

When are Inferences too Fragile to be Believed?

John Aldrich

Economics Division
School of Social Sciences
University of Southampton
Southampton
SO17 1BJ
UK

Fax (0)(+44) 23 80593858
e-mail: john.aldrich@soton.ac.uk

Abstract

The use of sensitivity analysis is routine in some fields of empirical econometrics, although econometric theorists have generally taken a critical attitude towards it. This paper presents a framework in which arguments for and against such analysis can be evaluated. It shows that sensitivity is not necessarily a bad nor sturdiness necessarily a good thing.

Acknowledgments

Some of the ideas in this paper came out of joint research I did with Janne Rayner about ten years ago. The present paper is a revised version of one given at the Macromodels 2003 conference. I am grateful to David Hendry and Søren Johansen who were there and made comments.

1 Introduction

Sensitivity analysis is a common response to the existence of several inferences on the same issue, e.g. estimates of the returns to schooling, and this paper considers its rationale. Sensitivity analysis can mean many things and I should first explain how it will be understood here. *Sensitivity* will mean the sensitivity of an inference from a *given* data set to change of assumption, where assumption will generally mean model specification. A *fragile* conclusion or inference is one that does not survive a change of assumption. Sensitivity analysis will be the investigation of the fragility status of inferences with a view to condemning fragile inferences and commending sturdy–non-fragile–ones.

In the 25 years since Leamer brought sensitivity analysis to the fore in his *Specification Searches* (1978) modellers have debated it on at least three occasions—in *Econometric Reviews* (1982), in the *American Economic Review* (1983-6) and in the *Scandinavian Journal of Economics* (1991). There was also a mini-symposium on “replication in economics” in *History of Political Economy* (1992). In some areas of empirical economics, e.g. in growth regressions (see Levine & Renelt (1992) and Temple (1999)) or VAR analysis (see Faust (1998)) sensitivity analysis is now well-established. Yet for all this practical experience and reflection the basic issue remains obscure: why should fragility tell against an inference?

The proponents and critics of sensitivity analysis often seem to be talking past one other. Critics may wonder at the yearning for sturdy inferences when new assumptions/specifications are introduced because the old are thought to lead to erroneous inferences—the aim is to change the inferences! Proponents, like Cooley and LeRoy (1981, p. 827), seem to start from entirely different presuppositions:

[A]ssume that it is concluded that the parameter restrictions implied by the theory to not appear to be satisfied in the majority of the conducted regression tests. In that case one of two conclusions *must follow*. The first is *of course* that the theory is incorrect. The second possible explanation ... is that the least squares projections do not provide a suitable analogue of the controlled experiments of the natural sciences

The phrases I have italicised—*must follow* and *of course*—indicate that these authors consider they are in the realm of the obvious. Yet the critic’s response to the “majority” of tests is—so what?

It is uncontroversial that there are situations in which fragility and its converse, sturdiness, matter. In the “no-data” situations of economic theory conclusion are generated from (uncertain) assumptions alone and the investigation of fragility is routine. The situations envisaged in this paper involve conclusions based on a combination of assumptions and data. In such “with-data” situations the role of sensitivity analysis is more debateable.

The aim of the present paper is to identify the presuppositions of both the supporters and critics of sensitivity analysis to show why their positions seem so obviously right or obviously wrong. Some of the points have been made before but these are easily lost because of the “occasional” nature of the contributions and because the debates were mostly over a particular form of sensitivity analysis—Leamer’s extreme bounds analysis—applied to a particular data set.

§2 provides the basis for a rational reconstruction of the arguments and procedures that have appeared in the literature, though, as with any such reconstruction, there is a danger that the point has been missed. In later sections arguments for and against sensitivity are presented in this framework. Specially interesting are the fragile inferences that will embarrass nobody (§7) and the sturdy inferences that embarrass everybody (§8).

2 Framework and inequalities

Statistical theory admits several reasons why we might consider more than one inference on the same issue, say estimates of the same thing. Outside the tradition of Fisherian estimation theory which seeks a method of estimation “uniquely superior to all possible alternatives”—Fisher (1950, 10.308a)—classical decision theory makes the inference depend on the loss function and Bayesian decision theory makes it depend on the prior as well. Thus a researcher may present different estimates to convince readers with different losses and priors, or even readers of different inference persuasions. The present analysis will not be concerned with these matters.

A Bayesian framework seems most convenient for formulating the issues. However, although I use a Bayesian framework for considering econometric arguments, the analysis is more Bayesian philosophy of science, as in Howson & Urbach (1989), than Bayesian econometrics. Assumptions take many forms. They may concern model specifications, values of parameters, values of future variables, etc. The typical “assumption” in the sensitivity literature and in the discussion below is a regression model with a particular selection of regressors. Leamer (1978) and the contributors to Kadane (1984) treat robustness of inference to changes in prior. Box & Tiao (1982) treat robustness to changes in the model or likelihood. The formalism here will follow Box & Tiao but the argument can be recast in terms of robustness to change of prior. The development of inequalities most resembles the argument of Manski (1995).

