Cross-validation using the t statistic

Kleijnen, J.P.C.

Published in:
European Journal of Operational Research

Publication date:
1983

Cross-validation using the t statistic *

Jack P.C. KLEIJNEN
Department of Business and Economics, Tilburg University, 5000 LE Tilburg, Netherlands

Received September 1981
Revised February 1982

One application area of regression analysis is simulation where the regression model may explain the relationship between the simulation model's inputs and outputs.

However, whether or not the regression model is used in a simulation context, its validity can be tested by comparing the model's forecast to one or more new observations not used in the estimation of the model's parameters. The familiar Student or t statistic is proposed for this comparison, combined with a Bonferroni approach accounting for the presence of multiple, dependent validation observations.

A 'trick' is used to obtain as many validation observations as possible. This trick is also known as cross-validation.

Several Monte Carlo experiments are performed to study the α and β errors of the proposed validation procedure. The experimental results suggest that the procedure is worthwhile.

1. Introduction

In various publications I have discussed how the response of a simulation model to changes in its parameters can be explained through a regression (meta)model; see Kleijnen (1981). In symbols, let the simulation model denoted by f_1, have responses y, input factors x, and random number seed r:

$$y = f_1(x_1, x_2, \ldots, x_k, r). \tag{1}$$

Then the regression metamodel is

$$y = f_2(x_1, x_2, \ldots, x_k) + e \tag{2}$$

where e represents noise and f_2 is a much simpler function than f_1, for instance, f_2 equals

$$E(y) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j \tag{3}$$

where β_j ($j = 0, 1, \ldots, k$) are regression parameters. To test the validity of the (meta)model I proposed the following t statistic:

$$t = \frac{y - \hat{y}}{s_{y-\hat{y}}} \tag{4}$$

where $\hat{y}$ is the regression forecast and $s_{y-\hat{y}}$ denotes the estimated standard deviation of $y - \hat{y}$. The present paper investigates the statistical behaviour of the proposed t statistic in more detail, using a Monte Carlo approach.

Note that the t statistic of eq. (4) may also be used in contexts different from metamodeling, e.g. in the selection of the appropriate degree of a polynomial regression model when constructing confidence intervals in simulation; see Heidelberger and Welch (1980).

2. Cross-validation

Let n simulation runs be available, yielding

$$\{x_{ij}, y_i, s_{y_i}\}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k \tag{5}$$

where s_{y_i} (or briefly s_i) denotes the estimated standard deviation of y_i. This s_i is based on simulation run i using a technique like batch-means, spectral analysis, renewal analysis, etc.; see Fishman (1978). If each simulation run uses different seeds r, then we *know* that the y_i are independent. Hence, the covariance-matrix $\boldsymbol{\Omega}_y$ of $\boldsymbol{y}' = (y_1, \ldots, y_n)$ is a diagonal matrix, say $\boldsymbol{D}$ with main-diagonal elements $\sigma_i^2 = E(s_i^2)$.

* This research was done while the author was visiting the IBM Thomas Watson Research Center in Yorktown Heights, New York.
My visit to Yorktown Heights was financed by IBM Netherlands, IBM Europe, and IBM Research. Computer time for the Monte Carlo experiments was made available by IBM Research. My knowledge of APL was quickly refreshed through the kind assistance of drs. P. Heidelberger, S. Lavenberg and P. Welch, all in Yorktown Heights. For the experiments' analysis I could use a new graphical package, called GRAFSTAT, developed by drs. L. Wu and P. Welch in Yorktown Heights.

North-Holland Publishing Company
European Journal of Operational Research 13 (1983) 133–141

Standard regression analysis results in the regression metamodel's forecast

$$\hat{y}_i = \boldsymbol{x}_i' \cdot \hat{\boldsymbol{\beta}} \tag{6}$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}' \cdot \boldsymbol{y}. \tag{7}$$

Defining $\boldsymbol{W} \equiv (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}$ yields

$$\boldsymbol{\Omega}_{\hat{\beta}} = \boldsymbol{W} \cdot \boldsymbol{\Omega}_y \cdot \boldsymbol{W}'. \tag{8}$$

Hence

$$\text{var}(\hat{y}_i) = \boldsymbol{x}_i' \cdot \boldsymbol{\Omega}_{\hat{\beta}} \cdot \boldsymbol{x}_i. \tag{9}$$

So the denominator in eq. (4) becomes

$$s_{y-\hat{y}} = \left\{ s_i^2 + \hat{\text{var}}(\hat{y}_i) \right\}^{1/2} \tag{10}$$

where the right-hand estimators follow from eqs. (5) and (9). Eq. (10) assumes that y and $\hat{y}$ are independent. This statistical requirement is automatically met if the regression model is *validated* in the following traditional scientific way:

(i) Estimate the (regression) model from, say v observations (in the above equations $v = n$ but below I shall propose $v = n - 1$).

(ii) Use the estimated model to forecast the response y at a *new* set of simulation inputs $\boldsymbol{x}$.

