KM 22(1986)
pag 93 - 112

REGRESSION ANALYSIS OF SIMULATION EXPERIMENTS: FUNCTIONAL SOFTWARE SPECIFICATION

Jack P.C. Kleijnen*)

ABSTRACT

Regression models are often used to analyze simulation models. However, simulation data have special problems and opportunities which are not easily handled by current regression software. This paper investigates the following five assumptions: (1) Non-collinearity of the matrix of independent variables $\underset{\sim}{X}$. In simulation, experimental design is used, possibly in a sequentially way, so that collinearity is no acute problem. (2) Constant variances ($\sigma_i^2 = \sigma^2$): Simulation yields variance estimates, which often differ substantially. These estimates can be used (a) to Correct the covariance matrix of the ordinary Least Squares point estimator $\hat{\beta}$ (CLS), or (b) to compute Estimated Weighted Least Squares (EWLS), or (c) to derive a Sequential Least Squares design (SLS). (3) Independence: Common random-number seeds destroy the independence. Again CLS, EWLS or SLS applies. (4) Normality: Outliers can be eliminated after running the simulation program with new seeds. Rank regression is another useful option. (5) No specification error: The validity of the regression model can be tested through cross-validation, accounting for non-constant variances which results in a maximum absolute Studentized forecast error. For deterministic simulation the relative error $\hat{y}/y$ is proposed as criterion.

ADDITIONAL KEYWORDS: transformations, robustness, nonparametric, optimization, applications, case studies.

*) Professor of Simulation and Information Systems, Department of Information Systems and Auditing (ISA), School of Business and Economics (FEW), Catholic University Brabant (KUB), P.O. Box 90153, 5000 LE Tilburg, The Netherlands, Phone 013-662029.

1. INTRODUCTION

There are many statistical packages (such as MINITAB, SAS, TSP) with modules for linear regression analysis and Analysis of Variance. These packages have provisions for econometric models (including autocorrelation, simultaneous equations, errors in variables, and so on) and experimental designs (including blocking, split plots, etc.). However, there is no software tailored to the needs of simulation analysts! And simulation experiments do have special problems and opportunities, as we shall see, and simulation is applied in many disciplines. Our experience shows that when simulation analysts apply standard regression software to simulation data, they are easily led to erroneous interpretations, for example, they try to interpret a non-significant Durbin-Watson statistic, without realizing that in their simulation experiment such a statistic makes no sense.

Before proceeding we shall demonstrate why current regression software is inadequate for use by simulation practitioners. In Section 4 we shall see that in simulation there are unbiased estimators $\hat{\sigma}_h^2$ of the response variances σ_h^2, which may differ substantially ($\sigma_h^2 \neq \sigma^2$). We shall propose a combination of the Ordinary Least Squares estimator $\hat{\underset{\sim}{\beta}}$ (see eq. 4) with the correct estimated covariance matrix $\hat{\underset{\sim}{\Omega}}_{\hat{\beta}}$ (eq. 9 differs from the ordinary formula $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\hat{\sigma}^2$). This combination is an option not available in current software (we checked SAS, TSP and MINITAB manuals). Actually this option is not even mentioned in textbooks; the option is discussed in our own publications - for example, Kleijnen et al. (1979), Kleijnen (1983) - and in Nozari (1984). Of course, new formulas (such as eq. 9) can be added to any "open ended" package (that is, software with either an interface to a general purpose language such as Fortran, or its own language; for example, SAS has an APL-like language). Practitioners, however, want to use existing options only! We note that some issues we shall discuss, are easily implemented in certain software but not in other software; for example, sorting - needed in eq. (13) - is easy in SAS, not in TSP. In this paper we shall discuss the characteristics of simulation experiments. The "Summaries" at the end of sections 3 through 7 provide the functional specification of a package

for Regression Analysis of Simulation Experiments (RASE). We plan to implement RASE in SAS, unless a software house announces that it will offer a package with the facilities of RASE.

In the present paper we do not discuss in detail, why regression analysis of simulation data is advantageous. Suffices it to say that regression analysis is used for validation of simulation models, optimization, what - if questions, and so on; see Kleijnen (1979, 1986). We use the following terminology; also see eq. (1). The simulation model f_1 has k parameters (or factors) $z_1,\ldots,z_k$. A random (or stochastic) simulation model has an additional input, namely the random-number seed (or initial value) r_0. We shall concentrate on random simulation, and only when necessary we shall discuss deterministic simulation separately. A simulation model yields several time series, which are summarized by a few measures such as the average. We concentrate on a single measure y per simulation run. We may solve the multivariate problem, applying univariate regression analysis (per response y) in combination with the Bonferroni inequality; see Miller (1966, pp. 189-210). In summary, we represent the simulation model through the following function:

$$y = f_1 (z_1, z_2, \ldots, z_k, r_0) \tag{1}$$

This simulation model is approximated by a regression model; see the next section.

