

Sensitivity Analysis of Empirical Studies

Hans Nyquist¹

Abstract: Results from an empirical study depend on data, the statistical model, and the statistical techniques used. The statistical techniques are in turn constructed according to the model and a set of principles of statistical inference. Since statistical models are ultimately false, a source of uncertainty in the results is introduced. This uncertainty is distinguished from the sampling error, which traditionally is of primary concern in the statistics literature. Here, sensitivity analysis is defined as the investigation of how model misspecification and anomalous data points influence results.

Included in sensitivity analysis are methods for assessing the influence of a particular source on the results, and suggestions about how to reduce unacceptably large influences. The purpose of this paper is to define some concepts of sensitivity analysis, illustrate the concepts in a few examples, and, perhaps most importantly, to emphasize the need for further research within this area.

Key words: Assessing sensitivity; data penetration; model building; model formulation; principles of statistical inference; robustness; sensitivity function.

1. Introduction

In an empirical study of a particular subject matter it is often convenient to use a statistical model and statistical techniques for analysing the data. However, being a simplification of reality, a statistical model is ultimately false if one looks closely enough. Assumptions such as normally distributed observations, independent replications, and linear functional forms are often made but are rarely true. In addition, statistical techniques are often selected according to their properties, but these properties are valid only if the model is true, implying a noticeable risk of selecting an inferior technique for the analysis. This does not mean, however, that the use of statistical models and statistical techniques is useless.

What matters is whether the model is true enough and whether the properties of the statistical techniques used are valid enough according to the purpose of the study. The particular statistical techniques used in the analysis, the model we consider, and the data on which the conclusions are based are components of considerable importance for the results in an empirical study. In order to determine the applicability of the results it is necessary to analyse the influence on the results from particular components of the study. The main purpose of this paper is to discuss a general approach for analysing the results' dependence on a specific model specification or a specific set of data. Specifically, the results' sensitivity to alterations in the model specification and alterations in the data set is assessed. This type of analysis is here called sensitivity analysis. It is also

¹Department of Statistics, University of Umeå, S-907 87 Umeå Sweden.

the aim of the paper to point out the need for further research in this area.

An extensive discussion on different parts of sensitivity analysis has existed for a long time. For example, a rich literature on influential data points in linear regression is available. The most well known textbooks in this field include Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), and Chatterjee and Hadi (1988). In addition to this, a large number of research papers have been published. On the other hand, very few results about influential data points in fields other than linear regression are published. Pregibon (1981) and Chatterjee and Hadi (1988, ch. 8) are two important exceptions treating the analysis of generalized linear models. The effects of a misspecified model are again rather well known in the case of linear regression (many results appear in standard textbooks) but only few results are published for other models. An important exception here is the field of robust statistics (see e.g., Huber 1981 and Hampel, Ronchetti, Rousseeuw, and Stahel 1986) which can be said to have emerged from the problem of erroneous distributional assumptions.

While proposed assessments of sensitivity are often *ad hoc* we discuss a general strategy for performing a sensitivity analysis, applicable to a large class of problems.

The present paper is organized as follows. The role of sensitivity analysis and its relations to other steps in an empirical study is discussed in the next section. The problem of assessing the results' dependence on model assumptions and on data points are discussed in Sections 3 and 4, respectively. Two approaches to limit the effects of influential model assumptions or influential data points are presented in Section 5. The final section contains some concluding remarks.

2. Sensitivity Analysis

A paradigm for statistical analysis of empirical data is that discussed by Box (1979, 1980), as illustrated in Fig. 1. The process begins with formulating the problem of interest in terms of a probability model. The model consists of a set of assumptions, such as assumed functional relations between variables, assumptions about statistical independence, and assumptions about probability distributions. Some of these assumptions are derived from a theory about the subject matter under study. Examples of these kinds of assumptions include the existence of a functional relation between price and demand for a commodity in a model for consumer demand, and assumptions about the experimental units in a planned experiment leading to the use of a block design. In addition to such assumptions, the model often contains assumptions of another kind, here called auxiliary assumptions. Without auxiliary assumptions an analysis is in many situations impossible. In other situations, auxiliary assumptions facilitate the analysis considerably. Examples of auxiliary assumptions include assumptions about a specific functional form relating price and demand for commodities in a consumer demand model, and assumptions about independence, homogeneity, and normality of error terms in an analysis of data from a planned experiment. In an empirical study, it is usually desirable if results depend strongly on assumptions that are consistent with theories about the subject matter. On the other hand, results can be misleading if they depend too strongly on dubious model assumptions that are made to facilitate the analysis rather than being derived from theories about the subject matter.

The arrow from the left-hand box to the right-hand box indicates the procedure in which the model is confronted with data and

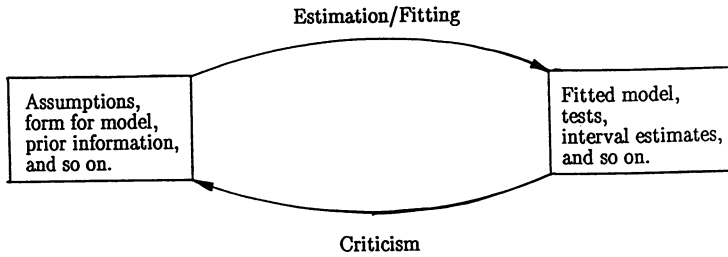


Fig. 1. A paradigm for statistical analysis according to Box (1979, 1980)

results from the analysis are derived. Selection of the particular technique actually used for the inference depends on the model. For example, the maximum likelihood estimator of the symmetry point under a model of i.i.d. normally distributed observations is the sample mean, while it is the sample median under a model of i.i.d. double exponentially distributed observations. However, the selection of a technique for statistical inference also depends on the principles of statistical inference the analyst believes are important as well as on the analyst's interpretation of probability. This is illustrated by noting that the maximum likelihood estimator of the symmetry point under a model of i.i.d. double exponentially distributed observations is the sample median. The least squares estimator and the method of moments estimator, which are based on other principles of inference, are the sample mean under the same model. Still other principles in combination with an additional assumption about a prior distribution lead to a Bayes estimator. Hence, the set of principles of statistical inference adopted has implications for the selection of statistical technique and, in this way, also influences the results from the empirical study. It should also be noted that seldom does a principle yield satisfactory results in all cases. This is of particular importance in a world with models that are not true: an estimator which is unbiased and has a minimum variance property under a specified

model, may be far from unbiased and far from the minimum variance property in the current situation.