(iii) Compare the result of step (ii), say $\hat{y}$, to the actual simulation observation y. Observe that y and $\hat{y}$ are statistically independent ($\hat{y}$ depends on $\hat{\boldsymbol{\beta}}$, i.e., on the v *old* observations which are independent of the new observation; remember $\boldsymbol{\Omega}_y = \boldsymbol{D}$).

Note that classical statistical tests for testing the adequacy of a postulated regression model, do not set apart one or more observations for validation; see Kleijnen (1975).

In preceding publications, I have mentioned a special 'trick' for obtaining validation observations; see Kleijnen (1981). This trick turns out to be the same as cross-validation discussed in a few statistical articles; Allen (1974) and Stone (1974). What is new in the present paper is the combination of cross-validation and the t statistic defined in eq. (4). The 'trick' runs as follows, supposing n observation vectors are available as specified in eq. (5):

(i) For the time being *delete* , say, the last observation n, which results in a set of $n-1$ multi-dimensional observations, denoted by

$$\left\{ \boldsymbol{X}_{(n)}, \boldsymbol{y}_{(n)}, \hat{\boldsymbol{D}}_{(nn)} \right\} \tag{11}$$

where

$$\boldsymbol{X}_{(n)} = \begin{bmatrix} x_{11}, & x_{12}, \ldots, & x_{1q} \\ \vdots & & \\ x_{n-1,1}, & x_{n-1,2}, \ldots, & x_{n-1,q} \end{bmatrix} \tag{12}$$

with the dummy variables $x_{11} = \cdots = x_{n-1,1} \equiv 1$ and with $q \equiv k + 1$. Further

$$\boldsymbol{y}_{(n)}' = (y_1, \ldots, y_{n-1}) \tag{13}$$

and $\hat{\boldsymbol{D}}_{(nn)}$ is an $(n-1)$ by $(n-1)$ diagonal matrix obtained from $\hat{\boldsymbol{D}}$ by deleting row n and column n.

(ii) Estimate $\boldsymbol{\beta}$ from the remaining $(n-1)$ observations:

$$\hat{\boldsymbol{\beta}}_{(n)} = \left(\boldsymbol{X}_{(n)}' \cdot \boldsymbol{X}_{(n)} \right)^{-1} \cdot \boldsymbol{X}_{(n)}' \cdot \boldsymbol{y}_{(n)} \tag{14}$$

with

$$\boldsymbol{\Omega}_{\hat{\beta}(nn)} = \boldsymbol{W}_{(nn)} \cdot \boldsymbol{D}_{(nn)} \cdot \mathrm{W}_{(nn)}' \tag{15}$$

where

$$\boldsymbol{W}_{(nn)} \equiv \left(\boldsymbol{X}_{(n)}' \cdot \boldsymbol{X}_{(n)} \right)^{-1} \cdot \boldsymbol{X}_{(n)}'. \tag{16}$$

(iii) Now forecast the deleted observation! Hence y_n is forecasted by

$$\hat{y}_n = \boldsymbol{x}_n' \cdot \hat{\boldsymbol{\beta}}_{(n)} \tag{17}$$

where $\boldsymbol{x}_n' = (x_{n1}, \ldots, x_{nq})$. The standard error of the forecast $\hat{y}_n$ follows from

$$\text{var}(\hat{y}_n) = \boldsymbol{x}_n' \cdot \boldsymbol{\Omega}_{\hat{\beta}_{(nn)}} \cdot \boldsymbol{x}_n. \tag{18}$$

(iv) The seriousness of the forecast error is measured by

$$t_n = \frac{y_n - \hat{y}_n}{\left\{ s_n^2 + \text{var}(\hat{y}_n) \right\}^{1/2}}. \tag{19}$$

(v) Next the role of observation n – deleted in step i – is taken over by one of the other observations i' ($i' = 1, \ldots, n-1$). All together n dependent observations result for t defined in eq. (4) or (19). (This dependence can be illustrated as follows. Suppose there is one 'wild' observation y_3. This y_3 affects both $\hat{\boldsymbol{\beta}}_{(1)}$ and $\hat{\boldsymbol{\beta}}_{(2)}$, and makes t_1 and t_2 dependent.) Since the postulated regression model

should hold at all n observation points, the regression model is rejected whenever *any* of the n observations on t is 'significant'. Hence define the null-hypothesis

$$H_0: \quad E(\hat{y}_i) = E(y_i) \quad (i = 1,\ldots,n) \tag{20}$$

and reject H_0 if

$$\left\{ \max_i |t_i| \right\} > t^{\alpha'} \tag{21}$$

where $t^{\alpha'}$ is defined by

$$P(t > t^{\alpha'}) = 1 - P(t < t^{\alpha'}) \tag{22}$$

where $\alpha' = \alpha_C/2$ since a *two*-sided test is in order and α_C denotes the value of the 'per comparison' error rate, i.e., the error rate used in an individual test. The Bonferroni approach means that $\alpha_C = \alpha_E/n$ where α_E denotes the value of the 'experimentwise' error rate, i.e., the α error rate which holds over the *whole* experiment, i.e., under the *composite* hypothesis H_0 (in eq. (20) the index i assumes more than a single value); see Kleijnen (1975). For instance, if $n = 8$ and $\alpha_E = 20\%$, then $\alpha' = 1.25\%$. In summary

