Regression analysis

Kleijnen, J.P.C.

Publication date:
1983

Citation for published version (APA):
Kleijnen, J. P. C. (1983). *Regression analysis: Assumptions, alternatives, applications*. (Research memorandum / Tilburg University, Department of Economics; Vol. FEW 135). Unknown Publisher.

faculteit der economische wetenschappen

RESEARCH MEMORANDUM

[redacted]

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

[redacted]

REGRESSION ANALYSIS:
ASSUMPTIONS, ALTERNATIVES, APPLICATIONS

Jack P.C. Kleijnen

Department of Business and Economics
Katholieke Hogeschool Tilburg (Tilburg University)
Post Box 90153
5000 LE Tilburg
Netherlands

December 1983

REGRESSION ANALYSIS:

ASSUMPTIONS, ALTERNATIVES, APPLICATIONS

Jack P.C. Kleijnen

Department of Business and Economics
Katholieke Hogeschool Tilburg (Tilburg University)
Post Box 90153
5000 LE Tilburg
Netherlands

December 1983

ABSTRACT:

Are the assumptions of regression analysis realistic; how can they be verified; if an assumption is violated, are there alternative regression techniques? Recent developments are surveyed, emphasizing practical aspects and using only elementary statistical formulas. The specific assumptions are: (i) a non-singular matrix of independent variables (ii) a regression model linear in its parameters (iii) responses with constant variances (iv) independent responses (v) normally distributed responses (vi) a valid or correctly specified regression model. More than fifty selected references to the recent literature are included.

CR CATEGORIES AND SUBJECT DESCRIPTORS: G3 [Probability and Statistics] - statistical software;

I.6 [Simulation and Modelling] - model validation and analysis

GENERAL TERMS: Experimentation, Measurement

ADDITIONAL KEY WORDS AND PHRASES: regression, least squares, transformations, variance heterogeneity, common random numbers, normality, outliers, robustness, nonparametric regression, rank regression, validation, optimization, applications.

1. INTRODUCTION

Regression analysis is a statistical technique frequently used in many practical applications and scientific disciplines. In this paper we shall examine six assumptions of classical regression analysis, i.e., we shall try to answer questions such as: Is it reasonable to use this particular assumption; how can we test whether this assumption holds in a specific situation; are alternative assumptions accommodated by other regression techniques? When answering these questions, we shall refer to recent developments in statistics. Because our survey is meant for practitioners, we shall use only elementary statistical formulas.

Because regression analysis is applied in so many different fields, it would be impractical to cover all applications. We shall concentrate on a special type of application with which we are familiar, namely the use of a regression model to summarize the reaction of the output of a simulation program to changes in the input. (Such a regression model facilitates sensitivity analysis, validation and optimization of the simulation model; see [26, 29]. However, we emphasize that most of the material in our survey is also relevant to applications outside the simulation field.

2. BASIC REGRESSION ANALYSIS

In the present section we shall present the basic ideas and formulas of regression analysis. This section may serve as a refresher for the reader; if the reader is familiar with regression analysis he may immediately proceed to the conclusion of the present section.

There are n observations or (simulation) runs with $n \geq 1$. There are q independent (regression) variables x, including the dummy variable x_0 equal to one: $x_{i0} = 1$ for $i = 1,...,n$ with $1 \leq q \leq n$. Moreover, the independent variable x may be a binary variable, for instance, $x_{i1} = 1$ if in run i the qualitative factor "queuing discipline" equals first-come-first-served, and $x_{i1} = 0$ if the queuing discipline equals shortest-jobs-first. Each run yields one observation on the output: the regression model's dependent variable (if there are multiple outputs we apply the analysis per dependent variable, applying the Bonferroni inequality; see [34] and Section 8.2). Least Squares is a mathematical and not a statistical problem formulation: Given the n observations ($y = y_i$ when $x = x_i$ where $i = 1,...,n$) and given a family of curves with parameters β (e.g., the family of linear curves $\beta_0 + \beta_1 x$) we wish to determine the parameter values $\hat{\beta}$ that minimize the sum of squared deviations ($\sum_1^n (y_i - \hat{y}_i)^2$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$). If the curve is linear in the parameters β (see later) then the Least Squares values $\hat{\beta}$ can be found in any textbook on regression analysis (see [5, 11]):

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

where we follow the matrix notation traditional in regression analysis. The following expression in scalar notation results for the case of a single independent variable x_1 (besides the dummy variable x_0, that is, $q = 2$):

$$\hat{\beta}_1 = \frac{\Sigma(x_i-\bar{x}).(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1.\bar{x} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the familiar averages $\Sigma x_i/n$ and $\Sigma y_i/n$ respectively.

The estimation of β becomes a statistical problem if we assume that given the independent variables x, there is a population of possible response values y. The simplest statistical model specifies that the random variable y is normally distributed, with expected value equal to $E(y\ X) = X\beta$. The simplest statistical model further specifies a constant variance: $\mathrm{var}(y|X) = \sigma^2$. Moreover, the n responses are assumed to be independent, so that the covariance matrix of y is given by $\Omega_y = \sigma^2 I$, where the symbol I denotes the identy matrix. In other words, the errors (noise, disturbance) e defined by

$$y = X.\beta + e \qquad (3)$$

satisfy the Classical Assumptions, i.e., the errors are normally and independently distributed (NID) with zero mean and constant variance σ^2: $e \sim \mathrm{NID}(0,\ \sigma^2.I)$.

If the errors have zero expectation, then we can prove that the (mathematical) Least Squares algorithm leads to estimators of the regression parameters β that are unbiased (a statistical property): $E(\hat{\beta}) = \beta$ where $\hat{\beta}$ was given by eq. (1); in this equation y is now a random variable. We further observe that $\hat{\beta}$ is a linear estimator, i.e., it is a linear transformation of the responses y. If we further assume that the errors are independent with common variance ($\Omega_e = \sigma^2 I$) then we can prove that the Least Squares estimator $\hat{\beta}$ is the unbiased linear estimator with the smallest variance: Best Linear Unbiased Estimator (BLUE) where "best" means minimum variance. The values of these (minimal) variances can be derived from the following general formula: if a random vector y_2 is a linear transformation of y_1, i.e., $y_2 = Ay_1$, and y_1 has covariance matrix Ω_1, then y_2 has a covariance matrix Ω_2 given by

$$\Omega_2 = A.\Omega_1.A' \tag{4}$$

Applying eq. (4) to eq. (1) using $\Omega_y = \sigma^2 I$ yields the covariance matrix of $\hat{\beta}$:

$$\Omega_{\hat{\beta}} = (X'.X)^{-1}.\sigma^2 \tag{5}$$

In textbooks on regression analysis σ^2 is estimated through the Mean Squared Residuals (MSR):

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

where q denotes the number of estimated regression parameters and $\hat{y}_i$ is the i-th component of the predicted observations $\hat{y} = X\hat{\beta}$. In random simulation we have estimators of σ^2 different from eq. (6); see Section 5. The main-diagonal elements of $\hat{\Omega}_{\hat{\beta}}$ defined by eqs. (5) and (6) are the estimated variances of the regression parameter estimators, and their square roots s_j are the estimated standard deviations or "standard errors". To test whether β_j equals zero - or more generally equals the value β_j^0 - we use the t statistic:

$$t_{v,j} = (\hat{\beta}_j - \beta_j^0)/s_j \qquad (j = 0,1,\ldots,q-1) \tag{7}$$

where the degrees of freedom of t equal the degrees of freedom of $\hat{\sigma}^2$, i.e., if we use eq. (6) then $v = n-q$. If we accept the hypothesized value β_j^0,

then we may re-estimate the remaining parameters $\beta_{j'}$ $(j' \neq j)$. The resulting values will not differ from the old values, if the independent variables j and j' are orthogonal, i.e., if $\sum_i x_{ij} \; x_{ij'} = 0$ where $j,j' = 0,1,\ldots,q-1$