A more detailed scheme of the upper arrow would then be as in Fig. 2. The inference can be regarded as a process whose inputs are the model, the technique, and the data, each being dependent on other factors, and whose output is the results from the empirical study.

The right-hand box in Fig. 1 denotes the output from the inference, here called results. In many applications, results are point estimates, interval estimates of particular parameters, or conclusions from tests of hypotheses formulated in the model. Other examples of results from an inference are predictions, forecasts, and decisions about realizing particular actions selected from larger sets of feasible actions.

In the outline of Box's paradigm illustrated in Fig. 1, there is also an act of critical assessment of the model, given the actual results from the inference. This is labeled criticism and is indicated by the arrow from the right-hand box to the left-hand box. As illustrated in Fig. 3 the aims of this criticism are threefold. First, from the subject matter point of view it is important to compare results with what are expected from theory and to explain and interpret results that do not agree with implications from the theory. Second, it is important to assess the influence of the auxiliary assumptions on the results, here called analysis of model

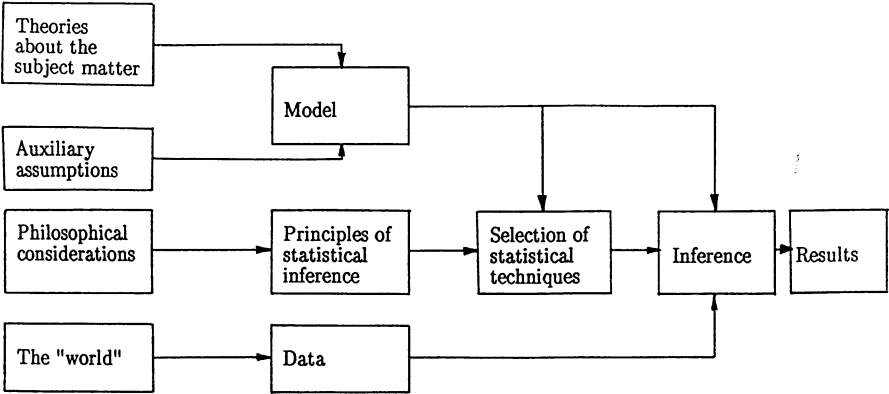


Fig. 2. A more detailed scheme of the upper arrow in Fig. 1

sensitivity. Questions like “how important for the estimate on price elasticity is the assumption about independent error terms” in a model for consumer demand, and “how important for the conclusions about treatment effects is the assumption about normally distributed errors” in an analysis of a planned experiment are typical questions to raise in this context. The third aim of the criticism is related to the general idea in empirical studies that we do not want a small subset of data, extreme in some respect, to have a large effect on the results. In many applications we pursue balanced data so that each data point has the same influence on the results. It is therefore important

to assess the sensitivity to particular observations (in the light of the model) and thereby identify influential data points.

The term sensitivity analysis is here used to denote the investigation of how a model specification and particular data points influence the results of an empirical study (i.e., attention is restricted to the second and the third aims of the criticism). It also contains a discussion about how sensitivity can be reduced.

3. Assessing Model Sensitivity

The models used in statistics are often termed statistical models, or models of

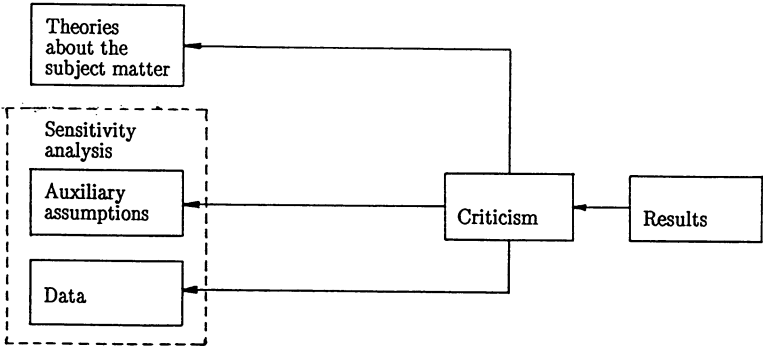


Fig. 3. A more detailed scheme of the lower arrow in Fig. 1

random events. These terms highlight the random nature of the phenomenon that is modelled. In many cases the model is a parametric model in the sense that a family of probability models is used; each member of the family is indexed by a value at a finite dimensional parameter θ . Formally, the possible outcome y of a model M is a specified set, the sample space, $\mathcal{S} = \{y\}$. In a parametric model each of the possible probability distributions of Y is indexed by a value at θ in a specified set, the parameter space, $\Omega = \{\theta\}$, and is represented by a probability function $f(y, \theta)$. According to M the probability of an outcome y in a measurable subset S of \mathcal{S} is

$$P(S; \theta) = \int_S f(y, \theta) d\mu(y)$$

where μ is a specified (σ -finite) measure on \mathcal{S} . A parametric model is thus characterized by $\langle \Omega, \mathcal{S}, f, \mu \rangle$.

Example 3.1. As an example of a parametric model, we consider the simple linear regression model

$$y_i = \theta_1 + \theta_2 x_i + u_i, \quad i = 1, \dots, n$$

x_i are known constants (without measurement errors)

u_i are stochastically independent error terms, all with the same probability distribution

u_i are normally distributed with mean zero and variance θ_3 .

In this example $\mathcal{S} = R^n$, $\Omega = R^2 \times R^+$,

$$f(y; \theta) = (2\pi\theta_3)^{-n/2} \times \exp \left\{ - \sum_{i=1}^n (y_i - \theta_1 - \theta_2 x_i)^2 / (2\theta_3) \right\}$$

and μ is the Lebesgue measure on R^n . \square

A parametric model is necessary when the inference is based on likelihood arguments or on sufficiency arguments for the con-

struction of maximum likelihood estimators, likelihood ratio tests, MVUE, best tests in the Neyman–Pearson sense, etc.

