Open access • Journal Article • DOI:10.1177/003754978504400504

On the interpretation of variables — Source link

Jack P. C. Kleijnen

Institutions: Tilburg University

Published on: 01 May 1985 - Simulation (SAGE Publications)

Topics: Risk analysis, Variables and Terminology

On the interpretation of variables

Kleijnen, J.P.C.

Publication date:
1983

faculteit der economische wetenschappen

RESEARCH MEMORANDUM

[redacted]

TILBURG UNIVERSITY
DEPARTMENT OF ECONOMICS
Postbus 90153 - 5000 LE Tilburg
Netherlands

FEW

136

ON THE INTERPRETATION OF VARIABLES

Jack P.C. Kleijnen

October 1983

Department of Business and Economics

Tilburg University

P.O. Box 90153

5000 LE Tilburg

Netherlands

CONTENTS

ABSTRACT

The input of a computer program, say a simulation program, specifies parameters, variables, and behavioral relationships. Parameters are not directly observable. Variables can be specified through enumeration, mathematical functions, and scenarios. In regression models the scenarios correspond to binary variables. Regression models accept different measurement scales: nominal, interval, ratio, absolute scales. The interpretation of interval variables may be misleading if there are interactions between regression variables. The interpretation of quantitative and qualitative variables (in regression versus ANOVA models) is different. The user distinguishes between environmental and controllable variables. Environmental variables involve validation, risk analysis, and sensitivity analysis. Controllable variables lead to optimization, control, and what-if questions.

1. INTRODUCTION

Terms like parameters, scenarios, variables are often used in practice and in the literature without definition. In this contribution we shall define these terms and interpret their role in various types of models, concentrating on simulation and regression models. (These two model types are very popular in management practice; moreover, regression models may be used in the interpretation of simulation models; see Kleijnen (1981)). The user's view of the model leads to questions of validation, risk analysis, sensitivity analysis, optimization, etc. Table 1 summarizes our contribution; we shall refer to that table as we proceed.

2. TYPES OF VARIABLES

As the first column of Table 1 shows, we translate the simulation model into a simulation program (using a general programming language or a special simulation language). The simulation program has input and output.

The output of a simulation program may be called the simulation model's response (second column of Table 1). This output comprises one or more time series. We can characterize a time series through one or more measures, e.g., we may capture the waiting times of consecutive customers by their average or by their .90 quantile, i.e., we may summarize W_t with $t = 1,2,...,T$ by $\bar{W}$ and $W_{.90}$ where $W_{.90}$ is defined by $P(W_t < W_{.90}) = 0.90$. In this article we shall concentrate on a single response, i.e., a single time series characterized by a single measure,

TABLE 1

Types of Variables

Computer program	Simulation model	Regression model	User view
Output	Response	Dependent variable (y)	Result
Input	1. Parameter	Independent variable	1. Environmental
	2. Variable	(x)	
	(i) Enumeration	(i) Continuous	(i) Validation (ii) Risk analysis
	(ii) Function	(ii) Discrete	2. Controllable
	(iii) Scenario	(iii) Binary	(i) Optimization (ii) Goal output (control) (iii) Satisficing (what-if)
	3. Behavioral relationship		

say, the average waiting time $\bar{W}$. In the terminology of regression analysis the response or output is called the dependent variable, usually denoted by the symbol y (third column). If we had multiple responses, we would speak of multivariate regression analysis.

The input of the simulation program can specify the simulation model's parameters, variables, and behavioral relationships. (Note that the input of a computer program might also be called the program's parameters, i.e., the program is a big subroutine or procedure that is called specifying the actual values of its parameters.) In a modeling context - to be distinguished from a programming context - we differentiate between variables and parameters. For instance, in a queuing model we may introduce a sequence of observed or actual service times $s_1, s_2, \ldots$. (We may use these actual values when validating the model.) Alternatively, we may sample the consecutive service times $s_1, s_2, \ldots$ from a statistical distribution (like the exponential distribution) with its "parameter" (say, λ); this distribution forms a submodel (for the service process) of the total queuing model. The difference between the variables and the parameters of a model is as follows: a variable is directly observable whereas a parameter requires statistical inference. Another difference is that a parameter remains constant during a simulation run; a variable may change during the run.

In the above terminology the number of service stations is a variable, not a parameter. And a sequence of actual service times is a sequence of variables. Actually we have several techniques for specifying such a time series:

(i) Enumeration: we list the individual values of the sequence. We use such enumeration when validating the simulation model. Note that the sequence may consist of a single value, e.g., the number of servers.

(ii) Functional specification: we specify a mathematical function like $x_t = a_1 + a_2t$. In deterministic simulations we often specify a trend, a sinus, a ramp function, etc. In random simulations we specify the sequence of variables by their distribution function (e.g., the exponential distribution with parameter λ) plus the random number seed. The parameters (like a_1, a_2, λ) of the functional specification should be inferred from historical data.

(iii) Scenarios: this term we reserve for complicated, dynamic specifications. An example is: "a single server is available as long as individual queuing times do not exceed five minutes; a second server becomes available whenever that server has not been active during the last ten minutes and ..." A different scenario changes the values (five, ten) and possibly the rules ("... and more than one hour until closing time remains"). The word "scenario" is currently very popular and often used without definition. For a more general discussion of scenarios we refer to Becker (1981).

A behavioral relationship specifies the model's reaction to changes in its parameters and variables. Mathematically, a behavioral relationship is a function (such as $y = 2x + 5$) excluding tautologies, i.e., mathematical identities (such as $\bar{W} \equiv \sum_1^n W_i/n$). In our queuing example an interesting behavioral relationship is the priority rule or queuing discipline (first-in-first-out, shortest-jobs-first, etc.). A different priority rule may be evoked by calling a computer subroutine using different input values. For example, Hellerman and Conroy (1975,

p. 113) execute priority rules by finding the customer with the minimum value for the "service variable" SV; hence if the queuing discipline is first-come-first-served then SV equals arrival time AT; however, if small-jobs-first is the rule then SV equals service time ST. Consequently, we may specify the actual priority rule by a binary variable, say X and a programming instruction like "if X = 0 then SV = AT else SV = ST". (Obviously other programming styles are possible.) We shall return to this example when discussing binary variables in regression analysis.