The output of inference is a conclusion and the inputs data, assumptions and prior beliefs. Conclusions also take many forms. In Leamer’s (1978, 1983, etc.) extreme bounds analysis the conclusion is a point estimate of a scalar focus parameter, like the returns to years of schooling. The conclusion for Pötzelberger & Polasek (1991) is a highest probability region for a parameter value. The conclusion of the present section is that some statement, S , is true with probability exceeding some value. The formulae are quite general but the most natural interpretation is that S is an interval statement about a scalar parameter θ and the conclusion is that $P(S|y) \geq c$. Here S states that $\theta \in I$ where I is a specified interval; ideally the interval is small and c is large.

Sensitivity analysis—the investigation of the sensitivity of an inference to change of specification or assumption—exists in relation to, and reaction against, two other forms of inference. In the first, the inference is based on a single specification, while in the second, alternative specifications are entertained *and* appraised. In sensitivity analysis different specifications are entertained but they are not appraised.

The first form of inference is based on entertaining a single assumption, model or specification, A say. The conclusion is based on the quantity $P(S|A, y)$. Although the assumption is chosen because its prior probability, $P(A)$, is greatest, this prior probability is not unity and it is generally agreed that it is usually foolish to base inferences on a single conditional

analysis when the underlying A may well be false. It may be that the final inference rests on a single conditional analysis but only after the assumption that generates it has been evaluated against other assumptions.

Suppose A_1, \dots, A_m are the alternative (sets of) assumptions that can be entertained. (An obvious change in notation could accommodate a continuum of assumptions, as when the assumptions concern the value of a parameter—cf. Box & Tiao’s (1972, p. 164) treatment of non-normality.) In a full Bayesian analysis assumption uncertainty would be managed by combining the different conditional analyses according to the law of total probability

$$P(S|y) = \sum_{i=1}^m P(S|A_i, y)P(A_i|y), \quad (2.1)$$

assuming that one of the assumptions is actually true, i.e. that

$$\sum_{i=1}^m P(A_i|y) = 1.$$

The weights $P(A_i|y)$ are complicated objects. These posterior probabilities depend on the density of y given by, say, $f((y|A_i, \theta, \xi)$, the prior distribution of the parameters θ and ξ (the nuisance parameters that invariably appear in these models) and the prior probabilities of the assumptions $P(A_i)$. The details will not be needed here but can be found in Zellner (1971)

Averaging across assumptions as in (2.1) is rare and it is much more usual to investigate the plausibility of the assumptions via $P(A_i|y)$ and if possible find one with a value close enough to unity so that

$$P(S|y) = \sum P(S|A_i, y)P(A_i|y) \approx P(S|A_1, y). \quad (2.2)$$

This scheme covers both informal specification searching using diagnostic tests and organised specification searching such as that advocated by Hendry (1995).

Two inequalities follow from (2.1). The inequality for *sturdy* inference states

$$\text{If } P(S|A_i, y) \geq c \text{ for all } i, \text{ then } P(S|y) \geq c. \quad (2.3)$$

If, on *all* possible assumptions, the statement is probably true, then the statement is probably true. Of course there is the proviso that it is known that one of the assumptions is actually true, although this proviso could be dispensed with and $P(S|y)$ replaced by $P(S|\bigcup_1^m A_i, y)$. Perhaps *all* assumptions similar to Bishop Ussher’s dated the beginning of the world to around 4004 BC but such sturdiness does not make the dating correct.

An inequality for *fragile* inference also follows from (2.1). This states that the value of $P(S|y)$ lies between the smallest and the largest of the $P(S|A_i, y)$:

$$\min_i \{P(S|A_i, y)\} \leq P(S|y) \leq \max_i \{P(S|A_i, y)\}.. \quad (2.4)$$

If these bounds are far apart, the inference is fragile and Leamer’s dictum applies, “when an incredibly narrow set of assumptions is required to produce a usefully narrow set of conclusions, inferences from the given data set are ... too fragile to be believed.”

3 Sensitivity analysis

Sensitivity analysis compares inferences from different assumptions, the $P(S|A_i, y)$, *without* using the posterior probabilities of the assumptions, the $P(A_i|y)$. Inferences based on the inequalities (2.3) and (2.4) are weaker than inferences based on (2.1) or (2.2) but if the posterior probabilities are either not available or are such that the stronger forms of inference are not possible then the inequalities may be the best that can be done.