$$\alpha' = \alpha_C/2 = (\alpha_E/n)/2. \tag{23}$$

Note that if the n observations on t were independent, then the Bonferroni inequality would not be needed. The Bonferroni inequality is conservative, because it guarantees that α_E equals or is smaller than $n \cdot \alpha_C$. For n independent t statistics it is easy to guarantee the experimentwise error rate exactly by solving $(1 - \alpha_E) = (1 - \alpha_C)^n$, and using the corresponding α' in eq. (21). Moreover, for independent t statistics alternatives for $\max|t_i|$ are possible (but not necessarily better), for instance, $\Sigma|t_i|$; see Reynolds and Deaton (1981).

3. Monte Carlo experiment I

This paper is devoted to the study of the statistical behaviour of the t statistic defined above. The two classical quantitative measures of this behaviour are the type α and type β errors. Eqs. (20) and (21) imply that the error of the first kind is

$$\alpha = P\left(\max_{1 \leq i \leq n} |t_i| > t^{\alpha'} \,|H_0 \right), \tag{24}$$

The (conservative) Bonferroni approach should yield $\alpha \leq \alpha_E$. Obviously, in practice H_0 never holds exactly. To study the exact α error I *make* H_0 hold exactly, i.e., I make f_1 defined in eq. (1) identical to f_2 in eq. (2)! Experience shows that f_2 specified as a function linear in the regression parameters $\boldsymbol{\beta}$, gives good results; see Kleijnen et al. (1979). Hence at this stage of the investigation

$$f_1(\boldsymbol{X}) \equiv f_2(\boldsymbol{X}) = \boldsymbol{X} \cdot \boldsymbol{\beta}. \tag{25}$$

To quantify the α error the following experiment is done:

(i) Select some arbitrary $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{D}$. Note that these matrices imply specific values for n and q.

(ii) Next, obtain independent samples of the errors e_i from $N(0,\sigma_i^2)$, where σ_i^2 is a main-diagonal element of $\boldsymbol{D}$.

(iii) From steps (i) and (ii) compute

$$\boldsymbol{y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{e}. \tag{26}$$

(iv) To obtain $\boldsymbol{y}_{(i)}$, $\boldsymbol{X}_{(i)}$ and $\boldsymbol{D}_{(ii)}$ delete specific values, as explained in eqs. (11) through (13); next, compute $\hat{\boldsymbol{\beta}}_{(i)}$ resulting in $\hat{y}_i$ and $\hat{\text{var}}(\hat{y}_i)$; finally, compute the 'normalized' prediction error t_i. Execute this procedure n times, namely for $i = 1,\ldots,n$ (cross-validation). Determine the maximum of the absolute values of t_i and using eq. (24) find whether the procedure incorrectly rejects H_0.

Note that for simplicity's sake I assume that $\boldsymbol{D}$ is known so that the individual t are distributed as $z \sim N(0,1)$; in Section 5 $\boldsymbol{D}$ becomes unknown.

(v) To reduce the noise in this experiment repeat steps (i) through (iv) m times, the only difference each time being the use of different random number seeds. I selected $m = 100$.

(vi) Repeat steps (i) through (v) for a few different $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{D}$, to study the Monte Carlo result's sensitivity to the parameters of the experiment.

Steps (i) through (v) yield $n \times m$ values of t, say, t_{ij} with $i = 1,\ldots,n$ and $j = 1,\ldots,m$.[1] Actually, the proposed procedure does not use the individual t values but their maximum absolute value:

$$t\max_j \equiv \max_{1 \leq i \leq n} |t_{ij}| \quad (j = 1,\ldots,m). \tag{27}$$

[1] The m replications of the experiment make t_{ij} and $t_{ij'}$ $(j \neq n')$ independent. However, dependence does exist within the same replication j, so that t_{ij} and $t_{i'j}$ $(i, i' = 1,\ldots,n)$ are dependent. Upon inspection of a randomly selected number of plots the dependence among the t within the same replication does not seem to be strong.