We may also hypothesize that more than one parameter is zero. For example, we may hypothesize that the input x has no effect, i.e., if the original "full" model was a second-degree polynomial $\beta_0 + \beta_1 x +$ $+ \beta_2 x^2$, then our null-hypothesis (H_0) becomes: $\beta_1 = \beta_2 = 0$. We can test such a (composite) hypothesis using the ANOVA F statistic. Briefly, this procedure runs as follows. The observations on the dependent variable y_i yield the "total sum of squares": $SS_{TOTAL} = \sum_1^n (y_i - \bar{y}_i)^2$ with degrees of freedom equal to n-1 (minus one because of the restriction $\bar{y}_i = \Sigma y_i/n$). We can split this fixed total into two components, namely the variation explained by the regression model and the unexplained, residual portion: $SS_{total} = SS_{explained} + SS_{error}$ where the latter term corresponds to the numerator of eq. (6), and the former term can be easily computed by subtracting the error term from the total sum of squares. When we wish to test a hypothesis like the one above, we can compute two different values for the sum of squared errors, namely one value for the "full" model - i.e., the model without the restriction specified by H_0 - and one value for the "reduced" model, i.e., the model including that restriction. (Obviously the restricted model cannot yield a smaller sum of squared errors.) Intuitively, when the reduced model results in a drastic increase in the sum of squared residuals we reject the null-hypothesis. More precisely, let the upper-indices F and R correspond to the full and restricted model respectively; let p denote the number of effects hypothesized to be zero (in the above H_0 we have p = 2); then compute

$$F_{p,n-q} = \frac{\{SS^{(R)}_{error} - SS^{(F)}_{error}\}/p}{SS^{F}_{error}/(n-q)} \qquad (8)$$

where the denominator is an independent and unbiased estimator of σ^2; see eq. (6).

Summary: The Least Squares estimator of eq. (1) is BLUE and can be easily tested, if the following assumptions hold: (1) The X matrix is non-singular. (2) The regression model is linear in its parameters β. (3) The y have a constant variance σ^2. (4) The y are independent. (5) The y are normally distributed. (6) The regression model is valid. We shall discuss these assumptions in separate sections.

3. NON-SINGULAR MATRIX X

In the social sciences the analyst cannot fix the independent variables x. He can only observe those variables; their values are fixed by the environment. Consequently X, the n×q (with $n \geq q$) matrix of independent variables, may be singular or nearly-singular, i.e., the inverse $(X'.X)^{-1}$ may not exist or this inverse may have very bad numerical qualities, i.e., minor changes in one or more elements of X may result in completely different values for the corresponding inverse. Statistically, a near-singular or ill-conditioned matrix X means highly correlated independent variables and, hence, large standard errors for the parameter estimators $\hat{\beta}$. This problem is also known under the name multicollinearity (one or more columns of X can be expressed as a linear combination of the remaining columns). Note that the sum of residuals may differ from zero if X is ill-conditioned.

We note in passing, that the lack of experimental control in the social sciences also implies that replication of specific experimental conditions - specified by the row vector $X_i = (1,x_{i1},\ldots,x_{ij},\ldots,x_{i,q-1})$ - is virtually impossible. Therefore the experimental error variance σ_i^2 is estimated from the residuals $y-\hat{y}$ assuming a common variance $\sigma_i^2 = \sigma^2$; see eq. (6).

In the social sciences the observed values of the independent variables x are often modeled as observations on random variables, i.e., the dependent variable y and the independent variables x have a joint distribution function. Consequently the independent variables may show strong correlation and their observed values may result in a (nearly) singular X matrix. In the regression analysis of such data the results are usually presented conditionally on the observed values X, i.e., the independent variables are treated as deterministic variables; see [48].

In the "hard" sciences, e.g., computer science, the experimental conditions can be better controlled. In the 1930's statistical theory was developed for experimentation in agriculture. In subsequent decades the theory of experiments was applied to other areas, e.g., chemical experiments. More recently the technique of simulation has been applied in both hard and soft sciences. In simulation experiments the theory of experimental design can certainly be applied because all factors are controllable; also see [39]. In the design of the simulation experiment we purposefully fix the values of the independent variables. Consequently X is not singular, in general (often X will be orthogonal). However, by accident X may turn out to be (nearly) singular, e.g., after we have designed and run the simulation experiment we may decide to use new independent variables in the analysis; these new variables were not controlled and they may create (near) singularity. We may either add some new runs to the old design or we may analyse the old design applying special analysis techniques. One of these special techniques is ridge regression which we shall briefly discuss next.

In ridge regression the estimators of the parameters β are no longer unbiased: however, this bias may be outweighted by a decrease in variance attained through a proper choice of the ridge algorithm parameter, say r (for more general definitions see the literature below):

$$\hat{\beta}_r = (X'X + r.I)^{-1}.X'.y \tag{8}$$

Unfortunately the optimal r, which minimizes the Mean Squared Error of the estimators $\hat{\beta}_r$, depends on the unknown true parameters β. There are several methods for the estimation of the optimal ridge parameter r. Confidence intervals for the ridge regression estimators were discussed by Obenchain [38], who proposed to use the classical confidence intervals, which are centered around the OLS point estimators of the regression parameters. More than two hundred publications on ridge regression were presented by Hoerl and Kennard [20] in an annotated bibliography. We refer to these references for more details on ridge regression and other techniques.

Summary: In the social sciences singularity of X may be a problem. In the hard sciences and in simulation, we can always specify a "good" matrix X. However, the ad hoc introduction of new independent variables may lead to (near) singularity. A "bad" matrix X may then be handled through ridge regression.

4. LINEAR MODEL

The linearity assumption does not mean that the regression model is necessarily linear in the independent variables. For instance, the regression model may be a second degree polynomial in x. Another example is:

$$y = \beta_0 + \beta_1 . \log z + e \tag{9}$$

so that in the notation of Section 2 we have $x_{i1} = \log z_i$. The last equation is equivalent to

$$y^* = \beta_0^* . z^{\beta_1} . e^* \tag{10}$$

where $y = \log y^*$, $\beta_0 = \log \beta_0^*$ and $e = \log e^*$. In the linear regression analysis of eq. (9) we assume that the (additive) noise e is normally distributed, or equivalently that e^* in eq. (10) is lognormally distributed. The lognormal distribution has the following properties; see [1]:

$$E(e^*) = \exp\{E(e) + \mathrm{var}(e)/2\} = \exp(\sigma^2/2) \tag{11}$$

and

$$\mathrm{var}(e^*) = \{E(e^*)\}^2 . \{\exp[\mathrm{var}(e)]-1\} = \exp(\sigma^2) . \{\exp(\sigma^2)-1\} \tag{12}$$

We shall return to transformations later on.

In general, a model not linear in its parameters can sometimes be transformed into a model which is linear in its parameters. However, if we cannot find such a transformation then we have to apply nonlinear regression analysis. In our experience linear regression analysis is flexible enough for the summarization of simulation models. Nonlinear regression is applied to, e.g., data from chemical experiments where enough theoretical knowledge is available to suggest a specific family of nonlinear models; see [9, 32].

5. CONSTANT VARIANCES

The Classical Assumptions imply that y and e have the same variance namely σ^2 . The assumption of a "homogeneous" variance is unrealistic in general. If the random variable y has an expected value that depends on x, then it seems logical to assume that y has a variance that also varies with x, i.e., we introduce $var(y_i) = \sigma_i^2$ with $i = 1,\ldots,n$. Moreover, in random simulation we obtain not only the point estimator y_i but also the standard error of y_i, denoted by $\hat{\sigma}_i$; in [27] we surveyed different techniques for the estimation of σ_i^2 in simulation; it is our experience that the variance estimates $\hat{\sigma}_i^2$ differ greatly, say, by a factor 100 and more. So for logical and empirical reasons the assumption of a constant variance seems unrealistic, certainly in random simulation.

Before we proceed we add some notes:

(i) Sometimes we can transform the original output y such that the transformed output y^* has an approximately constant variance. We emphasize that the interpretation of the data should be in terms of the original observations.

(ii) If we assumed a common variance σ^2 then we could pool the n estimators $\hat{\sigma}_i^2$ (each with degrees of freedom equal to, say, v_i) in order to obtain a more accurate estimator, with degrees of freedom equal to Σv_i whereas the Mean Squared Residual estimator of eq. (6) has degrees of freedom equal to n-q.

(iii) For a discussion of the assumption of a constant variance σ^2 in the regression modelling of deterministic simulation, we refer to [28] and note 1.

(iv) There are a number of tests for comparing n variances; see [15] In simulation the variance estimators $\hat{\sigma}_i^2$ differ so much that a formal statistical test is superfluous.