The particular formulation of a model depends on our knowledge about the subject matter under consideration, as it is formulated in the theories applied. Several assumptions in the model can often be regarded as auxiliary assumptions. For example, the assumption about no measurement error in the explanatory variable x_i above depends on how the readings on x_i are made. If x_i is the number of customers arriving to a queue, it seems that there is a good chance to get an exact measurement on x_i . If, on the other hand, x_i is a latent variable (such as “intelligence”) which can only be observed via a set of indicator variables, it is impossible to get an exact measurement on x_i .

The model we use is in many respects a simplified representation of the subject matter under consideration. In particular, it must be consistent with the subject matter with respect to the characteristics under study. However, we never require the model to be true. Although, when selecting a technique for statistical inference, and when performing the inference, we act as if the model were true. It is here often argued that “small” deviations from the model assumptions only cause “small” deviations from the theoretical properties of the inference technique selected. Being aware of the approximate nature of the model the analyst claims he or she is acting approximately in accordance with the principles of inference adapted, and is getting an approximately correct answer to the problem. Unfortunately, this argument may be completely wrong. In some applications deviations from the model assumptions can have very large effects on the results, and the analyst merely gets an approximate answer to the wrong problem.

For assessing the influence of a questionable model assumption on a property of a statistical technique, suppose that T is the statistic considered and $A_T(M_0)$ is the specified property of T (for example, the (population) value of T , the expected value of T , or the variance of T) under the model M_0 . Suppose further that M_0 is embedded into a larger class, say \mathcal{M} , of models. It is then often the case that \mathcal{M} itself constitutes a (augmented) parametric model. Letting ϕ be the parameter indexing the members in \mathcal{M} such that M_0 is obtained for $\phi = 0$ and M_ϕ is an arbitrary member of \mathcal{M} , a possible assessment of the sensitivity with respect to changes in ϕ is obtained by comparing $A_T(M_0)$ with $A_T(M_\phi)$. In particular, we can compute the relative change of A_T and take the limit as ϕ tends to zero, thus defining a sensitivity function

$$\begin{aligned} \text{SF}(\mathcal{M}; T, M_0, A) \\ = \lim_{\phi \rightarrow 0} \{A_T(M_\phi) - A_T(M_0)\}/\phi \end{aligned}$$

provided that the limit exists. If \mathcal{M} defines a linear space and A_T is an operator defined on \mathcal{M} , the sensitivity function can be interpreted as a directional derivative of A_T at M_0 .

Example 3.2. To illustrate the application of the sensitivity function approach for assessing model sensitivity, we consider the case of measurement errors in a simple linear regression model. The explanatory variable x_i in Example 3.1 is now assumed to be observed indirectly through the variable z_i related to x_i as

$$z_i = x_i + v_i, \quad i = 1, \dots, n.$$

The least squares estimator of θ_2 is biased when z_i is substituted for x_i in the regression. If the measurement errors v_i all have expectation zero, variance ϕ , and are independent, the bias of the least squares estimator of the slope θ_2 is known to be $-\theta_2\phi/$

$(\sigma_x^2 + \phi)$, where σ_x^2 is the variance of x_i . Using the definition of the sensitivity function we obtain

$$\begin{aligned} \text{SF} &= \lim_{\phi \rightarrow 0} \{-\theta_2\phi/(\sigma_x^2 + \phi) - 0\}/\phi \\ &= -\theta_2/\sigma_x^2 \end{aligned}$$

as the sensitivity of bias due to measurement errors in the explanatory variable. Thus, the relative change in bias properties for the least squares estimator when a measurement error with infinitesimal variance is added to the explanatory variable is proportional to the slope parameter and inversely proportional to the variance of the explanatory variable. If σ_x^2 is large compared to θ_2 this kind of model deficiency would probably not be harmful. On the other hand, the sensitivity of bias is unbounded (positive or negative depending on the sign of θ_2) as σ_x^2 tends to zero. Similar arguments applied to a multiple linear regression model suggest that the errors-in-variable problem is particularly severe in cases with collinearities among the explanatory variables. \square

Example 3.3. As a second illustration of how the sensitivity function can be applied for assessing model sensitivity, we consider an inference for analysing matched pairs based on the permutation approach. Thus, we consider a finite population in which each subject receives a treatment at dose Z and exhibits a response r . Each subject has also a vector x of observed pretreatment covariates and a scalar unobserved covariate u . The purpose of the analysis is to test the null hypothesis of no treatment effect.

For evaluating the effect of the unobserved covariate u Rosenbaum (1989) proposes a (semiparametric) model for the dose assignments under the null hypothesis. Some notable special cases of the model include the logit model for two doses, linear models for

normally distributed doses, and certain log-linear models for doses taking several discrete values. The model contains a parameter ϕ with the property that if $\phi = 0$ adjustments for x suffice in testing the null hypothesis, while $\phi \neq 0$ implies that adjustments for (x, u) would have sufficed had u been observed.

Based on n matched pairs $i = 1, \dots, n$, we wish to test the null hypothesis. Test statistics considered are of the form $T = \sum q_i s_i [V_i d_i]_+$, where $q_i \geq 0$ are scores depending on $(|d_1|, \dots, |d_n|)$, $d_i = r_j - r_k$, $s_i \geq 0$ are scores depending on (t_1, \dots, t_n) , $t_i = |Z_j - Z_k|$, $V_i = Z_j - Z_k$ if pair i consists of the pair of subjects (j, k) and $[a]_+ = 1$ if $a > 0$ and $[a]_+ = 0$ otherwise. Particular cases of this test statistic include versions of Wilcoxon's signed rank statistic and a statistic that resembles, but is different from, the Spearman rank correlation.