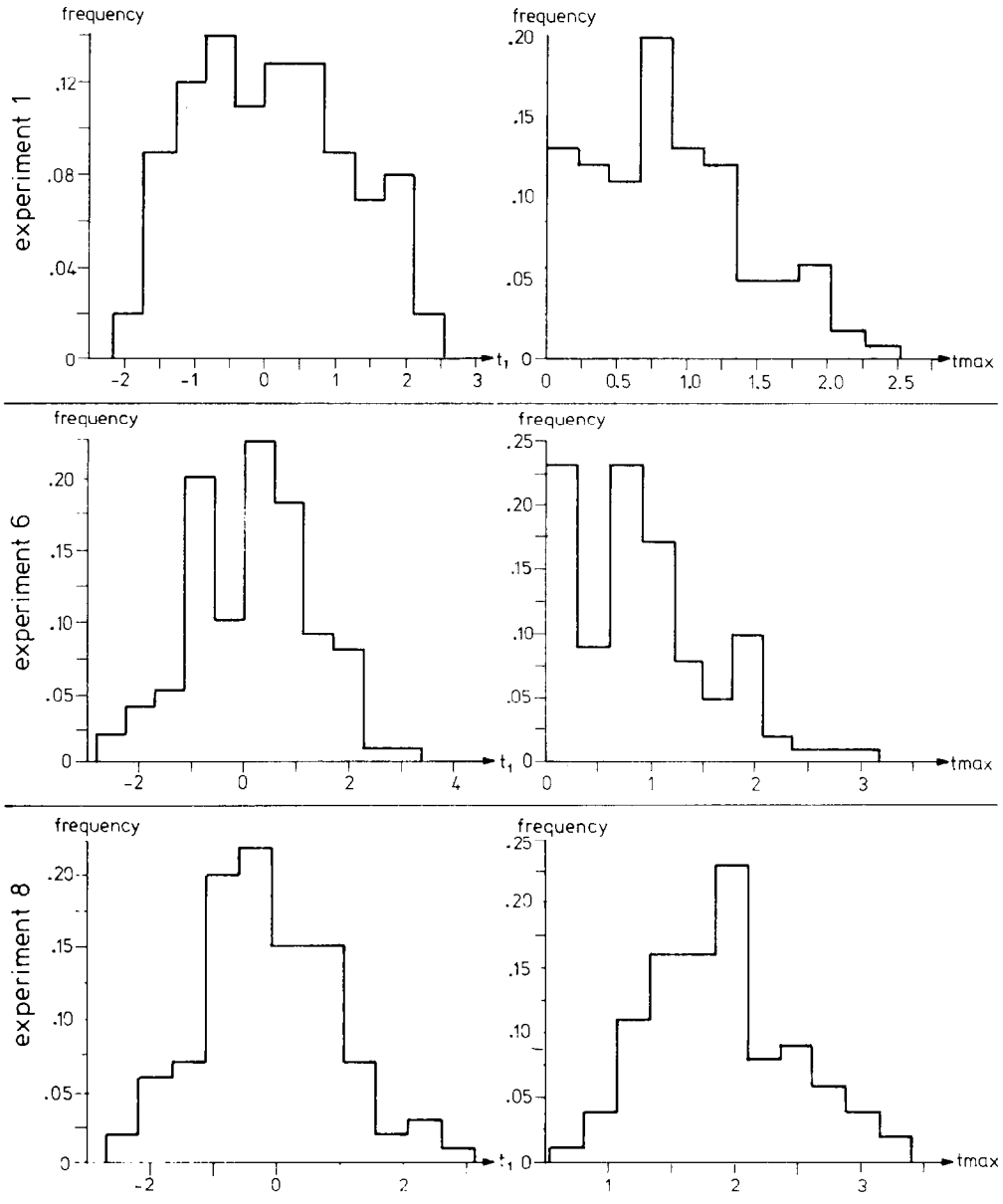

Fig. 1. Sample histograms of t_1 and $t\max$. (Experiments are characterized in Table 1.)

One could test the hypothesis that this $t\max$ has a specific distribution. Statistics related to $t\max$ are surveyed in chapter VB of Kleijnen (1975). Sample distributions of t_1 and $t\max$ are displayed in Fig. 1. However, as eqs. (20) through (22) show, it is the *tails* of the $t\max$ distribution that really matter. Hence define a new null-hypothesis H_{00} (to be distinguished from the 'lower level' null-hypothesis H_0 in eq. (20) as follows. Let

$$x = \begin{cases} 1 & \text{if } t\max < t^{\alpha'}, \\ 0 & \text{if } t\max \geq t^{\alpha'} \end{cases} \tag{28}$$

so that

$$E(x) \equiv p = P(t\max < t^{\alpha'}). \tag{29}$$

Then the new null-hypothesis is

$$H_{00}: \quad q \equiv 1 - p \leq \alpha_E. \tag{30}$$

H_{00} can be easily tested through the binomial distribution (and its normal approximation): Estimate p through

$$\hat{p} = \sum_{j=1}^{m} x_j/m. \tag{31}$$

The variance of $\hat{p}$ is

$$\text{var}(\hat{p}) = p \cdot (1-p)/m. \tag{32}$$

Hence reject H_{00} if

$$1 - \hat{p} \equiv \hat{q} > \alpha_E + z^{0.10} \cdot \{\alpha_E \cdot (1-\alpha_E)/m\}^{1/2}. \tag{33}$$

Appendix 1 shows that this test results in an α error not exceeding 10%. Since the reader may prefer a value different from 10%, Table 1 displays z' defined as follows:

$$z' = \frac{\hat{q} - \alpha_E}{\{\alpha_E \cdot (1-\alpha_E)/m\}^{1/2}}. \tag{34}$$

So H_{00} is rejected if $z' > z^{\alpha}$ where some familiar values are $z^{0.10} = 1.282$, $z^{0.05} = 1.645$, $z^{0.01} = 2.327$. Note that negative values of z' never lead to rejection of H_{00}.