A first possibility is that the $P(A_i|y)$ values do not point to any clear winner among the assumptions; the use of (2.2) is blocked, though not that of (2.1). One possibility is that the data examined has turned out to be indecisive about the truth of the assumptions: the $P(A_i|y)$ have been calculated but none has been found to be overwhelmingly large. Here the data has turned out to be uninformative about the assumptions but another possibility is that *any* data would be uninformative. This is the case of observationally equivalent assumptions. It is an important case and is discussed in §5 below. In this section we will consider the possibility that the $P(A_i|y)$ are *not* available, and so inference based on (2.2) or (2.1) cannot be done.

Cooley (1982) thought that there was no effective way of finding the $P(A_i|y)$. This was a limitation of existing statistical theory rather than something inherent in the problem and the critics argued that there numerous situations in which assumptions can be eliminated, i.e. their posterior probabilities can be shown to be negligible, thus making possible inference around (2.2).

Leamer (e.g. 1983 or 1986) paid particular attention to situations where disagreement is deeply entrenched—e.g. between “bleeding heart liberals” and “right wingers” on the causes of crime or between Keynesians and Monetarists on the effectiveness of fiscal policy. The parties to radical disagreement will not agree about the values to be given to the posterior probabilities $P(A_i|y)$ because they cannot agree on the prior probabilities $P(A_i)$. One could proceed to a higher level where the impartial spectator imposes a super-prior on the conflicting priors but this prior will not recommend itself to the parties in conflict.

It is in these circumstances that Leamer recommends sensitivity analysis. However sensitivity analysis is no answer to Patinkin’s lament of the 1960s (quoted in Mayer (1980))

I will begin to believe in economics as a science when out of Yale there comes an empirical Ph.D. thesis demonstrating the supremacy of monetary policy in some historical episode and out of Chicago, one demonstrating the supremacy of fiscal policy.

For sturdy inferences are only possible on side-issues. A flat earther and a round earther, asked the way to the city centre, may well give the same directions.

4 Analogies?

I have represented sensitivity analysis in a particular way but perhaps there are better ways of thinking about it. I will consider two based on what I consider a *false* analogies and hence a dangerous fund of intuition. The first analogy has been discussed by Cartwright (1992) whose approach is from the philosophy of science literature.

There is a principle that we believe more strongly in a statement when independent bits of evidence supporting it accumulate; the principle is familiar from statistical inference and it

is discussed in works on induction such as Keynes (1921). The principle may be formalised as follows. A piece of evidence y supports a statement S rather than its negation \bar{S} when

$$P(S|y) > P(S), \text{ i.e. when } P(y|S) > P(y).$$

Pieces of evidence y_1 and y_2 are independent in relation to S and \bar{S} when

$$\begin{aligned} P(y_1 \cap y_2 | S) &= P(y_1 | S)P(y_2 | S) \text{ and} \\ P(y_1 \cap y_2 | \bar{S}) &= P(y_1 | \bar{S})P(y_2 | \bar{S}) \end{aligned}$$

Granted independence and using the rules for Bayesian updating, it follows that if y_1 and y_2 separately support the statement S then together they will support it and

$$P(S|y_1 \cap y_2) > P(S|y_1) > P(S).$$

Of course it is easy to find examples where the independence assumption does *not* hold: 4 heads in 4 tosses is evidence for the double-sidedness of a coin, 4 tails in a row would also be evidence but *together* they rule out double-sidedness; their occurrence would decisively refute that hypothesis. If we put $y_1 = \hat{\theta}_1$ and $y_2 = \hat{\theta}_2$, where the hats denote estimates of θ , from analysing the same data on different assumptions, the obvious presumption is that $\hat{\theta}_1$ and $\hat{\theta}_2$ will *not* be independent; indeed misspecification analysis and the encompassing principle are founded on this presumption. Not only may additional estimates add no support, they may take away support as in the coin example; see §6 for an example. This scheme of accumulating evidence from different statistics could be developed but serious encompassing-like conditional probability calculations would be involved which takes the analysis out of the simple sensitivity analysis set-up.

The second analogy is based on likening the inferences/estimates from different assumptions to noisy messages and to be comforted when the same message comes through regardless of the noise. This analogy can be formalised as the measurement expression of the last. Again assumptions can be found to justify the application of this analogy but they are unlikely to

apply to the (mis-)specifications found in economics.

5 No relevant data

The critics of sensitivity analysis usually take as typical situations in which post-data evaluations of the assumptions, given by $P(A_i|y)$, weed out most specifications as incredible. However the data may be *uninformative* about the assumptions. Already in the Cowles era Koopmans (1952, p. 205) thought sensitivity analysis the only resort in this eventuality: “If doubt remains about a basic specification not subject to conclusive statistical test, the only remaining line of defense is a study of the effect on policy conclusions of presumably possible degrees of departure from the specification in question.”