What are the alternatives if we conclude that the assumption of a constant variance σ^2 does not hold? Intuitively, if a response has a high standard error, that response should receive less weight when fitting a curve. Formally, if the variances σ_i^2 were known then the transformation $y_i^* = y_i/\sigma_i$ would result in constant variances: var (y_i^*) $= \text{var}(y_i)/\sigma_i^2 = 1$; next we could apply (Ordinary) Least Squares (OLS) to the transformed output y. This approach would result in the Weighted Least Squares (WLS) estimator

$$\tilde{\beta} = (X'.\Omega_y^{-1}.X)^{-1}.X'.\Omega_y^{-1}.y \qquad (13)$$

with covariance matrix

$$\Omega_{\tilde{\beta}} = (X'.\Omega_y^{-1}.X)^{-1} \tag{14}$$

WLS would yield the BLUE (Best Linear Unbiased Estimator); the WLS algorithm minimizes $\Sigma w_i(y_i-\hat{y}_i)^2$ with weights $w_i = 1/\sigma_i^2$.

In practice we do not know Ω_y. Therefore one possibility is to replace Ω_y by an estimator $\hat{\Omega}_y$. In simulation the estimator $\hat{\Omega}_y$ equals a diagonal matrix with elements on the main diagonal equal to $\hat{\sigma}_i^2$; we discussed the variance estimator $\hat{\sigma}^2$ at length in [27]. For example, in case of m replicated runs we have $\hat{\sigma}_i^2 = \hat{var}(\bar{y}_i) = \hat{var}(y_i)/m$.
Other estimators used outside simulation assume that the regression model is correct (valid) and are based on the residuals $\hat{e}$; see [22].

If we replace Ω_y in eq. (13) by its (unbiased) estimator $\hat{\Omega}_y$, then <u>Estimated Weighted Least Squares</u> (EWLS) result. Schmidt [40] proved that - under mild technical assumptions - EWLS yield unbiased estimators of β with an asymptotic covariance matrix following from eq. (14). Unfortunately, we cannot derive the small-sample behavior of EWLS analytically (EWLS yield a non-linear estimator because in eq. (13) both y and Ω_y become random). In a Monte Carlo study [29] we found: (i) The EWLS estimators of β are unbiased. (ii) The asymptotic covariance formula - see eq. (14) - also holds in small samples, provided at least five independent replicates are used to estimate σ_i^2. (iii) The EWLS estimators have smaller standard errors than the Ordinary Least Squares estimators $\hat{\beta}$ have (provided the actual variances σ_i^2 do differ, and their estimators $\hat{\sigma}_i^2$ are based on more than two observations). Unfortunately another Monte Carlo experiment [37] showed that with fewer than ten replicates the "coverage" is too small, i.e., the confidence interval misses the true β value more often than the nominal α fraction specifies. More Monte Carlo experimentation seems necessary.

One alternative to EWLS is: use OLS to obtain the unbiased point estimators $\hat{\beta}$ but base the standard errors upon the correct formula. In other words, we continue to use eq. (1) but we replace eqs. (5) and (6) by the unbiased estimator of the covariance matrix $\Omega_{\hat{\beta}}$ obtained by applying eq. (4):

$$\Omega_{\hat{\beta}} = W.\Omega_{y}.W' \text{ with } W \equiv (X'.X)^{-1}.X' \qquad (15)$$

A final alternative simply ignores the heterogeneity of variance. In general, however, the correct covariance formula - eq. (15) - differs from the classical formula, eq. (5). How much effect this difference has on the confidence intervals and tests was investigated by several authors. They tend to reject reliance on the insensitivity of the classical regression analysis (including ANOVA) to heterogeneity of variance; see [8, 37].

If the number of observations (n) equals the number of regression parameters (q) then (Estimated) Weighted Least Squares and Ordinary Least Squares become identical.

One practical advice is: apply several statistical techniques to the same data and see if they result in similar conclusions. If the conclusions are similar, then we are lucky. Otherwise, we may turn to a professional statistician for expert advice.

Summary: In general the assumption of constant variances is unrealistic. Weighted Least Squares with estimated variances $\hat{\sigma}_i^2$ yield more accurate estimators of β. Some Monte Carlo studies on the resulting confidence intervals and tests suggest that a better alternative might be the Ordinary Least Squares estimators $\hat{\beta}$ with the corrected covariance matrix $\Omega_{\hat{\beta}}$ of eq. (15).

6. INDEPENDENCE

We refer to the econometrics literature for a discussion of dependence over time (autocorrelation); see [17]. We concentrate on dependence in simulation. In simulation we can force the responses y (and hence the errors e) to be independent by sampling the random number seeds independently. However, practitioners often use common random number streams, and then the independence assumption is violated and Ω_y is no longer a diagonal matrix. Common random numbers may increase the efficiency (see below) but they also complicate the regression analysis. If we use OLS then $\Omega_{\hat{\beta}}$ is given by eq. (15). We can also use a generalization of Weighted Least Squares, namely, eq. (13) with Ω_y no longer diagonal: Generalized Least Squares.

The estimation of Ω_y involves not only σ_i^2 but also $\sigma_{ii'} \equiv \mathrm{cov}(y_i, y_{i'})$ where $i, i' = 1, \ldots, n$. This estimation is simplest if we replicate each run i a number of times, say, $m_i = m$ times. Hence if, $y_i \equiv \bar{y}_i \equiv \sum_r y_{ir}/m$ and $\sigma_{ii} \equiv \sigma_{i'}^2$, then

$$\hat{\sigma}_{ii'} = \frac{\sum_{r=1}^{m} (y_{ir}-y_i).(y_{i'r}-y_{i'})}{(m-1).m} \qquad (i,i' = 1,...,n) \qquad (16)$$

; also see [28] Note that the estimated covariance matrix $\hat{\Omega}_y$ may be nearly singular when common random numbers are used.

We can prove that common random numbers decrease the variances of the estimators of β_j ($j = 1,...q-1$) and increase the variance of the β_0 estimator; see [42] and note 2. If we were interested in the estimated effects $\hat{\beta}_j$ ($j > 0$) only, and not in the estimated response $\hat{y}$, then we would certainly use common random numbers. Actually we are also interested in the response itself. One reason is that before we test the individual effect estimators $\hat{\beta}_j$ ($j > 0$) we want to know whether the regression model as a whole is valid. To test the validity of the regression model we compare the predictor $\hat{y}$ to the actual response y (see Section 8). And the predictor $\hat{y}$ depends on $\hat{\beta}_0$. So common random numbers may yield better estimators of β_j but a bad estimator of y itself; see [28, 41]. If we use common random numbers then we should perform at least two experiments so that overestimated responses can compensate underestimated responses.

If we use common random numbers then we can analyze the results through Ordinary Least Squares (OLS) or Estimated Generalized Least Squares (EGLS). If Ω_y were known then GLS would yield the Best Linear Unbiased Estimator (BLUE). Under specific conditions - see below - EGLS and OLS yield identical estimators. We have already seen that common random numbers give better estimators of β_j ($j > 0$) even when analyzed by OLS; and β_0 and hence E(y) are systematically overestimated or underestimated whatever technique we use to analyze a single experiment.

GLS and OLS give identical estimators if

- the design matrix is saturated: n = q; see the preceding section;
- the covariance matrix has a very specific structure; see [42, p. 512].

Since this structure involves a quite complex mathematical relationship, we suggest that the practitioner do not check this relationship a priori but that he compare the values of the EGLS and OLS estimates aposteriori to see if the estimates are identical. Moreover, since in practice we use estimated values for Ω_y, the chance of realizing this specific mathematical relationship seems negligible.

Summary: If we want the simplest analysis, then we should make the responses y independent and use the results of the preceding section. If we are prepared to estimate the covariances among the n responses (see eq. 16) then we should use common random numbers and perform at least two independent experiments (to reduce the chance of a systematic over- or underestimation in the estimator of y itself). We may then analyze the results through Ordinary or Estimated Generalized Least Squares.

