

When are inferences too fragile to be believed?

John Aldrich

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 appears that sensitivity is not necessarily a bad, nor sturdiness necessarily a good.

Keywords: sensitivity, robustness, fragility

1 INTRODUCTION

Sensitivity analysis is a frequent 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. As sensitivity analysis can mean many things, I should first explain how it will be understood here. ‘Sensitivity’ will mean the sensitivity of an inference from a *given* dataset to change of assumption, where assumption will generally mean model specification. A *fragile* conclusion is one that does not survive a change of assumption; a *sturdy*, or robust, conclusion is one that does. Sensitivity analysis is the investigation of the fragility status of inferences with a view to condemning fragile inferences and commending sturdy ones.

Since Leamer (1978) brought sensitivity analysis to the fore there have been three set-piece debates, in *Econometric Reviews* (1982), in the *American Economic Review* (1983–6) and in the *Scandinavian Journal of Economics* (1991); there has also been a Mini-Symposium on ‘replication in economics’ in *History of Political Economy* (1992). The balance of these discussions seems to have been against sensitivity analysis and yet in some areas of empirical economics, e.g., in growth regressions (see Levine and Renelt 1992 and Temple 1999) or VAR analysis (see Faust 1998), sensitivity analysis is well-established. In the growth literature there has been debate about its appropriateness; see Hoover and Perez (2004) and the literature cited there. Despite all this reflection and practical experience, the basic issue remains

obscure: why should sturdiness support a conclusion and fragility tell against one?

The present paper tries to identify the presuppositions of both the supporters and the critics of sensitivity analysis to discover why their own position should seem obviously right and the opposed position obviously wrong. The paper is *not* about how to do sensitivity analysis or what to do instead of it. Most of the discussions have been about a particular form of sensitivity analysis, Leamer's extreme bounds analysis, and on alternatives to it. The earlier literature produced some general points which I think are worth reiterating. Hopefully the points will be preserved and not lost because of the 'occasional' nature of the earlier contributions.

Section 2 provides the basis for a rational reconstruction of the arguments and procedures that have appeared in the literature. In later sections the framework is used to consider different situations. Among these situations two cases are of special interest: the fragile inferences that give no grounds for concern (section 7) and the sturdy inferences that give every ground for concern (section 8).

2 FRAMEWORK AND INEQUALITIES

One tradition in statistical theory looks for a method of estimation 'uniquely superior to all possible alternatives'—Fisher (1950, 10.308a), but other traditions admit more than one estimate, or more generally, more than one inference on the same issue. 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 in the spirit of Bayesian philosophy of science, as in Howson and Urbach (1989), than of 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 and Tiao (1972) treat robustness to changes in the model or likelihood. The formalism here will follow Box and Tiao but the argument can be recast in terms of robustness to change of prior. The idea of developing inequalities has been used by Manski (1995) in a different but related context. The present argument is perhaps closest to Sims (1988).

The output of inference is a conclusion and the inputs are data, assumptions and prior beliefs. Conclusions can take many forms. In Leamer's (1978, 1983, 1985, 1991) 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 and Polasek (1991) is a highest probability region for a parameter value. The conclusion treated in the present formalization 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 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 and 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:

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

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) can be done (see e.g. Hoeting *et al.* (1999)) but it is rare in econometrics where it is 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 organized specification searching such as that advocated by Hendry (1995) and in mechanized form by Krolzig and Hendry (2001). Hoover and Perez (2004) apply these techniques to cross-country growth regressions.

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 vital proviso that it is known that one of the assumptions is actually true. Perhaps *all* assumptions similar to Bishop Ussher's would date the beginning of the world to around 4004 BC but such sturdiness does not make the dating correct. Formally the proviso could be dispensed with and $P(S|y)$ replaced by $P\left(S \bigg| \bigcup_1^m A_i, y\right)$.

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 seems to apply, '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' (Leamer 1985: 308).