In theoretical debate conclusions are generated from (uncertain) assumptions alone and the investigation of fragility, i.e. sensitivity analysis, is routine. Of course the possibilities of inference can be described in the framework of the equations and inequalities of §2 by rubbing out y . Policy simulation is a half-way house between such no-data situations and the no-relevant data situations. Here sensitivity analysis is also routine.

Suppose the model to be used is

$$y_t = \delta x_t + u_t, \quad u_t \sim IN(0, \sigma^2)$$

and there is data to estimate the parameters. The inference concerns $\theta = \delta x^*$ the expected value of y associated with policy value x^* . To make this inference, using an estimate $\hat{\delta}$ we need a value for the policy x^* . Suppose there are two equally reasonable values x_1^* and x_2^* (corresponding to A_1 and A_2) with associated predictions $\hat{\theta}_1$ and $\hat{\theta}_2$. If these are close together, there is no difficulty in choosing one or the other or some combination of them. If they are not close together, there is a problem of choice. To resolve this problem one might investigate the $P(A_i)$ or one might report the two estimates and comment on the fragility.

If we call the quantity θ *unidentified* we can tap into a major theme in econometric theory. Each of the ‘classical’ models of econometrics—the linear regression model, the errors in variables model and the simultaneous equations model—generates an “identification problem”

although there is only a substantial literature on sensitivity analysis for the errors in variables model; see for example Klepper & Leamer (1984). More recently, however, the VAR modelling movement, which generates empirically indistinguishable models almost as a matter of principle, has found a role for sensitivity analysis; see Sims (1981) and Faust (1998).

Leamer (1978, chapter 5) presents extreme bounds analysis in close proximity to his discussion of multicollinearity and identification, suggesting the following scenario. There is a set of potential regressors including X_1 and X_2 as well as x . The regression equation containing all the variables is not identified. However two plausible sets of identifying restrictions produce the estimable models

$$A_1 : y = \theta x + X_1\beta_1 + u_1$$

$$A_2 : y = \theta x + X_2\beta_2 + u_2$$

where the matrices (x, X_1) and (x, X_2) have full rank. If the columns of these matrices span the same space, the two models will be empirically indistinguishable; see Rayner & Aldrich (1992). Estimates of θ can be obtained but there is no way of using the data to choose between the models that underpin these estimates. The spectator can investigate whether there is a consensus regarding the “effect of x on y .” The spectator is satisfied if the estimates of θ are similar. If they are very different, he will agree with Leamer & Leonard’s (1983, p. 306) comment, “inference from these data are too fragile to be useful.”

Cooley & LeRoy (1981) proposed using sensitivity analysis where the search for support for a theory has produced a proliferation of regressions. The theory, which relates y to x under some unspecified *ceteris paribus* assumptions, makes a prediction about the value of θ , e.g. that the effect of the interest rate on the demand for money is negative. The search is for a specification which, when fitted to data, produces an estimate satisfying this prediction. Researchers work through a number of models which make different assumptions about the importance of other variables and choose an A_i for which $P(S|A_i, y)$ is respectably large, i.e. the restriction is satisfied. Cooley & LeRoy (p. 825) argue against this way of proceeding:

if the restrictions indicated by the theory are satisfied in some projections but not

in others that have equal claim to represent implications of the theory, one cannot conclude that the theory has been confirmed.

In other words, it not sufficient that, for some choices of i , $P(S|A_i, y)$ should be appreciable.

Cooley & LeRoy (p. 827) elaborate this point with the argument quoted in the Introduction with its reference to the “majority ” of tests. A majority view promises a resolution of the problem of what to believe when there is incomplete agreement. Thus in (2.1)

$$P(S|y) = \sum P(S|A_i, y)P(A_i|y)$$

suppose $P(A_i|y)$ is the same for all i (“equal claims”) and suppose $P(S|A_i, y)$ is unity for a majority of assumptions and zero for the remainder. With these assumptions

$$P(S|y) > \frac{1}{2} > P(\bar{S}|y).$$

This may be an over-literal transcription of the argument of Cooley & LeRoy but by juggling with different majority rules it is possible to consider how big a majority is needed to compensate for values of $P(S|A_i, y)$ that are not unity.

In interpreting “equal claims” as the condition that $P(A_i|y)$ is the same for all assumptions I may be misconstruing the authors’ meaning. A more natural assumption is that the prior weights, $P(A_i)$, are equal; the investigator can contemplate these without analysing the data. However the posterior weights are the ones that matter for inference. Cooley and LeRoy’s proposals make most sense for situations in which the alternative specifications are indistinguishable, i.e., $P(A_i) = P(A_i|y)$. Otherwise it is hard to understand their use of “of course”.