So the input of the simulation program specifies the simulation model's parameters, variables, and behavioral relationships. In the terminology of regression analysis, these inputs are called the independent variables, usually denoted by x. If we have more than a single x then we speak of multiple regression analysis. How do the simulation model's parameters, variables and relationships correspond to the regression variables?

(i) Consider a simulation parameter (like the exponential distribution's parameter λ or the trend parameter a_2) which specifies a sequence of variables. This simulation parameter may correspond to a continuous regression variable, e.g., $x_1 = \lambda$. The independent variables x may also correspond to functions of the simulation parameters, e.g., $x_2 = \lambda^2$.

(ii) A single variable like "number of servers" NS is handled in regression analysis exactly as parameters are handled. For instance, x_1 = NS or $x_1 = (NS)^2$, etc. Discrete variables (like x = NS) and continuous variables (like $x = \lambda$) are treated identically in regression analysis.

(iii) Differences among scenarios cannot be quantified so simply. Similar problems arise when we use different enumerations or behavioral relationships in the simulation model. These differences are qualitative rather than quantitative. We can represent qualitative differences among simulation models by using a regression model with binary variables, i.e., x is zero or one: see the example in the preceding paragraph with the service variable SV. Note that sometimes binary variables are called dummy variables, but we reserve the term "dummy" for a variable that remains constant, i.e., in regression analysis the constant β_0 corresponds with $x_0 = 1$. Next we shall examine the differences between "qualitative" and "quantitative" in more detail.

3. MEASUREMENT SCALES AND REGRESSION

Qualitative phenomena are measured on a nominal scale whereas quantitative phenomena are measured on an interval, a ratio or an absolute scale. We consider these four scales in more detail, because their differences are important when using regression analysis (the literature gives more types of scales; Hauser and Shugan (1980), Sprent (1981)):

(i) Nominal scale, for instance, machine type A, B or C; priority rule 1 (first-in-first-out) or rule 2 (last-in-first-out). In these examples the letters A, B, C and the numbers 1 and 2 are short-hand notations (mnemonics) used to distinguish priority rules; they imply no ranking.

(ii) Interval scale: this scale does rank objects, but it has an arbitrary zero point so that an object with value 2x is "better" than an object with value x but it is not twice as good. Examples are:

- Intelligence measured by the intelligence quotient (IQ): a person with an IQ (according to a specific test) of 140 is not twice as smart as one with an IQ of 70.
- Temperature measured in Celsius (x^*) or Fahrenheit (x), related by the equation

$$x^* = (x - 32).(5/9) \qquad (1)$$

Note that 20° C is not twice as warm as 10° C, which is clearly demonstrated when we choose a different scale such as Fahrenheit.

(iii) Ratio scale: this scale implies a ranking among objects; moreover it has a meaningful zero point such that 2x means "twice as much as x". Examples are length, measured in centimeters or inches (and the derived measures for surface and volume); angles measured in degrees or radians; richness measured in U.S. dollars or Dutch guilders. Different ratio scales are related by a linear transformation like eq. (1) but with a zero intercept, e.g., centimeters (x^*) and inches (x) are related by

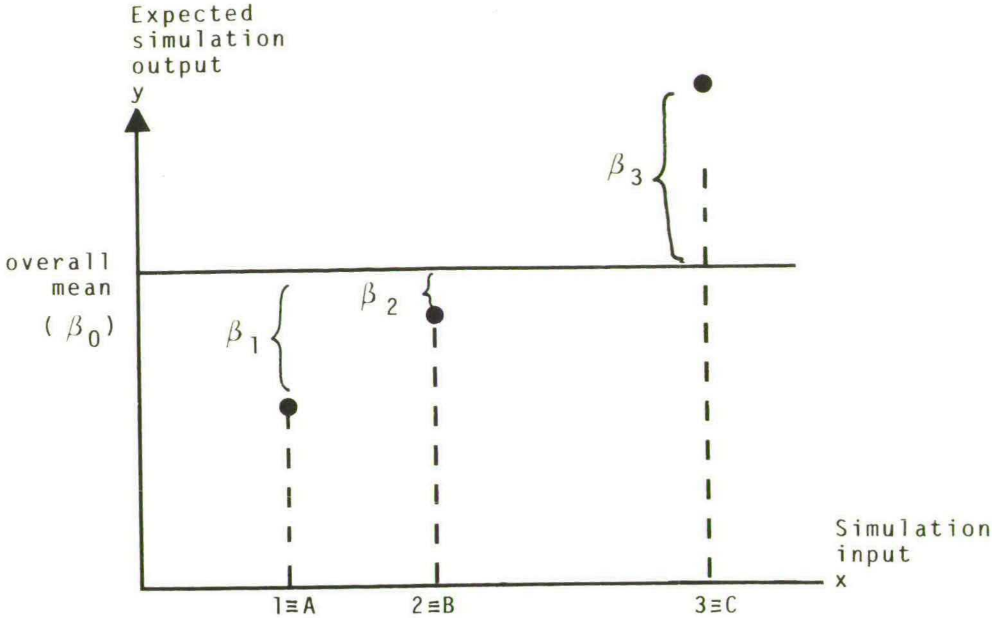
$$x^* = 2.56\ x \qquad (2)$$

(iv) Absolute scale: no transformulation is applicable. Examples are provided by the counting of the number of servers, or the number of customer arrivals. Counting results in integer values: 0,1,2,3,... Note that counting processes may be the object of statistical laws like the binomial and the Poisson distributions.