The null distribution of T depends on u and therefore cannot be evaluated unless $\phi = 0$. However, Rosenbaum (1989) shows that, under some conditions, the null distribution of T is asymptotically normal with expectation ranging between $\sum p_i g_i$ and $\sum (1 - p_i) g_i$ depending on u , and with variance $\sum p_i (1 - p_i) g_i^2$, where

$$p_i = \frac{\exp(\phi t_i/2)}{\exp(-\phi t_i/2) + \exp(\phi t_i/2)}$$

and $g_i = q_i s_i$.

The sensitivity analysis suggested by Rosenbaum (1989) is, in our notation, defined by comparing the standardized values of the test statistic, $A_T(\phi) = (T - E(T))/\sqrt{\text{var}(T)}$ for a range of values on ϕ and using the upper as well as the lower bound for the expectation $E(T)$. The test is considered sensitive if one value of $A_T(0)$ and $A_T(\phi)$ suggest a rejection of the null hypothesis while the other cannot support a rejection. This sensitivity analysis is global in the sense that a range of models is con-

sidered. A local sensitivity analysis is obtained by considering the sensitivity function of A_T defined as the derivative of A_T with respect to ϕ at $\phi = 0$. Straightforward calculus yields

$$\lim_{\phi \rightarrow 0} \frac{A_T(\phi) - A_T(0)}{\phi} = \pm \frac{1}{2} \frac{\sum t_i g_i}{\{\sum g_i^2\}^{1/2}}$$

where the sign depends on which of the two extremes of expectation is used.

Rosenbaum (1989) provides a numerical example adapted from Kelsey et al. (1978) where the effects of oral contraceptives on various breast diseases are studied. The test statistic considered is defined by $q_i = 1$ and $s_i = t_i$ (i.e., T is the sum of the dose scores for pairs in which the case had the higher dose). With the data presented Rosenbaum (1989) finds the test to be sensitive to effects from a possible unobserved covariate.

The sensitivity function for the test statistic in this case reduces to $\pm (\sum t_i^2)^{1/2}/2$. Thus, a less sensitive test appear if smaller absolute dose differences are used. \square

When \mathcal{M} is parametric it is often possible to formally test the model for a respecification by testing the hypothesis $\phi = 0$. These tests, often referred to as specification tests, are by now the subject of growing research interest in statistics. Some recent developments in this area, primarily directed toward dependent error terms in linear regression models, are collected in King and Giles (1987). One well-known example of a specification test is the Durbin-Watson test for testing the hypothesis of no first-order autoregression in the error terms of a linear regression model. Here, M_0 is augmented by a first-order autoregressive error process with ϕ as the parameter and the hypothesis $\phi = 0$ is tested.

The property A_T studied in specification tests assesses how the model fits the data, e.g., the log likelihood value at the fitted

model. Notice, however, that other properties of the same statistic tell other stories. For example, although the introduction of an additional explanatory variable is not significant (specification test) it can have large effects on estimates of parameters already included in the model.

If it is not possible to formulate the larger class of models \mathcal{M} as a parametric model, it is more complicated to assess the sensitivity to variations in the model. One approach is to restrict attention to parametric subclasses of \mathcal{M} . Of course, this approach yields only insight about model sensitivity to variations within the subclasses considered.

When analysing sensitivity to deviations from distributional assumptions various nonparameterizable classes of models are, for obvious reasons, often employed. One type of subclasses of models that is frequently used is formulated as follows: Suppose that a randomly selected observation has probability $1 - \phi$ of being generated by the model M_0 and probability ϕ of being generated by a second model, say M_1 . We can think of $(1 - \phi)M_0 + \phi M_1$ as the relevant model. This model fits particularly well to the contamination neighborhood, to be described in Section 4, but is also useful for other nonparameterizable classes of models.

The sensitivity function in this case is

$$\begin{aligned} \text{SF}(M_1; T, M_0, A) \\ = \lim_{\phi \rightarrow 0} [A_T\{(1 - \phi)M_0 + \phi M_1\} \\ - A_T(M_0)]/\phi. \end{aligned}$$

In the particular case where A_T is the population value of the statistic T under consideration and the alternative model M_1 assigns pointmass 1 at the point z in the sample space (represented by its probability distribution function Δ_z) the sensitivity

function reduces to

$$\begin{aligned} \text{IF}(z; T, F_0) = \lim_{\varepsilon \rightarrow 0} [T\{(1 - \varepsilon)F_0 + \varepsilon\Delta_z\} \\ - T(F_0)]/\varepsilon \end{aligned}$$

which is recognized as the influence function, introduced by Hampel (1968, 1974) and extensively discussed in the robustness literature (e.g., Hampel et al. 1986 and Huber 1981). The influence function can thus be interpreted as a collection of directional derivatives (in the direction determined by Δ_z) of T when T is interpreted as a functional defined on a space of probability distribution functions.

Under the model of i.i.d. $N(\mu, \sigma)$ observations, simple calculus shows that the influence function for the sample mean \bar{X} is $z - \mu$. Thus, the influence function is unbounded, indicating that in a large sample the relative change of the sample mean can be arbitrarily large when the sample is contaminated with an infinitesimal proportion of observations from M_1 (the population in which all observations take the value z).

The change of variance function (Rousseeuw 1981, 1982) is another tool for assessing an estimator's sensitivity to deviations in the distributional assumption. For an estimator T of a location parameter μ in a symmetric distribution, the change of variance function can be viewed as a sensitivity function with A_T being the (asymptotic) variance of T ; the models are represented by their probability distribution functions and M_1 is the probability distribution function that assigns pointmass 1/2 at $\mu - z$ and at $\mu + z$ in the sample space.

4. Assessing Data Sensitivity

Model evaluation is traditionally concerned with questions such as (1) How well does this model fit to data? (2) Are there any data points that are not well described by the

fitted model? (3) Are there any data points that highly influence the results of the analysis? Questions 1 and 2 are more directed toward an analysis of results while question 3 asks for an assessment of the sensitivity of results to particular data points. Applications of methods for assessing the influence on results of particular data points are therefore an important area common to both model evaluation and sensitivity analysis. There exists a large body of literature on this topic, including Belsley et al. (1980) primarily concerned with applications in econometrics; Cook and Weisberg (1982) and Chatterjee and Hadi (1988) discuss influential observations in regression analysis; Atkinson (1985) and Cook and Wang (1983) deal with the influence of transformations in linear regression; Hadi and Wells (1990) and Wang and Nyquist (1991) discuss influential observations in eigenstructures of data matrices; Mason and Gunst (1985), Hadi (1988), and Nyquist (1988) discuss observations creating or hiding collinearities in linear regression; and Pregibon (1981) deals with generalized linear models. Since this literature covers most of the relevant aspects of sensitivity due to data, we give only a few remarks here.