3 SENSITIVITY ANALYSIS

Sensitivity analysis compares inferences from different assumptions, the $P(S|A_i, y)$, *without* assessing the assumptions by using their posterior probabilities, 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 have 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 have 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 sections 5 and 6 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 1985) 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 (Mayer 1980: 166).

For sturdy inferences are only possible on side-issues. I expect a flat earther and a round earther, asked the way to the city centre, would 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 as *false* analogies and hence a dangerous fund of intuition. The first analogy has been discussed by Cartwright (1991) whose approach is from the philosophy of science literature.

There is a principle that we believe more strongly in a statement when independent pieces 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 formalized 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:

$$P(y_1 \cap y_2 | S) = P(y_1 | S)P(y_2 | S) \text{ and}$$

$$P(y_1 \cap y_2 | \bar{S}) = P(y_1 | \bar{S})P(y_2 | \bar{S})$$

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: four heads in four tosses is evidence for the double-sidedness of a coin, four 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 section 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 formalized as the measurement version of the analogy just described. 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: 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'.

Another situation where sensitivity analysis is routine is in making 'projections' based on alternative 'scenarios' or forecasting when assumptions are made about exogenous variables. The uncertainty is in the specification of the values of the exogenous variables. Suppose the model 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 x^* , a future value of x . To make this inference, using an estimate $\hat{\delta}$ we need a value for x^* . Suppose there are two plausible values x_1^* and x_2^* (corresponding to A_1 and A_2) with associated predictions $\hat{\theta}_1$ and $\hat{\theta}_2$. If the predictions are close together, there is no difficulty; either one – or some combination of them – will do. If they are not, there is a problem. 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 describe the quantity θ as *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 and 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, ch. 5) presents extreme bounds analysis in close proximity to his discussion of multicollinearity and identification, suggesting the following situation. 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 and 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 and Leonard's (1983: 306) comment, 'inference from these data are too fragile to be useful'.

In their work illustrating the use of sensitivity analysis Cooley and LeRoy (1981) envisage a situation in which a 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. They criticize researchers who produce evidence for S by working through a number of regression 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. They 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. (Cooley and LeRoy: 825)

In other words, it not sufficient that, for some choices of i , $P(S|A_i, y)$ should be appreciable. When they develop their point they introduce the idea of a majority verdict:

Assume that it is concluded that the parameter restrictions implied by the theory do 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 (827)

It is instructive to put the argument for a majority verdict into the framework of section 2. Suppose in (2.1):

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

that $P(A_i|y)$ 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. That is an exaggeration of the decisiveness of the tests but let us continue. With these assumptions:

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

and the theory is 'confirmed'.

However, a more natural interpretation of 'equal claims' is the condition that the prior probabilities $P(A_i)$, are equal; the investigator can contemplate these without analysing the data. As the posterior weights $P(A_i|y)$ are the ones that matter for inference, their proposals make most sense for situations in which the alternative specifications are indistinguishable, i.e., $P(A_i) = P(A_i|y)$.

6 ASKING DIFFERENT QUESTIONS?

It has been argued that sensitivity analysis is a solution to a problem that does not exist. There is no point comparing inferences when they are answering different questions. Johansen (1991) makes this point with reference to the specifications:

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

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

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.

To discuss sameness and difference, the population/structure conceptualization of Koopmans and Reiersøl (1950) is useful; see Aldrich (1994) for a discussion of the ideas associated with it. Versions of this conceptualization appear throughout the econometrics literature; perhaps it is an element in what is meant by 'econometrics'. It is behind Spanos's (1986) distinction between 'statistical parameters' and 'theoretical parameters' and Leamer's (1978: 298) characterization of two sources of uncertainty about a parameter θ : 'misspecification uncertainty' arising from the existence of more than one algorithm relating θ to a population parameter π and 'sampling uncertainty' associated with the need to estimate π .