2. BASIC REGRESSION ANALYSIS

In this section we present the basic formulas of regression analysis, in order to define our symbols and terminology. The regression model is linear in its parameters $\underset{\sim}{\beta}$ and has additive errors $\underset{\sim}{e}$:

$$\underset{\sim}{y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{e} \tag{2}$$

where $\underset{\sim}{y} = (y_1,\ldots,y_N)'$ since there are N simulation runs ($N > 1$); $\underset{\sim}{X}$ is an N x q matrix of independent (regression) variables x_{ij} ($i=1,\ldots,N$; $j=1,\ldots,q$; $1 \leq q < N$; see eq. 6 and Section 7); $\underset{\sim}{\beta} = (\beta_1,\ldots,\beta_q)'$ and $\underset{\sim}{e} = (e_1,\ldots,e_N)'$. (As we shall see in Section 4, these N runs may con-

sist of n factor combinations h replicated m_h times such that $\sum_{h=1}^{n} m_h = N$.) Obviously, this linear model is not necessarily linear in the simulation factors $z_1,\ldots,z_k$; for example, the regression model may be a second-degree polynomial in z (so that $q = (k+1)(k+2)/2$ or the variable x may equal log z. In our experience, linear regression is flexible enough to summarize simulation models (non-linear regression is applied to, for example, chemical experiments where enough theoretical knowledge is available to suggest a specific family of nonlinear models).

The Classical Assumptions for the regression model are: the errors $\underset{\sim}{e}$ are Normally and Independently Distributed (NID) with zero means and constant variances σ^2, or

$$\underset{\sim}{e} \sim N(\underset{\sim}{0}, \sigma^2\underset{\sim}{I}). \tag{3}$$

Under these assumptions the Ordinary Least Squares (OLS) estimator

$$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}' \underset{\sim}{X})^{-1}\underset{\sim}{X}' \underset{\sim}{y} \tag{4}$$

has several attractive properties: $\hat{\underset{\sim}{\beta}}$ is the Maximum Likelihood (ML) estimator, and it is the Best Linear Unbiased Estimator (BLUE; this property does not require normally distributed errors); confidence intervals for $\underset{\sim}{\beta}$ can be based on the F statistic; and so on. The covariance matrix of $\hat{\underset{\sim}{\beta}}$ is

$$\underset{\sim}{\Omega}_{\hat{\beta}} = (\underset{\sim}{X}' \underset{\sim}{X})^{-1}\sigma^2. \tag{5}$$

The unknown parameter σ^2 in eq. (5) is estimated through the Mean Squared Residuals (MSR)

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N - q} \tag{6}$$

where $\hat{y}_i$ is the ith element of $\hat{\underset{\sim}{y}} = \underset{\sim}{X}\hat{\underset{\sim}{\beta}}$. The individual regression parameters β_j are tested through the t statistic:

$$t_j = \frac{\hat{\beta}_j - \beta_j}{s_j} \qquad (j = 1,\ldots,q) \tag{7}$$

where t_j is a Student t statistic with N-q degrees of freedom and $s_j = (\omega_{jj})^{\frac{1}{2}}$ where ω_{jj} denotes the j th diagonal element of $\hat{\Omega}_{\hat{\beta}}$; see eqs. (5) and (6). We shall discuss each assumption of OLS in a separate section, concentrating on the characteristics of simulation data.

Summary: The OLS estimator $\hat{\beta}$ has attractive properties, provided the following assumptions hold: (1) X is not collinear; hence $(X' X)^{-1}$ exists. (2) The error variances are constant: $\sigma_i^2 = \sigma^2$. (3) The errors e are independent. (4) The errors are normally distributed. (5) The errors have zero expected values (the regression model is correctly specified so that it is a valid model).

3. NON-COLLINEAR MATRIX X

In simulation there is usually no problem of collinearity, as we shall see in the next paragraph. In the social sciences, however, the analyst cannot fix the independent variables, but can only observe them. Consequently X may be exactly or "nearly collinear", i.e., minor changes in X yield major changes in $\hat{\beta}$. Actually the independent variables are random and may be highly correlated. Replication (of specific environmental conditions) is virtually impossible. Therefore the error variances are assumed constant ($\sigma_i^2 = \sigma^2$) and estimated from the residuals ($\hat{y}_i - y_i$); see eq. (6). And if X is nearly collinear, ridge regression may be useful; see Hoerl and Kennard's (1981) annotated bibliography with over 200 references.

In other sciences (such as agriculture and chemistry) the analyst may proceed from passive observation to active experimentation. In simulation (in both soft and hard sciences) all factors are controllable; and the theory of experimental design should be applied. Consequently X is not collinear in general; often X is orthogonal (that is, $X'X = N I$, for example, 2^{k-p} designs have that property; see Kleijnen, 1974/1975, 1986). Sometimes, X may turn out to be collinear, namely if

we design and run the simulation experiment and later on we add new independent variables (like two-factor interactions in a 2^{k-p} design). If such a collinearity problem arises, then - assuming we have computer time left to make extra runs with the simulation model - we should not use ridge regression analysis but we should add some new runs to the old design. Modern regression packages (such as SAS) make the addition of data simple. (The sofware might even suggest which factor combinations to add to the old combinations; see Kleijnen, 1986.)