7. NORMALITY

In case of nonnormality we may be interested in distribution-free and robust procedures. Detecting nonnormality includes the detection of outliers. The responses y may indeed be nonnormal. For instance, if in eq. (11) y^* were normal then the transformed response $y = \log y^*$ in eq. (9) would be nonnormal (usually we assume that $y = \log y^*$ in the linear model of eq. (9) is normal and consequently y^* is "lognormal"). In general, we may apply transformations to obtain normally distributed responses (see later). Transformations may not be necessary if the simulation response is an average (e.g. average queuing time) so that a limit theorem (for either independent or autocorrelated variables) explains normality. So in practice nonnormality may be no serious problem. But let us see what the consequences of (serious) nonnormality can be.

If the n responses are nonnormal then Least Squares - Ordinary or Generalized - still yield unbiased estimators, and the standard errors are still specified by eqs. (14) and (15), but it may be wrong to use the t and F tests of eqs. (7) and (8). When we test a single regression parameter, we use the t statistic. The t statistic is quite robust. When we test several parameters simultaneously, we apply an (ANOVA) F statistic. This F statistic is also quite robust, especially if we replicate each experimental point an equal number of times (we observe that the F statistic for comparing two variances is <u>not</u> robust). We refer to the literature for more details; [8, 43].

Let us return to the relationship between the <u>mathematical</u> Least Squares algorithm and the <u>statistical</u> BLUE property; see Section 2. The mathematical problem is to determine the "best" fit between the observations and the (regression) function $\hat{y}$. To solve this mathematical problem we quantify what we mean by "best". The criterion that results in a simple mathematical solution, is the Least Squares criterion: We can minimize the quadratic expression $\Sigma(y_i-\hat{y}_i)^2$ by solving a set of linear equations (so called normal equations). If we add the classical statistical assumptions for the errors then we can prove that the Least Squares estimator is BLUE, and that the t statistic gives confidence intervals for $\hat{\beta}$ and $\hat{y}$ (moreover, $\hat{\beta}$ is then the maximum likelihood estimator).

Now we consider other mathematical criteria for fitting a curve. Mathematicians have introduced one class of criteria, namely the class of L_p norms: minimize $\sum_{i=1}^{n} |y_i-\hat{y}_i|^p$ where p need not be an integer. Some interesting members of this class are: (i) p = 2: squared deviations, (ii) p = 1: absolute deviations, (iii) $p = \infty$: maximum absolute deviation, $\max|y_i-\hat{y}_i|$(Chebychev norm). A practitioner's criterion may be: minimize the sum of relative absolute errors ($|y-\hat{y}|/ y$). Of no practical relevance seems the L_∞ norm because this norm considers only the maximum deviation. The L_1 norm has one pleasant property when compared to the L_2 norm: extreme deviations (outliers) have less effect on the fitted line. Unfortunately, other criteria than Least Squares lead to estimators of β with statistical properties that are not well-known at present. For the OLS, WLS and GLS estimators we know the asymptotic properties and we have many small-sample experimental results. Note that the other criteria require other algorithms, e.g., the sum of absolute errors can be minimized through Linear Programming; see [50].

There are more criteria. For instance, statistical reasoning leads to more complex criteria, e.g., criteria including discontinuities (see robust regression later on).

When we discussed ridge regression (eq. 8), we noticed that we might calculate a point estimate using one criterion and a confidence interval (centered around a different point estimate) based on, say, OLS formulas. Of course when the former point estimate lies outside the latter confidence interval, this approach is not attractive.

Above we saw how the mathematical criterion of Least Squares (L_2 norm) fits together with the statistical assumption of normality. We shall proceed as follows: (i) How can we detect and reduce nonnormality? (ii) Are there distribution-free regression procedures? (iii) Are there robust procedures? (iv) Miscellaneous.

Sub (i): Detection of nonnormality

Because the responses y_i have non-constant means (determined by the independent variables x_{ij}) we examine the errors $e_i = y_i - E(y_i)$. We assume initially that the remaining classical assumptions are satisfied, i.e., the errors have zero means, constant variances σ^2, and they are independent. Then we estimate the vector of errors e by the vector

$$\hat{e} = y-\hat{y} = y-X.\hat{\beta} = y-X.(X'.X)^{-1}.X'.y = \{I-X.(X'.X)^{-1}.X'\}.y = (I-H).y \quad (17)$$

where the "hat matrix" H is defined by the last equality. Can we use standard techniques such as normality plots and the χ^2 goodness-of-fit statistic? Indeed (older) software often produces plots of the estimated errors $\hat{e}$. However, recent publications have emphasized that even if the true errors e are independent with common variance σ^2, then the estimated errors $\hat{e}$ are dependent with non-constant variances: Because the hat matrix H is symmetric and idempotent ($H^2 = H$) eq. (17) yields:

$$\Omega_{\hat{e}} = \{I-X.(X'.X)^{-1}.X'\}.\sigma^2 \quad (18)$$

To remove the effect of non-constant variances we may "Studentize" the estimated errors: $t_i = \hat{e}_i/\sqrt{\hat{\text{var}}(e_i)}$ where the numerator and denominator follow from eqs. (17), and (18) and (6). Unfortunately, the dependence among transformed errors does not permit a simple test; see [14].

If the true errors show heterogeneity of variance (and possibly dependence) so that we may apply Weighted (or Generalized) Least Squares then we replace eq. (18) by $\Omega_{\tilde{e}} = \Omega_y\{I-X(X'\Omega_y^{-1}X)^{-1}X'\Omega_y^{-1}\} = \Omega_y\{I-H_2\}$.

It is good practice to compute the estimated residuals and to plot them. Several plots are traditional in the older literature and the older software, e.g., the empirical distribution of $\hat{e}$; plots of $\hat{e}_i$ versus y_i (heterogeneity of variance may show up if e increases with y where y is determined by x); plots of y versus x_j, see the bibliography in [49]. These plots may signal problems such as a wrong regression model specification, heterogeneity of variance, nonnormal distributions with heavy tails (kurtosis) resulting in many outliers (the topic of the present discussion). Statistical tests for outliers are difficult when the number of outliers is unknown (the most extreme outlier may look reasonable when there is another outlier which "masks" the former outlier) and when the estimated errors show dependence and heterogeneity of variance (see eq. 18). There is a sizable statistical literature on outliers. However, its relevance for practitioners is limited because the literature assumes constant variances, etc. Note that in regression analysis outliers are important only in so far as they result in drastic changes in the parameter estimates $\hat{\beta}$ and the predicted responses $\hat{y}$; see [2, 4].

In simulation it is much easier to check whether an extreme observation y is due to pure chance: in random simulation the computer program can again be executed with a different random stream. In real-life experiments it is often difficult to realize true replication (i.e., to observe the response for the same input condition); difficulties are time trends, learning effects, changing environments, etc. In real-life experiments an outlier may also be caused by a measurement error. In simulation an outlier is caused by a programming error ("bug") or an "extreme" random number stream. We recommend to replicate a suspicious observation more than once; if the suspicious observation is more extreme than all its replicates, we throw away the outlier and add the replicates to the regression material.

Outliers may occur not only in the responses but also in the independent variables x. Outliers in x are a problem indeed in the social sciences (see Section 3). In well-designed experiments, however, no outliers occur unless we make a mistake. We can signal outliers in x by computing the hat matrix H of eq. (17), provided we use Ordinary Least Squares. Outlying values of x are indicated by "large" values of the diagonal elements h_{ii} of H, say, $h_{ii} > 2\ q/n$. In well-designed experiments all diagonal elements h_{ii} are equal. Unfortunately, in Generalized (and Weighted) Least Squares the hat matrix is more complicated; see [5, 19].

Most robust regression estimators (see later) are insensitive to changes in the dependent variable y, but they are not robust relative to the independent variables x.

Summary: We should study the estimated residuals because they may signal problems such as nonnormality. Exact tests are difficult. In simulation we can identify outliers by replicating suspicious observations using different random number streams.

We can reduce the effect of extreme observations by using the median instead of the mean (for symmetric distributions both location measures coincide). If we have a number of replicates m_i, then we may compute the sample median per combination i provided $m_i \geq 3$. When we use these medians then our regression model predicts the population median, not the population mean. For instance, y may represent the median waiting time. Academic studies often concentrate on the mean. Practical studies may measure the mean because of statistical tradition; actually the median may be more relevant. Of course we may report both quantities (mean and median) to the user.

Suppose we are interested in the population mean, not the median. To reduce the effects of outliers we may still compute the (sample) average but only after we have removed extreme observations, i.e. per combination we automatically eliminate a certain percentage of the extremely small and extremely large observations. Strictly speaking, we cannot analyze the remaining observations using the familiar formulas; special formulas were presented by Tiku [45].