We are in the car park looking for our friend's new car; which he has said is 'the best'. I look for the fastest car, you for the most expensive. We each have two answers to the question, what are you looking for? One answer is common, 'the best car in the eyes of our friend'. 'The fastest' car and 'the most exclusive' are our answers to the population question. Our different algorithms derive from our different ways of structuring our friend's tastes. There is no conflict in my indicating this car as the fastest and your indicating that as the most expensive. Conflict would come when we apply our algorithms to identify our friend's car. To inject some sampling uncertainty, suppose it is dark in the car park!

In Johansen's example the θ of the two specifications is *not* the same population parameter in the sense of being the same functional of the joint distribution of x , y and z . However, these distinct functionals may represent 'the effect of x on y ' in different structural models. Suppose there are two structural models involving the three variables y , x and z and, associated with them, two ways of construing θ , 'the effect of x on y '. 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:

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

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

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

where v , w and x are generated as follows:

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

In A_1 the 'effect of x of 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 parameters of both models are identified. The population densities or reduced forms are tri-normal with variance matrices:

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

Neither A_1 nor A_2 implies any testable restrictions on the joint distribution of the observables, x , y and z . The models are observationally equivalent and thus empirically indistinguishable. 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 behind the estimates. We are in the car park or in the no-relevant data situation of the last section.

The fact of fragility has no implications for the credibility of the assumptions but 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. 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 the condition expressed 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 other indistinguishable contexts there will be a condition for consensus. In the prediction problem of section 4, $\delta x_1^* = \delta x_2^*$ implies $\delta=0$.

We can wish that the conclusion is the same whichever of the assumptions is true, that the fastest car is the most expensive, but there is no reason to suppose that truth manifests itself as a sturdy conclusion.

7 UNEMBARRASSINGLY FRAGILE INFERENCES

In the situations discussed in the last two sections there has been disagreement about structure but none about population. Here and in the next section there will be disagreement about both; the assumptions will no longer be observationally equivalent. In general, the posterior probability of an assumption changes according to the success of the predictions it makes. In particular, the predictions may concern whether the inferences agree or disagree. This section takes a first look at predicted disagreement.

McAlear *et al.* (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 the Johansen example from section 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 change it so that 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. Indeed, one of the fragility measures proposed by Leamer is the Hausman (1978) statistic for testing θ is 0.

In terms of the constructions of section 2, we find in (2.1):

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

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 $P(A_2|y)$. The inequality (2.4):

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

is not violated, only $P(S|y)$ is close to $\max_i \{P(S|A_i, y)\}$, which is $P(S|A_1, y)$.

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,' (Leamer 1985: 308) 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 nor can one be surprised by the inferences the other model makes for they are implied by one's own inferences. 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 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 section 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. The possibilities of a plague on all our models are considered in the next section.

8 EMBARRASSINGLY STURDY INFERENCES

To illustrate the over-sturdy inference, suppose I am trying to determine the rank of a churchman, whether priest or bishop. I have two authorities. The first uses the colour of the garment. According to him, bishops wear white and priests black; as all clergy wear crosses, he does not ask about that. The second asks about the cross, for, according to her, priests wear crosses and bishops do not; she does not consider the colour of the robes because everyone wears black. I report to the first authority that the robes are white and to the second that there is no cross. They both say the person is a bishop. On reflection, their agreement does not help me for they must be

wrong about the dress code as they both forbid the combination observed; there is no such person as the other authority's bishop. The robustness of the conclusion has not undermined the conclusion by showing it to be false, but it has undermined two arguments leading to that conclusion. I know that the right dress code has to admit the combination, white without a cross. I can imagine codes where this is the costume of a priest but also codes where it is the costume of a bishop.