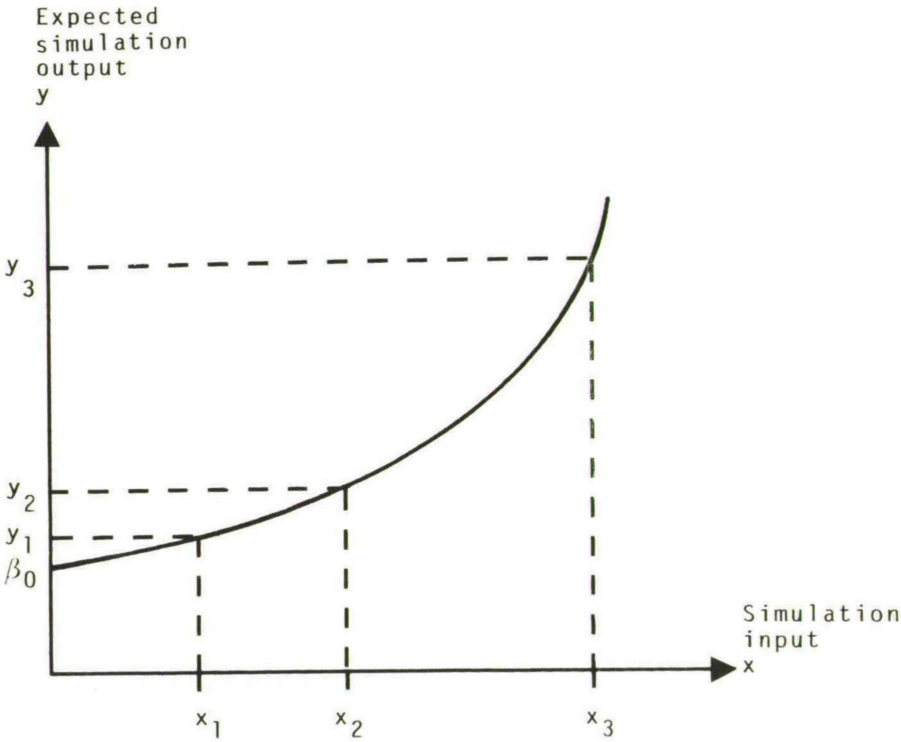
As we acquire more operational knowledge about a problem, we proceed from a nominal to an interval and next to a ratio scale. In science we have acquired a good grip on certain topics such as measuring length and monetary richness, for which we have ratio scales (even temperature we can now measure on Kelvin's ratio scale). Other topics still have an arbitrary zero: intelligence, utility, etc. In mathematical statistics we may quantify the type of distribution through the parameter value of a family of distributions. For example, the exponential is a member of the Erlang family which is a member of the Gamma family.

A qualitative variable is measured on a nominal scale, whereas a quantitative variable is measured on one of the remaining three scales (interval, ratio, absolute scale). Regression analysis handles all scales in the same way, i.e., the regression model has independent variables x and some x may be binary, representing qualitative variables, and some other x may be discrete or continuous. However, in the interpretation of the regression results we have to be more careful: If we can measure a variable on a ratio scale then no interpretation problems arise, e.g., it does not matter whether we measure length in centimeters or inches. But, if we use an interval scale and there are <u>interactions</u> among variables, then we have to be more careful. Our practical advice is: measure the variable on the scale to which the user is accustomed. For instance, if the user measures temperature in Fahrenheit then the analyst should use that scale too; the regression coefficient β then represents the effect on the response when changing temperature by one degree Fahrenheit. In Appendix 1 we discuss standardization of variables in detail.

<u>Interpolation</u> (and extrapolation) make no sense for qualitative variables. A regression model which has only qualitative variables is known as <u>Analysis of Variance</u> or ANOVA; see FIG. 1. In ANOVA we test whether a factor has <u>any</u> effect at all; the effect at "value" i (with i = 1,2,3) of the factor is denoted by β_i in the figure (we can derive that $\Sigma\beta_i = 0$). In regression analysis, we are more ambitious: we quant-



Analysis of Variance model.



Regression model

FIG.1. Analysis of Variance versus regression model.

ify <u>how much</u> the response reacts to changes in the factor; this quantification implies that we can test whether the factor has no effect: is the regression curve "flat" (zero regression parameters β except for the y-intercept β_0)? Note that if both quantitative and qualitative variables are present in the model then we speak of Analysis of Covariance (ANCOVA).

Consider an independent variable x_j which is quantitative (and assume that the regression curve is a straight line; for nonlinear models we refer to Appendix 1, eq. (1.7)). Then β_j measures the change of the expected output <u>per unit</u> change of x_j. The total change as x_j varies over its whole domain (L_j, U_j) equals the product $\beta_j\ (U_j - L_j)$; see FIG. 2. Consequently if the independent variables x_1 and x_2 have different ranges ($R = U-L$) then the total effect of x_1 may be larger than the total effect of x_2 even if $\beta_1 < \beta_2$; Also see Fiacco and Ghaemi (1982, pp. 17-19). Obviously if β is zero then the size of the range has no effect. Fortunately, significance tests can detect the unit effect β easier as the range of the (original) variable is larger: We can prove that the variance of the estimated effect decreases as the range increases, and consequently the significance test has more power. For qualitative variables β does not measure a "unit" effect.

4. USER VIEW: RISK AND SENSITIVITY ANALYSIS

The <u>user</u> of the simulation model (the manager, government agency, commanding officer) makes a different distinction among variables: environmental or exogenous variables versus controllable or instrumental variables; see Table 1 and FIG. 3.