One more way to reduce the effects of nonnormality is provided by transformations like the logarithmic transformation of eq. (9). A more general transformation is the "power" transformation which should result in a regression model with errors that satisfy the Classical Assumptions (including normality and constant variances): $y^* = (y^\lambda - 1)/\lambda$

if $\lambda \neq 0$, and $y^* = \log y$ if $\lambda = 0$, where λ is estimated from the regression data using maximum likelihood estimation; see [3].

Sub (ii): Distribution-free regression analysis

The recent statistical literature gives several nonparametric techniques. For instance, a procedure may consider the ranks of the residuals. These procedures yield asymptotically valid confidence intervals for the regression parameters β. Unfortunately, these methods are more complex, conceptually and computationally. Further they assume a symmetric distribution for the errors and common variance (more strictly, they assume identically distributed errors); see [12].

Conover and Iman's rank regression may interest practitioners because it combines the well-known regression analysis procedures with the simple rank transformation; see [10]. In rank regression we replace the original observations (y_i, x_{ij}) by their ranks, i.e., we explain the rank of y_i as a function of the ranks of x_{ij}. For instance:

$$R(y_i) = \beta_0 + \beta_1 . R(x_{i1}) + \beta_2 . R(x_{i2}) + \beta_{12} . R(x_{i1}) . R(x_{i2}) + e_i \qquad (19)$$

where $\beta_1 = 0$ and $\beta_{12} = 0$ if factor 1 has no effect, etc. The response y (not its rank) is estimated by linear interpolation; see [10]. Iman and Conover applied their procedure in the analysis of several simulation models. The method works well if y is a monotonic function in x_j; it does not work for hill-shaped response functions where different x values (different ranks) yield the same y values (same rank). We emphasize that a rank-transform model like eq. (19) can tell whether the response is affected by a factor x, but it does not help much in explaining how the response is affected. For example, if y denotes consumption and x denotes income then a model in y and x results in the marginal income effect β_1, whereas a model in the ranks R(y) and R(x) has no such interpretation. In practice we may analyze the data using both a classical parametric technique and the rank transformation; if the two analyses give different conclusions then we should look for outliers and the like.

Sub (iii): Robust regression analysis

Robust procedures take a middle position between parametric and nonparametric procedures: A parametric procedure assumes one specific type of distribution, e.g., the normal distribution with parameters μ and σ^2 $(-\infty < \mu < \infty, \sigma^2 > 0)$. A nonparametric procedure makes extremely weak assumptions, e.g., the class of all symmetric distributions. A robust procedure assumes a smaller class of distributions, e.g., the class of "contaminated" normal distributions: $y = py_1 + (1-p)y_2$ with $y_1 \sim N(\mu_1, \sigma_1^2)$ and $y_2 \sim N(\mu_2^2, \sigma_2)$ where $0 < 1-p << p < 1$. In other words, a small percentage (1-p) of the observations comes from a normal distribution with much different parameters (e.g., $\mu_1 = \mu_2 = 0$ but $\sigma_1^2 << \sigma_2^2$) resulting in outliers.

Robust procedures automatically give less weight to outlying responses (whereas classical regression analysis tries to detect and remove such outliers, as we saw). We feel that the practitioner will find robust procedures too complicated. Therefore we do not discuss these procedures further, but refer to the literature; [6, 21, 23].

Sub (iv): Miscellaneous

We have excluded Bayesian and decision-theoretic methods (prior probabilities, loss functions, minimizing expected or maximum loss). Dempster et al. [13] examined no less than fifty-seven different regression estimators in an extensive simulation experiment (160 data sets). Instead of selecting an appropriate regression algorithm, we may select a matrix of independent variables X such that the sensitivity of the regression estimates to outliers is minimized.

Summary: We discussed several mathematical criteria for curve fitting, e.g., Least Absolute Deviations (L_1 norm). However, Least Squares (L_2 norm) combined with the statistical assumption of normality, yield the familiar t and (ANOVA) F statistics. Nonnormality may have little effect on these statistics. Detecting serious nonnormality is based on the estimated residuals $\hat{e}$ but exact tests are difficult. In simulation we can detect outliers by replicating runs with new random number seeds. Effects of outliers can be reduced by regressing on sample medians instead of sample means, by transforming the response (power transformation), by distribution-free procedures and by robust procedures.

8. SPECIFICATION AND VALIDATION

8.1. Introduction

We shall discuss how we get to a specification of the regression model; how we can statistically test the validity of the specified model; and if the model was misspecified how we can improve the original model.

The specification of the regression model depends on (i) general principles of science,and (ii) specific statistical principles. For instance, common sense tells that in a computer system the response of interest y may be waiting time of jobs and idle time of the CPU, and the independent variables x may be arrival rate, service rate, and computer configuration. Deductive reasoning leads to queuing theory which suggests that the ratio λ of the arrival rate and service rate is a fundamental independent variable. Analytical solutions of simplified queuing models yield additional insight. Measurements performed on the existing system and on the simulated system may result in specific insight, possibly after statistical manipulation. Statistical theory suggests that the response variable may be the 90% quantile rather than the average waiting time.

We emphasize that statistical theory does not specify which variables are of interest, let alone the form of the relationships among variables. Which variables may be important, is clearly indicated in the "hard" sciences like computer science, and is fuzzily indicated in the "soft" sciences like management information systems theory. In the hard sciences we know which independent variables to study and we may even postulate specific forms (non-linear in the regression parameters). In econometrics, however, we wish to forecast demand for a particular product and we do not know which products are really competitive (so that their prices should be included in the regression model); obviously the shape of the relationship is even more obscure: maybe we should make a logarithmic transformation of y and x such that the regression parameters β can be interpreted as elasticity coefficients (a classical concept in economics). Logarithmic scales emphasize relative magnitudes. Other popular scale transformations are: y^2, $\sqrt{y}$ and $-1/y$; see [46]

8.2. Statistical technique

Our approach applies to the validation of any model, be it a chemical or an econometric model, a queuing simulation, a regression model, etc. (Other statistical approaches - such as the lack of fit F test - have less appeal to practitioners and are limited by more statistical assumptions.) Our approach comprises the following steps: (i) Devise the model's general form. (ii) "Calibrate" the model, i.e., determine the values of its parameters. (iii) Use the model to forecast a new situation, i.e., a situation not used in the preceding two steps. (iv) Compare the model's forecast to the actual response. In the case of regression models the procedure runs as follows. (i) We postulate a regression model, plus a statistical submodel; see the Classical Assumptions. (ii) From the sample of n observations we estimate the regression parameters β, using (say) Least Squares. (iii) We define a new situation $x'_{n+1} \equiv (1, x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,q-1}) \neq x'_i$ $(i = 1, \ldots, n)$, and forecast the reponse: $\hat{y}_{n+1} = x'_{n+1}\hat{\beta}$. (iv) We observe or simulate that new situation x_{n+1} and obtain the response y_{n+1}. Obviously the forecast $\hat{y}_{n+1}$ and the actual response y_{n+1} will not be exactly equal. Large deviations are acceptable if the statistical submodel (see i) specified large variability σ_i^2. Therefore we compute the Studentized deviation:

$$z_{n+1} = \{y_{n+1} - \hat{y}_{n+1}\} / \{\hat{\mathrm{var}}(y_{n+1}) + \hat{\mathrm{var}}(\hat{y}_{n+1})\}^{\frac{1}{2}} \qquad (20)$$

In eq. (20) $\hat{\mathrm{var}}\ (y_{n+1})$ follows from the analysis of the (simulation) run n+1; see [27]. In deterministic simulation we have: $\mathrm{var}(y_{n+1}) = 0$. If we do not simulate then we have to obtain replicated observations for situation n+1. The term $\hat{\mathrm{var}}(\hat{y}_{n+1})$ follows from eq. (4): $\hat{\mathrm{var}}(\hat{y}_{n+1}) =$

$= x'_{n+1} \; \hat{\Omega}_{\hat{\beta}} \; x_{n+1}$ where $\Omega_{\hat{\beta}}$ was given in eq. (5). Note that y_{n+1} and $\hat{y}_{n+1}$ are independent: $\hat{y}_{n+1}$ depends on $\hat{\beta}$ and $\hat{\beta}$ depends on $y_1,\ldots,y_n$ but not on y_{n+1}. Our Monte Carlo experiment suggest that we may test the significance of the Studentized forecast error by comparing z_{n+1} of eq. (20) to the standard normal variable z; see [26].

