data analysis by resampling: concepts and applications*. Duxbury Press, Pacific Grove, California

[34] Rao, C.R. (1959), Some problems involving linear hypothesis in multivariate analysis. *Biometrika*, 46, pp. 49-58

[35] Rao, C. R. (1967), Least squares theory using an estimated dispersion matrix and its application to measurement of signals. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* I, pp. 355-372

[36] Ruud, P. A. (2000), *An introduction to classical econometric theory.* Oxford University Press, New York

[37] Salibian-Barrera, M. (2006), Bootstrapping MM-estimators for linear regression with fixed designs. *Statistics & Probability Letters*, in press

[38] Saltelli, A. and I.M. Sobol (1995), About the use of rank transformation in sensitivity analysis of model output. *Reliability Engineering and System Safety*, 50, pp. 225-239

[39] Simpson, T. W., Booker, A. J., Ghosh, D., Giunta, A. A., Koch, P. N., and Yang, R.-J. (2004), Approximation methods in multidisciplinary analysis and optimization: a Panel discussion. *Structural and Multidisciplinary Optimization*, 27, no. 5, pp. 302-313

[40] Van Beers, W.C.M. and J.P.C. Kleijnen (2006), Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. Working Paper, Tilburg University, Tilburg, Netherlands