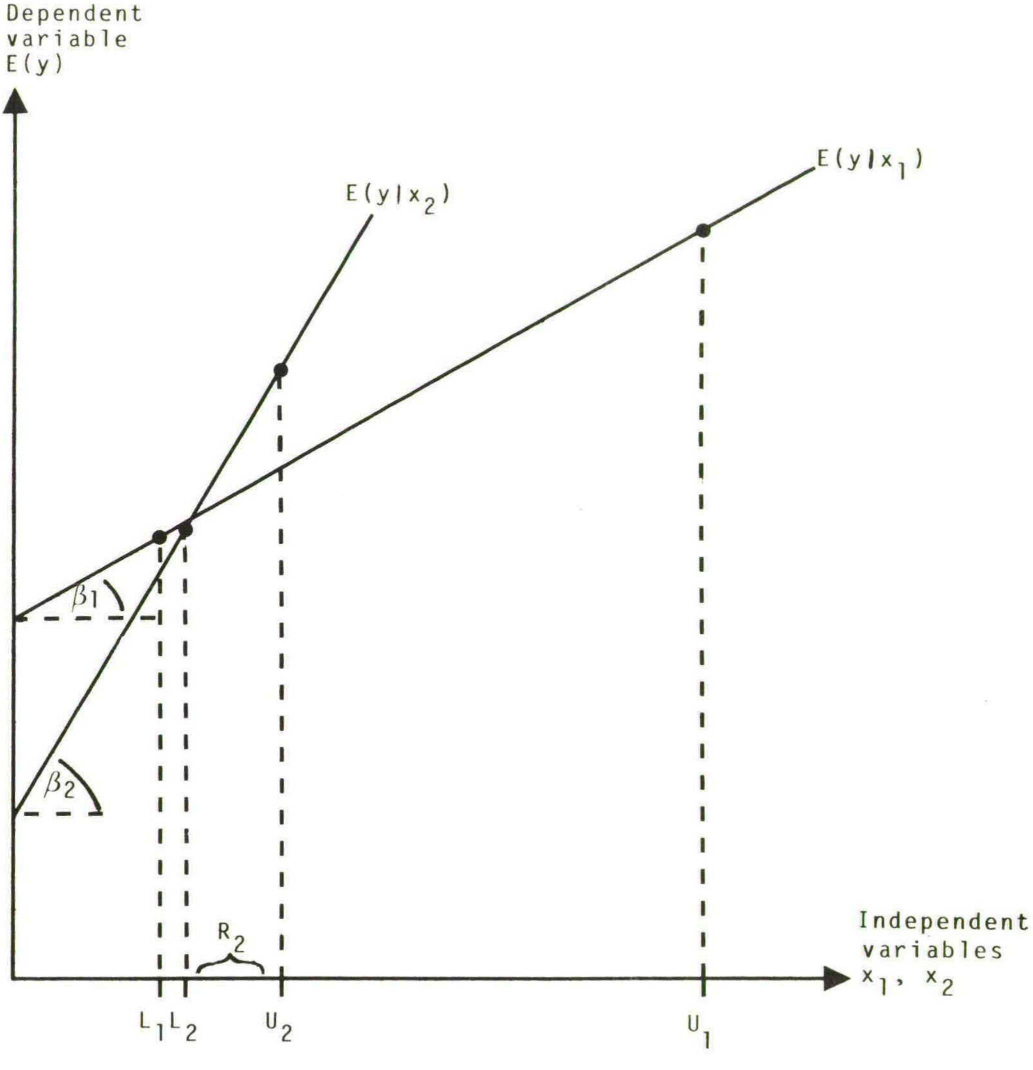


FIG.2. Effect of range R of independent variable

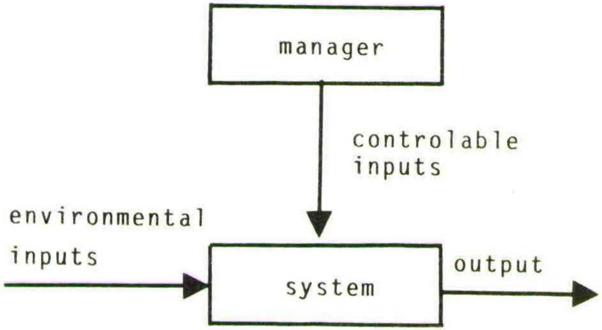


FIG.3. User's view of modeled system

Environmental variables are not under the user's (immediate) control; an example is the arrival rate of customers (we ignore long-range effects via marketing, etc.). Therefore we must try to find out how sensitive the output is to the environmental variables. If that sensitivity is high then we may try to obtain information about the exact value of this input. Sometimes accurate information can indeed be obtained, e.g., more interarrival times may be collected by going back to historical data or by collecting more recent data. Anyhow, in so far as a model represents future behavior of the system, we need information on future input. If the environment shows little variation ("placid" environment) then we may analyse historical data and ignore the uncertainty of future input. If the output, however, is sensitive to the environmental input and information about that input is not accurate then the validity of te model is questionable. If the environmental input corresponds to a qualitative variable (e.g., type of scenario) then we may formulate several model variants; and if we cannot say which variant is valid then the user has to rely on his intuition when implementing recommendations based on different model variants. If the environmental input, however, represents a quantitative variable (such as an arrival rate) then we may specify a distribution of possible values for that variable based either on (objective) historical data or on (subjective) expert opinion. Next we can estimate the probability of a specific output by (Monte Carlo) sampling from the distribution of the input; so-called Risk Analysis. Examples of Risk Analysis, including case studies, can be found in Kleijnen (1980); also see Appendix 2.