The more realistic assumption of non-constant variances means that we use Estimated Weighted Least Squares with its (approximate) covariance matrix - see eq. (14) - or Ordinary Least Squares with the corrected covariance matrix of eq. (15).

The above discussion assumes a single validation run, namely run n+1. There is a trick, however, to obtain many runs for the validation of the regression model, provided there are more runs than there are regression parameters: If $n > q$ then one run can be deleted (say, run 1) and the regression parameters can still be estimated from the remaining n-1 runs. The deleted run (run 1) can next be forecasted and the Studentized forecast error can be computed using eq. (20). The trick continues as follows: Now a different run is deleted (say, run 2 is deleted and run 1 is again added to the data available for estimation of the regression parameters). And so on. This permutation or cross-validation approach yields n validation runs resulting in n dependent forecast errors; also see [3, 5, 19].

The postulated regression model should hold at all n observation points. Consequently we reject the regression model, whenever any of the n values of the Studentized forecast errors is significant. Now a statistical complication arises: If we have, say, one hundred observations (n = 100) and we test the forecast error of eq. (20) with a significance level of 5% then we expect five false alarms (remember the definition of the type I or α error). In symbols: Our null-hypothesis is that the regression metamodel is valid, or $H_0 : E(\hat{y}_i) = E(y_i)$ with $i = 1,\ldots,n$. We reject this null-hypothesis if any z_i value defined by eq. (20) is significant, or $\max | z_i | > z^{\alpha}$ where $\alpha = \alpha_C/2$ and α_C is the "per comparison" error rate, i.e., α_C is the error rate used in an individual test, and the Bonferroni approach - see [33] - means that $\alpha_C = \alpha_E/n$ where α_E denotes the "experimentwise" error rate, i.e., the error rate that holds over the whole experiment (under the composite null-hypothesis, the experiment comprises n observations). Obviously the factor 2 in $\alpha = \alpha_C/2$ corresponds to a two-sided test: both overestimation and underestimation are unacceptable. For instance, if n = 8 and $\alpha_E = 20\%$ then $\alpha = 1.25\%$.

What is the effect of nonnormality on the validation test? A recent Monte Carlo experiment showed that in these types of tests tails heavier than "Gaussian" lead to a chance higher than the nominal α value of finding extreme values; see [34]. This result agrees with the general idea that a test based on a maximum of certain statistics is not robust; see [33]. In simulation we can correct a false alarm by replicating the suspicious input combination a number of times using new random number seeds. For additional comments see [28].

8.3. Related issues

(i) Our validation procedure concentrates, not on the individual estimated parameters $\hat{\beta}$, but on the resulting (single) forecast $\hat{y}$. Concentrating on $\hat{\beta}$ would result in the following norm. When we delete run i then we reestimate the regression parameters β from the remaining (n-1) runs; let this estimator be denoted by the vector $\hat{\beta}^{(i)}$ (with i = 1,...,n). Ideally the n vectors $\hat{\beta}^{(i)}$ would remain constant (and equal to the true parameter vector β). Drastic changes in $\hat{\beta}^{(i)}$ indicate outliers in the dependent variable y or in the independent variables x_j. We may characterize changes in the vector $\hat{\beta}^{(i)}$ by a single number c_i:
$c_i = \sum_{j=1}^{q} (\hat{\beta}_j^{(i)} - \hat{\beta}_j)^2/q$ with i = 1,...,n; also see [11].

(ii) Our test considers the absolute magnitude of the deviation $\hat{y}-y$ whereas in practice we tend to concentrate on relative deviations $\hat{y}/y$. Unfortunately we do not know a simple statistic for this relative forecast error, although the variance of the ratio of two random variables can be approximated. Note that Least Squares minimizes squared residuals $\hat{y}-y$, not relative residuals $\hat{y}/y$.

(iii) Textbooks and standard software present the traditional R^2 criterion:

$$R^2 \equiv \Sigma(\hat{y}_i-\bar{y})^2/\Sigma(y_i-\bar{y})^2 \equiv 1 - \Sigma(y_i-\hat{y}_i)^2/\Sigma(y_i-\bar{y})^2$$

1

which shows that the regression model gives an adequate explanation when R^2 approaches the value one. However, R^2 always improves whenever we add more explanatory variables; if $q = n$ then $y_i = \hat{y}$ and $R^2 = 1$. There is no statistical criterion for testing whether R^2 is large enough, given n and q.

(iv) The statistical literature includes other techniques for the selection of an appropriate regression model. We shall be short on these techniques because they ignore the knowledge the analyst must have about the system under investigation; also see Section 8.1. In stepwise regression we begin with the independent variable that shows the highest correlation with the dependent variable. Next we introduce the remaining independent variable that has the highest correlation with the dependent variable, etc. So in each step we introduce one new variable. In backwards elimination we start with the "largest" model and eliminate nonsignificant individual parameters. As an alternative to these sequential procedures statisticians have proposed to compute all subsets of regression models, i.e., consider the single model with all q independent variables (including the dummy variable x_0); next the q-1 different models obtained by deleting variable 1, variable 2,...,variable q-1 respectively. And so on. All together there are $2^{q-1}-1$ possible subsets. See [19, 44].

(v) Only if the regression model is valid the errors e have zero expectation and the estimators of the regression parameters β are unbiased. Consequently confidence intervals for the individual parameters β should not be derived before the regression model as a whole has been tested. We also have to decide whether we want to test each regression parameter individually or whether we test some parameters jointly. As an illustration we consider a regression model representing the effects of k parameters: $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + e$. We consider each parameter individually, i.e., the interpretation of the experiment does not hinge on the joint results of the individual t tests of eq. (7). Now we consider a different example where we study only two parameters but a more complicated model seems necessary: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x\ x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + e$. Suppose we find that all estimated regression parameters are significant (using eq. 7 with, say, $\alpha = 0.05$) except for β_1. Nevertheless $\hat{\beta}_1$ is an unbiased estimator (if certain assumptions hold then $\hat{\beta}_1$ is even the minimum variance estimator: BLUE). We would not replace $\hat{\beta}_1$ by zero, unless we have strong reasons to postulate such a zero value. A different question is: can we replace the second-order polynomial by a first-order polynomial in x? We can estimate the first-order model and validate this simpler model, using eq. (20). If we have to reject this simpler model in favor of the second-order model then we do not know whether this rejection is caused by a large value of β_{12}, β_{11} or β_{22}. A more detailed analysis runs as follows. Estimate the more complicated model and test the composite hypothesis $H_0 : \beta_{12} = 0,\ \beta_{11} = 0,\ \beta_{22} = 0$. We can test this hypothesis by testing the individual regression parameters β_{12}, β_{11} and β_{22} combined with the Bonferroni approach, i.e., we use $\alpha = \alpha_E/3$ in the individual t tests of eq. (7). Instead of this conservative (but robust) approach we might apply the exact ANOVA F statistic of eq. (8).

Summary: We first discussed general principles used to specify a regression model (including the gamut from black-box to white-box sciences). Next we presented a statistical test which compares the regression forecast $\hat{y}$ to the actual response y, accounting for inherent variability. Cross-validation yields many validation points. We discussed related issues, e.g., the R^2 criterion, stepwise regression, testing individual regression parameters $\hat{\beta}$.

9. REVISING FALSE REGRESSION MODELS

Next we shall investigate the alternatives if we reject the (initial) regression model, i.e. can we revise the model such that it becomes valid?

(i) Transformations: Before we postulate any regression model, we should think hard about the fundamental variables in the regression model. For instance, queuing theory proves - albeit for simplified analytical models - that the fundamental variable is not the arrival rate or the service rate, but their ratio λ, i.e., the traffic load. Consequently it is probably better to use a model with that ratio λ. In general the correct specification of the regression model may be inspired by the known solution for a simplified model, e.g., the steady-state solution of a Poisson queuing model. And in a harbor simulation examination of several plots revealed that the response curve became linear when the mean interarrival time was replaced by its reciprocal, the interarrival rate. We repeat that another reason for transformations is that we wish to satisfy statistical assumptions like constant variances and normality.