To fix our ideas, suppose $T = T(F)$ is a statistic under consideration and suppose z_1, z_2, \dots, z_n are observations drawn from the model distribution F . Then, the observed value on the statistic is

$$t = t(z_1, z_2, \dots, z_n) = T(F_n)$$

where F_n is the empirical distribution function. A simple and natural way to assess the influence of a particular observation, say z_i , on t is to compare t with $t_{(i)}$, the value on t with z_i deleted from the data

$$t_{(i)} = t(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n).$$

In particular, the difference

$$t - t_{(i)} = (n - 1)^{-1} \text{SIC}(z_i; t)$$

is related to the so called sample influence function (SIC) (Mallows 1975), which is the jack knife estimate of the influence function of T evaluated at z_i (Efron 1982).

An alternative assessment of the sensitivity is obtained through the empirical influence function (EIC) defined as the influence function evaluated at the empirical distribution function F_n (Mallows 1975). We now give two examples illustrating how the empirical influence function can be applied.

Example 4.1. Coursey and Nyquist (1988) report on the performance of a single-equation log-linear demand model with first order autoregressive error terms. The model considered is $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$, $u_t = \phi u_{t-1} + v_t$, with different distributional assumptions on the disturbance terms v_t , where y_t is the logarithm of per capita quantity of a good purchased at time t , x_{1t} the logarithm of per capita income, and x_{2t} the logarithm of the price of the commodity relative to a price index of all prices. The model was estimated using nondurable and services commodity time series data for Sweden and the United States. The model is estimated using a Cochrane-Orcutt iteration scheme in which each iteration consists of two linear regressions. In the first regression $y_t^* = y_t - \phi^{(k)} y_{t-1}$ is regressed on $x_{1t}^* = x_{1t} - \phi^{(k)} x_{1t-1}$ and $x_{2t}^* = x_{2t} - \phi^{(k)} x_{2t-1}$ for a given value of $\phi^{(k)}$. This regression results in the estimates $(1 - \phi^{(k)})\beta_0^{(k)}$, $\beta_1^{(k)}$, and $\beta_2^{(k)}$. In the second regression $\hat{u}_t = y_t - \beta_0^{(k)} - \beta_1^{(k)} x_{1t} - \beta_2^{(k)} x_{2t}$ is regressed on \hat{u}_{t-1} yielding the estimate $\phi^{(k+1)}$ and used in the subsequent iteration. The iterations are stopped in accordance with a convergence criterion on the criterion function.

The first two rows of Table 1 report parameter estimates and t -values when the model is applied to annual data on consumption of gas and oil in the United States from 1946 to 1982. The estimates are all as

Table 1. Estimated demand equation for consumption of gas and oil in the United States (t-values in parentheses)

	β_0	β_1	β_2	φ
Complete data	0.800 (0.71)	0.637 (4.70)	-0.220 (-6.72)	0.943 (148)
Reduced data*	-1.811 (-2.31)	0.929 (9.18)	-0.207 (-7.16)	0.899 (84.9)

For the reduced data estimates are computed after deleting the data point $(x_{1,5}^, x_{2,5}^*, y_5^*)$ in the first regression and the data point (\hat{u}_4, \hat{u}_5) in the second regression.

expected: The income elasticity β_1 is significantly positive and the price elasticity β_2 is significantly negative.

Detection of influential observations can now be applied to each of the two regressions. In particular, Table 2 reports observations with the five largest values on EIC in the second regression, i.e., an assessment of the observation's influence on the estimate of φ . Obviously, observation 5 has an unusually large influence on the estimate of φ . For illustrative purposes we estimate the model with the data point $(x_{1,5}^*, x_{2,5}^*, y_5^*)$ deleted in the first regression and the data point (\hat{u}_4, \hat{u}_5) deleted in the second regression. The results obtained are reported in the last two rows of Table 1. Estimates of the elasticities are still as expected, although they differ from the complete data estimates. In particular, the income elasticity estimates differ greatly by showing a 46% increase.

This analysis suggests that the observations from 1949 and 1950 obey a different dynamic behavior than the other observa-

tions. Whether this is due to the data collection (changes in definitions, usually large measurement errors, etc.) or to an overly restrictive model remains to be explained. □

Example 4.2. This example considers the problem of tracing observations that highly influence the condition number of a data matrix Z of size $n \times p$. Writing the eigenvalues $l_1 > l_2 > \dots > l_p$ of $Z^T Z$ as functionals of F_n , $l_1 = \lambda_1(F_n)$, \dots , $l_p = \lambda_p(F_n)$, the condition number k of Z is the functional

$$k = \kappa(F_n) = \{\lambda_1(F_n)/\lambda_p(F_n)\}^{1/2}$$

and the influence function of κ is given by

$$\text{IF}(z; F, \kappa) = \frac{\kappa}{2} \left\{ \frac{(\gamma_1^T z)^2}{\lambda_1} - \frac{(\gamma_p^T z)^2}{\lambda_p} \right\}$$

where γ_1 and γ_p are the population eigenvectors associated to the population eigenvalues λ_1 and λ_p , respectively (Nyquist 1988). A measure of the relative change of the condition number when adding one observation z to a very large sample would then be

$$H(z) = \frac{|\text{IF}(z; F, \kappa)|}{\kappa}.$$

A finite sample approximation to $H(z)$ is obtained if the sample influence function is substituted for the influence function. The resulting diagnostic