Mroz (1987: 775) finds an 'unexpected similarity' between the estimates in two labour supply functions and connects this finding to another of his findings, that both functions are misspecified. To see how this can happen, consider the following simple but contrived set-up:

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

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

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

The least squares estimates associated with each specification are given by:

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

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

It is easy to calculate that when A_1 is true, the expectation of $\hat{\theta}_2$ is zero, whatever the true value of θ , and similarly that if A_2 is true, the expectation of $\hat{\theta}_1$ is zero:

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

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

When $\hat{\theta}_1$ and $\hat{\theta}_2$ are close to 0 we have a case of 'expected similarity' because both specifications imply that θ is close to 0 and imply that the other specification will get this answer. This is satisfying because, when θ is 0, the two models are the same. 'Unexpected similarity' (the 'bishop' case) obtains when $\hat{\theta}_1$ and $\hat{\theta}_2$ are close to each other *away from* 0. This is not expected on either specification.

The unexpected similarity undermines the authority of both A_1 and A_2 . It does not indicate what would be a better specification and thus how a better estimate of θ could be obtained. The 'unexpected' would be expected if the

true specification were:

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

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

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

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

If A_3 were the true specification, then the best 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. However, with more imagination and work, I may find a specification which accounts for the behaviour of $\hat{\theta}_1$ and $\hat{\theta}_2$ and which produces the same value for θ .

Of course, the case of unexpected sturdiness has a mirror image in the case of unexpected fragility, the situation where both specifications predict a similarity, which fails to materialize. Consider estimating θ in:

$$y = \theta x + u$$

using instrumental variables. There are two instrumental variables available, z and w . The choice between them is based on whether A_1 or A_2 holds:

$$A_1 : x = \pi_{xz}z + v_{xz} : z = \pi_{zw}w + v_{zw} : Ewv_{zw} = 0 = Euv_{zw}$$

$$A_2 : x = \pi_{xw}w + v_{xw} : w = \pi_{wz}z + v_{wz} : Ezv_{wz} = 0 = Euv_{wz}.$$

Each modeller concedes that the other's procedure is consistent, albeit based on an unnecessarily noisy instrument. So, provided the sample is large enough, each will expect the two estimates to be close together. It is bad news, evidence of misspecification, if they are *not* close together, i.e., the inference is fragile. This situation is quite unlike that described in section 4, where different authorities gave different judgements; there, one does not know what to believe because one does not know which authority to believe. Here, the fact that the authorities disagree undermines their authority; one does not know what to believe because the authorities have been discredited.

These examples are designed to make the point that similarity (and fragility) may be unexpected and a symptom that both specifications are unsatisfactory. The formal framework of section 2 does not admit the possibility of universal misspecification: the posterior probabilities $P(A_1|y)$

and $P(A_2|y)$ must sum to one if those assumptions are the only ones entertained. But once a suitable A_3 is admitted those posterior probabilities 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. Examples are given in sections 5 and 6.

In situations in which the models are distinguishable the analysis of sturdiness and fragility is more complicated. 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 and saves a lot of work. If the assumptions entertained do not cover the truth then sturdiness may signal this, although there would be other methods of discovering the misspecification Disagreement among the inferences, i.e., fragility, might be a reason for condemning all the inferences, if fragility comes as a surprise on all of the assumptions, but it may be nothing to worry about if one of the assumptions predicts fragility.

John Aldrich
University of Southampton
john.aldrich@soton.ac.uk

ACKNOWLEDGMENTS

Some of the ideas in this paper came out of work 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. I am also grateful to the referees for their comments and suggestions.

REFERENCES

- Aldrich, J. (1994) 'Haavelmo's identification theory', *Econometric Theory* 10: 198–219.
- Box, G.E.P. and Tiao, G.C. (1972) *Bayesian Inference in Statistical Analysis*, London: Addison-Wesley.
- Breusch, T.S. (1990) 'Simplified extreme bounds', in C.W.J. Granger (ed.) *Modelling Economic Series*, Oxford: Oxford University Press, pp. 72–81.