If we have no clues as to the form of the model, then we have to rely on "raw" experimentation: In the preliminary phase of the experiment we vary one variable, say x_1, and keep all other variables constant. Next we repeat this procedure for a different variable, say x_2. Then we change the first two factors, x_1 and x_2, simultaneously in order to check the presence of interactions (see below). We may study the absolute output y or the marginal output $\partial y/\partial x_j$.

(ii) Higher order models: Mathematically speaking we can formulate the regression model as a Taylor series approximation to the true model. Consequently if we reject the first-order approximation then we may proceed to a second-order approximation, i.e., we add k "pure quadratic" effects β_{jj} (where $j = 1,...,k$) and $k(k-1)/2$ "two-factor interactions" $\beta_{jj'}$ ($j < j'$ where $j' = 2,...,k$). The interpretation of these additional parameters β is as follows.

A regression model with interactions implies that the response curves are not parallel, i.e., the marginal output of an independent variable is not constant but depends on the values of the other variables. A positive interaction ($\beta_{12} > 0$) means that the two inputs x_1 and x_2 are "complementary", i.e., an increase of x_1 has an extra effect on the output when accompanied by an increase of x_2. A negative interaction means that the marginal output of x_1 is much smaller when more of x_2 is available which can be substituted for x_1. Several authors have emphasized the need to consider interactions when analyzing simulation models or utility models); see [47] resp. [16, 24].

We might introduce interactions among more than two variables. Although including such high-order interactions is traditional in ANOVA we do not recommend it. The main reason is that we can define such interactions mathematically but it is hard to interpret these interactions. Moreover the addition of independent variables (like $x_1x_2x_3$) increases the variance of the predicted response (except for "pathelogical" cases); of course such additional variables may decrease the bias of the regression predictions. Finally the addition of variables may require more runs: A necessary (but not sufficient) condition on X is that $n \geq q$ (see Section 3) and q increases with the addition of high-order interactions.

Pure quadratic effects mean that the response model shows curvature. If the first-order model is not valid and if the independent variables are quantitative then pure quadratic effects may provide a good model. The larger the area is over which we let the independent variables range, the more desirable it is to proceed from a first-order to a second-order approximation. Also see the following discussion.

(iii) Smaller domain: The Taylor series argument suggests that an approximation may become valid if we reduce the domain of the function. Of course alternative (iii) limits the generality of the regression model. This limitation is no problem if the objective is not to obtain a general understanding but to search for the optimum values of the (quantitative) parameters x; see the next section.

Summary: We may improve the validity of the regression model through: (i) transformations, (ii) addition of interactions and quadratic effects, (iii) restriction to a smaller domain.

10. OPTIMIZATION OF SYSTEMS

Optimization may use Response Surface Methodology (RSM), or several other approaches. We shall concentrate on RSM because this approach fits in nicely with our regression modeling approach and RSM seems not inferior - to say the least - to other approaches; see [28, 31, 32, 35]:

Step 1: We start in a subdomain of the full experimental area. In such a small area a first-order model may very well be valid.

Step 2: We use the fitted (calibrated) first-order model to find the direction of improvement. If we fix $\hat{y}$ to a specific value, say $y^{(1)}$, then many combinations of x can yield that response: equi- or iso-response lines. Suppose for illustration purposes that $\hat{\beta}_1 > \hat{\beta}_2 > 0$. If we wish to maximize the response then we should add more of x_1 and x_2, and it is efficient to increase x_1 more than x_2. We can prove that the "path of steepest ascent" is perpendicular to the fitted first-order model. This path is realized if we change the variables such that $\Delta x_j / \Delta x_{j'} = \beta_j / \beta_{j'}$.

Step 3: Along the path of steepest ascent we again experiment. As soon as a run does not yield a higher response, we explore the new area by fitting a (local) first-order model. The new estimates of the parameters β of the first-order approximation in that new area, will yield a new direction for the steepest ascent path.

Step 4: We repeat step 3 a number of times until we apparently reach the optimum region: a first-order approximation (hyperplane) cannot represent a "hill". So if we reject the first-order model then we proceed to the next step.

Step 5: We estimate a second-order model in the optimum area. Because such a model has more parameters β than has a first-order model, we must add some extra runs. We can use special designs to specify those extra runs.

Step 6: Taking derivatives $\partial/\partial x$ of the second-order regression model, and solving $\partial/\partial x = 0$ we estimate the optimal values of x, say x^*. It is possible that x^* does not correspond to a unique maximum but to a saddle-point or a ridge. The shape of the optimum response surface is revealed by a mathematical technique called canonical analysis.

Step 7: We may check whether x^* is indeed optimal, by experimenting with some other input combinations, both close to x^* and far away from x^* (the latter option checks whether we have become stuck on a local "hill").

RSM is a heuristic approach, i.e., it does not guarantee a truly optimal solution. For example, we have to use intuition to decide on the size of the "local" experimental area, and on the size of the steps we take along the path of steepest ascent. And we may end with a local maximum instead of a global maximum. And when we follow the steepest ascent path, then we might stop prematurely: a lower response may be due to random error; see [36]. Other problems arise if the maximalization is restricted by side conditions, as in mathematical programming, or if there are multiple responses. RSM and alternative approaches all employ completely automated search procedures. However, modern interactive computer systems combining computer speed with human pattern recognition may perform better.

Summary: We can optimize a (real or simulated) system applying RSM, although it does not guarantee an overall optimum.

11. MISCELLANEOUS

(i) Regression analysis explains how the output reacts to the input: sensitivity analysis. Sometimes, however, we are first of all interested in the absolute value of the output for the various inputs. For instance, we simulated a computerized inventory control system for various inputs (e.g., different cost parameters) and tested whether the realized service was significantly lower than the desired service percentage. Only after simulated service turned out to be too low in certain situations, we raised the question "which factors cause this disservice?" and we applied regression analysis; see [30].

(ii) We have ignored the numerical aspects of regression computations. For example, computing the inverse in $\hat{\beta} = (X'X)^{-1}X'y$ can be avoided using numerical algorithms due to Choleski, Gram-Smidt, Householder, Givens, etc. Software packages use such algorithms, which results in smaller numerical inaccuracies and improved computational speed and memory size; see [7]. Recent statistical publications on cross-validation, also discuss numerical aspects; [18].

12. APPLICATIONS IN SIMULATION

The regression model summarizes in an explicit form the relationship between input and output of the simulation program. This metamodel can guide the user in the validation of the simulation model, in optimization, and so on. Applications of regression modeling in simulation have started to appear. These applications concern steel plants, medical services, harbors, computers, job shops, ecological systems, inventory control, statistical procedures, etc. The simulations were performed by industrial and academic analysts. Most simulation models were random; a few were deterministic. We give more applications in [28].

13. CONCLUSIONS

We can perform regression analysis using Ordinary Least Squares (OLS) or Estimated Weighted Least Squares (EWLS), accounting for possibly strong heterogeneity of variance. We can test the validity of the resulting regression model. We can base optimization on Response Surface Methodology.

Obviously regression analysis should not be used mechanically. For instance, the specification of regression models requires more than a bag of statistical tricks. The form of the model and the values specified in null-hypotheses have to come from nonstatistical sources such as computer and management science. Subjective elements remain in the selection of the α values and in the evaluation of the statistical technique's sensitivity to assumptions like normality.

We can apply regression analysis to reduce the ad hoc character of simulation. The resulting metamodel helps us to interpret the simulation results, including validation, optimization, etc.

NOTES

1. In deterministic simulation we have a completely fixed response y, given the values for the simulation parameters or the independent variables x, i.e., we have var $(y|x) = 0$. Because the regression model is only an approximation, errors e will remain. Since infinitely many combinations of simulation parameters are possible, we have infinitely many errors e_i $(i = 1,2,\ldots,\infty)$. The population of these errors has a variance denoted by σ^2. We might assume that the errors do not show a systematically different behavior in certain areas of the space of the simulation parameters. We sample the simulation parameter values randomly or more or less systematically. So, in the regression model of the deterministic simulation the independent variables x become random variables. Consequently, the $\hat{\beta}$ being a function of x (see eq. 1) become random, and so does $\hat{y}$ so that $e = y-\hat{y}$ is random too. In order to detect a systematic behavior in e (including heterogeneity of variance) we can make plots of the estimated errors or "residuals" $\hat{e}$. For instance, we may plot $\hat{e}_i$ versus x_i $(i = 1,\ldots,n)$ or $\hat{e}_i$ versus y_i; also see Section 7. In the statistical literature we find situations where x is deterministic and y is random (called "regression" situation) and situations where both x and y are random ("correlation" situation). We introduced a third situation, namely x is random and y is deterministic; this area deserves more research.