$$H_i^s = (n - 1) \frac{|k - k_{(i)}|}{k}$$

Table 2. Five largest values of EIC when \hat{u}_t is regressed on \hat{u}_{t-1}

t	EIC
5	0.158
6	-0.089
7	-0.057
16	0.067
18	0.051

has a nice interpretation as the relative change in the condition number due to the deletion of the i th observation. Here, $k_{(i)}$ is the condition number of the data matrix with z_i removed. This measure is suggested by Hadi (1988) and discussed in Chatterjee and Hadi (1988). Using the empirical influence function we alternatively obtain

$$H_i^E = \frac{1}{2} \left| \frac{(g_1^T z_i)^2}{l_1} - \frac{(g_p^T z_i)^2}{l_p} \right|$$

where g_1 and g_p are the observed sample eigenvectors corresponding to l_1 and l_p , respectively. An immediate advantage of H_i^E over H_i^S is that H_i^E is much easier to compute. Computation of H_i^S requires computation of eigenvalues of $n + 1$ matrices, each of size $p \times p$, while it suffices to compute the largest and the smallest eigenvalue and corresponding eigenvalues of $Z^T Z$ when computing H_i^E . Note however, that Hadi (1988) also suggests a quite accurate approximation to H_i^S which requires only computation of eigenvalues of $Z^T Z$ and has therefore the same computational advantage as H_i^E . □

More generally, define

$$t_{(i)}(w) = t(z_1, \dots, w z_i, \dots, z_n)$$

so that $t_{(i)}(1) = t$ and $t_{(i)}(0) = t_{(i)}$. This definition allows assessing the effect of infinitesimal perturbations of z_i by differentiation of $t_{(i)}(w)$ with respect to w , thus defining the empirical sensitivity function

$$\begin{aligned} \text{ESF}(z_i, w, t) \\ = \lim_{s \rightarrow 0} \{t_{(i)}(w + s) - t_{(i)}(w)\}/s \end{aligned}$$

provided the limit exists. It is here interesting to note that

$$\begin{aligned} \text{ESF}(z_i, 1, t) \\ = \lim_{s \rightarrow 0} \left\{ T \left(\frac{n}{n + s} F_n \right. \right. \end{aligned}$$

$$\begin{aligned} & \left. + \frac{s}{n + s} \Delta_i \right) - T(F_n) \Big\} / s \\ & = \lim_{r \rightarrow 0} (1 - r)n^{-1} [T\{(1 - r)F_n \\ & \quad + r\Delta_i\} - T(F_n)]/r \\ & = n^{-1} \text{EIC}(z_i; t) \end{aligned}$$

where Δ_i is the distribution function that assigns pointmass 1 at z_i . Thus, $\text{ESF}(z_i, 1, t)$ is related to the empirical influence function. $\text{ESF}(z_i, 1, t)$ describes local changes in t at the fitted model while $\text{ESF}(z_i, 0, t)$ describes local changes in t after the i th observation has been removed. $\text{ESF}(z_i, 0, t)$ has as yet been little discussed in the literature. Its properties and its potential use as an assessment of a statistic's sensitivity to data remain to be clarified. It should also be noted that, due to the mean value theorem, we have that

$$\begin{aligned} (n - 1)^{-1} \text{SIC}(z_i; t) &= \frac{t_{(i)}(1) - t_{(i)}(0)}{1 - 0} \\ &= \text{ESF}(z_i, \bar{w}; t) \end{aligned}$$

for some \bar{w} , $0 < \bar{w} < 1$. That is, the sample influence curve is related to the derivative of $t_{(i)}(w)$ evaluated at a point \bar{w} lying inside the unit interval. Furthermore, SIC can be interpreted as a compromise between the empirical influence function (equal to $n\text{ESF}(z_i, 1; t)$) and $(n - 1)\text{ESF}(z_i, 0; t)$.

Assessing the effects on estimates is perhaps the most common type of analysis of sensitivity to particular data points. It should, however, be emphasized that it may be relevant to consider other aspects of an empirical study. In fact, the sensitivity of the condition number, discussed in Example 4.2, is an illustration where the conditions for a regression analysis rather than particular estimates of regression parameters are considered.

A very interesting approach to assessing sensitivity to data is that of local influence

provided by Cook (1986). Of key importance in this approach is the curvature of the likelihood displacement defined as

$$LD(w) = 2\{L(\hat{\theta}) - L(\hat{\theta}_w)\}$$

where $L(\theta)$ is the log-likelihood function, $\hat{\theta} = \hat{\theta}_{w=1}$, and $\hat{\theta}_w$ is the maximum likelihood estimate of the model parameter θ when the i th observation has weight w . By considering the log-likelihood function as the statistic of interest, in our notations $t_{(i)}(w) = L(\hat{\theta}_w)$, we find that the local influence approach fits into the framework for assessing sensitivity to data presented in this section.

5. Approaches to Limit Model Sensitivity

The two previous sections have been concerned with assessing the sensitivity of results to model assumptions and particular data points, respectively. This section is devoted to a presentation of two different approaches to take care of questionable, influential model assumptions. The first approach, the model formulation approach, is based on an embedding of the model M_0 into a larger class \mathcal{M} such that the model assumption under consideration is weakened in \mathcal{M} , i.e., the model is reformulated to accommodate possible deviations from M_0 . In the second approach, the sensitivity function approach, the model M_0 is kept unchanged, but the inference technique is constructed to have a bounded sensitivity function.

5.1. Model formulation approach

In this approach the approximative nature of the model is emphasized by making the model larger. If \mathcal{M} itself constitutes a (augmented) parametric model, the members in \mathcal{M} being indexed by a parameter ϕ , then θ and ϕ can be estimated simultaneously usually yielding an estimator of θ that is less sensitive to the model assumption under

consideration. In small samples this approach would of course lead to an increased variance. However, this loss of precision should be contrasted to the possible deterioration of results that can appear when using the smaller model M_0 .

Methods derived for parametric inference are usually not applicable when the augmented model is not parameterized. A reason for this is that some principles of inference cannot be interpreted in the augmented model. The variance of a statistic, for example, may depend on the distribution from which the sample is drawn. If the distribution is not known (up to a finite dimensional parameter) the notion of minimum variance for an estimator loses its meaning. However, it is often the case that a slight modification of the principles of inference, essentially retaining their interpretation, admit application to the augmented model, and hence, new inference methods with desirable properties can be derived.