- Cartwright, N. (1991) 'Replicability, reproducibility, and robustness – comments on Harry Collins', *History of Political Economy* 23: 143–55.
- Collins, H. (1991) 'Replicability, reproducibility, and robustness', *History of Political Economy* 23: 23–42.
- Cooley, T.F. (1982) 'Specification analysis with discriminating priors: an application to the profits concentration debate', *Econometric Reviews* 1: 97–128.
- Cooley, T.T. and LeRoy, S.F. (1981) 'Identification and estimation of money demand', *American Economic Review* 71: 825–44.
- Faust, J. (1998) 'The robustness of identified VAR conclusions about money', *Carnegie-Rochester Conference on Public Policy* 49: 207–44.
- Fisher, R.A. (1950) Author's Note to 'On the mathematical foundations of theoretical statistics', in W.A. Shewhart (ed.) *Contributions to Mathematical Statistics*, New York: Wiley.
- Granger, C.W.J. (ed.) (1990) *Modelling Economic Series*, Oxford: Oxford University Press.
- Hausman, J. (1978) 'Specification tests in econometrics', *Econometrica* 46: 1251–72.
- Hendry, D.F. (1995) *Dynamic Econometrics*, Oxford: Oxford University Press.
- Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999) 'Bayesian model averaging: a tutorial', *Statistical Science* 14: 382–401.
- Hoover, K.D. and Perez, S.J. (2004) Truth and robustness in cross-country growth regressions', *Oxford Bulletin of Economics and Statistics* 66: 765–98.
- Howson, C. and Urbach, P. (1989) *Scientific Reasoning: the Bayesian Approach*, La Salle, IL: Open Court.
- Johansen, S. (1991) Comment on E. E. Leamer 'a Bayesian perspective on inference from macroeconomic data', *Scandinavian Journal of Economics* 93: 249–51.
- 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. and Leamer, E.E. (1984) 'Consistent sets of estimates for regressions with errors in all variables', *Econometrica* 52: 163–84.
- Koopmans, T.C. (1952) 'Toward partial redirection of econometrics: comment', *Review of Economics and Statistics* 34: 200–05.
- Koopmans, T.C. and Reiersøl, O. (1950) The identification of structural characteristics, *Annals of Mathematical Statistics* 21: 165–81.
- Krolzig, H.-M. and Hendry, D.F. (2001) 'Computer automation of general-to-specific model selection model procedures', *Journal of Economic Dynamics and Control* 25: 831–66.
- Leamer, E.E. (1978) *Specification Searches*, New York: John Wiley.
- Learner, E.E. (1983) 'Let's take the con out of econometrics', *American Economic Review* 73: 31–44. Reprinted in Granger (1990).
- Learner, E.E. (1985) 'Sensitivity analysis would help', *American Economic Review* 75: 308–13. Reprinted in Granger (1990).
- Learner, E.E. (1991) 'A Bayesian perspective on inference from macroeconomic data, with comments by S. Johansen, K. Juselius and S. Yitzhak', *Scandinavian Journal of Economics* 93: 225–48.
- Leamer, E.E. and Leonard, H. (1983) 'Reporting the fragility of regression estimates', *Review of Economics and Statistics* 65: 306–17.
- Levine, R. and Renelt, D. (1992) 'A sensitivity analysis of cross-country growth regressions', *American Economic Review* 82: 942–63.
- McAleer, M., Pagan, A.R. and Volker, P.A. (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. and Polasek, W. (1991) 'Robust HPD regions in Bayesian regression models', *Econometrica* 59: 1581–90.
- Rayner, J. and Aldrich, J. (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, pp. 283–327.
- Sims, C.A. (1988) 'Uncertainty across models', *American Economic Review* 78: 163–67.
- Spanos, A. (1986) *Statistical Foundations of Econometric Modelling*, Cambridge: Cambridge University Press.
- Temple, J. (1999) 'The new growth evidence', *Journal of Economic Literature* 37: 112–56.
- Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*, New York: John Wiley.