6 Asking different questions?

It can be argued that sensitivity analysis is a solution to a non-existent problem. for the conflicting inferences are answers to different questions and so the conflict is only apparent.

Johansen (1991) makes this point for the set-up

$$A_1 : y = \theta x + \xi z + u_1, \quad u_1 \sim N(0, \sigma^2 I)$$

$$A_2 : y = \theta x + u_2, \quad u_2 \sim N(0, \omega^2 I)$$

where it is assumed that x and z are fixed in repeated sampling. He observes that the two θ 's are answers to two “different questions”: one measures a partial effect, the other a total effect.

The population/structure conceptualisation of Koopmans & Reiersøl (1950) is useful here. θ in A_1 is *not* the same population parameter as θ in A_2 in the sense of being the same functional of the joint distribution of x , y and z . But these distinct functionals may represent the effect of x on y in different structural models. Leamer (1978, p. 298) describes two sources of uncertainty about a parameter θ : the “misspecification uncertainty” arising from the existence of more than one algorithm relating θ to a population parameter π and the “sampling uncertainty” associated with the need to estimate π . The reason why there is more than one algorithm relating θ to π is that there is uncertainty about the appropriate model.

Suppose we are looking in the car park for our friend’s new car which is “simply the best.” I look for the fastest car, you for the most exclusive. We each have two answers to the question, what are you looking for? We both answer “our friend’s car” for that is our common answer to the structural question but while the “fastest car” is my answer to the population question, the “most exclusive” is your answer. Even if we agree on which car is the fastest and which the most exclusive our answers to the structural question will be different unless the fastest car is the most exclusive. If there were a third friend with another interpretation of “very special,” which had an equal claim to be the correct one, the majority principle of Cooley and LeRoy might be appropriate. Of course, on seeing the car park full of very special cars we might conclude that inferences from this data set are too fragile to be believed.

Returning to the Johansen example, consider two complete structural models of the three variables y , x and z . The two assumptions lead to two the different ways of estimating θ , the “effect of x on y ”. Suppose A_1 is a block recursive specification with an ordering from x and

z to y

$$y = \theta x + \xi z + u,$$

as part of a complete model of the three variables y , x and z where u , z and x are generated as follows

$$\begin{pmatrix} u \\ z_t \\ x_t \end{pmatrix} \sim IN(0, \Sigma) \text{ where } \Sigma = \begin{pmatrix} \sigma_u^2 & 0 & 0 \\ 0 & \sigma_z^2 & \sigma_{xz} \\ 0 & \sigma_{xz} & \sigma_x^2 \end{pmatrix}.$$

A_2 is another complete system where the ordering is from x to y to z

$$\begin{aligned} y &= \theta x + v_1, \\ z &= \delta y + w \end{aligned}$$

where v , w and x are generated as follows

$$\begin{pmatrix} v_t \\ w_t \\ x_t \end{pmatrix} \sim IN(0, \Omega) \text{ where } \Omega = \begin{pmatrix} \omega_v^2 & \omega_{vw} & 0 \\ \omega_{vw} & \omega_w^2 & 0 \\ 0 & 0 & \omega_x^2 \end{pmatrix}.$$

A_1 and A_2 are observationally equivalent and thus empirically indistinguishable. Although the parameters of both models are identified there are no testable restrictions on the joint distribution of the observables, x , y and z . The population densities or reduced forms are

tri-normal with variance matrices

$$\begin{aligned}
 A_1 : \text{var} \begin{pmatrix} y_t \\ z_t \\ x_t \end{pmatrix} &= \begin{pmatrix} \theta^2 \sigma_x^2 + \xi^2 \sigma_z^2 + 2\theta\xi\sigma_{xz} + \sigma_u^2 & \cdot & \cdot \\ \theta\sigma_{xz} + \xi\sigma_z^2 & \sigma_z^2 & \sigma_{xz} \\ \theta\sigma_z^2 + \xi\sigma_{xz} & \sigma_{xz} & \sigma_x^2 \end{pmatrix}; \\
 A_2 : \text{var} \begin{pmatrix} y_t \\ z_t \\ x_t \end{pmatrix} &= \begin{pmatrix} \theta^2 \omega_x^2 + \omega_w^2 & \cdot & \cdot \\ \delta\theta^2 \omega_x^2 + \delta\omega_w^2 + \omega_{vw} & (\delta\theta)^2 \omega_x^2 + \delta^2 \omega_w^2 + \omega_w^2 + 2\omega_{vw} & \cdot \\ \theta\omega_x^2 & \delta\theta\omega_x^2 & \omega_x^2 \end{pmatrix}.
 \end{aligned}$$

In A_1 the “effect of x on y ”, θ , is represented by the coefficient of x in the conditional expectation of y given x and z because z is a cause to be controlled for; in A_2 it is represented by the coefficient of x in the conditional expectation of y given x alone for it is not appropriate to control for an effect. The data on x , y and z provides no way of choosing between the estimates because there is no way of choosing between the models that justify the estimates.