As an example, we consider the case where an assumption about normally distributed error terms in a linear regression model is put into question. The distributional assumption is then replaced by an assumption that the error distribution belongs to a certain class \mathcal{F} of distributions, in which the normal distribution is a particular member. Examples of classes of distributions whose members cannot be indexed by a finite dimensional parameter include the Levy neighborhood and the contamination neighborhood. For two probability distribution functions F and G define a distance function $\rho(F, G)$ as the infimum of all $h > 0$ such that

$$F(u - h) - h \leq G(u) \leq F(u + h) + h$$

for all u in the sample space. Then, ρ defines a metric in the space of probability distributions. The Levy neighborhood to the

normal distribution is defined by

$$\mathcal{F} = \{F; \rho(F, \Phi) \leq \lambda\}$$

for a fixed and prespecified $\lambda > 0$, where Φ is the probability distribution function for the normal distribution. Since a sequence $\{F_n\}$ of probability distributions converges to a probability distribution F if and only if $\rho(F_n, F) \rightarrow 0$ (see e.g., Feller 1971, ch. 8) the use of Levy neighborhoods seems to be an attractive alternative to an assumption about normality based on Central Limit Theorem arguments.

The contamination neighborhood to the normal distribution is defined by

$$\mathcal{F} = \{F; F(u) = (1 - \lambda)\Phi(u) + \lambda G(u)\}$$

for a fixed and prespecified λ ($0 < \lambda < 1$), where G is any distribution function symmetric about zero. An interpretation of this class of distributions is that the probability of obtaining an observation from the normal distribution equals $1 - \lambda$, while λ is the probability of obtaining a "wild" observation.

Usually no estimator dominates all other estimators in terms of variance, when the distribution is varied over the members in a Levy or a contamination neighborhood. Thus, it may happen that the smallest possible variance is attained at different estimators when the distribution is varied. Hence, the principle of selecting the estimator that gives the smallest variance is not applicable when it is not known from which distribution in \mathcal{F} the observations are drawn. The notion of variance can, however, be extended to a notion of sup variance, the supremum of an estimator's variance where the supremum is taken over all distributions in \mathcal{F} . The principle of inference is accordingly changed to one of selecting the estimator that gives the smallest sup variance. Alternatively, the choice of estimator can be formulated in a decision theoretical

framework. The set of available actions to the analyst (decision maker) is the set of all estimators in some specified class \mathcal{J} of estimators and the set of possible states of Nature is the class \mathcal{F} of possible error distributions. The loss received if the analyst selects the estimator $T \in \mathcal{J}$ and the Nature selects the distribution $F \in \mathcal{F}$ is the (asymptotic) variance $V(T, F)$. An application of the minimax principle as the decision rule implies that the estimator that minimizes (over estimators in \mathcal{J}) the maximal (over distributions in \mathcal{F}) (asymptotic) variance is to be selected. This strategy was originally suggested by Huber (1964).

It is here interesting to note that if \mathcal{J} is the class of M -estimators and if \mathcal{F} is a contamination neighborhood to the normal distribution, the minimax chosen estimator $\hat{\theta}$ of the symmetry point in a univariate sample is the Huber estimator, implicitly defined by

$$\sum_{i=1}^n \psi_c \left(\frac{y_i - \hat{\theta}}{\sigma} \right) = 0$$

where $\psi_c(s) = \max(-c, \min(c, s))$; σ is the standard deviation; and $c > 0$ is a constant related to λ , the probability of obtaining an observation from the contamination distribution G . A small value on λ corresponds to a large value on c and vice versa. In this estimator observations in the range $(\hat{\theta} - \sigma c, \hat{\theta} + \sigma c)$ are considered as being drawn from the normal distribution and are kept unchanged, while observations outside this range are moved to $\hat{\theta} - \sigma c$ or $\hat{\theta} + \sigma c$ depending on the size of the observation. After this conditional transformation of observations $\hat{\theta}$ is computed as the ordinary sample mean.

Unfortunately, only a few theoretical results are known for the Levy neighborhood. In a recent article Collins and Wiens (1989) summarize and extend the minimax

theory for estimation of a location parameter. Since its potential use in model building is evident from a subject matter point of view, it is important to continue the research about the Levy neighborhood in order to clarify its implications.

The theory for this approach is originally derived for the case of unknown error distributions. Its extension to cover problems with other model assumptions seems, however, to be straightforward in many cases. Further research to formulate the details is therefore welcome.

5.2. Sensitivity function approach

The technique used for statistical inference in a particular application is often selected according to some principles of inference. Many of these principles are formulated as optimality conditions in different respects. For example, in the traditional theory of hypothesis testing, the test statistic used is selected so as to maximize power under the alternative hypothesis, and, not to forget, given the remaining assumptions in the model. When estimating parameters, consistency is often of primary concern. Additionally, in many applications unbiasedness and minimum variance are desirable properties. As a consequence, estimators that are consistent, unbiased and, in combination with that, have the smallest variance given a particular model are developed.

It should now be recognized that the derived techniques are optimal only if the model, including the auxiliary assumptions, is true. Generally, there is no guarantee that a technique which is optimal with respect to a set of principles of inference and under a model still is optimal if an auxiliary assumption in the model is violated. This is the starting point for the second approach to take care of questionable influential model assumptions. In this approach the principles

of inference selected are supplemented by a condition about a bounded sensitivity function. When estimating parameters then, we rather look for an estimator that minimizes variance among estimators that are consistent, unbiased, and have bounded sensitivity functions (with respect to the model assumptions from which we wish to bound the influence). Hence, in this approach the approximate nature of the stated model is emphasized in the principles of inference used and reflected in a more cautious selection of a statistical technique.