Table 1 tests $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{D}$ as follows:

(i) All matrices $\boldsymbol{X}$ (except for experiment 11) are experimental designs, i.e., they consist of plus one's and minus one's. For instance, $\boldsymbol{X} = \boldsymbol{2}^2$ means

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}.$$

Readers unfamiliar with the basis of experimental design are referred to Chapter IV in Kleijnen (1975).

(ii) The diagonal elements of $\boldsymbol{D}$ are shown as a vector $\boldsymbol{\sigma}^2$. For example, $(\boldsymbol{\sigma}^2)' = (1,1,1,1)$ means that all four diagonal elements of $\boldsymbol{D}$ equal one.

(iii) If only a single line is given for an experiment it means that the preceding $\boldsymbol{X}$ and $\boldsymbol{\beta}$ apply, e.g., experiment 10 uses $\boldsymbol{X} = 2^{9-5}$, $\boldsymbol{\beta}' = (1,10,\ldots,10)$, $(\boldsymbol{\sigma}^2)' = (3,6,9,\ldots,45,48)$.

Table 1 shows that in the experiments the experimentswise error rate is indeed smaller than 20%. I could not find a clear-cut relationship between the estimated error rate $\hat{q}$ and the experiment's parameters $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{D}$. The experiments do *not* confirm the conjecture that the Bonferroni

Table 1
Testing the α error in H_{00}: $q \leq \alpha_E = 0.20$

Experiment	$\hat{q}$	(z')
1) $\boldsymbol{X} = \boldsymbol{2}^2$		
$\boldsymbol{\beta}' = (1,10,10)$		
$(\boldsymbol{\sigma}^2)' = (1,1,1,1)$	0.05	(-3.75)
2) $(\boldsymbol{\sigma}^2)' = (1,2,3,4)$	0.03	(-4.25)
3) $(\boldsymbol{\sigma}^2)' = (10,20,30,40)$	0.02	(-4.5)
4) $\boldsymbol{X} = \boldsymbol{2}^2$		
$\boldsymbol{\beta}' = (1,10,20)$		
$(\boldsymbol{\sigma}^2)' = (1,1,1,1)$	0.05	(-3.75)
5) $(\boldsymbol{\sigma}^2)' = (1,2,3,4)$	0.03	(-4.25)
6) $(\boldsymbol{\sigma}^2)' = (10,20,30,40)$	0.07	(-3.25)
7) $\boldsymbol{X} = 2^{9-5}$ [a]		
$\boldsymbol{\beta}' = (1,10,\ldots,10)$		
$(\boldsymbol{\sigma}^2)' = (1,1,\ldots,1)$	0.21	(0.25)
8) (second replicate)	0.14	(-1.5)
9) $(\boldsymbol{\sigma}^2)' = (i/2)$ $(i = 1,\ldots,6)$	0.14	(-1.5)
10) $(\boldsymbol{\sigma}^2)' = (3i)$ $(i = 1,\ldots,6)$	0.16	(-1.0)
11) $\boldsymbol{X}' = \begin{bmatrix} 1,1,\ldots \\ 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 \\ 1,2,3,1,3,6,1,1,8,5,8,3,7,3,3,11 \end{bmatrix}$	0.09	(-2.75)
$\boldsymbol{\beta}' = (1,10,10)$		
$(\boldsymbol{\sigma}^2)' = (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)$		

[a] Generators: 6 = 23, 7 = 24, 8 = 25, 9 = 34, 10 = 35; see Kleijnen (1975).

approach becomes more conservative (smaller q) as n increases. As σ_i^2 increases outliers become more likely, but the t statistic corrects for these 'wild' observations through its denominator.

Because Monte Carlo experimentation requires much computer time, I stopped the investigation of the α error rate at this point, and switched to the study of the β error.

4. Monte Carlo experiment II: β error

The values of the parameters $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{D}$ of the preceding experiments, executed to study the α error of the t statistic, are also used to examine the β error of the same statistic. From eqs. (20), (21), and (28) through (32) it follows that the estimated β error is

$$\hat{\beta} = \hat{p}_1 \equiv P\left(t\max < t^{\alpha'} | \mathrm{H}_1\right) \tag{35}$$

where H_1 denotes the alternative to H_0. Hypothesis testing was appropriate when studying the α error; see eq. (30). Now there is no reason to hypothesize a specific β value, and therefore a confidence interval approach becomes appropriate:

$$P\left(\beta \in \hat{\beta} \pm t_{m-1}^{\alpha/2} \cdot s(\hat{\beta})\right) = 1 - \alpha \tag{36}$$

where

$$s^2(\hat{\beta}) = \hat{\beta} \cdot (1 - \hat{\beta})/m \quad \text{if } 0 < \hat{\beta} < 1. \tag{37}$$

If $\hat{\beta}$ is one or zero no confidence interval is given since the experimental result is then obvious. Because $m - 1 = 99$ the value $t_{m-1}^{\alpha/2}$ can be replaced by $z^{\alpha/2}$.

There is only one way to satisfy H_0, but there are infinitely many ways to deviate from H_0. Actually, H_1 refers to f_1 given in eq. (1). Many specifications of f_1 could be studied, e.g., a queuing model could be specified. However, the simplest deviation of H_1 from H_0 I can imagine, is in terms of *interactions*: Let H_0 be specified by eq. (3) or (25):

$$\mathrm{H}_0: \quad E(\boldsymbol{y}) = \boldsymbol{X} \cdot \boldsymbol{\beta} \tag{38}$$

so that $\hat{\boldsymbol{\beta}}_{(n)}$ of eq. (14) is unbiased under H_0. Then specify

$$\mathrm{H}_1: \quad E(\boldsymbol{y}) = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{X}_2 \cdot \boldsymbol{\beta}_2 \tag{39}$$

where $\boldsymbol{\beta}_2$ denotes the vector of interactions. There are $q_2 \equiv k \cdot (k-1)/2$ interactions (k was defined below eq. (12)). Hence $\boldsymbol{X}_2$ is a matrix with n rows and q_2 columns. The formation of its elements can be illustrated as follows: For $k = 3$ eq. (38) yields

$$E(y_i|\mathrm{H}_0) = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} \quad (i = 1, \ldots, n) \tag{40}$$

so that eq. (39) results in

$$E(y_i|\mathrm{H}_1) = E(y_i|\mathrm{H}_0) + \beta_{12} \cdot x_{i1} \cdot x_{i2} + \beta_{13} \cdot x_{i1} \cdot x_{i3} + \beta_{23} \cdot x_{i2} \cdot x_{i3} \quad (i = 1, \ldots, n). \tag{41}$$

Next consider Table 2. Since $\boldsymbol{D}$ did not have an important effect in Table 1, I decided to save computer time by studying the β error (or its complement, the power of the test) only for $\sigma_i^2 = 1$. The numbers (1, 4, 7, 11) in the first column of Table 2 refer to the corresponding experiments in Table 1; corresponding experiments use the same random numbers. I made the size of the interaction equal to the size of the smallest effect, if $k = 2$ (see also Fig. 2). If $k = 9$ then there are 36 interactions. It seems most difficult for the test procedure to detect the presence of interactions when all interactions except one, are zero. Therefore I selected $\beta_{12} = 10$ and all other interactions zero in the experiment numbered 7. In experiment 7′ all interactions except two are equal to zero. The results of experiments 7 and 7′ (namely $\hat{\beta} = \hat{\alpha}$) are explained by the particular X matrix: $\boldsymbol{X} = 2^{9-5}$

Table 2
Estimating the power $1 - \beta$ ($\sigma_i^2 = 1, \alpha_E = 0.20$)

Experiment	$1 - \hat{\beta}$	$s(\hat{q})$
1) $\boldsymbol{X} = 2^2$ $\boldsymbol{\beta}' = (1,10,10)$ $\beta_{12} = 10$	1.0	–
4) $\boldsymbol{X} = 2^2$ $\boldsymbol{\beta}' = (1,10,20)$ $\beta_{12} = 10$	1.0	–
7) $\boldsymbol{X} = 2^{9-5}$ $\boldsymbol{\beta}' = (1,10,\ldots,10)$ $\boldsymbol{\beta}_2' = (10,0,0,\ldots,0,0)$	0.21	0.041
7′) $\boldsymbol{\beta}_2' = (10,10,0,\ldots,0)$	0.21	0.041
11) $\boldsymbol{X}$: see 11) in Table 1 $\boldsymbol{\beta}' = (1,10,10)$ $\beta_{12} = 10$	1.0	–

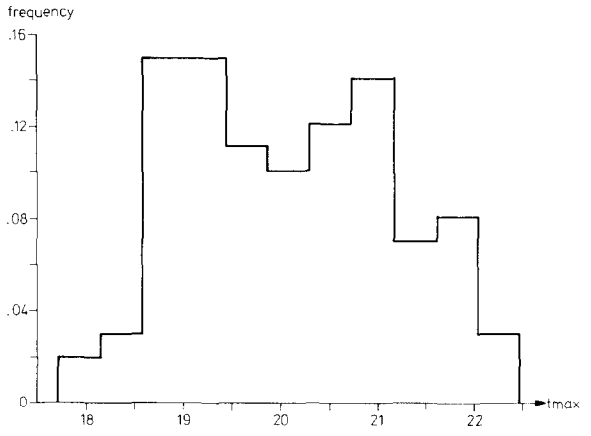

Fig. 2. Histogram of t max for experiment 1 in Table 2.

results in $E(\hat{\beta}_5) = \beta_5 + \beta_{12}$ and $E(\hat{\beta}_6) = \beta_6 + \beta_{13}$, so that – as Kleijnen et al. (1979) already mentioned – an experimental design can be self-defeating when the model on which it is based is misspecified.[2] Finally, experiment 11 represents an X matrix not taken from the experimental design literature. In summary, the results are that the power of the proposed procedure is perfect, except for the experiments involving the 2^{9-5} design.

5. Replacing the variances σ_i^2 by their estimators s_i^2

In the preceding experiments I assumed for simplicity's sake that $\boldsymbol{D}$ was known. Actually $\boldsymbol{D}$ is estimated from the estimators s_i^2; see eq. (5). Replacing $\boldsymbol{D}$ by $\hat{\boldsymbol{D}}$ affects the following variables: $\hat{\boldsymbol{\Omega}}_{\hat{\beta}}$, $\hat{\text{var}}(\hat{y})$, $\hat{\text{var}}(y)$, and t.

[2] If $x_{i5} = (x_{i1}) \cdot (x_{i2})$ for all values of i, then experimental design theory shows that $E(\hat{\beta}_5) = \beta_5 + \beta_{12}$. Hence when the false first-order model

$$E(y_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \cdot x_{ij}$$

is used when estimating β_j then

$$E(\hat{y}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \cdot x_{ij} + \beta_{12} \cdot x_{i5}$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j \cdot x_{ij} + \beta_{12} \cdot x_{i1} \cdot x_{i2}.$$

So the expectation of the estimated false model equals the true model!

Table 3
Estimated α and β errors when using s_i^2 instead of σ_i^2 ($\boldsymbol{X} = \boldsymbol{2}^2$, $\boldsymbol{\beta}' = (1,10,10)$, $\sigma_i^2 = 1$)

Degrees of freedom of s^2: L	Estimated α error: $\hat{q}$	Testing $q \leq \alpha_L$: z'
100	0.03	4.25
10	0.06	−3.5
2	0.11	−2.25
L, interaction	Estimated power: $\hat{q}$	
$L = 10$, $\beta_{12} = 10$	1.0	

Assume that s_i^2 is estimated from L subruns y_{il} ($l = 1,\ldots,L$). Hence in each replication j of the Monte Carlo experiment s_i^2 is sampled from a χ^2 distribution with $L-1$ degrees of freedom; see Appendix 2. Where in the preceding sections σ_i^2 was used, now s_i^2 is used.

Since a Monte Carlo experiment with sampling of s^2 requires much computer time, I studied only the situation corresponding to experiment 1 in Table 1: $\boldsymbol{X} = \boldsymbol{2}^2$, $\boldsymbol{\beta}' = (1, 10, 10)$ and $\sigma_i^2 = 1$. (All experiments in Table 3 use different random numbers, when compared among each other and when compared to Tables 1 and 2.) Note that even with L as low as two, t was compared to the critical values $z^{\alpha'}$ instead of $t^{\alpha'}$ thus avoiding the issue of determining the proper degrees of freedom. The results of Table 3 are close to those of Tables 1 and 2, i.e., replacing s^2 by σ^2 in the Monte Carlo experiments in order to save computer time, seems justified.

6. Summary and future research

The procedure investigated in this paper is based on the following approach, generally used in science when validating a model:

(i) Estimate the model;

(ii) Use the estimated model to forecast new observations;

(iii) Compare the forecasted values to the actual new observations. The proposed t statistic incorporates the inherent variability of the forecasted and the actual (simulation) observations.

The t statistic is combined with a 'trick' for obtaining many 'new' observations for validation,

namely cross-validation. To keep the experimentwise error rate under control, a Bonferroni approach is used.

Monte Carlo experimentation shows that the α error (erroneously rejecting a true model) is smaller than the value α_E specified for the experimentwise error rate. Fortunately, this low α error is not obtained at the price of an unacceptable β error, i.e., in the Monte Carlo experiments ignored interactions were detected in most situations. However, in one situation – namely a 2^{9-5} design – the ignored effects destroyed the otherwise attractive properties of well-balanced experiments.

Further research may concentrate on the following aspects:

(i) If in practice a significant t value is found at one or more (cross)validation points, then simulation offers a special advantage: After changing the random number seed (and keeping all other parameters constant) it can be checked whether the significance was due to pure chance (remember the definition of the α error).

(ii) If the number of observations n is very large, then cross-validation may be restricted to less than a complete permutation. For instance, the (meta)model may be validated at a randomly selected set of observation points. Besides saving computer time and analysis time, the Bonferroni approach becomes less conservative. However, potential information is ignored.

(iii) Kleijnen et al. (1981) found that in simulation Weighted Least Squares (WLS, using the weights $1/s_i^2$) gives more accurate estimators of β. Hence the t statistic may be based on WLS instead of Ordinary Least Squares.

Appendix 1. The one-sided binomial test

Consider the null-hypothesis

$$H_0: \quad p \leq p_0 (< 0.50). \tag{A1.1}$$

The point estimator of p is $\hat{p}$ with variance

$$\text{var}(\hat{p}) = p \cdot (1-p)/m. \tag{A1.2}$$

Note that $\text{var}(\hat{p})$ increases when p increases from zero to a half. Assuming a normal distribution for $\hat{p}$, the test procedure is: Reject H_0 if

$$\hat{p} > p_0 + z^{0.10} \cdot \{p_0 \cdot (1-p_0)/m\}^{1/2}. \tag{A1.3}$$

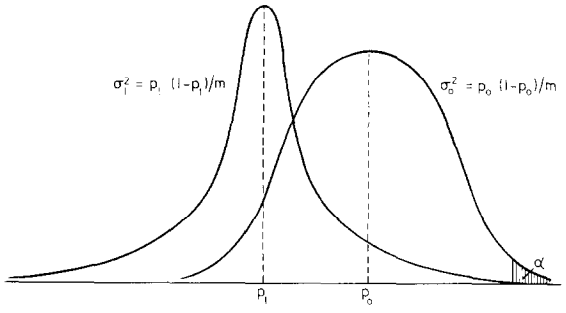

Fig. A1. The α error in a one-sided binomial test

To find the size of the α error, first assume that p is exactly equal to p_0. Then eq. (A1.3) implies an α error exactly equal to 10%. Next suppose p is smaller than p_0. Then Fig. A1 shows that the α error is smaller than 10%. In summary, the test procedure results in an α error not exceeding 10%.