Summary: In simulation, as opposed to the social sciences, collinearity of $\underset{\sim}{X}$ is no problem in general, since the theory of experimental design can be applied so that $\underset{\sim}{X}$ may even be orthogonal. Addition of extra variables (like interactions) may create collinearity, which can be eliminated adding a few extra runs. So the regression software should not resort to special analysis techniques such as ridge regression; instead the software should enable the addition of extra runs.

4. CONSTANT VARIANCES

Our experience is that the error variances σ_i^2 differ substantially in random simulation, for example, Kleijnen et al. (1979, p. 60) report a simulation experiment in which the estimated variances range between 64 and 93,228 (the simulation represents part of the Rotterdam harbor). Apart from these empirical results, it seems strange to assume that the expected responses do depend on the factors z (or $\underset{\sim}{X}$) but the response variances do not.

In random simulation the estimation of σ_i^2 is easy, when compared to experiments with real-life technical systems (chemical plants, agricultural plots) and socio-technical systems (organizations like business companies); see also Dykstra (1959, p. 63). In random simulation we can execute the same simulation program m_1 times, using m_1 different random number seeds r_0 ($m_1 \geq 2$). Next we change the simulation program (f_1 or its inputs z; see eq. 1) and run that program m_2 times ($m_2 \geq 2$), and so on. So there are n ($\geq q$) combinations of simulation parameters z, each replicated m_h times ($h = 1,\ldots,n$) where $m_h \geq 2$. Hence $\underset{\sim}{X}$ has N rows where

$N = \sum_{1}^{n} m_h$, but only n rows are different. We may rearrange the vector $\underset{\sim}{y}$ into a two-way table (used in Analysis of Variance): $y_1 = y_{11}$, $y_2 = y_{12}, \ldots, y_N = y_{nm_n}$. Then we compute

$$\hat{\sigma}_h^2 = \frac{\sum_{r=1}^{m_h} (y_{hr} - \bar{y}_h)^2}{m_h - 1} \qquad (h = 1, \ldots, n) \tag{8}$$

where $\bar{y}_h = \sum_r y_{hr}/m_h$. These $\hat{\sigma}_h^2$ are unbiased estimators of $\sigma_h^2 = \text{var}(y_{hr_h})$; obviously $\text{var}(y_{11}) = \text{var}(y_{12}) = \ldots = \text{var}(y_{1n_1}) = \sigma_1^2$, and so on.

Current regression software does not contain eq. (8). Instead OLS assumes constant variances, $\sigma_h^2 = \sigma^2$, and estimates σ^2 from the estimated residuals $\hat{\underset{\sim}{e}}$. Another option assumes known variances and computes Weighted Least Squares (WLS); also see eqs. (10) and (11). Software for simulation should allow the user <u>either</u> to supply the responses y_{hr} whereupon the software applies eq. (8) <u>or</u> to supply the variance estimates $\hat{\sigma}_h^2$ or the standard errors $\hat{\sigma}_h$ which may be the output of the simulation program.

Readers familiar with simulation know that in the simulation of steady-state models (for example, certain queuing models) other estimators of σ_h^2 are possible, using subruns, spectral analysis, and so on; see Kleijnen (1974/1975, 1986). These estimators are more complicated, and may be biased. Anyhow, in random simulation we should obtain not only the responses y but also their standard errors $\hat{\sigma}$. In deterministic simulation the assumption of constant variances might be realistic, i.e., the deviations $\underset{\sim}{e}$ between the simulation responses $\underset{\sim}{y}$ and the linear model $\underset{\sim}{X}\,\underset{\sim}{\beta}$ have a common distribution with variance σ^2 (and zero mean if the linear model is valid; see Section 7), even though the conditional variances var(y/x) are zero; see Kleijnen (1986, pp. 163-164). We do not discuss variance stabilizing transformations, since the interpretation of the experimental data should be in terms of the original (non-transformed) responses; see Scheffé (1964, pp. 364-368) and the references in Hoyle (1973) and Kleijnen (1986).

We distinguish the following options in case of variance heterogeneity (comments follow in the next paragraph).

(a) Ordinary Least Squared (OLS): We simply ignore variance differences; see eqs. (4) through (7).

(b) Corrected Least Squared (CLS): The OLS point estimator of eq. (4) is combined with the corrected estimated covariance matrix

$$\hat{\Omega}_{\hat{\beta}} = (\underset{\sim}{X}' \underset{\sim}{X})^{-1} \underset{\sim}{X}' \hat{\underset{\sim}{\Omega}}_y \underset{\sim}{X} (\underset{\sim}{X}' \underset{\sim}{X})^{-1} \tag{9}$$

where $\hat{\underset{\sim}{\Omega}}_y$ is an N x N diagonal matrix with the first m_1 diagonal elements equal to $\hat{\sigma}_1^2$, the next m_2 elements equal to $\hat{\sigma}_2^2$, and so on; see eq. (8).