As an example we consider the errors-in-variables model discussed in Example 3.2. Under this model (assuming no measurement errors) the least squares estimator is optimal according to several sets of principles of statistical inference. However, the bias of the least squares estimator is sensitive to the appearance of measurement errors in the explanatory variable. It would therefore be interesting to contrast the least squares estimator to an estimator that is (i) consistent, (ii) has bounded sensitivity with respect to ϕ , and (iii) has the highest efficiency within a specified class of estimators. Unfortunately, such an estimator has not yet been developed. Research in this field is therefore warmly welcome from a sensitivity analysis point of view.

One of the approaches to robust estimation (often referred to as the influence function approach) is to use estimators with bounded influence functions. An optimally B-robust estimator (Hampel et al. 1986) is defined as an estimator that is (i) consistent, (ii) has a specified finite value on the gross error sensitivity $\gamma^*(T, F_0) = \sup_z |IF(z; T, F_0)|$, and (iii) has the highest efficiency within a specified class of estimators. In the class of M-estimators, the optimally B-robust estimator of the expectation μ in a $N(\mu, \sigma)$ distribution is the Huber estimator defined in Section 5.1, but in this approach the con-

stant c is related to a prespecified gross error sensitivity. A large value on $\gamma^*(T, F_0)$ corresponds here to a large value on c and vice versa.

6. Conclusions

Results from an empirical study depend on data, the statistical model, and the statistical techniques used. The statistical techniques are, in turn, constructed according to the model and a set of principles of statistical inference. It has been argued that statistical models used in empirical studies are ultimately false. Therefore, the techniques selected for the statistical inference often work under other conditions than those assumed, and consequently, a new source of uncertainty is introduced. Note here that this uncertainty is distinguished from, e.g., sampling error, which traditionally is the error of primary concern in the statistical literature and in empirical studies. Note also that the uncertainty stems from the application of the principles of statistical inference and not the principles themselves.

To assert that the results from an empirical study are sensitive to a particular auxiliary model assumption is not equivalent to asserting that the assumption is not satisfied. It claims only that a small departure from the assumption may drastically change the results and eventually hide a parent pattern. An empirical study cannot be dismissed only because its results are found to be sensitive in some respect. However, the appearance of sensitive results should lead to an increased caution in forming conclusions, interpretations, policy recommendations, etc., from the study.

The discussion in this paper has as its primary concern to emphasize the importance of further research within the area of sensitivity analysis. There is certainly a need

for further theoretical development in order to systematize the concepts of sensitivity analysis and to clarify the relations between the concepts. There is also a need to practice sensitivity analysis in applications in order to increase experience and to implement sensitivity analysis as a step that is routinely taken in empirical studies.

7. References

- Atkinson, A.C. (1985). *Plots, Transformations, and Regression*. Oxford: Oxford University Press.
- Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley.
- Box, G.E.P. (1979). Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics*, ed. R.L. Launer and G.N. Wilkinson, New York: Academic Press.
- Box, G.E.P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society, Ser. A*, 143, 383–430.
- Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression*. New York: John Wiley.
- Cook, R.D. (1986). Assessment of Local Influence. *Journal of the Royal Statistical Society, Ser. B*, 48, 133–169.
- Cook, R.D. and Wang, P.C. (1983). Transformations and Influential Cases in Linear Regression. *Technometrics*, 25, 337–343.
- Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.
- Collins, J. and Wiens, D. (1989). Minimax Properties of M -, R - and L -Estimators of Location in Levy Neighbourhoods. *Annals of Statistics*, 17, 327–336.
- Coursey, D. and Nyquist, H. (1988).

- Applications of Robust Estimation Techniques in Demand Analysis. *Applied Economics*, 20, 595–610.
- Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. CBMS Regional Conference Series in Applied Mathematics 38, Society for Industrial and Applied Mathematics, Philadelphia.
- Feller, W. (1971). An Introduction to Probability Theory and Its Applications, 2nd ed. New York: John Wiley.
- Hadi, A.S. (1988). Diagnosing Collinearity – Influential Observations. *Computational Statistics and Data Analysis*, 7, 143–159.
- Hadi, A.S. and Wells, M.T. (1990). Assessing the Effects of Multiple Rows on the Condition Index of a Matrix. *Journal of the American Statistical Association*, 85, 786–792.
- Hampel, F.R. (1968). Contributions to the Theory of Robust Estimation. Ph.D. thesis. University of California, Berkeley.
- Hampel, F.R. (1974). The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, 69, 383–393.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics*. New York: John Wiley.
- Huber, P.J. (1964). Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics*, 35, 73–101.
- Huber, P.J. (1981). *Robust Statistics*. New York: John Wiley.
- Kelsey, J., Holford, T., White, C., Mayer, E., Kilty, S., and Acheson, R. (1978). Oral Contraceptives and Breast Disease. *American Journal of Epidemiology*, 107, 236–244.
- King, M.L. and Giles, D.E.A. (1987). *Specification Analysis in the Linear Model*. London: Routledge and Kegan Paul.
- Mallows, C.L. (1975). On Some Topics in Robustness. Unpublished Bell Telephone Laboratories Report, Murray Hill, N.J.
- Mason, R.L. and Gunst, R.F. (1985). Outlier-Induced Collinearities. *Technometrics*, 27, 401–407.
- Nyquist, H. (1988). On Diagnosing Collinearity – Influential Points in Linear Regression. Discussion paper, Department of Statistics, University of Umeå.
- Pregibon, D. (1981). Logistic Regression Diagnostics. *Annals of Statistics*, 9, 705–724.
- Rosenbaum, P.R. (1989). Sensitivity Analysis for Matched Observational Studies with Many Ordered Treatments. *Scandinavian Journal of Statistics*, 16, 227–236.
- Rousseeuw, P.J. (1981). A New Infinitesimal Approach to Robust Estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 56, 127–132.
- Rousseeuw, P.J. (1982). Most Robust M-estimators in the Infinitesimal Sense. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61, 541–555.
- Wang, S.-G. and Nyquist, H. (1991). Effects on the Eigenstructure of a Data Matrix when Deleting an Observation. *Computational Statistics and Data Analysis*, 11, 179–188.

Received May 1989
Revised March 1992