Although the fact of fragility has no implications for the credibility of the assumptions, it is informative about the world in other respects. In only certain states of the world is consensus possible. That this is one of those states is a testable hypothesis. Although this is not a hypothesis about the truth of either assumption, it can be interpreted in terms of the parameters associated with each assumption. Thus here $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimating the same population quantity when the partial correlation of y and z given x is zero. This is to express the condition in a structurally neutral language: interpreted through A_1 the condition is that $\xi = 0$, interpreted through A_2 it is that $\delta\omega_w^2 = -\omega_{vw}$. In the prediction problem of §4, $\delta x_1^* = \delta x_2^*$ implies $\delta = 0$. Of course we hope that our world is sympathetic to sturdiness.

We can wish that the conclusion is the same whichever of the assumptions is true but there is no reason to suppose that truth manifests itself as a sturdy conclusion. Thinking of the regression model—discussed in the next section—conventional omitted variable analysis shows that the effect on the estimation of the coefficient of x depends on the correlation between the regressors and the true values of the parameters. Sturdiness will only obtain for very particular configurations of these quantities; the most easily understood of these is when x is orthogonal to the other regressors. But of course conclusions based on (2.1), (2.2) or (2.3) are

all equally acceptable provided the assumptions underlying them are valid.

7 Unembarrassingly fragile inferences

In the situations discussed in the last two sections there has been disagreement about structure but none about population. In this section and the next there will be disagreement about both; the assumptions will no longer be observationally equivalent. The posterior probability of an assumption will change according to the success of the predictions it makes. The predictions can be about agreement or disagreement of the inferences.

McAleer, Pagan & Volker (1985), Pagan (1987) and Breusch (1990) describe a situation where the reasonable reaction to an inference's fragility is not to suspend belief but to reject one of conflicting inferences. The fact of fragility can be a guide in choosing the best of the conflicting inferences.

Consider again A_1 and A_2 from §4

$$A_1 : y = \theta x + \xi z + u_1, \quad u_1 \sim N(0, \sigma^2 I)$$

$$A_2 : y = \theta x + u_2, \quad u_2 \sim N(0, \omega^2 I)$$

but where the force of A_2 is that y and z are conditionally independent given x . In this case belief in A_1 and an S based on $\hat{\theta}_1$ is not weakened by finding a very discrepant value for $\hat{\theta}_2$. For such a value would be treated as evidence against A_2 and therefore $\hat{\theta}_2$ would be discounted. On the other hand, if we had started with A_2 and perturbed it to A_1 then we would have revised our belief, *not*, however, to a suspension of belief but to belief based on A_1 .

What is happening is that, under A_2 , the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ are both estimators of the same quantity but, under A_1 , there is no expectation of similarity for $\hat{\theta}_1$ is a more or less biased estimate of θ . Under A_1 agreement *is* possible, when the value of θ is close to 0, but disagreement is unsurprising, indicating only that θ is not close to 0. Here fragility does not justify suspension of belief. Actually one of the fragility measures proposed by Leamer is the Hausman statistic for testing θ is 0.

In terms of (2.1) and the analysis of §2 we find in

$$P(S|y) = P(S|A_1, y)P(A_1|y) + P(S|A_2, y)P(A_2|y)$$

that when A_1 generates an S that θ is not close to 0 the same data will generate a large value of $P(A_1|y)$ and a small value of $P(A_2|y)$.

In Leamer’s dictum, “when an incredibly narrow set of assumptions is required to produce a usefully narrow set of conclusions, inferences from the given data set are ... too fragile to be believed,” the phrase “incredibly narrow” refers to the posterior probabilities. Assumptions that begin as incredibly narrow retain their narrowness but as evidence comes in they may lose their incredibility.

In the case of indistinguishable models there is nothing in the data to move the probability that an assumption is correct but in cases like the present one these probabilities move: fragility does not paralyse inference because there is an assumption—and only one—which predicts it. Similar reasoning suggests some answers to the question, when are inferences too fragile are to be believed? When fragility is expected on all assumptions, fragility does nothing to undermine any of the assumptions and so the disparate inferences stand. When fragility is unexpected on all of the assumptions *all* are undermined. However the set-up of §2 does not admit the possibility that all assumptions are undermined because the data only redistributes the probabilities of the assumptions which must sum to one. Nor does the framework admit the more interesting possibility that inferences can be too sturdy to be believed when sturdiness is unexpected on all the assumptions. This possibility is considered in the next section.