2. For illustration purposes we consider the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ given in eq. (2). The estimate of β_1 does not change if we transform x_i and y_i such that $\bar{x} = 0$ and $\bar{y} = 0$. We assume constant variances $\sigma_i^2 = \sigma^2$. Inde-

pendent runs yield: $\mathrm{var}(\hat{\beta}_1) = \sigma^2 \Sigma a_i^2$ with $a_i \equiv x_i/\Sigma(x_i^2)$. Dependent runs yield: $\mathrm{var}(\hat{\beta}_1) = \sigma^2 \Sigma a_i^2 + \underset{i \neq i'}{\Sigma\,\Sigma}\; a_i a_{i'} \mathrm{cov}(y_i, y_{i'})$ where $\mathrm{cov}(y_i, y_{i'})$ is positive if common random numbers "work" We assume that the correlations or covariances are constant. Because $\bar{x} = 0$ the sum of cross-products $\Sigma\,\Sigma\, a_i\, a_{i'}$ is negative (expand $(\Sigma x_i)^2$). So common random numbers decrease the variance of the OLS estimator of the slope β_1. For the intercept eq. (2) yields $\mathrm{var}(\hat{\beta}_0) = \mathrm{var}(\bar{y}) + (\bar{x})^2\ \mathrm{var}(\hat{\beta}_1) - 2\bar{x}\ \mathrm{cov}(\bar{y}, \hat{\beta}_1)$ where common random numbers increase $\mathrm{var}(\bar{y})$; this increase may or may not be compensated by the remaining terms; if $\bar{x} = 0$ then there is no compensation.

REFERENCES

1. Aitchison, J. and J.A.C. Brown (1966). The Lognormal Distribution. Cambridge University Press, London.

2. Andrews, D.F. and D. Pregibon (1978). Finding the outliers that matter.
Journal Royal Statistical Society. Series B, 40, no. 1: 85-93.

3. Atkinson, A.C. (1982). Regression diagnostics, transformations and constructed variables. Journal Royal Statistical Society. Series B, 44, no. 3: 1-36.

4. Barnett, V. and T. Lewis (1978), Outliers in Statistical Data. John Wiley & Sons, Inc., New York.

5. Belsey, D.A., E. Kuhn and R.E. Welsch (1980). Regression Diagnostics, Identifying Influential Data and Sources of Collinearity. John Wiley and Sons, New York.

6. Bickel, P.J. (1976). Another look at robustness: a review of reviews and some new developments (+ discussion). Scandinavian Journal of Statistics, 3: 145-168.

7. 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.

8. Bradley, J.V. (1980). Nonrobustness in Z, t and F tests at large sample sizes. Bulletin Psychonomic Society. 16, no. 5: 333-336.

9. Bunke, H. (1980). Parameter estimation in nonlinear regression models. Handbook of Statistics, Volume I, edited by P.R. Krishnaiah, North-Holland Publishing Company, Amsterdam.

10. 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.

11. Cook, R.D. and S. Weisberg (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

12. De Kroon, J and P. van der Laan (1983). A generalization of Friedman's rank statistic. *Statistica Neerlandica*. *37*, no. 1: 1-4.

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

14. Dempster, A.P. and M. Gasko-Green (1981). New tools for residual analysis. *Anuals of Statistics*. *9*, no. 5: 945-959.

15. Games, P.A. and G.S. Wolfgang (1983). A review of six multifactor tests for homogeneity of spread. *Computerational Statistics & Data Analysis*. *1*, no. 1: 41-52.

16. Grochow, J.M., (1972). A utility theoretic approach to valuation of a time-sharing system. In: *Statistical Computer Performance Evaluation*, edited by W. Freiberger, Academic Press, Inc., New York.

17. Hendry, D.F. (1983). Monte Carlo experimentation in econometrics. *Handbook of Econometrics. Volume II*. Edited by Z. Griliches and M.D. Intrilligator, North-Holland Publishing Company, Amsterdam.

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

19. Hocking, R.R. and O.J. Pendleton (1983). The regression dilemma. Communications in Statistics, Theory and Methods. 12, no. 5: 497-527.

20. 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.

21. Hogg, R.V. (1977). Robustness. Special Issue of Communications in Statistics - Theory and Methods, Volume A6, no. 9: 789-894.

22. Horn, S.D., R.A. Horn and D.B. Duncan (1975). Estimating hetero scedastic variances in linear models. Journal American Statistical Association. 70, no. 350: 380-385.

23. Huber, P.J. (1981). Robust Statistics. John Wiley & Sons, Inc., New York.

24. Keeney, R.L. and H. Raiffa (1976). Decisions with Multiple Object ives: Preferences and Value Tradeoffs. John Wiley & Sons, Inc., New York.

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

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

27. Kleijnen, J.P.C. (1984). Statistical analysis of steady-state simulations: survey of recent progress. European Journal Operational Research.

28. Kleijnen, J.P.C. (1985). Statistical Tools for Simulation Practitioners. Marcel Dekker, Inc. New York.

29. Kleijnen, J.P.C., R. Brent and R. Brouwers (1981). Small-sample behavior of weighted least squares in experimental design applications. Communications in Statistics, Simulation and Computation. B10, no. 303-313.

30. Kleijnen, J.P.C. and P.J. Rens (1978). IMPACT revisited: a critical analysis of IBM's inventory package "IMPACT". Production and Inventory Management. 19, no. 1: 71-90 (first quarter).

31. Kumar, B. and E.S. Davidson (1980). Computer system design using a hierarchical approach to performance evaluation. Communications ACM, 23, no. 9: 511-521.

32. Mead, R. and D.J. Pike (1975). A review of response surface methodology from a biometric viewpoint. Biometrics, 31: 803-851.

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

34. 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.

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

36. Myers, R.H. and A.I. Khuri (1979). A new procedure for steepest ascent. Communications in Statistics, Theory and Methods, A8, no. 14: 1359-1376.

37. Nozari, A. (1983). Dealing with Unequal Variances in Analysis of Experimental Design Applications. School of Industrial Engineering, University of Oklahoma, Norman (Oklahoma). (Submitted for publication.)

38. Obenchain, R.L. (1977). Classical F tests and confidence regions for ridge regression. Technometrics. 17, no. 4: 429-439.

39. Schatzoff, M. and C.C. Tillman (1975). Design of experiments in simulation validation. IBM Journal of Research and Development. 19, no. 3: 252-262.

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

41. 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.

42. Schruben, L.W. and B.H. Marjolin (1978). Pseudorandom number assign ment in statistically designed simulation and distribution sampling experiments. Journal American Statistical Association. 73, no. 363: 504-525.

43. Tan, W.Y. (1982). Sampling distributions and robustness of t, F and variance-ratio in two samples and ANOVA models with respect to departure from normality. Communications in Statistics, Theory and Methods. 11, no. 22: 2485-2511.

44. Thompson, M. (1978). Selection of variables in multiple regression: Part I. A review and evaluation. International Statistical Review. 46, no. 1: 1-19.

45. Tiku, M.L. (1978). Linear regression model with censored observations. Communications in Statistics, Theory and Methods. A7, no. 13: 1219-1232.

46. Tukey, J.W. (1977). Exploratory Data Analysis. Addison-Wesley Publishing Company, Reading (Massachusetts).

47. Weeks, J.K. (1979). A simulation study of predictable due-dates. Management Science. 25, no. 4: 363-373.

48. White, H. (1980). Using least squares to approximate unknown regression functions. *International Economic Review*. *21*, no. 1: 149-170.

49. Wilson, S.R. (1979). Examination of regression residuals. *Australian Journal of Statistics*. *21*, no. 3: 18-29.

50. Zanakis, S.H. and J.S. Rustagi, editors (1982). *Optimization in Statistics*. North-Holland Publishing Company, Amsterdam.

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.