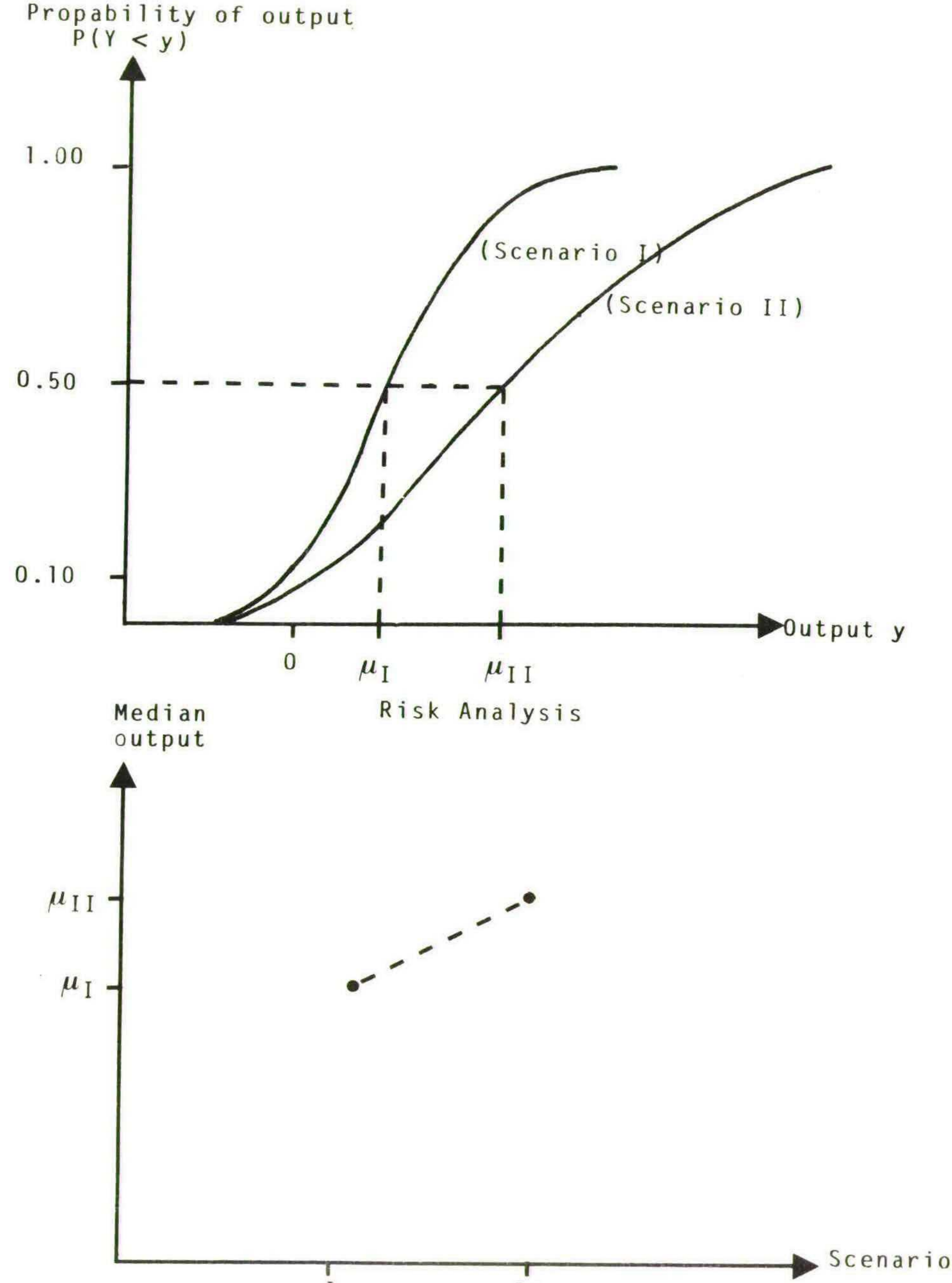


FIG.4. Risk Analysis and sensitivity Analysis.

Note that we might subject Risk Analysis to sensitivity analysis. FIG. 4 illustrates that we can repeat the Risk Analysis exercise with different qualitative environmental inputs, say, different scenarios.The Risk Analysis output may be summarized by a single measure, e.g., the median (if the distribution is symmetric then the mean μ and the median $y_{.50}$ coincide; for simplicity's sake we denote the median by μ in FIG. 4). Other measures of interest may be the 0.10 quantile or the probability of negative output (see the point 0.10 on the vertical axis and the point 0 on the horizontal axis). In Risk Analysis we sample the input values whereas in sensitivity analysis we change the input systematically (one exception is sensitivity analysis based on random experimental designs).

5. USER VIEW: OPTIMIZATION AND WHAT-IF

The remaining elements of Table 1 refer to inputs which are under the user's control:

(i) Our most ambitious goal is to find the optimal solution, and a technique for maximalization of the output is Response Surface Methodology; see Kleijnen (1985) and Myers (1971).

(ii) Sometimes the output level is prespecified, and our goal becomes to find a solution that yields this specified value: control problem. Instead of a trial-and-error approach to this control problem, we may follow a procedure based on regression modeling; see Appendix 3.

(iii) Simon (1960) emphasized that in practice users are not interested in the optimal solution and that they settle for an acceptable or "satisficing" solution. This attitude corresponds to the "what if" approach: make a change in the model (qualitative or quantitative change) and see what happens to the output (also see the Industrial or System Dynamics approach to socio-technical problems). It is ideal if our solution is robust, i.e., our advise to the user is not sensitive to the precise specification of our model.

6. SUMMARY

We distinguish output (response) and input of the system. The input of a simulation model are parameters, variables and behavioral relationships. In regression models the output is called the dependent variable; the inputs are the independent variables, some of which may be binary variables. The user's "model" distinguishes variables that are under his control or that are environmental. Environmental variables must be accurately specified to achieve model validity. Quantitative environmental variables (parameters) may be sampled: risk analysis. Controllable variables may be optimized or they may be selected such that a specific output level is realized. In the what-if approach we determine what happens to the output if we change one or more inputs.

APPENDIX 1. CODING OF VARIABLES

Consider the regression model with main effects β_j and interactions $\beta_{jj'}$:

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j . x_j + \sum_{j=1}^{k-1} \sum_{j'=j+1}^{k} \beta_{jj'} . x_j . x_{j'} + e \qquad (1.1)$$

where the x are standardized variables, i.e., each x_j ranges between minus one and plus one with an average value of zero. We can standardize the original variables, say, z_j ranging between L_j and U_j, by the following linear transformation:

$$x_j = a_j + b_j . z_j \text{ with } a_j = \frac{L_j + U_j}{L_j - U_j}$$

and

$$b_j = -2/(L_j - U_j) \qquad (1.2)$$

The eqs. (1.2) and (1.1) yield the regression model in the original variables z with regression parameters γ:

$$y = \gamma_0 + \sum_j \gamma_j . z_j + \sum_j \sum_{j'} \gamma_{jj'} . z_j . z_{j'} + e \qquad (1.3)$$

where the following relations hold among the "standardized" effects β and the "original" effects γ:

$$\gamma_0 = \beta_0 + \sum_j a_j . \beta_j + \sum_j \sum_{j'} a_j . a_{j'} . \beta_{jj'} \qquad (1.4)$$

$$\gamma_j = b_j . \beta_j + b_j . \sum_{j'} a_{j'} . \beta_{jj'} . \qquad (1.5)$$

$$\gamma_{jj'} = b_j . b_{j'} . \beta_{jj'} \qquad (1.6)$$

Consequently, if there are no interactions ($\gamma_{jj'} = \beta_{jj'} = 0$; see eq. 1.6) then zero main effects of the standardized variables ($\beta_j = 0$) imply zero main effects of the original variables; see eq. (1.5). However, if there are interactions then zero main effects β_j do <u>not</u> imply zero main effects γ_j. We can compute the marginal responses $\partial y/\partial z_j$ from the regression model for the standardized variables or from the regression model in the original variables. For instance, $\partial y/\partial z_j$ if $z_{j'} = 0$ with $j' \neq j$ is computed from eq. (1.3) as γ_j; from eqs. (1.1) and (1.2) it follows that for $z_{j'} = 0$ or $x_{j'} = a_{j'}$:

$$\frac{\partial y}{\partial z_j} = \frac{\partial y}{\partial x_j} . \frac{\partial x_j}{\partial z_j} = (\beta_j + \sum_{j'} \beta_{jj'} . a_{j'}) . b_j = \gamma_j \qquad (1.7)$$

More about coding can be found in Mendenhall (1968, pp. 221-229, 251-257) and Mihram (1972, pp. 359-360).

APPENDIX 2. RISK ANALYSIS ON REGRESSION MODELS

Consider the following simplistic model (related but more realistic models are used in, e.g., econometrics):

$$y_t = \beta_0 + \beta_1 . x_t + e_t \qquad (2.1)$$

where $e_t \sim NID(0,\sigma^2)$. Obviously, given this specification of the model, the unknown parameters are the β's and σ. These unknown parameters can be estimated through the regression analysis of, say, T historical data points. The standard errors - or more generally the covariance matrix -

of the $\hat{\beta}$'s are given by $(X'X)^{-1}\sigma^2$. It can further be proved - see, e.g., Johnston (1972, p. 26) - that $\hat{\sigma}^2$ has a χ^2- distribution with T-2 degrees of freedom and that $\hat{\sigma}^2$ is independent of the β estimators.

Suppose that we computed the estimated parameters $\hat{\beta}$ and $\hat{\sigma}^2$ for the above regression model, using (historical) data from the (sampling) period $t = 1,\ldots,T$. And next we wish to use the estimated model for "forecasting". Several alternatives seem reasonable:

Case 1(a): Predict the most likely value for the next period $t = T+1$, given the independent variable x_{T+1}. An estimator is:

$$\hat{y}_{T+1} = \hat{\beta}_0 + \hat{\beta}_1 . x_{T+1} \tag{2.2}$$

This value is also an unbiased estimator of the expected value for the period T+1:

$$E(\hat{y}_{T+1}) = E(\hat{\beta}_0) + E(\hat{\beta}_1).x_{T+1} = \beta_0 + \beta_1 . x_{T+1} = E(y_{T+1}) \tag{2.3}$$

Case 1(b): We can derive the probability of values different from the most likely or expected value as follows: The estimated analogue of eq. (2.1) is

$$\tilde{y}_{T+1} = \hat{\beta}_0 + \hat{\beta}_1 . x_{T+1} + \hat{e}_{T+1} \tag{2.4}$$

where $\hat{e}_{T+1} \sim N(0,\hat{\sigma}^2)$. Since $\tilde{y}_{T+1}$ is a linear combination of the normally distributed variables $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{e}_{T+1}$ we know that $\tilde{y}_{T+1}$ is normally distributed. Its mean was given by eq. (2.3); its variance is

$$\mathrm{var}(\tilde{y}_{T+1}) = \mathrm{var}(\hat{\beta}_0) + (x_{T+1})^2.\mathrm{var}(\hat{\beta}_1) + $$
$$+ 2x_{T+1}.\mathrm{cov}(\hat{\beta}_0,\hat{\beta}_1) + \mathrm{var}(\hat{e}_{T+1}) \qquad (2.5)$$

Note that $\mathrm{cov}(\hat{e}_{T+1},\hat{\beta}) = 0$ because $\hat{\beta}$ depends on $(y_1,\ldots,y_T)$ and not on y_{T+1}, and the error terms are serially independent. In the first paragraph of this appendix we referred to estimators for the terms in eq. (2.5). Finally, using the table for the standard normal variable N(0,1), we can compute the probability of values $\tilde{y}_{T+1}$ different from the most likely value $\hat{y}_{T+1}$.

Case 2(a): Predict several periods ahead, e.g., predict the response for T+1 and T+2. Now the example of eq. (2.1) is too simple to illustrate what is at stake. Therefore we introduce a slightly more complicated example, including a lagged dependent variable: eq. (2.1) is replaced by

$$y_t = \beta_0 + \beta_1.x_t + \beta_2.y_{t-1} + e_t \qquad (2.6)$$

Consequently the most likely value or the expected value for period T+1 is no longer estimated by eq. (2.2) but by the unbiased estimator

$$\hat{y}_{T+1} = \hat{\beta}_0 + \hat{\beta}_1.x_{T+1} + \hat{\beta}_2.y_T \qquad (2.7)$$

where y_T is an observed (sample) value. However, when we extrapolate for more than one period ahead, we obtain:

$$\hat{y}_{T+2} = \hat{\beta}_0 + \hat{\beta}_1.x_{T+2} + \hat{\beta}_2.\hat{y}_{T+1} \qquad (2.8)$$

where the last term uses the estimator given by eq. (2.7). Unfortunately eq. (2.8) is biased: for the last term of eq. (2.8) the following inequality holds because the random variable $\hat{y}_{T+1}$ depends on the random variable $\hat{\beta}_2$:

$$E(\hat{\beta}_2 . \hat{y}_{T+1}) \neq E(\hat{\beta}_2).E(\hat{y}_{T+1}) = \beta_2 . E(y_{T+1}) \qquad (2.9)$$

Therefore we may resort to simulation: for $t > T$ we sample $\hat{e}_t$ from $(0, \hat{\sigma}^2)$; this $\hat{e}_t$ yields $\tilde{y}_t$ (see eq. 2.4); etc.

Case 2(b): It is straightforward to compute the probability of values different from the most likely value or the expected value in period T+1, but it is complicated to compute this probability for period T+2:

$$\tilde{y}_{T+1} = \hat{\beta}_0 + \hat{\beta}_1 . x_{T+1} + \hat{\beta}_2 . y_T + \hat{e}_{T+1} \qquad (2.10)$$

and

$$\tilde{y}_{T+2} = \hat{\beta}_0 + \hat{\beta}_1 . x_{T+2} + \hat{\beta}_2 . \tilde{y}_{T+1} + \hat{e}_{T+2} \qquad (2.11)$$

Again simulation provides the answer.

We emphasize that some values of y_{T+2} are "impossible", given that y_{T+1} has a particular value (under the normality assumption theoretically all values are possible; however the probability of "extreme" values is virtually zero). So some time paths are virtually impossible.

Summarizing, in cases 1(a) and 2(a) we estimated the expected value one period ahead and two periods ahead respectively. If we forecast several periods ahead, we may use simulation. In cases 1(b) and

2(b) we were interested in the probability of deviations from these values. The latter probabilistic element entered through the random noise e, estimated by $\hat{e}$.

A different chance element enters our analysis, if we realize that the model itself may be incorrect! More specifically, even if we assume that we specified the correct <u>form</u> (i.e., a linear model) then we may still use the wrong <u>parameter</u> values: the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are not precisely equal to β_0 and β_1, and $\hat{\sigma}^2$ is not exactly equal to σ^2. Therefore we may apply <u>risk analysis</u>: we can sample $\hat{\beta}_0$ and $\hat{\beta}_1$ in the eqs. (2.2) through (2.11) from a bivariate normal distribution (with mean and covariance matrix given by the standard regression analysis of the historical time series) and we can sample $\hat{\sigma}^2$ from the χ^2_{T-2} distribution; see the first paragraph of this appendix.

We can combine each of the (say n_1) sampled triplets $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$ of the risk analysis with each of the (say n_2) simulated time paths obtained by sampling the disturbances $(\hat{e}_{T+1}, \hat{e}_{T+2})$, and this combination results in an $n_1 \times n_2$ table. From the n_2 observations per row we might estimate the conditional probabilities $P(y_{T+2}\ \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$. However, the purpose of the risk analysis is to estimate the unconditional probabilities $P(y_{T+2})$ which incorporates noise around the expected value plus noise in the estimation of the model's parameters. From these unconditional probabilities we can compute the mean and median response, the probability of negative responses, etc.

Note that it would be incorrect to use the following risk analysis procedure (which at first sight might look reasonable):

(i) Sample $\hat{\beta}_0$ and $\hat{\beta}_1$.

(ii) Comupute the corresponding historical values

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 . x_t \qquad (t = 1,...,T) \qquad (2.12)$$

(iii) Compute the corresponding historical residuals

$$u_t = y_t - \hat{y}_t \qquad (t = 1,...,T) \qquad (2.13)$$

These residuals no longer satisfy the Least Squares properties such as $\Sigma\, u_t = 0$.

(iv) Compute the corresponding $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \Sigma\, u_t/(T-2) \qquad (2.14)$$

This estimator is no longer an optimal estimator of σ^2; see the comment on step (iii).

APPENDIX 3. INVERSE REGRESSION IN CONTROL PROBLEMS

In the control problem we have a target value for the response, and we wish to estimate the values for the instrumental variables, given specific values for the environmental variables. The simplest solution is to proceed "as usual", i.e., estimate E(y) as a function of the k independent variables ($k = k_1 + k_2$ where k_1 denotes the number of instrumental variables and k_2 is the number of environmental variables; the latter variables are shown in parentheses):

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 . x_1 + \ldots + \hat{\beta}_{k_1} . x_{k_1} (+ \ldots + \hat{\beta}_k . x_k) \quad (3.1)$$

The simplest situation arises if we have a single controllable variable ($k_1 = 1$). Then we can estimate the required value of x_1, say x_1^*, from eq. (3.2) where y_G denotes the goal value:

$$x_1^* = \frac{1}{\hat{\beta}_1} . \{y_G - \hat{\beta}_0 - (\hat{\beta}_2 . x_2 + \ldots + \hat{\beta}_k . x_k)\} \quad (3.2)$$

However, there is an alternative estimator: We can regress x_1 on y (and on the prespecified values of the k_2 environmental variables where $k_2 \geqq 0$):

$$x_1 = \gamma_0 + \gamma_1 . y (+ \gamma_2 . x_2 + \ldots + \gamma_{k-1} . x_{k-1}) + e \quad (3.3)$$

A second estimator, say x_1^{**}, results if we use the estimators $\hat{\gamma}$ and substitute y_G for y. Which estimator is best, is not clear, even in the simplest situations ($k_2 = 0$; $k_2 = 1$; Classical Assumptions for e); see Turiel et al. (1982) for a recent survey of the various statistical problems of "inverse" regression.

In simulation the situation is more complicated. The response is probably sensitive to several environmental variables ($k_2 > 1$) and there are several controllable variables ($k_1 > 1$). If $k_1 > 1$ then the second estimator x^{**} results in a system of simultaneous equations (see the third term in the following equations):

$$x_1 = \gamma_0 + \gamma_1 \cdot y + \gamma_2 \cdot x_2 (+ \ldots + \gamma_{k-1} \cdot x_{k-1}) + e_1 \qquad (3.4)$$

$$x_2 = \delta_0 + \delta_1 \cdot y + \delta_2 \cdot x_1 (+ \ldots + \delta_{k-1} \cdot x_{k-1}) + e_2 \qquad (3.5)$$

The statistical problems of simultaneous regression equations are discussed in econometric handbooks. Fortunately, we can take advantage of the peculiarities of simulation, i.e., after we have used the regression metamodel to estimate the required value (or values) of the instrumental variable (or variables), we can check this solution by performing a simulation run with the indicated values for the instrumental variables and verifying whether the resulting output does not deviate significantly from the target value.

One more approach we would suggest is to transform the control problem into an optimization problem, i.e., select the control variables such that the deviation between the target value and the realized value of the dependent variable y is minimial. Such minimization problems can be approached through Response Surface Methodology.

REFERENCES

Becker, H.A. (1981). The Role of Gaming and Simulation in Scenario Projects. International Institute of Applied Systems Analysis (IIASA), Laxenburg (Austria).

Fiacco, A.V. and A. Ghaemi (1982). Sensitivity analysis of a nonlinear water pollution control model using an Upper Hudson River data base. Operations Research. 30, no. 1: 1-28.

Hauser, J.R. and S.M. Shugan (1980). Intensity measures of consumer preference. Operations Research. 28, no. 2: 278-320.

Hellerman, H. and T.F. Conroy (1975). Computer System Performance. McGraw-Hill Book Company, New York.

Kleijnen, J.P.C. (1980). Computers and Profits; Quantifying Financial Benefits of Information. Addison-Wesley Publishing Company, Reading (Massachusetts).

Kleijnen, J.P.C. (1981). On hierarchical modeling. Communications ACM, 24: 774-775.

Kleijnen, J.P.C. (1985). Statistical tools for simulation practitioners, Marcel Dekker, Inc., New York

Mendenhall, W. (1968). Introduction to Linear Models and the Design and Analysis of Experiments. Wadsworth Publishing Company, Inc., Belmont (California).

Mihram, G.A. (1972). Simulation: Statistical Foundations and Methodology. Academic Press, New York.

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

Simon, H.A. (1960). The New Science of Management Decision, Harper & Row, New York.

Sprent, P. (1981). Quick Statistics; an Introduction to Non-Parametric Methods. Penguin Books, Ltd., Harmondsworth (England).

Turiel, T.P., G.J. Hahn and W.T. Tucker (1982). New simulation results for the calibration and inverse median estimation problems. Communications in Statistics, Simulation and Computation. 11, no. 6: 677-713.

IN 1982 REEDS VERSCHENEN

107 A.J. de Zeeuw
Hierarchical decentralized optimal control in econometric policy models.

108 A. Kapteyn en T. Wansbeek
Identification in Factor Analysis.

109 G. van der Laan en A.J.J. Talman
Simplical Algorithms for finding Stationary Points, a unifying description.

110 P. Boot
Ekonomiese betrekkingen tussen Oost en West Europa.

111 B.B. van der Genugten
The asymptotic behaviour of the estimated generalized least squares method in the linear regression model.

112 J. P.C. Kleijnen en A.J. van Reeken
Principle of computer charging in a university-like organization.

113 H. Tigelaar
The informative sample size for dynamic multiple equation systems with moving average errors.

114 Drs. H.G. van Gemert, Dr. R.J. de Groof en Ir. A.J. Markink
Sektorstruktuur en Economische Ontwikkeling.

115 P.H.M. Ruys
The tripolar model: a unifying approach to change.

116 A.J. de Zeeuw
Policy solutions for Mini-Interplay, a Linked Model for two Common Market Countries.

117 F.J.M. van Doorne en P.H.M. Ruys
Die Struktur einer Sprechhandlung in Habermas' Forschungsprogramm. Formale Analyse mit den Mitteln des tripolaren Modells.

118 A.J.J. Talman en G. van der Laan
Simplicial Approximation of Solutions to the Non-Linear Complementarity Problem.

119 P. Boot
The East German Planning System Reconsidered.

120 J.P.C. Kleijnen
Regression Metamodel Summarization of Model Behaviour.

121 P.G.H. Mulder en A.L. Hempenius
Evaluating life time in a competing risks model for a chronic desease.

IN 1982 REEDS VERSCHENEN (vervolg)

122 A.J.J. Talman en G. van der Laan
From Fixed Point to Equilibrium.

123 J.P.C. Kleijnen
Design of simulation experiments.

124 H.L. Theuns en A.M.L. Passier-Grootjans
Internationaal toerisme; een gids in de algemene basisliteratuur en het bronnenmateriaal.

125 J.H.F. Schilderinck
Interregional Structure of the EUROPEAN community.
Part I: Imports and Exports, Sub-divided by Countries Aggregated According the Branches of the European Community Interregional Input-Outputtables 1959, 1965, 1970 and 1975.

IN 1983 REEDS VERSCHENEN:

126 H.H. Tigelaar
Identification of noisy linear systems with multiple arma inputs.

127 J.P.C. Kleijnen
Statistical Analysis of Steady-State Simulations: Survey of Recent Progress.

128 A.J. de Zeeuw
Two notes on Nash and Information.

129 H.L. Theuns en A.M.L. Passier-Grootjans
Toeristische ontwikkeling - voorwaarden en systematiek; een selectief literatuuroverzicht.

130 J. Plasmans en V. Somers
A Maximum Likelihood Estimation Method of a Three Market Disequilibrium Model.

131 R. van Montfort, R. Schippers, R. Heuts
Johnson S_U-transformations for parameter estimation in arma-models when data are non-gaussian.

132 J. Glombowski en M. Krüger
On the Rôle of Distribution in Different Theories of Cyclical Growth.

133 J.W.A. Vingerhoets en H.J.A. Coppens
Internationale Grondstoffenovereenkomsten.
Effecten, kosten en oligopolisten.

134 W.J. Oomens
The economic interpretation of the advertising effect of Lydia Pinkham.

135 J.P.C. Kleijnen
Regression analysis: assumptions, alternatives, applications.