8 Embarrassingly sturdy inferences

The vulnerable assumption behind the sturdiness inequality (2.2) and the case for sturdiness is that one of the specifications being entertained is the true one. A demonstration of sturdiness may only be as good as the coverage of the assumptions. A more interesting possibility of that of over-sturdy inference where the unexpected similarity of the conditional inferences removes the ground for their common conclusion. Mroz (1987, p. 775) finds such an “unexpected

similarity” between the estimates in two labour supply functions; the finding is connected to another of his findings that both functions are misspecified.

To illustrate the possibility that sturdiness can count against specifications consider this simple but contrived set-up

$$A_1 : y = \theta x - \theta z + u_1, u_1 \sim N(0, \sigma^2 I)$$

$$A_2 : y = \theta x + \theta z + u_2, u_2 \sim N(0, \omega^2 I)$$

with x and z fixed in repeated sampling. Also assume that x and z are orthogonal with $\sum x^2 = \sum z^2$.

From the values

$$A_1 : \hat{\theta}_1 = \frac{\sum y(x-z)}{\sum (x-z)^2} = \frac{\sum y(x-z)}{2 \sum x^2}$$

$$A_2 : \hat{\theta}_2 = \frac{\sum y(x+z)}{\sum (x+z)^2} = \frac{\sum y(x+z)}{2 \sum x^2}$$

it is concluded that when A_1 is true, $\hat{\theta}_2$ is expected to be close to zero, whatever the true value of θ and similarly that if A_2 is true, $\hat{\theta}_1$ is expected to be close to zero.

$$E_{A_1} \hat{\theta}_2 = \frac{\theta \sum (x^2 - z^2)}{2 \sum x^2} = 0$$

$$E_{A_2} \hat{\theta}_1 = \frac{\theta \sum (x^2 - z^2)}{2 \sum x^2} = 0$$

“Expected similarity” obtains when θ is close to 0. This is nice because when θ is 0 the two models are the same. “Unexpected similarity” obtains when $\hat{\theta}_1$ and $\hat{\theta}_2$ are close to each other *away from* 0. This is an unexpected outcome on either specification.

The unexpected similarity suggests something is wrong with both specifications and that the range of alternatives be expanded beyond A_1 and A_2 . It does not indicate what would be a better specification and thus what would be a better estimate of θ . However this “unexpected

similarity” would be expected were true specification

$$A_3 : y = \theta x + u_3$$

with θ large and the coefficient of z equal to 0 for

$$E_{A_3} \hat{\theta}_1 = \theta \frac{\sum x(x-z)}{2 \sum x^2} = \frac{\theta}{2}$$

$$E_{A_3} \hat{\theta}_2 = \theta \frac{\sum x(x+z)}{2 \sum x^2} = \frac{\theta}{2}.$$

If A_3 were the true specification, then the best value estimate of θ would be $\hat{\theta}_3$ where

$$\hat{\theta}_3 = \frac{\sum yx}{\sum x^2}$$

which is twice the common value of $\hat{\theta}_1$ and $\hat{\theta}_2$. The sturdy value is not a good guide, being half what it should be.

This example is designed to make the point that similarity may be unexpected and a symptom of misspecification. The formal framework of §2 does not admit the possibility of misspecification: the posterior probabilities $P(A_1|y)$ and $P(A_2|y)$ sum to one if those assumptions are the only ones entertained. But if A_3 is admitted those posterior probabilities will drop away to zero.

9 Conclusion

Sensitivity analysis is an issue on which intuitions run high. The aim here has been to bring intuitions down to earth. The role of sensitivity analysis is secure in no-data situations—there is nothing better. This paper has considered its role in various kinds of with-data situations. One of these situations, is essentially the same as the no-data situation. It is a kind of identification failure and examples of the phenomenon are easy to find in econometrics.

In with-data situations the need for sensitivity analysis is less obvious. By focussing on conditional analyses, information on the plausibility of different specifications/assumptions

does not have to be collected. If the assumptions entertained contain the truth and all assumptions lead to the same conclusion, then sturdiness is a good. If the truth is not among the assumptions entertained, then sturdiness can count for nothing. Disagreement among the inferences—fragility—is no reason for condemning all the inferences and concluding that the truth lies elsewhere. It is a reason for doing more work and is in that respect bad news.