Appendix 2. Sampling s_i^2 from D

Obviously, the following relations hold:

$$\sigma_i^2 \equiv \text{var}(y_i) = \text{var}(e_i) \quad (i = 1, \ldots, n). \tag{A2.1}$$

Suppose y_i is the average of L simulation subrun responses y_{il}:

$$y_i = \sum_{l=1}^{L} y_{il}/L. \tag{A2.2}$$

(Note that y_{il} may be an average, a quantile, etc.) Hence a more explicit notation is:

$$y_i \equiv \bar{y}_i. \tag{A2.3}$$

Assuming independent subruns y_{il} yields

$$\sigma_i^2 \equiv \text{var}(\bar{y}_i) = \text{var}(y_{il})/L. \tag{A2.4}$$

The corresponding variance estimators are

$$\hat{\text{var}}(\bar{y}_i) \equiv \hat{\sigma}_i^2 \equiv s_i^2 \tag{A2.5}$$

and

$$\hat{\text{var}}(y_{il}) = \sum_{l=1}^{L} (y_{il} - \bar{y}_i)^2/(L-1). \tag{A2.6}$$

The estimator $\hat{\text{var}}(y_{il})$ can be sampled using

$$\hat{\text{var}}(y_{il}) = \sum_{l=1}^{L} (e_{il} - \bar{e}_i)^2/(L-1). \tag{A2.7}$$

Summarizing:

(i) Sample e_{il} L times from $N(0, L \cdot \sigma_i^2)$.

(ii) Compute the average

$$e_i \equiv \bar{e}_i = \sum_{l=1}^{L} e_{il}/L \tag{A2.8}$$

to be added to $E(y_i) = \boldsymbol{x}_i' \cdot \boldsymbol{\beta}_i$ in the Monte Carlo experiment.

(iii) Compute

$$s_i^2 \equiv \hat{\text{var}}(y_i) = \left\{ \sum_{l=1}^{L} (e_{il} - \bar{e}_i)^2/(L-1) \right\}/L$$

$$= \left\{ \sum_{l=1}^{L} e_{il}^2 - L \cdot \bar{e}_i^2)/(L-1) \right\}/L. \tag{A2.9}$$

Note that $\bar{e}_i$ and $\hat{\text{var}}(e_{il})$ are independent so that the same statistical distribution results if eq. (A2.7) would be replaced by the following computationally more efficient formula:

$$\hat{\text{var}}(y_{il}) = \sum_{l=1}^{L} e_{il}^2/(L-1). \tag{A2.10}$$

In a specific realization eqs. (A2.7) and (A2.10) yield different results.

Note further that compared to the Monte Carlo experiments with no sampling of σ_i^2, more random numbers are needed, namely L times more. [3)] One consequence is that if two experiments use the same seed, the two random number streams get out of step (assuming no special programming tricks are applied). Consequently, all experiments in Table 3 use different random number streams.

[3)] After I finished the experiment, I discovered an APL subroutine permitting the sampling of s^2 from a χ^2 distribution using a *single* random number.

References

Allen, D.M. (1974), The relationship between variable selection and data augmentation and a method of prediction, *Technometrics* 16, 125–127.

Fishman, G.S. (1978), *Principles of discrete event simulation*, Wiley-Interscience, New York.

Heidelberger, P. and P.D. Welch, (1980) On the statistical control of simulation run length, RC 8571 (No. 37365), IBM Research, Yorktown Heights, NY.

Kleijnen, J.P.C. (1975), *Statistical techniques in simulation* (in two volumes), Marcel Dekker, New York. (Russian translation: Publishing House "Statistics", Moscow, 1978.)

Kleijnen, J.P.C. (1981), Regression analysis for simulation practitioners, *J. Operational Res. Soc. 32*, 35–43.

Kleijnen, J.P.C., R. Brent and R. Brouwers (1981), Small-sample behavior of weighted least squares in experimental design applications, *Comm. Statist. B – Simulation Comput. 10* (3), 303–313.

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 J. Operational Res. 3*, 50–64.

Reynolds, M.R. and M.L. Deaton (1981), Comparisons of some tests for validation of stochastic simulation models, *Comm. Statist. B – Simulation Comput.*, to appear.

Stone, M. (1974), Cross-validation choice and assessment of statistical predictions, *J. Roy. Statist. Soc. Ser. B 36*, 111–147.