(c) Estimated Weighted Least Squares (EWLS): The estimated response variances of eq. (8) yield the unbiased nonlinear point estimator

$$\tilde{\underset{\sim}{\beta}} = (\underset{\sim}{X}' \hat{\underset{\sim}{\Omega}}_y^{-1} \underset{\sim}{X})^{-1} \underset{\sim}{X}' \hat{\underset{\sim}{\Omega}}_y^{-1} \underset{\sim}{y} \tag{10}$$

with asymptotic covariance matrix

$$\underset{\sim\sim}{\Omega}_{\tilde{\beta}} = (\underset{\sim}{X}' \underset{\sim}{\Omega}_y^{-1} X)^{-1}, \tag{11}$$

provided certain mild technical assumptions hold; see Schmidt (1976, p. 71).

(d) Sequential Least Squares (SLS): First we take a pilot sample of m_h^0 observations, which yield a first estimate of σ_h^2; see eq. (8) with m_h replaced by m_h^0. Next we take more observations for the experimental conditions with high variability. If we take

$$m_h = c\, \hat{\sigma}_h^2 \quad (h = 1,\ldots,n) \; (c > 0) \tag{12}$$

then we can fit the regression model to the average responses $\bar{y}_h$ with (approximately) constant variance 1/c.

OLS (option a) yields conservative tests; EWLS gives valid tests provided the number of replications is large, say $m_h \geq 25$; CLS gives valid tests; see Kleijnen et al. (1985). SLS may give conservative tests; see Kleijnen and Van Groenendaal (1986). We recommend that a user

apply several analysis techniques (EWLS and CLS, not OLS) to the same data. For example, Kleijnen et al. (1979) apply both CLS and EWLS; these techniques give different quantitative results (point estimates of $\underset{\sim}{\beta}$) but identical qualitative results (which factors are really important?). SLS is a design, not only an analysis technique. The software should at least provide the EWLS and the CLS options.

Summary: In simulation the response variances σ_h^2 (h = 1,...,n) may differ substantially. Simulation experiments provide estimates of σ_h^2 (besides the responses y_i). These estimated variances yield several point estimates and standard errors for the regression parameters $\underset{\sim}{\beta}$: Corrected Least Squares (CLS) and Estimated Weighted Least Squares (EWLS). A possible design uses Sequential Least Squares (SLS).

5. INDEPENDENCE

In the non-experimental sciences most data form time series; consequently autocorrelation is a major problem. Simulation yields many time series, each characterized by a single or a few statistics (see Section 1; in other words, in simulation there is an information overload problem, not a dirth of data). Simulation yields perfectly independent responses y_i if the random number seeds are independent. (In steady-state simulations the responses y_{hr} and $y_{hr'}$ with r, r' = 1,..., m_h are auto-correlated, if subruns are used; see the paragraph below eq. 8; the averages per combination h can be independent.)

Practitioners often use the same random number seed for all n factor combinations and then some responses are dependent, namely y_{hr} and $y_{h'r}$ are dependent where h, h' = 1,...,n and r = 1,...,m; obviously $m_h = m$. In other words, common random numbers yield a non-diagonal matrix $\underset{\sim}{\Omega}_y$. This practice may increase the efficiency of the simulation experiment (the variances of $\hat{\beta}_j$ are reduced), but it also complicates the analysis, as we shall see. [Common random numbers resemble blocking in experiments with non-simulated systems. However, the standard analysis of a blocked design assumes a special covariance matrix $\underset{\sim}{\Omega}_y$, namely constant correlations within blocks. Empirical results show that the

assumption of constant correlations is unrealistic. See Kleijnen (1986, p. 317), Nozari et al. (1984), Schruben (1979).]

The estimation of the response covariances is simple, if all n combinations of simulation parameters are run with a common seed, and this is repeated with m different seeds [obviously the covariances remain constant over replications: cov $(y_{hr}, y_{h'r}) = \sigma_{hh'}$]:

$$\hat{\sigma}_{hh'} = \frac{\sum_{r=1}^{m} (y_{hr} - \bar{y}_h)(y_{h'r} - \bar{y}_{h'})}{(m-1)}$$

$$(h, h' = 1,\ldots,n)(m \geq 2) \qquad (12)$$

Obviously $\hat{\sigma}_h^2 = \hat{\sigma}_{hh'}$ with h = h'; see eq. (8). In steady-state simulation there are other estimators for the covariances which, however, are more difficult; see Kleijnen (1986, p. 171).

The analysis of simulation with common seeds should use the estimated covariances of eq. (12). We have the same options as in the preceding section (CLS. EWLS, SLS). In the present section, however, $\hat{\Omega}_{\sim y}$ is block-diagonal, i.e. on the main diagonal of the N x N matrix $\hat{\Omega}_{\sim y}$ (with N=nm) there are m equal submatrices of size n x n. The estimated covariance matrix $\hat{\Omega}_{\sim y}$ may be singular; singularity occurs if common seeds result in perfect linear correlation coefficients ($\rho_{hh'} = 1$). Then EWLS does not exist; see Kleijnen (1986b). In option (d) we fit the regression model to the averages $\bar{y}_h$ with approximately constant variance, and then $\hat{\Omega}_{\sim y}$ is an n x n non-diagonal matrix.

Current regression analysis handles dependence as follows. In econometric packages $\Omega_{\sim y}$ is estimated through k-stage least squares using estimated residuals. Estimation of residuals makes the results dependent on the regression model specification (also see Section 7). Multivariate regression analysis may be used to analyze simulation experiments with common seeds. Unfortunately, multivariate statistics is not part of the basic training most practitioners receive. Therefore we recommend independent seeds. Common seeds may increase efficiency but it also increas-

es the analysis complexity; the confidence interval lengths of the β estimators may also be decreased by adding simulations runs (which means that cheap computer power is substituted for expensive human resources).

Summary: If and only if the simulation uses common random numbers, then the responses y are dependent. The resulting covariances $\sigma_{hh'}$ should then be estimated; see eq. (12). These estimated covariances can be used in several options (CLS, EWLS, SLS; see the preceding section).

6. NORMALITY

Nonnormality may be a smaller problem in simulation than in other areas. For example, the simulation response y may be the average waiting time of a simulation run; such an average may be approximately normal as explained by a limit theorem for autocorrelated variables (see Janssens 1982); replication of the long simulation run is needed to estimate the response variances σ_h^2, and to derive confidence intervals for $\underset{\sim}{\beta}$. Nonnormality in simulation does not show special problems. Consequently the standard options of modern regression analysis apply, such as detection and removal of nonnormality including outliers, Least Absolute Deviation regression analysis, robust regression analysis, distribution - free regression analysis. For details we refer to modern regression software and literature. This literature has been growing dramatically over the past decade. Atkinson (1985), Beckman and Cook (1983), Hocking (1983/1984) and Kleijnen (1986) give many references; for more references we refer to journals like Technometrics and Communications in Statistics.

Simulation has one special possibility: it is easy to check if an extreme response is due to pure. We can execute the simulation program using a new random number seed (in non-simulated systems it is often difficult to get a new replication). We recommend to replicate a suspicious response more than once. If the suspicious response is more extreme than all its replicates, then we add the new replicates to the data set and eliminate the outlier since the outlier influences the regression results too much, especially in Least Squares analysis. For a

case study (involving a computer center) we refer to Keyzer et al. (1981). Current software makes it easy to add data.

Outliers in the independent variables $\underset{\sim}{X}$ do not occur in simulation, if the simulation experiment is well designed. Modern regression software signals possible outliers in $\underset{\sim}{X}$; see Gray and Ling (1984).

There is one distribution-free regression procedure that has been applied to several simulations, and that is simple, both conceptually and computationally. Conover and Iman (1981) replace the original observations (y_i, x_{ij}) by the ranks, i.e., they explain the rank of y_i as a function of the ranks of x_{ij}; for example,

$$R(y_i) = \beta_0 + \beta_1 R(x_{i1}) + \beta_2 R(x_{i2}) + \beta_{12} R(x_{i1}) R(x_{i2}) + e_i \qquad (13)$$

where, if x_1 has no effect, $\beta_1 = 0$ and $\beta_{12} = 0$. The response y (not its rank) is estimated by linear interpolation. Interpolation and ranking (or sorting) are standard procedures, so that rank regression remains simple. Rank regression may work well, provided y is a monotonic function of x_j. Rank regression may show whether a factor is important; it does not explain how the response is affected, since the original scale is replaced by the ordinal scale. Rank regression combined with OLS should be added to the options presented in the two preceding sections (CLS, EWLS, SLS). (We do not recommend to apply as many as 57 different regression estimators; see Dempster et al. (1977)'s Monte Carlo experiment.)

Summary: Nonnormality may be a smaller problem in simulation than in other areas. If the fitted model shows outliers, then we may use new random number seeds and add these new data to the old data, possibly removing suspected responses. The diagnostic messages and the options of modern regression software may also be helpful in the analysis of simulation data. Rank regression should be added to the options, provided the response is a monotonic function of the inputs.

7. VALIDATION

Kleijnen (1986, pp. 185-186) discusses how to obtain a correct specification of the regression model. Once the regression model is specified, it remains to test the validity of that model; specification error implies $E(e) \neq 0$. The experimental design literature concentrates on the lack-of-fit F-test (for references see Kleijnen, 1986, pp. 229-233). The older regression literature discusses several other tests; see, for example, Rao (1959). The modern literature recommends cross-validation, defined below. We too recommend cross-validation, because it uses an approach, that is also followed in simulation, not only in regression analysis (briefly, the approach uses the model to predict a response; next it compares the predicted response to the actual response; the latter response does not depend on the model to be validated). We adjust cross-validation of modern regression texts, since we wish to account for variance heterogeneity and for replication; see eq. (15).

In cross-validation we delete factor combination h ($h = 1,\ldots,n$) and from the remaining $N-m_h$ observations we obtain the estimator $\hat{\underset{\sim}{\beta}}_{(-h)}$ (or some other estimator, like the EWLS estimator $\tilde{\underset{\sim}{\beta}}_{(-h)}$). Then we predict the response for combination h:

$$\hat{y}_h = \underset{\sim}{x}_h' \hat{\underset{\sim}{\beta}}_{(-h)} \qquad (h = 1,..,n) \qquad (14)$$

where $\underset{\sim}{x}_h'$ is the hth row of $\bar{\underset{\sim}{X}}$ obtained from $\underset{\sim}{X}$ by deleting identical rows ($\bar{\underset{\sim}{X}}_{(-h)}$ will denote $\bar{\underset{\sim}{X}}$ excluding $\underset{\sim}{x}_h$). The predictor $\hat{y}_h$ is compared to the actual average simulation response $\bar{y}_h$. We reject the model, if the prediction error is large, accounting for the variability of the simulation responses:

$$z_h = \frac{\bar{y}_h - \hat{y}_h}{\{\hat{var}\,(\bar{y}_h) + \hat{var}\,(\hat{y}_h)\}^{\frac{1}{2}}} \qquad (15)$$

where $\hat{var}\,(\bar{y}_h)$ is part of the simulation data(see the comments on eq. 8: $\hat{var}\,(\bar{y}_h) = \hat{\sigma}_h^2/m_h$) and eq. (14) yields

$$\hat{\text{var}}\,(\hat{y}_h) = \underset{\sim}{x}_h' \, \hat{\underset{\sim}{\Omega}}_{\hat{\beta}_{(-h)}} \, \underset{\sim}{x}_h \qquad (16)$$

where $\hat{\underset{\sim}{\Omega}}_{\hat{\beta}_{(-h)}}$ follows from eq. (9) (or eq. 11 if $\tilde{\beta}$ replaces $\hat{\beta}$ in eq. 14). [If common random numbers are used, then the denominator of eq. (15) should be expanded with the covariance term $-2\,\hat{\text{cov}}\,(\bar{y}_h, \hat{y}_h)$; we do not know whether this term can be neglected; we might apply the multivariate procedure due to Rao (1959); common random numbers indeed complicate the analysis; see Section 5.] In eqs. (14) through (16) h ranges from 1 to n. Obviously, if $\bar{\underset{\sim}{X}}_{(-h)}$ is collinear, we cannot compute $\hat{\underset{\sim}{\beta}}_{(-h)}$; a necessary (but not sufficient) condition is $n > q$ (where q denotes the number of regression parameters β_j). This cross-validation approach yields n "Studentized" prediction errors z_h if no matrix $\bar{\underset{\sim}{X}}_{(-h)}$ is collinear. Otherwise we limit the cross-validation to n' ($1 \leq n' \leq n$) non-collinear matrixes $\bar{\underset{\sim}{X}}_{(-h)}$. Unfortunately the variables z_h are dependent. Kleijnen (1983) examines the following test (the literature assumes constant variances; see Atkinson, 1985, Beckman and Cook, 1983, Ghosh, 1983, Hocking, 1983).

The regression model should hold at all n observation points. Consequently we reject the model, whenever <u>any</u> prediction error z_h is significant. In order to keep the "experimentwise" type-I error below the value α_E, we use the Bonferroni inequality (see Miller, 1966), i.e., we reject the regression model if

$$\max_{1 \leq h \leq n'} |z_h| > z^{\alpha} \text{ with } \alpha = \frac{\alpha_E}{2n'} \qquad (17)$$

where z is the standard normal N(0,1) and $P(z > z^{\alpha}) = \alpha$. For example, if $n' = 8$ and $\alpha_E = 20\%$ then $\alpha = 1.25\%$. Also see Cook and Prescott (1981), Kleijnen (1986, p. 189).

The Monte Carlo experiment in Kleijnen (1983) suggests that the test of inequality (17) has good power. A Monte Carlo experiment by Miyashita and Newbold (1983) suggests that the statistic is sensitive to

nonnormality, i.e., tails heavier than Gaussian lead to a chance higher than the nominal α level of finding extreme values.

Deterministic simulation implies $\hat{var}(\bar{y}_h) = 0$ in eq. (15). To compute $\hat{var}(\hat{y}_h)$ we might estimate the common response variance σ^2 (see Section 4) from the Mean Squared Residuals (see eq. 6). However, an incorrect regression model inflates the MSR: the worse the model is, the smaller the power of our test becomes! Therefore we recommend to compute the relative prediction errors $\hat{y}_h/\bar{y}_h$ and to reject the model if these errors are too "big", where "big" depends on the actual use of the simulation model. We note that in rank regression $\hat{y}$ is computed through linear interpolation; hence it seems impossible to derive $\hat{var}(\hat{y})$ in eq. (15). Therefore we may again use $\hat{y}/y$ as a criterion.

Only if the regression model (as a whole) is valid, the estimators $\hat{\underset{\sim}{\beta}}$ and $\tilde{\underset{\sim}{\beta}}$ are unbiased. Therefore regression software should present validation results before standard errors and t values of individual regression parameters are presented. Kleijnen (1986, pp. 190, 193-194) discusses the investigation of individual parameters and subsets of parameters.

Summary: To test the validity of the regression model the model's forecast $\hat{y}_h$ is compared to the actual average simulation response $\bar{y}_h$ where $\hat{y}_h$ is computed from all runs excluding factor combination h. This cross-validation yields many validation points. The maximum absolute Studentized error is used as a test statistic. Deterministic simulation, however, may use the relative prediction errors $\hat{y}_h/\bar{y}_h$.

8. EPILOGUE

Modern regression software has many more capabilities that are also useful in the analysis of simulation data. For example, that software permits transformations of variables in order to obtain a linear model or a model which meets the statistical assumptions of constant variances and normality; it allows the addition of new independent variables, including interactions among factors and (purely) quadratic effects.

Regression packages should also be capable of serving as a big module within a larger suite of modules, i.e., the regression package gets its input from the simulation program and the regression may deliver its output to a next module, for example, a Response Surface Methodology (RSM) program. Briefly, RSM optimizes the simulated (or real) system; RSM is a heuristic, stepwise procedure that combines the steepest ascent technique with linear regression models fitted locally; see Kleijnen (1986), Myers (1971).

In this paper we have tried to specify which changes and additions should be made to standard regression software, in order to accommodate the special problems and possibilities of (random and deterministic) simulation. We have ignored the numerical aspects; see Beckman and Cook (1983, p. 141), Bock and Brandt (1980), Hoaglin and Welsch (1978), Wolach (1983).

Regression analysis of simulation data provides an explicit summary of the relationships between the inputs and outputs of the simulation computer program (or simulation model). The regression model is a metamodel that guides the simulation analyst in the validation of the simulation model, in what-if questions, and in optimization. Applications of regression analysis of simulation data have started to appear, in academia and in practice; Kleijnen (1986) gives a long list of references and two detailed case studies.

9. CONCLUSIONS

Correct applications of regression analysis will be stimulated by modern regression software tailored to the needs of simulation analysts (who are familiar with computers and mathematical modeling but not with advanced statistics):

(a) Simulation implies active experimentation instead of passive observation. Hence experimental designs (such as 2^{k-p} designs) are often used. If nevertheless near-collinearity arises, then the regression software should not resort to special analysis techniques like ridge regres-

sion; instead the software should permit addition of a few extra runs reducing collinearity.

(b) Simulation runs provide not only average responses $\bar{y}_h$ but also variance estimates $\hat{\sigma}_h^2$ ($h = 1,\ldots,n$). These $\hat{\sigma}_h^2$ may differ substantially. The $\hat{\sigma}_h^2$ can be used in Corrected Least Squares (CLS), Estimated Weighted Least Squares (EWLS), and Sequential Least Squares (SLS).

(c) Practitioners often use common random number seeds in the n combinations of simulation factors. Then estimates of the resulting covariances among responses should be obtained. These estimated covariances can again be used in CLS, EWLS, and SLS.

(d) Nonnormality may be less problematic in simulation, if the responses y_i ($i = 1,\ldots,N = \sum_1^n m_h$) are based on long (sub)runs. If, nevertheless, the fitted regression model shows outliers, then we may use new seeds to obtain new responses, which may replace the outliers. An alternative analysis is rank regression, provided the simulation output is a monotonic function of the inputs.

(e) To test the specification of the regression (meta)model, we use cross-validation. For deterministic models the criterion is the relative prediction errors $\hat{y}/y$. For random simulation the criterion is the maximum absolute Studentized forecast error, accounting for variance heterogenity.

We hope that this article provides a warning to simulation analysts using standard regression software, and a challenge to developers of regression software!

ACKNOWLEDGMENT

I thank the anonymous referees for their comments that helped to improve the presentation of this paper.

REFERENCES

Atkinson, A.C. (1985). Plots, Transformations, and Regression. Clarendon Press.

Beckman, R.J. and R.D. Cook (1983). Outlier ... s. (And discussion). Technometrics, 25, no. 2: 119-163.

Bock, R.D. and D. Brandt (1980). Comparison of some computer programs for univariate and multivariate analysis of variance. Handbook of Statistics, Volume I, edited by P.R. Krishnaiah, North-Holland Publishing Company, Amsterdam.

Conover, W.J. and R.L. Iman (1981). Rank transformations as a bridge between parametric and nonparametric statistics. (Including comments and rejoinder). The American Statistician. 35, no. 3: 124-133.

Cook, R.D. and P. Prescott (1981). On the accuracy of Bonferroni significance levels of detecting outliers in linear models. Technometrics, 23, no. 1: 59-63.

Dempster, A.O., M. Schatzoff and N. Wertmuth (1977). A simulation study of alternatives to ordinary least squares. Journal American Statistical Association. 72: 77-106.

Dykstra, O. (1959). Partial duplication of factorial experiments. Technometrics. 1: 63-75.

Ghosh, S. (1983). Influential observations in view of design and inference. Communications in Statistics, Theory and Methods, 12, no. 14: 1675-1683.

Gray, J.B. and R.F. Ling (1984). K-clustering as a detection tool for influential subsets in regression. (And discussion.) Technometrics, 26, no. 4: 305-320.

Hoaglin, D.C. and R.E. Welsch (1978). The hat matrix in regression and ANOVA. American Statistician. 32, no. 1: 17-22.

Hocking, R.R. (1983/1984). Developments in linear regression methodology: 1959-1982. (And discussion.) Technometrics, 25, no. 3: 219-249, and 26, no. 3: 297-301.

Hoerl, A.E. and R.W. Kennard (1981). Ridge regression - 1980; advances, algorithms and applications. American Journal Mathematical and Management Sciences. 1, no. 1: 5-83.

Hoyle, M.H. (1973). Transformations - an introduction and a bibliography. International Statistical Review. 14, no. 2: 203-223.

Janssens, G.K. (1982). References on stochastic processes, European Journal of Operational Research. 10, no. 4: 421-422.

Keyzer, F., J. Kleijnen, E. Mullenders and A. Van Reeken (1981). Optimization of priority class queues, with a computer center case study. American Journal Mathematical and Management Sciences, 1, no. 4: 341-358.

Kleijnen, J.P.C. (1974/1975). Statistical Techniques in Simulation. Volumes I and II. Marcel Dekker, Inc., New York. (Russian translation: Publishing House "Statistics", Moscow, 1978.)

Kleijnen, J.P.C. (1979). The role of statistical methodology in simulation. In: Methodology in Systems, Modeling and Simulation, edited by B. Zeigler, North-Holland Publishing Company, Amsterdam.

Kleijnen, J.P.C. (1983). Cross-validation using the t statistic. European Journal Operational Research. 13, no. 2: 133-141.

Kleijnen, J.P.C. (1986). Statistical Tools for Simulation Practitioners. Marcel Dekker, Inc., New York, (forthcoming).

Kleijnen, J.P.C. (1986b). Analyzing simulation experiments with common random numbers. Katholieke Universiteit Brabant.

Kleijnen, J.P.C., P. Cremers and F. Van Belle (1985). The power of weighted and ordinary lest squares with estimated unequal variances in experimental design. Communications in Statistics, Simulation and Computation, B14, no. 1: 85-102.

Kleijnen, J.P.C., A.J. Van den Burg and R.T. Van der Ham (1979). Generalization of simulation results: practicality of statistical methods. European Journal of Operational Research. 3: 50-64.

Kleijnen, J.P.C. and W. Van Groenendaal (1986). Regression analysis of factorial designs with sequential replication. Katholieke Universiteit Brabant.

Miller, R.G. (1966). Simultaneous Statistical Inference. McGraw-Hill Book Company, New York. Second edition. (Revised edition: Springer-Verlag, New York, 1981.)

Miyashita, H. and P. Newbold (1983). On the sensitivity to non-normality of a test for outliers in linear models. Communications in Statistics, Theory and Methods. 12, no. 12: 1413-1419.

Myers, R.H. (1971). Response Surface Methodology. Allyn and Bacon, Inc., Boston.

Nozari, A. (1984). Generalized and ordinary least squares with estimated and unequal variances. Communications in Statistics, Simulation and Computation, 13, no. 4: 521-537.

Nozari, A., S.F. Arnold and C.D. Pegden (1984). Statistical analysis under Schruben and Margolin correlation induction strategy. School of Industrial Engineering, University of Oklahoma. (Submitted for publication.)

Rao, C.R. (1959). Some problems involving linear hypotheses in multivariate analyses. Biometrika, 46: 49-58.

Scheffé, H. (1964). The Analysis of Variance. John Wiley & Sons, Inc., New York. Fourth printing.

Schmidt, P. (1976). Econometrics. Marcel Dekker, Inc., New York.

Schruben, L.W. (1979). Designing correlation induction strategies for simulation experiments. Current Issues in Computer Simulation, edited by N.R. Adam and A. Dogramaci Academic Press, Inc., New York.

Wolach, A.H. (1983). Basic Analysis of Variance Programs for Microcomputers. Wadsworth International Group, Belmont (California).