References

- Berger, J. O. (1984) The Robust Bayesian Viewpoint, Part II and pp. 63-104 of Kadane (1984).
- Blanchard O. J. & D. Quah (1989), The Dynamic Effects of Aggregate Demand and Supply Disturbances, *American Economic Review*, **79**, 655-673.
- Box, G. E. P. & G. C. Tiao (1972) *Bayesian Inference in Statistical Analysis*, London: Addison-Wesley.
- Breusch, T. S. (1990) Simplified Extreme Bounds. Chapter 4 and pp. 72-81 of Granger (1990).
- Cartwright, N. (1991) Replicability, Reproducibility, and Robustness—Comments on Harry Collins, *History of Political Economy*, **23**, 143
- Collins, H. (1991) Replicability, Reproducibility, and Robustness, *History of Political Economy*, **23**, 143
- Cooley, T. F. (1982) Specification Analysis with Discriminating Priors: An Application to the Profits Concentration Debate, *Econometric Reviews*, **1**, 97-128.
- & S. F. LeRoy (1981) Identification and Estimation of Money Demand, *American Economic Review*, **71**, 825-844.
- & S. F. LeRoy (1986) What will take the Con out of Econometrics? A Reply to McAleer, Pagan & Volker, *American Economic Review*, **76**, 504-507. Reprinted in Granger (1990).
- Dhrymes, P. (1982) Comment, *Econometric Reviews*, **1**, 129-132.
- Faust, J. (1998), The Robustness of Identified VAR Conclusions About Money, *Carnegie-Rochester Conference on Public Policy*, **49**, 207-244.
- Fisher, R. A. (1950), Author's Note to "On the mathematical foundations of theoretical statistics" in W. A. Shewhart (ed) (1950) *Contributions to Mathematical Statistics*, New York:

Wiley.

Granger, C. W. J. (ed.) (1990) *Modelling Economic Series*, Oxford: University Press.

Hausman, J. (1978) Specification Tests in Econometrics, *Econometrica*, **46**, 1251-1272.

Hendry D.F. (1995), *Dynamic Econometrics*, Oxford: Oxford University Press.

Howson, C. & P. Urbach (1989) *Scientific Reasoning: the Bayesian Approach*, Open Court, La Salle, Ill.

Kadane, J. H. (1984) *Robustness of Bayesian Analyses*, Amsterdam: North-Holland.

Keynes, J. M. (1921) *A Treatise on Probability*, London, Macmillan. Collected Writings Edition, 1973, London, Macmillan for the Royal Economic Society.

Klepper, S. E. & E. E. Leamer (1984) Consistent Sets of Estimates for Regressions with Errors in All Variables, *Econometrica*, **52**, 163-184.

Koopmans, T. C. (1952) Toward Partial Redirection of Econometrics: Comment, *Review of Economics and Statistics*, **34**, 200-205.

Koopmans, T. C. & O. Reiersøl (1950) The Identification of Structural Characteristics, *Annals of Mathematical Statistics*, **21**, 165-181.

Leamer, E. E. (1978) *Specification Searches*, New York: John Wiley.

----- (1983) Let's take the Con out of Econometrics, *American Economic Review*, **73**, 31-44. Reprinted in Granger (1990).

----- (1985) Sensitivity Analysis would Help, *American Economic Review*, **75**, 308-313. Reprinted in Granger (1990).

----- (1991) A Bayesian Perspective on Inference from Macroeconomic Data, with comments by S. Johansen, K. Juselius & S. Yitzhaki, *Scandinavian Journal of Economics*, **93**,

Leamer, E. E. & H. Leonard (1983) Reporting the Fragility of Regression Estimates, *Review of Economics and Statistics*, **65**, 306-317.

Levine, R. & D. Renelt (1992) A Sensitivity Analysis of Cross-Country Growth Regressions, *American Economic Review*, **82**, 942-963.

McAleer, M., A. R. Pagan & P. A. Volker (1985) What will take the Con out of Econometrics? *American Economic Review*, **75**, 293-307. Reprinted in Granger (1990).

Manski, C. F. (1995) *Identification Problems in the Social Sciences*, Cambridge: Harvard

University Press.

Mayer, T. (1980) Economics as a Hard Science: Realistic Goal or Wishful Thinking?, *Economic Inquiry*, **18**, 165-178.

Mroz, T. A. (1987) The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, *Econometrica*, **55**, 765-800.

Pagan, A. R. (1987) Three Econometric Methodologies: A Critical Appraisal, *Journal of Economic Surveys*, **1**, 3-24.

Pötzelberger, K. & W. Polasek (1991) Robust HPD Regions in Bayesian Regression Models, *Econometrica*, **59**, 1581-1590.

Rayner, J. & J. Aldrich (1992) Distinguishability and Identifiability, Southampton University Discussion Paper No. 92.

Sims C. A. (1981) An Autoregressive Index Model for the U.S. 1948-1975, in: J. Kmenta and J.B. Ramsey (eds.) *Large-Scale Macro-Econometric Models*, Amsterdam: North-Holland, 283-327.

----- (1988) Uncertainty across Models, *American Economic Review*, **78**, 163-167.

Temple, J (1999) The New Growth Evidence, *Journal of Economic Literature*, **37**, 112-156.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley.