Box, G. E. P. (1979): Robustness in the strategy of scientific model building. – In: Launer, R. L. & Wilkinson, G. N.: Robustness in Statistics. – S. 201-236, 14 Abb., 1 Tab.; New York (Academic Press).

Robustness in the Strategy of Scientific Model Building

G. E. P. Box

Robustness may be defined as the property of a procedure which renders the answers it gives insensitive to departures, of a kind which occur in practice, from ideal assumptions. Since assumptions imply some kind of scientific model, I believe that it is necessary to look at the process of scientific modelling itself to understand the nature of and the need for robust procedures. Against such a view it might be urged that some useful robust procedures have been derived empirically without an explicitly stated model. However, an empirical procedure implies some unstated model and there is often great virtue in bringing into the open* the kind of assumptions that lead to useful methods. The need for robust methods seems to be intimately mixed up with the need for simple models. This we now discuss.

*An example (1), (2) was the application in the 1950's of exponential smoothing for business forecasting and the wide adoption in this century of three-term controllers for process control. It was later realized that these essentially empirical procedures point to the usefulness of ARIMA time series models since both are optimal for disturbances generated by such models.

Copyright © 1979 by Academic Press, Inc.
All rights of reproduction in any form reserved.

ISBN 0-12-438150-2

THE NEED FOR SIMPLE SCIENTIFIC MODELS - PARSIMONY

The scientist, studying some physical or biological system and confronted with numerous data, typically seeks for a model in terms of which the underlying characteristics of the system may be expressed simply.

For example, he might consider a model of the form

$$y_u = f^{(p)}(\underset{\sim}{\xi}_u, \underset{\sim}{\theta}) + \varepsilon_u \qquad (u = 1,2,\ldots,n) \tag{1}$$

in which the expected value η_u of a measured output y_u is represented as some function of k inputs $\underset{\sim}{\xi}$ and of p parameters $\underset{\sim}{\theta}$, and ε_u is an "error". One important measure of simplicity of such a model is the number of parameters that it contains. When this number is small we say the model is parsimonious.

Parsimony is desirable because (i) when important aspects of the truth are simple, simplicity illuminates, and complication obscures; (ii) parsimony is typically rewarded by increased precision (see Appendix 1); (iii) indiscriminate model elaboration is in any case not a practical option because this road is endless*.

ALL MODELS ARE WRONG BUT SOME ARE USEFUL

Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do

*Suppose for example that in advance of any data we postulated a model of the form of (1) with the usual normal assumptions. Then it might be objected that the distribution of ε_u might turn out to be heavy-tailed. In principle this difficulty could be allowed for by replacing the normal distribution by a suitable family of distributions showing varying degrees of kurtosis. But now it might be objected that the distribution might be skew. Again, at the expense of further parameters to be estimated, we could again elaborate the class of distribution considered. But now the possibility might be raised that the errors could be serially correlated. We might attempt to deal with this employing, say, a first order autoregressive error model. However, it could then be argued that it should be second order or that a model of some other type ought to be employed. Obviously these possibilities are extensive, but they are not the only ones: the adequacy of the form of the function $f(\xi,\theta)$ could be called into question and elaborated in endless ways; the choice of input variables $\underset{\sim}{\xi}$ might be doubted and so on.

provide remarkably useful approximations. For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules.

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?".

ITERATIVE PROCESS OF MODEL BUILDING

How then is the model builder to know what aspects to include and what to omit so that parsimonious models that are illuminating and useful result from the model building process? We have seen that it is fruitless to attempt to allow for all contingencies in advance so in practice model building must be accomplished by iteration* the inferential stage of which is illustrated in Figure 1.

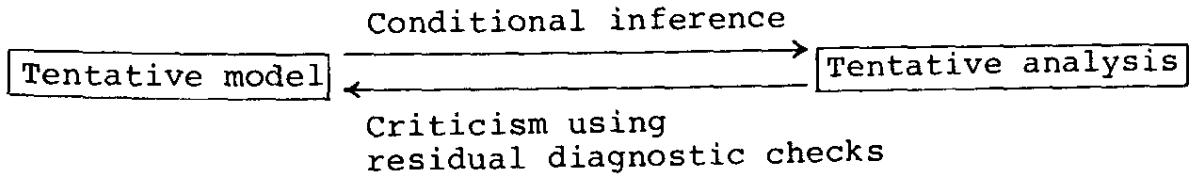


Figure 1. Interative Model Building

For example, preliminary graphical analysis of data, and careful thought about what is known about the phenomenon under

*The iterative building process for scientific models can take place over short or long periods of time, and can involve one investigator or many. One interesting example is the process of discovery of the structure of DNA described by J. D. Watson [3]. Another is the development by R. A. Fisher [4] of the theory of experimental design between 1922 and 1926. The recognition that scientific model building is an iterative process goes back to such classical authors as to Aristotle, Grossteste and Bacon. The suggestion that statistical procedures ought to be viewed in this iterative context was discussed for example in [1], [5], [6], [7], [22].

study, may suggest a first model worthy to be tentatively entertained. From this model a corresponding first tentative analysis may be made as if we believed it. The tentative inferences made, like all inferences are conditional on the applicability of the implied model, but the investigator now quickly switches his attitude from that of sponsor to that of critic. In fact the tentative analysis provides a basis for criticism of the model. This criticism phase is accomplished by examining residual quantities using graphical procedures and sometimes more formal tests of fit. Such diagnostic checks may fail to show cause for doubting the model's applicability, otherwise it may point to modification of the model leading to a new tentative model and, in turn, to a further cycle of the iteration.

WAYS TO ALLOW FOR MODEL DISCREPANCIES

How can we avoid the possibility that the parsimonious models we build by such an iteration might be misleading? There are two answers.

a) Knowing the scientific context of an investigation we can allow in advance for more important contingencies.
b) Suitable analysis of residuals can lead to our fixing up the model in other needed directions.

We call the first course model robustification the second iterative fixing.

JUDICIOUS MODEL ROBUSTIFICATION

Experience with data and known vulnerabilities of statistical procedures in a specific scientific context will alert the sensitive practitioner to likely discrepancies that can cause problems. He may then judiciously and grudgingly elaborate the model and hence the resulting procedure so as to insure against particular hazards in the most parsimonious manner. Models* providing for simple forms of

* It is currently fashionable to conduct robustness studies in which the normality assumption is relaxed (in favor of heavy tailed distribution or distributions containing outliers) but all other assumptions are retained. Thus it is still assumed that errors are independent that transformations are correctly specified and so on. This seems to be too naive and narrow a view.

autocorrelation in serial data, for simple transformation in data covering wide ranges, for outliers in the almost universal situation where perfect control of the experimental process is not available, are all examples of commonly needed parsimonious elaborations which can have major consequences.

ITERATIVE FIXING USING DIAGNOSTIC CHECKS

Once it is recognized that the choice of model is not an irrevocable decision, the investigator need not attempt to allow for *all* contingencies *a priori* which as we have said is in any case impossible. Instead, after appropriate robustification, he may look at residual quantities in an attempt to reveal discrepancies not already provided for.

To better appreciate such a process of iterative fixing, write the model (1) in the form

$$y_u = f(\underset{\sim}{\xi}_{1u}) + \varepsilon(\underset{\sim}{\xi}_{2u}) \tag{2}$$

where now the vector $\underset{\sim}{\xi}_{1u}$ previously denoted in (1) by $\underset{\sim}{\xi}_u$ represents those variables the investigator has specifically decided to study. The expression $\varepsilon(\underset{\sim}{\xi}_{2u})$ which replaces ε_u indicates explicitly that the error ε_u represents the joint influences on the output of all those other input variables $\underset{\sim}{\xi}_{2u}$ which are omitted from the model (usually because they are unknown). Many statistical procedures (in particular quality control, residual analysis and evolutionary operation) are concerned with discovering "assignable causes" - elements of $\underset{\sim}{\xi}_{2u}$ - which may be moved out of the unknown to the known as indicated by

$$y_u = f(\underset{\sim}{\xi}_{1u}) + \varepsilon(\underset{\sim}{\xi}_{2u}) \ . \tag{3}$$

Now let $\{a_t\}$ be a white noise sequence. That is a sequence of identically and independently distributed random variables having zero mean. If we now denote the n values of response and known inputs by $\underset{\sim}{y}$ and $\underset{\sim}{\xi}_1$ respectively then an ideal model

$$F_t\{\underset{\sim}{y}, \underset{\sim}{\xi}_1\} = a_t, \qquad t = 1,2,\ldots,n \tag{4}$$

would consist of a transformation of the data to white noise which was statistically independent of any other input.

Iteration towards such a model is partially motivated by diagnostic checking through examination of residuals at each stage of the evolving model.

Thus, patterns of residuals can indicate the presence of outliers or suggest the necessity for including specific new inputs. Also serial correlation of residuals shown for example by plotting $\hat{a}_{t+k}$ versus $\hat{a}_t$ (or more formally by examining estimates of sample autocorrelations for low lags k) can point to the need for allowing for serial correlation in the original noise model. Again as was shown by Tukey [8], dependence of $y - \hat{y}$ on $\hat{y}^2$ can indicate the need for transformation of the output. Examination of residuals at each stage parallels the chemical examination of the filtrate from an extraction process. When we can no longer discover any information in the residuals then we can conclude that all extractable information is in the model.*

MODEL ROBUSTIFICATION AND DIAGNOSTIC CHECKING

Robustification and iterative fixing following diagnostic checking of residuals are of course not rival but complementary techniques and we must try to see how to use both wisely.

Subject matter knowledge will often suggest the need for robustification by parsimonious model elaboration. For example, when models such as (2) are used in economics and business the output and input variables $\{y_u\}$, $\{\underset{\sim}{\xi}_{1u}\}$ are often collected serially as time series. They are then very likely to be autocorrelated. If this is so then the

---

* It should be remembered that just as the Declaration of Independence promises the pursuit of happiness rather than happiness itself, so the iterative scientific model building process offers only the pursuit of the perfect model. For even when we feel we have carried the model building process to a conclusion some new initiative may make further improvement possible. Fortunately to be useful a model does not have to be perfect.

In particular notice that, even though residuals from some model are consistent with a white noise error, this does not bar further model improvement. For example, this white noise error could depend on (theoretically, even, be proportional to) the white noise component of some, so far unrecognized, input variable.

components of $\{\xi_{2u}\}$ in the error $\varepsilon\{\xi_{2u}\}$ are equally likely to have this characteristic. It makes little sense, in this context, therefore, to postulate even tentatively a model of the form of (1) in which the ε_u are supposed independent. Instead the representation of ε_u by a simple time series model (for example a first order autoregressive process for which serial correlation falls off exponentially with lag) would provide a much more plausible starting place. Failure either, to robustify the model in this way initially, or, to check for serial correlation in residuals, resulted in one published example in t values (measuring the significance of regression coefficients) which were inflated by an order of magnitude [9,10]. We discuss this example in more detail later.

Again statistical analysis in an inappropriate metric can lead to wasteful inefficiency. For instance, textile data are presented in [11] where appropriate transformation would have resulted in a three fold decrease in the relative variance accompanied by reduction in the number of needed parameters from ten to four, for the expenditure of only one estimated transformation parameter.

In both examples discussed above a profound improvement in statistical analysis is made possible by suitable robustification of the model the need for which could have been detected by suitable diagnostic checks on residuals.

AVOIDANCE OF UNDETECTED MISSPECIFICATION

Unfortunately we cannot always rely on diagnostic checks to reveal serious model misspecification. The dangerous situation is that where initial model misspecification can result in a seriously misleading analysis whose inappropriateness is unlikely to be detected by diagnostic checks.

For example, the widely used model formulation (1) supposes its applicability for every observation y_u $(u = 1,2,\ldots,n)$ and so explicitly excludes the possibility of outliers. If such a model is (inappropriately) assumed in the common situations where occasional accidents possibly leading to outliers are to be expected, then any sensible method of estimation such as maximum likelihood if applied

using this inappropriate model must tend to conceal model inadequacy. This is because, in order to follow the mathematical instructions presented, it must choose parameters which make residuals even with this wrong model look as much as possible like white noise. That this has led some investigators to abandon standard inferential methods rather than the misleading model seems perverse.

As a further example of the hazard of undetected misspecification consider scientific problems requiring the comparison of variances. Using standard normal assumptions the investigator might be led to conduct an analysis based on Bartlett's test. However this procedure is known to be so sensitive to kurtosis that nonnormality unlikely to be detected by diagnostic checks could seriously invalidate results. This characteristic of the test is well known, of course, and it has long been recognized that the wise researcher should robustify initially. That is he should use a robust alternative to Bartlett's test ab initio rather than relying on a test of nonnormality followed by possible fix up.

The conclusion is that the role of model robustification is to take care of likely discrepancies that have dangerous consequences and are difficult to detect by diagnostic checks. This implies an ability by statisticians to worry about the right things. Unfortunately they have not always demonstrated this talent, see for example Appendix 2.

ROBUSTNESS AND ERROR TRANSMISSION

Since we need parsimonious models but we know they must be false we are led to consider how much deviation from the model of a kind typically met in practice will affect the procedure derived on the assumption that a model is exact.

The problem is analogous to the classical problem of error transmission. In its simplest manifestation that problem can be expressed as follows:

Consider a calibration function

$$\gamma = f(\beta) \tag{5}$$

which is used to determine γ at some value say $\beta = \beta_0$. Suppose that the function is mistakenly evaluated at some

other value of β, then the resulting error ε transmitted into γ is

$$(\gamma_0 + \varepsilon) - \gamma_0 = f(\beta) - f(\beta_0) \simeq (\beta - \beta_0) \times \rho \qquad (6)$$

where $\rho = \left.\frac{\partial \gamma}{\partial \beta}\right|_{\beta=0}$.

The expression for the transmitted error ε contains two factors β and ρ. The first is the size of the input error the second which we will call the specific transmission is the rate of change of γ as β is changed. It is frequently emphasized in discussing error transmission that both factors are important. In particular the existence of a large discrepancy $\beta - \beta_0$ does not lead to a large transmitted error ε if ρ is small. Conversely even a small error β can produce a large error ε if ρ is large.

Now consider a distribution of errors $p(\beta)$. Knowledge of the relation $\gamma = f(\beta)$ allows us to deduce the corresponding distribution $p(\varepsilon)$. In particular if the approximation (6) may be employed then $\sigma_\gamma = p\sigma_\beta$. The relevance of the above robustness studies is as follows. Suppose γ is some performance characteristic of a statistical procedure which it is desired to study. This characteristic might be some measure of closeness of an estimate to the true value, significance level, the length of a confidence interval, a critical probability, a posterior probability distribution, or a rate of convergence of some measure of efficiency or optimality. Also suppose β is some measure of departure from assumption such as a measure of nonnormal kurtosis or skewness or autocorrelation of the error distribution and suppose that $\beta = \beta_0$ is the value taken on standard assumptions. Then in the error transmission problem three features of importance are

(1) The distribution of β. This provides the probability distribution of deviations from assumption which are actually encountered in the real world. Notice this feature has nothing to do with mathematical derivation or with the statistical procedure used.

(ii) The specific transmission ρ. This is concerned with the error transmission characteristics of the statistical

procedure actually employed and may be studied mathematically. It is well known that different statistical procedures can have widely different ρ's. An example already quoted is the extreme sensitivity to distribution kurtosis of the significance level of likelihood ratio tests to compare variances (Bartlett's test) and the comparative insensitivity of corresponding tests to compare means (Analysis of variance tests).

(iii) If the data set is of sufficient size it can itself provide information about the discrepancy $\beta - \beta_0$ which occurs in that particular sample, thus reducing reliance on prior knowledge. Conversely if the sample size is small* or if β is of such a nature that a very large sample is needed to gain even an approximate idea of its value, heavier reliance must be placed on prior knowledge (whether explicitly admitted or not).

It seems to me that these three characteristics taken together determine what we should worry about. They are all incorporated precisely and appropriately in a Bayes formulation.

BAYES THEOREM AS A MEANS OF STUDYING ROBUSTNESS

From a Bayesian point of view given data $\underset{\sim}{y}$ all valid inferences about paramaters $\underset{\sim}{\theta}$ can be made from an appropriate posterior distribution $p(\underset{\sim}{\theta}|\underset{\sim}{y})$. To study the robustness of such inferences when discrepancies β from assumptions occur we can proceed as follows:

Consider a naive model relating data $\underset{\sim}{y}$ and parameters $\underset{\sim}{\theta}$. For example, it might assume that $p(\underset{\sim}{y}|\underset{\sim}{\theta})$ was a spherically normal density function, that $E(\underset{\sim}{y})$ was linear in the parameters $\underset{\sim}{\theta}$ and that before the data became available the desired state of ignorance about unknown parameters was expressed by suitable non-informative prior distributions leading to the standard analysis of variance and regression procedures. Suppose it was feared that certain discrepancies

*However even the small amount of information about β available from a small sample can be important. See for instance the analysis of Darwin's data which follows (Example 1).

from the model might occur (for example lack of independence, need for transformation, existence of outliers, non-normal kurtosis etc.). Two questions of interest are (A) how sensitive are inferences made about θ to these contemplated misspecifications of the model? (B) If necessary how may such inferences be made robust against such discrepancies as actually occur in practice?

QUESTION (A) SENSITIVITY

Suppose in all cases that discrepancies are parameterized by $\underset{\sim}{\beta}$. Also suppose the density function for $\underset{\sim}{y}$ given $\underset{\sim}{\theta}$ and $\underset{\sim}{\beta}$ is $p(\underset{\sim}{y}|\underset{\sim}{\theta},\underset{\sim}{\beta})$ and that $p(\underset{\sim}{\theta}|\underset{\sim}{\beta})$ is a non-informative prior for $\underset{\sim}{\theta}$ given $\underset{\sim}{\beta}$. Then comprehensive inferences about $\underset{\sim}{\theta}$ given $\underset{\sim}{\beta}$ and $\underset{\sim}{y}$ may be made in terms of the posterior distribution

$$p(\underset{\sim}{\theta}|\underset{\sim}{\beta},\underset{\sim}{y}) = k\, p(\underset{\sim}{y}|\underset{\sim}{\theta},\underset{\sim}{\beta})p(\underset{\sim}{\theta}|\underset{\sim}{\beta}) \tag{7}$$

where k is a normalizing constant. Sensitivity of such inferences to changes in $\underset{\sim}{\beta}$ may therefore be judged by inspection of $p(\underset{\sim}{\theta}|\underset{\sim}{\beta},y)$ for various values of $\underset{\sim}{\beta}$.

QUESTION (B) ROBUSTIFICATION

Suppose now that we introduce a prior density $p(\underset{\sim}{\beta})$ which approximates the probability of occurrence of $\underset{\sim}{\beta}$ in the real world. Then we can obtain $p(\underset{\sim}{\beta}|\underset{\sim}{y})$ from $\int p(\underset{\sim}{\theta},\underset{\sim}{\beta}|\underset{\sim}{y})d\underset{\sim}{\theta}$. This is the posterior distribution of $\underset{\sim}{\beta}$ given the prior $p(\beta)$ and given the data. Then

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}) = \int p(\underset{\sim}{\theta}|\underset{\sim}{\beta},\underset{\sim}{y})p(\underset{\sim}{\beta}|\underset{\sim}{y})d\underset{\sim}{\beta} \tag{8}$$

from which (robust) inferences may be made about $\underset{\sim}{\theta}$ independently of $\underset{\sim}{\beta}$ as required.*

Inference are best made by considering the whole posterior distribution however if point estimates are needed they can of course be obtained by considering suitable features of the posterior distribution. For example the posterior mean

*More generally the density function for y will contain nuisance parameters σ Equations (7) and (8) will then apply after these parameters have been eliminated by integration.

will minimize squared error loss. Other features of the posterior distribution will provide estimates for other loss functions (see for example [6], [12]).

It does seem to me that the inclusion of a prior distribution is essential in the formulation of robust problems. For example, the reason that robustifiers favour measures of location alternative to the sample average is surely because they have a prior belief that real error distributions may not be normal but may have heavy tails and/or may contain outliers. They evidence that belief covertly by the kind of methods and functions that they favour which place less weight on extreme observations. I think it healthier to bring such beliefs into the light of day where they can be critically examined, compared with reality, and, where necessary, changed. Some examples of this alternative approach are now given.

EXAMPLE 1 KURTOSIS AND THE PAIRED t TEST

This section follows the discussion by Box and Tiao [6], [13], [14] of Darwin's data quoted by Fisher on the differences in heights of 15 pairs of self and cross-fertilized plants. These differences are indicated by the dots in Figure 2. The curve labeled $\beta = 0$ in that diagram is a t distribution centered at the average $\bar{y} = 20.93$ with scale factor $s/\sqrt{n} = 9.75$. On standard normal assumptions it can be interpreted as a confidence distribution, a fiducial distribution or a posterior distribution of the mean difference θ. From the Bayesian view point the distribution can be written

$$p(\theta|\underset{\sim}{y}) = \text{const}\left\{1 + \frac{n(\theta - \bar{y})^2}{\nu s^2}\right\}^{-\frac{n}{2}} \tag{9}$$

and results from taking a non-informative prior distribution for the mean θ and the standard deviation σ. Alternatively we may write the distribution (9) in the form

$$p(\theta|\underset{\sim}{y}) = \text{const}[\Sigma(y - \theta)^2]^{-\frac{n}{2}} \tag{10}$$

and if

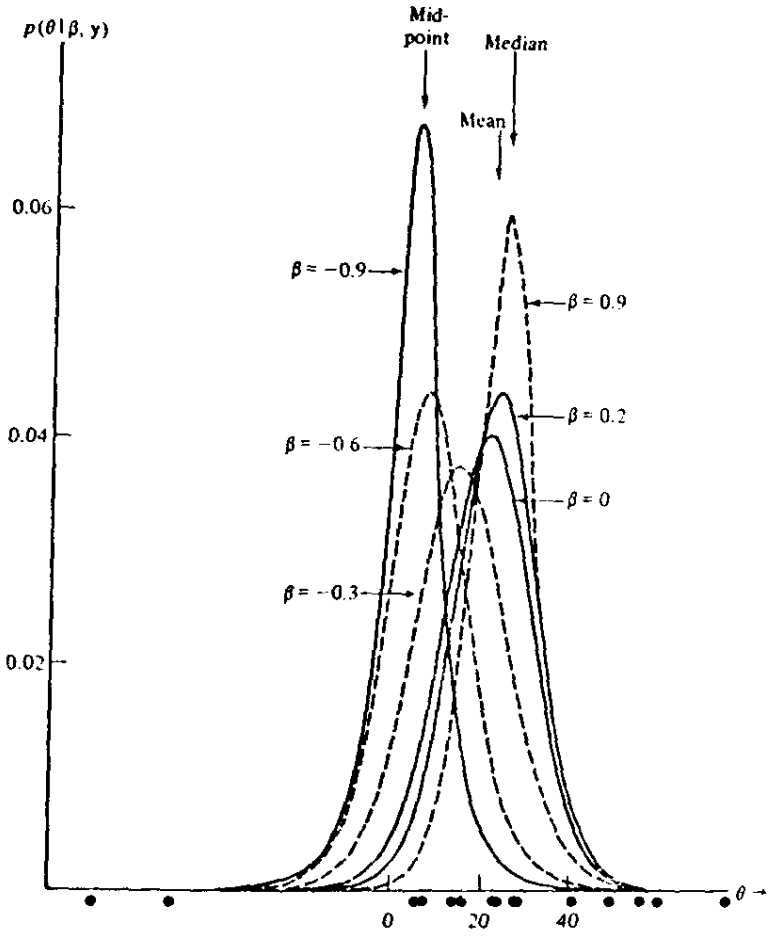


Figure 2. Posterior distributions $p(\theta|\beta,\underset{\sim}{y})$ of mean difference θ for parent distributions having differing amounts of kurtosis parameterized by β. Darwin's data.

$$M(\theta,q) = \Sigma|y_i - \theta|^q, \qquad \alpha \geq 1$$

(10) may be written as

$$p(\theta|y) = \text{const}\{M(\theta,2)\}^{-\frac{n}{2}} . \tag{11}$$

SENSITIVITY TO KURTOSIS

One way to consider discrepancies arising from non-normal kurtosis is to extend the class of density functions, using the exponential power family

$$p(y|\theta,\sigma,\beta) = \sigma^{-1} \exp\left\{-c(\beta)\left|\frac{y-\theta}{\sigma}\right|^{2/(1+\beta)}\right\} \tag{12}$$

where with $c(\beta) = \left\{\dfrac{\Gamma[\frac{3}{2}(1+\beta)]}{\Gamma[\frac{1}{2}(1+\beta)]}\right\}^{\frac{1}{1+\beta}}$ and θ and σ are the

mean and standard deviation as before. Then using the same noninformative prior distribution as before $p(\theta,\sigma|\beta) \propto \sigma^{-1}$ it is easily shown that in general

$$p(\theta|\beta,\underset{\sim}{y}) = \text{const } M\{\theta, 2/(1+\beta)\}^{-\frac{1}{2}n(1+\beta)} \tag{13}$$

and in particular if $\beta = 0$ (13) and (11) are identical.

The performance characteristic here is not a single quantity but the whole posterior distribution from which all inferences about θ can be made.

Sensitivity of the inference to changes in β is shown by the changes that occur in the posterior distributions $p(\theta|\beta,\underset{\sim}{y})$ when β is changed. Figure 2 shows these distributions for various values of β. Evidently, for this example, inferences* are quite sensitive to changes in the parent density involving more or less kurtosis.

ROBUSTIFICATION FOR KURTOSIS

As was earlier explained, high sensitivity alone does not necessarily produce lack of robustness. This depends also on how large are the discrepancies which are likely to occur, represented in (8) by the factor $p(\underset{\sim}{\beta}|\underset{\sim}{y})$. It is convenient to define a function $p_u(\underset{\sim}{\beta}|\underset{\sim}{y}) = p(\underset{\sim}{\beta}|\underset{\sim}{y})/p(\underset{\sim}{\beta})$ which fills the role of a pseudo-likelihood and represents the contribution of information about $\underset{\sim}{\beta}$ coming from the data. This factor is the posterior distribution of β when the prior is taken to be uniform. With this notation then for the present example

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}) = \int_{-1}^{1} p(\underset{\sim}{\theta}|\beta,\underset{\sim}{y})p_u(\beta|\underset{\sim}{y})p(\beta)d\beta = \int_{-1}^{1} p_u(\underset{\sim}{\theta},\beta|\underset{\sim}{y})p(\beta)d\beta. \tag{14}$$

For Darwin's data the distributions $p_u(\theta,\beta|y)$ and a particular $p(\beta)$ are shown in Figure 3. Figure 4 shows

---

* Notice however, the distinction that must be drawn between criterion and inference robustness [6], [14]. For example, for these data the significance level of the t criterion is changed hardly at all (from 2.485% to 2.388%) if we suppose the parent distribution is rectangular rather than normal.

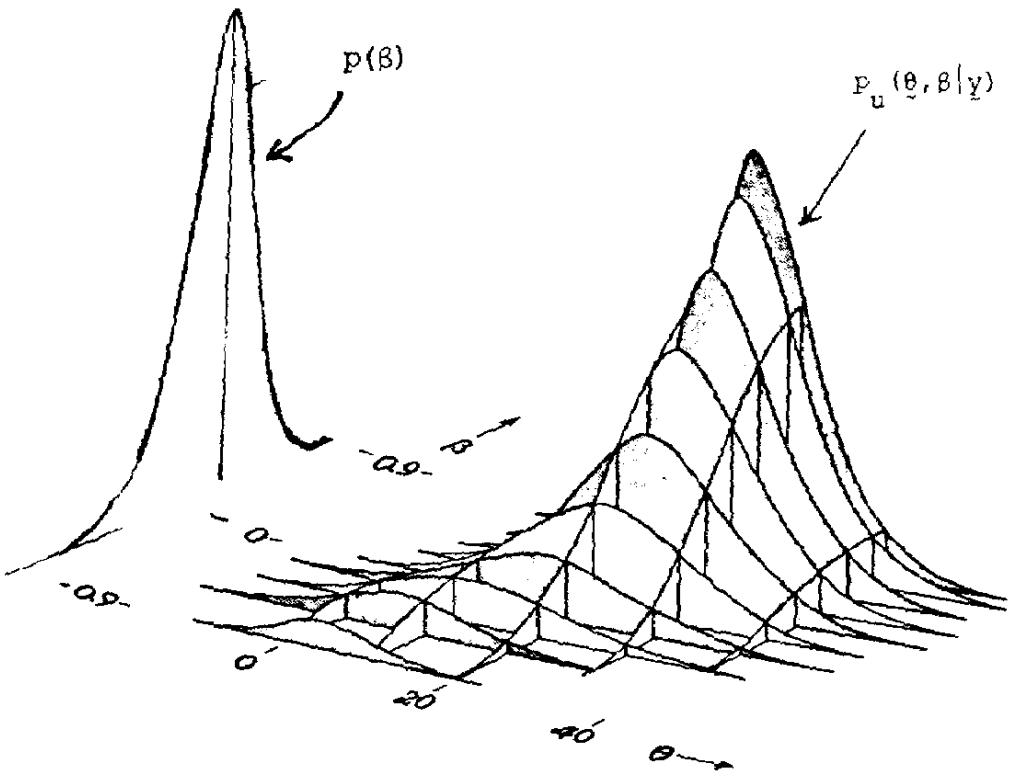


Figure 3. Joint posterior distribution $p_u(\theta,\beta|y)$ with a particular prior $p(\beta)$. Darwin's data.

$p(\beta|\underset{\sim}{y})$ for various choices of $p(\beta)$ while Figure 5 shows the corresponding distribution $p(\theta|\underset{\sim}{y})$.

a) Making $p(\beta)$ a delta function at $\beta = 0$ corresponds with the familiar absolute assumption of normality. It results in distribution (a) in Figure 5 which is the scaled t referred to earlier.

b) This choice for $p(\beta)$ is appropriate to a prior assumption that although not all distributions are normal; variations in kurtosis are such that the normal distribution takes a central role. For this particular example the resulting distribution (b) in Figure 5 is not very different from the t distribution.

c) Here, by making $p(\beta)$ uniform the modifier or pseudo-likelihood $p_u(\beta|\underset{\sim}{y})$ is explicitly produced which represents the information about kurtosis coming from the sample itself. For this extreme form of prior distribution, distribution (c) in Figure 5 is somewhat changed although not dramatically. The reason for this is that

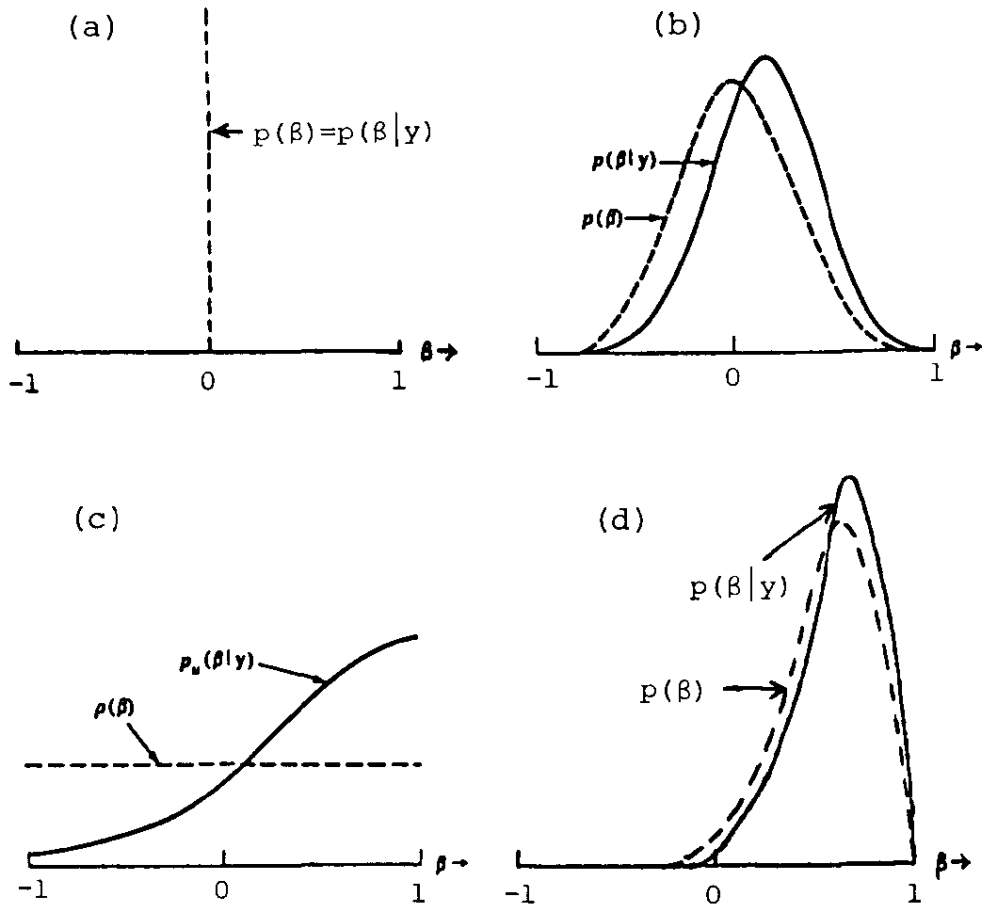


Figure 4. Posterior distributions $p(\beta|\underset{\sim}{y})$ for various choices of $p(\beta)$. Darwin's data.

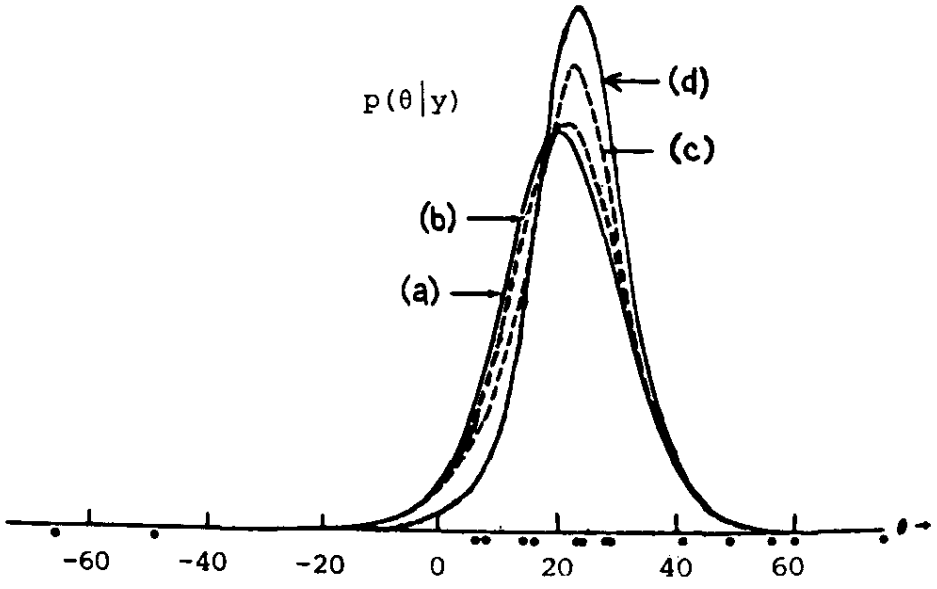


Figure 5. Posterior distributions $p(\theta|y)$ for various choices of $p(\beta)$. Darwin's data.

the widely discrepant distributions in Figure 2 for negative values of β are discounted by the information coming from the data.

d) This distribution is introduced to represent the kind of prior ideas which, following Tukey, many current robustifiers seem to have adopted. The resulting posterior distribution is shown in Figure 5(d).

This example brings to our attention the potential importance even for small samples of information coming from the data about the parameters $\underset{\sim}{\beta}$. In general if we compute the modifier $p_u(\underset{\sim}{\beta}|\underset{\sim}{y})$ from

$$p_u(\underset{\sim}{\beta}|\underset{\sim}{y}) = \frac{\int p(\theta, \underset{\sim}{\beta}|\underset{\sim}{y})d\theta}{p(\underset{\sim}{\beta})} \tag{15}$$

then we can write

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}) = \int p(\underset{\sim}{\theta}|\underset{\sim}{\beta},\underset{\sim}{y})p_u(\underset{\sim}{\beta}|\underset{\sim}{y})p(\underset{\sim}{\beta})d\underset{\sim}{\beta} \ . \tag{16}$$

Now even when the sensitivity factor is high that is when $p(\underset{\sim}{\theta}|\underset{\sim}{\beta},\underset{\sim}{y})$ changes rapidly as $\underset{\sim}{\beta}$ changes, this will lead to no uncertainty about $p(\underset{\sim}{\theta}|\underset{\sim}{y})$ if $p(\underset{\sim}{\beta}|\underset{\sim}{y})$ is sharp. This can be so either if $p(\beta)$ is sharp - there is an absolute assumption that we know β a priori - or if $p_u(\beta|\underset{\sim}{y})$ is sharp. In the common situation, the spread of $p_u(\beta|\underset{\sim}{y})$ will be proportional to $1/\sqrt{n}$ and for sufficiently large samples there will be a great deal of information from the sample about the relevant discrepancy parameters β. For small samples however this is not generally so. This amounts to saying that for sufficiently large samples it is always possible to check assumptions and in principle to robustify by incorporating sample information about discrepancies into our statistical procedure. For small samples we are always much more at the mercy of the accuracy of prior information whether we incorporate it by using Bayes theorem or not. Notice however, how a Bayes analysis can make use of sample data which would have been neglected by a sampling theory analysis. Comparison of Figures 2 and 5(c) makes clear the profound effect that the sampling information about β for only $n = 15$ observations has on the inferential situation. This sample information represented

by $p_u(\beta|\underset{\sim}{y})$ in Figure 4(c), although vague, is effective in discounting the possibility of platykurtic distributions which are the major cause of discrepancy in Figure 2. This effect accounts for the very moderate changes that occur in $p(\theta|\underset{\sim}{y})$ accompanying the drastic changes made in $p(\beta)$.

EXAMPLE 2: SERIAL CORRELATION AND REGRESSION ANALYSIS

Coen, Gomme and Kendall [9] gave 55 quarterly values of the Financial Times ordinary share index y_t, U.K. car production X_{1t} and Financial Times commodity index X_{2t}. They related y_t to the lagged values X_{1t-6} and X_{2t-7} by a regression equation

$$y_t = \theta_0 + \theta_1 t + \theta_2 X_{1t-6} + \theta_3 X_{2t-7} + \varepsilon_t \qquad (17)$$

which they fitted by least squares. As mentioned earlier they obtained estimates of θ_2 and θ_3 which were very highly significantly different from zero and concluded that X_1 and X_2 could be used as "leading indicators" to predict future share prices. Box and Newbold [10] pointed out that if allowance is made for the serial correlation which exists in the error ε_t then the apparently significant effects vanish and much better forecasts are obtained by using today's price to forecast the future. This is a case where inferences about $\underset{\sim}{\theta}$ are very non-robust to possible serial correlation.

In a recent Wisconsin Ph.D. thesis [15] Lars Pallesen reassessed the situation with a Bayesian analysis, supposing that ε_t may follow a first order autoregressive model $\varepsilon_t - \beta\varepsilon_{t-1} = a_t$, where a_t is a white noise sequence as earlier defined.

The dramatic shifts that occur in the posterior distributions of θ_2 and θ_3 when it is not assumed a priori that $\beta = 0$ are shown in Figures 6 and 7. The situation is further illuminated by Figures 8 and 9 which show the joint distribution of θ_2 and β and of θ_3 and β, together with the marginal distribution $p_u(\beta|y)$ based on non-informative prior distributions.

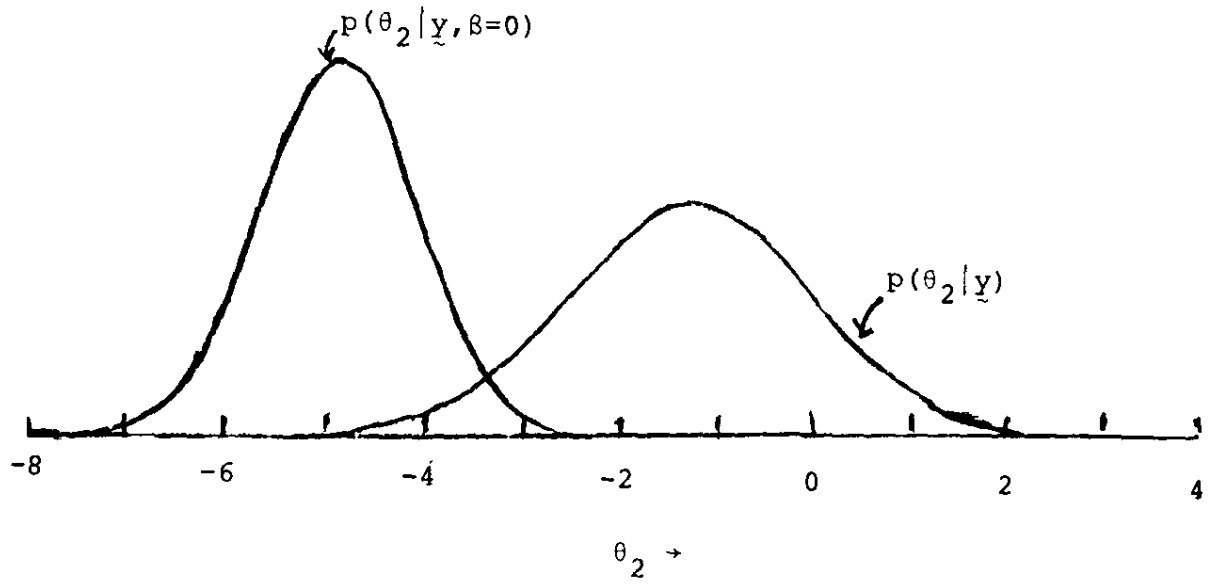


Figure 6. Effect of different assumptions on posterior distribution of θ_2. $p(\theta_2|\underset{\sim}{y})$ allows for possible autocorrelation of errors $p(\theta_2|y,\beta = 0)$ does not. θ_2 is regression coefficient of Share Index on car sales lagged six quarters.

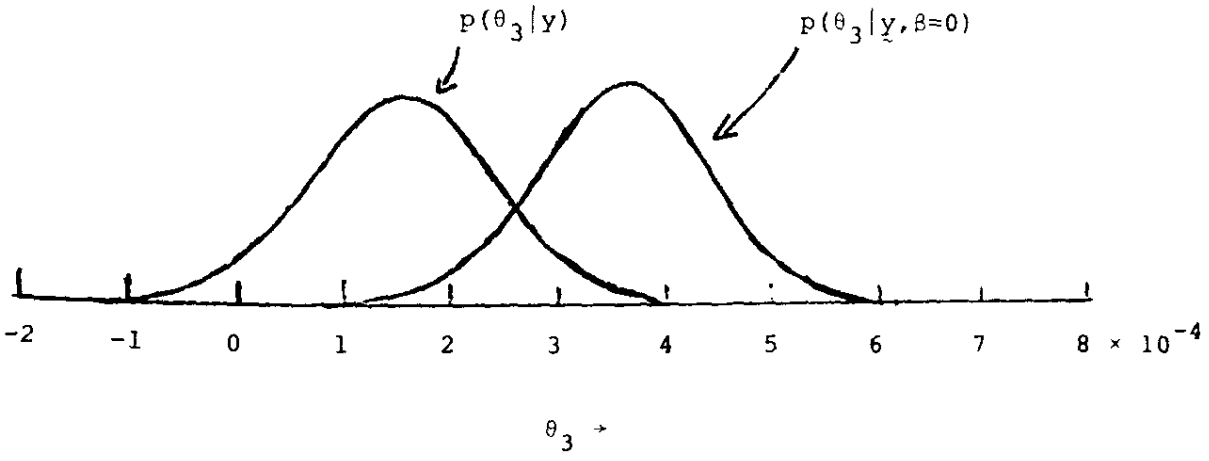


Figure 7. Effect of different assumptions on posterior distribution of θ_3. $p(\theta_3|\underset{\sim}{y})$ allows for possible autocorrelation of errors $p(\theta_3|y,\beta = 0)$ does not. θ_3 is regression coefficient of Share Index on Consumer Price Index lagged seven quarters.

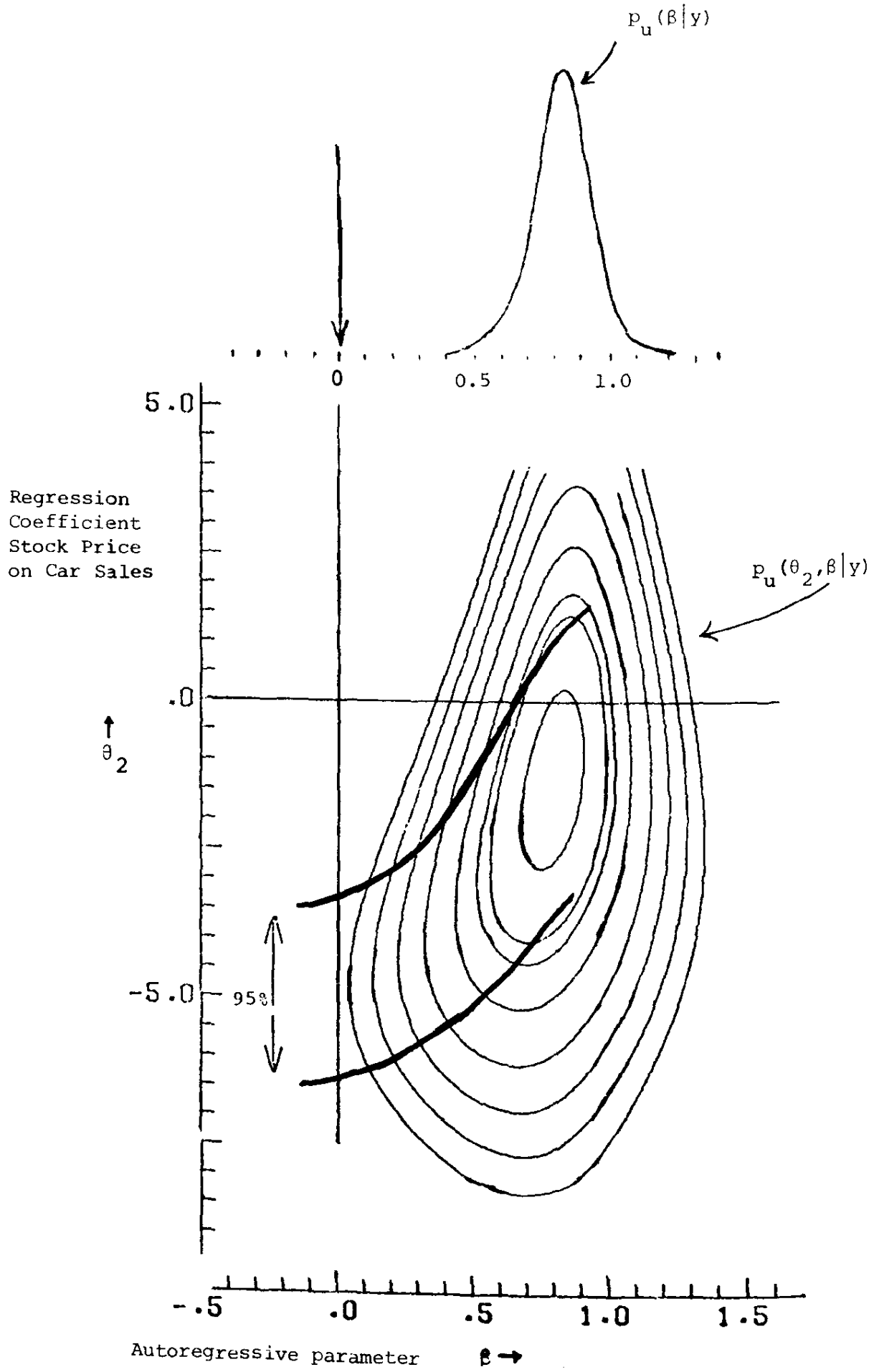


Figure 8. Joint posterior distribution of θ_2 and β and marginal posterior distribution of β. Note shift in 95% interval as β is changed.

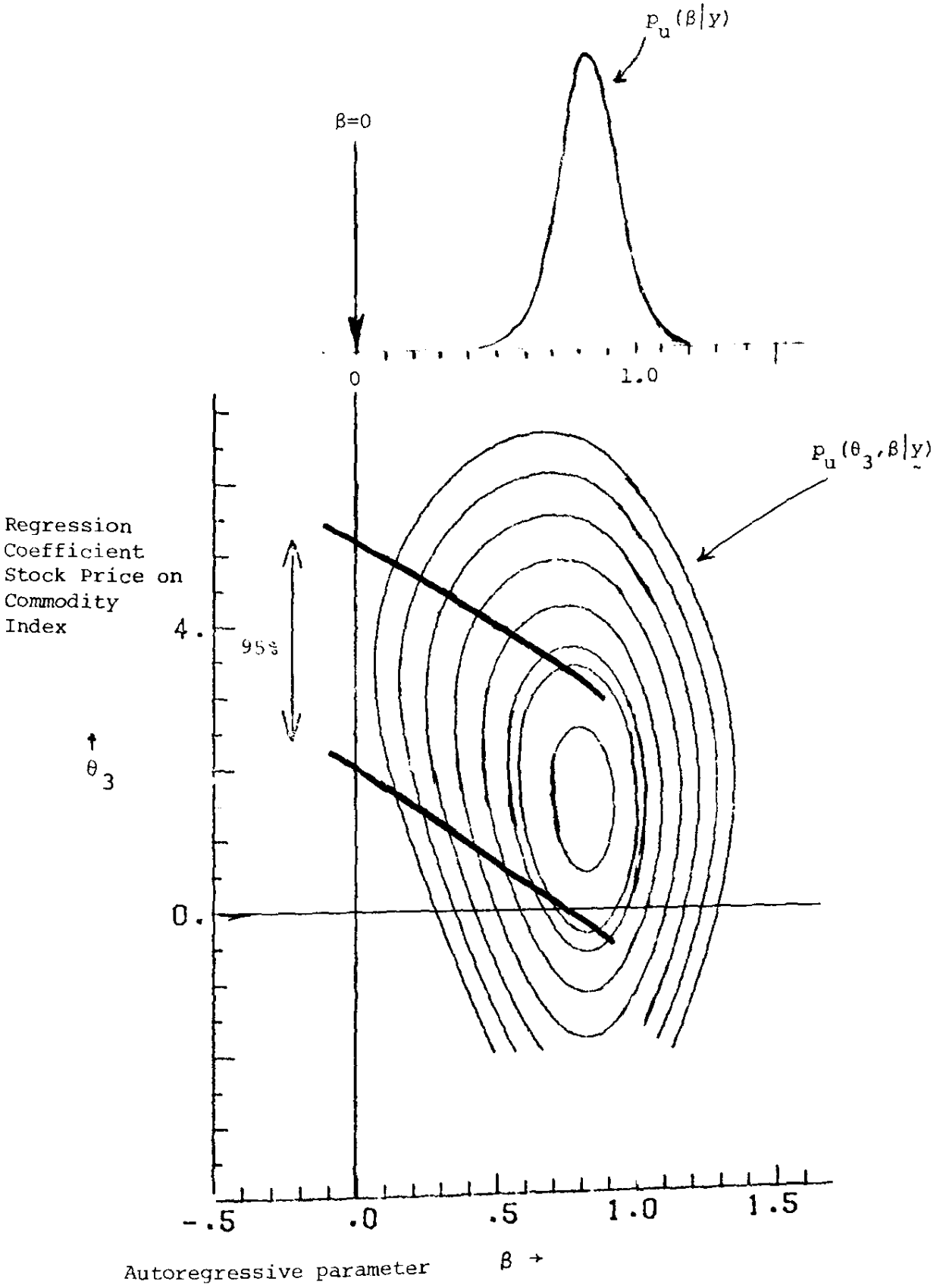


Figure 9. Joint posterior distribution of θ_3 and β and marginal posterior distribution of β. Note shift in 95% interval for θ_3 as β is changed.

EXAMPLE 3: OUTLIERS IN STANDARD STATISTICAL MODELS

Consider again the model form (1)

$$y_u = f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta}) + \varepsilon_u \qquad u = 1,2,\ldots,n \ . \tag{18}$$

With standard assumptions about ε_u with the restriction that the expectation function is linear in $\underset{\sim}{\theta}$ this is the widely used Normal linear model. The remarkable thing about this model is that it is ever seriously entertained even when assumptions of independence and normality seem plausible. For it specifically states that the model form is appropriate for $u = 1,2,\ldots,n$ (that is, for <u>every one</u> of the experiments run). Now anyone who has any experience of reality knows that data are frequently affected by accidents of one kind or another resulting in "bad values". In particular it is expecting too much of any flesh and blood experimenter that he could conduct experiments unerringly according to a pre-arranged plan. Every now and again at some stage in the generation of data, a mistake will be made which is unrecognized at the time. Thus a much more realistic statement would be that model like (18) applied, not for $u = 1,2,\ldots,n$, but in a proportion $1 - \alpha$ of the time and that during the remaining proportion α of the time some discrepant, imprecisely known, model was appropriate. Such a model was proposed by Tukey in 1960 [16]. We call observations from the first model "good" and those from the second model "bad".

This type of model was later used in a Bayesian context by Box and Tiao [17]. They assumed that the discrepant model which generated the bad values was of the same form as the standard model except that the error standard deviation was k times as large. The results are rather insensitive to the choice of k. A Bayesian analysis was later carried out by Abraham and Box [18] with a somewhat different version of the model which assumes that the discrepant errors contain an unknown bias δ.

Either approach yields results which are broadly similar in that the posterior distribution of the parameters $\underset{\sim}{\theta}$ appears as a weighted sum of distributions.

$$p(\theta|y) = w_o p_o(\theta|y) + \sum_{i=1}^{n} w_i p_i(\theta|y) + \sum_{i=j}^{n} w_{ij} p_{ij}(\theta|y) + \ldots \tag{19}$$

The distribution $p_o(\underset{\sim}{\theta}|\underset{\sim}{y})$ in the first term on the right would be appropriate if all n observations were good, i.e. generated from the central model. The distribution $p_i(\underset{\sim}{\theta}|\underset{\sim}{y})$ in the first summation allows the possibility that $n - 1$ observations are good but the ith is bad. The next summation allows for two bad observations and so on. The weights w are posterior probabilities, w_o that no observation is bad, w_i that only the ith observation is bad, w_{ij} that the ith and jth observations are bad and so on.

Strictly the series includes all 2^n possibilities but in practice terms after the first or second summation usually become negligible.

Figure 10 shows an analysis for the Darwin data mentioned earlier. In this analysis it is supposed that the error distribution for good values is $N(0,\sigma^2)$ and that for bad values is $N(0,k^2\sigma^2)$. The analysis is made, as before, using a non-informative prior for θ and σ with $k = 3$ and $\alpha = .05$. This choice of α is equivalent to supposing that with 20 observations there is a 63.2% chance that one or more observations are bad. The probability of at least one outlier for other choices of n and α are given below

	α		
n	0.10	0.05	0.01
10	63.2	39.2	9.5
15	77.7	52.8	13.9
20	86.5	63.2	18.1
40	98.0	86.5	33.0

The results [17] are very insensitive to the choice of k but are less insensitive to the choice of α. However

1) it should be possible for the investigator to guess this value of α reasonably well.
2) the calculation can be carried out for different α values and the effect of different choices considered [18].

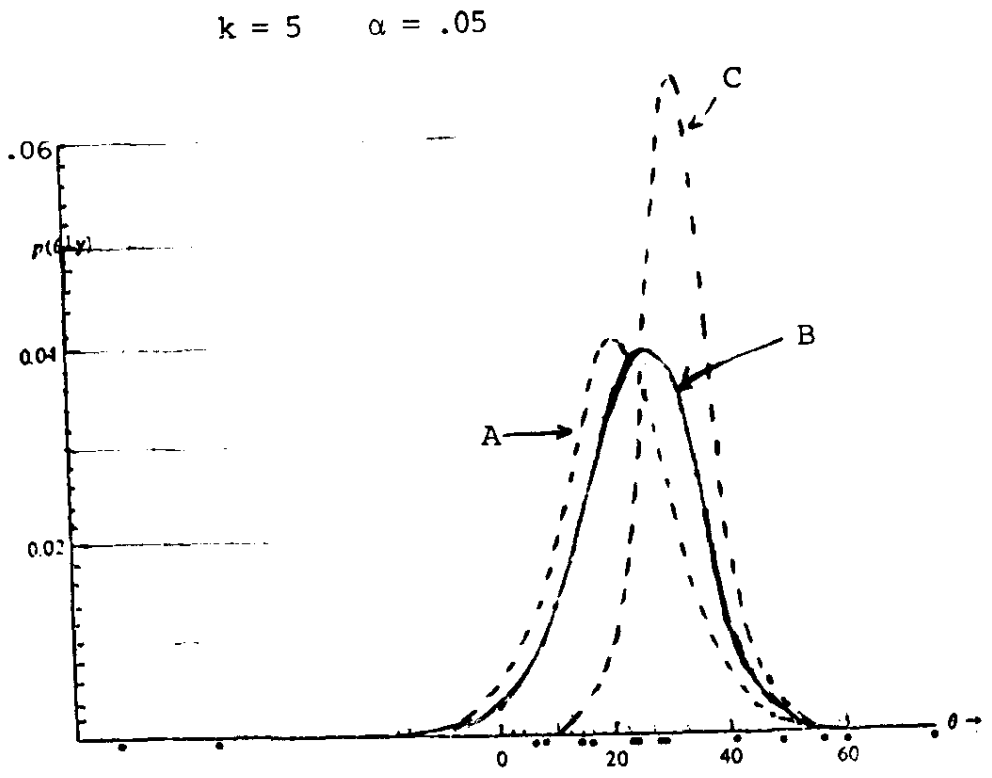


Figure 10. A. Assuming no outliers.
B. Allowing the possibility of outliers.
C. Assuming y_1 and y_2 are outliers.

Inspection of the weights w can also be informative in indicating possible outliers. For example [19] the time series shown in Figure 11 consists of 70 observations generated from the model:

$$y_t = \phi y_{t-1} + \delta_t + a_t$$

where

$$\delta_t = \begin{cases} 5 & \text{if } t = 50 \\ 0 & \text{otherwise} \end{cases} \qquad (20)$$

$\phi = .5$ and $\{a_t\}$ a set of independent normally distributed random variables with variance $\sigma_a^2 = 1$. The plot in Figure 12 of the weights w_i indicates the probability of each being bad and clearly points to discrepancy of the 50th observation.

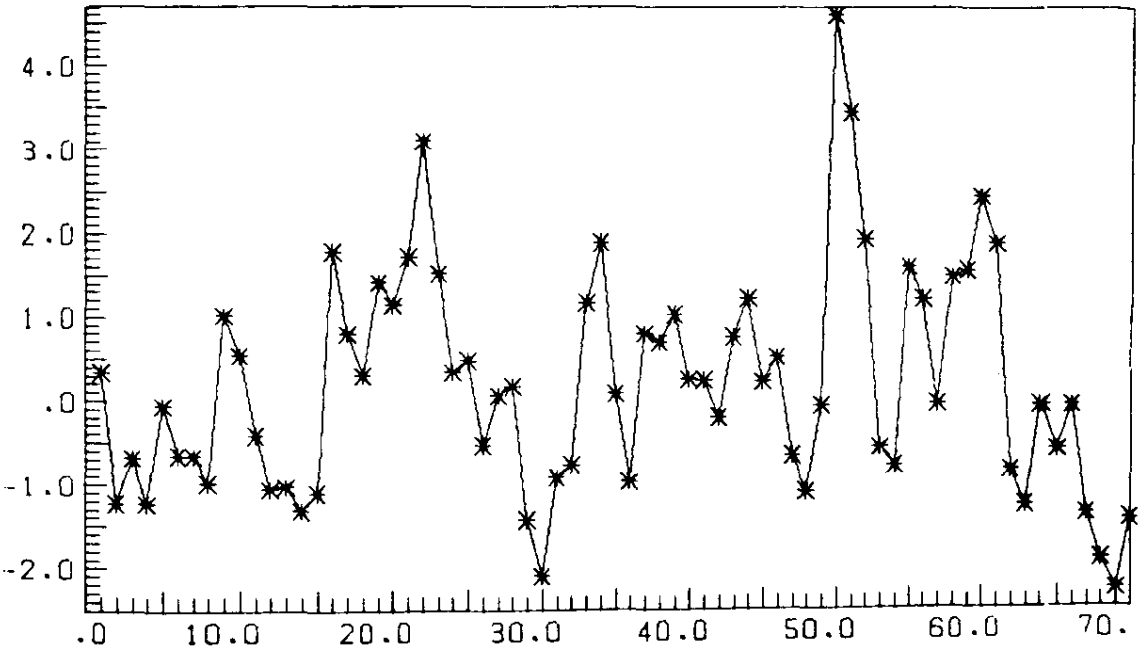


Figure 11. A time series generated from a first order autoregressive model with an outlier innovation at t = 50.

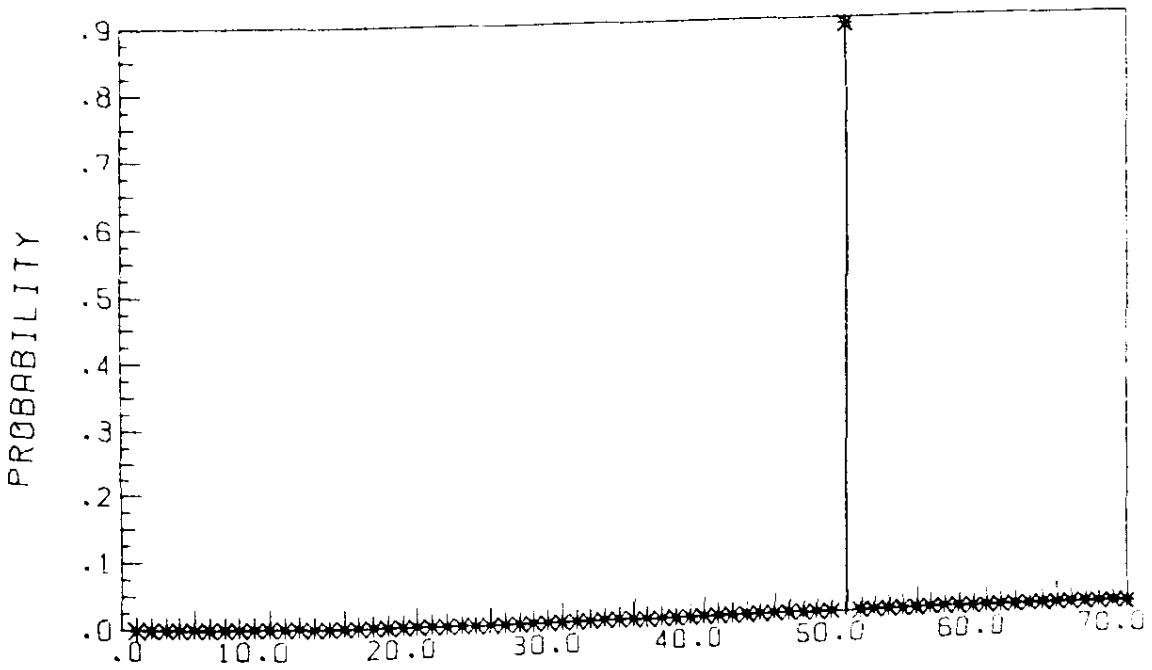


Figure 12. Posterior probabilities of bad values given that there is one bad value.

EXAMPLE 4: TRANSFORMATION OF THE DEPENDENT VARIABLE

Parsimony favors any device that expands model applicability with small expenditure of additional parameters. As was emphasized by Fisher, in suitable circumstances, parametric transformation can provide such a device. For example a suitable power transformation $Y = y^{\lambda}$ can have profound effect when y_{max}/y_{min} is not small.

In this application then the discrepancy parameter β measures the need for transformation. In particular for the power transformations no transformation corresponds to $\beta = \lambda = 1$.

The Bayes approach to parametric transformations was explored by Box and Cox [11]. One example they considered concerned a 3×4 factorial design with 4 animals per cell in which a total of $n = 48$ animals were exposed to three different poisons and four different treatments. The response was survival time. Since for this data $y_{max}/y_{min} = 12.4/1.8 \simeq 7$ we know a priori that the effect of needed transformation could be profound and it would be sensible to make provision for it in the first tentative model.

For this particular set of data, where there is a blatant need for transformation, an initial analysis with no transformation followed by iterative fix up would be effective also. Diagnostic checks involving residual plots of the kind suggested by Anscombe and Tukey [20], [21] certainly indicate [22] the dependence of cell variance on cell mean and less clearly non-additivity. Whatever route we take we are led to consider a transformation y^{λ} where λ approaches -1. As will be seen from the analysis of variance below this transformation not only eliminates any suggestion of an interaction between poisons and treatments but also greatly increases precision. This example seems to further illustrate how Bayesian robustification of the model illuminates the relation of the data to a spectrum of models. Using noninformative prior distributions Figure 13 shows posterior distributions for λ with different constraints applied to the basic normal, independent, model

$$y_{rci} = \mu_{rc} + \varepsilon_{rci} \qquad (21)$$

where the subscripts r, c, i apply to rows, columns and replicates.

Analyses of Variance of the Biological Data

	Mean squares × 1000			
	Degrees of freedom	Untransformed	Degrees of freedom	Reciprocal transformation (z form)
Poisons	2	516.3	2	568.7
Treatments	3	307.1	3	221.9
P × T	6	41.7	6	8.5
Within groups	36	22.2	36(35)	7.8(8.0)

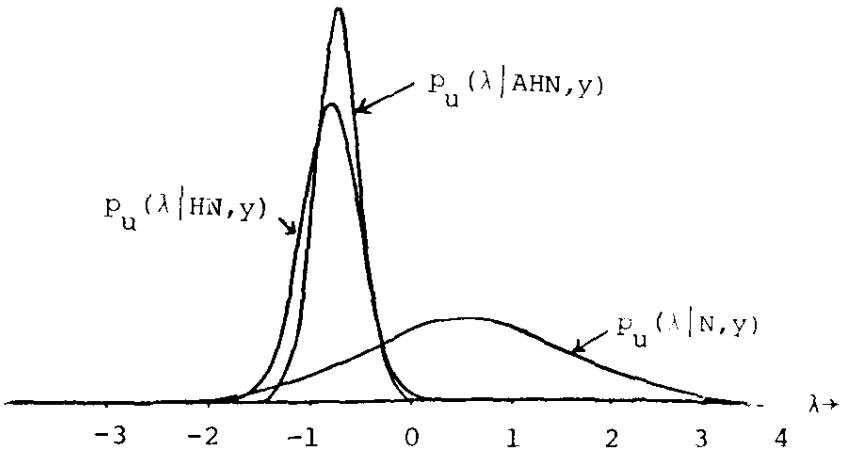


Figure 13. Posterior distributions for λ under various constraints N-Normality, Homogeneity of variance, A-Additivity.

The nature of the various distributions is indicated in the following table in which N, H, and A refer respectively to Normality, Homogeneity of Variance and Additivity and ρ_r and γ_c are row and column effects

Distribution	Constraint
$p_u(\lambda\|N,y)$	$V(\epsilon_{rci}) = \sigma^2_{rc}$
$p_u(\lambda\|HN,y)$	$\sigma^2_{rc} = \sigma^2$
$p_u(\lambda\|AHN,y)$	$\mu_{rc} = \mu + \rho_r + \gamma_c$ and $\sigma^2_{rc} = \sigma^2$.

The disperse nature of the distribution $p_u(\lambda|N,y)$ is to be expected since a sample of size $n = 48$ cannot tell us much about normality. The greater concentration of $p_u(\lambda|HN,y)$ arises because there is considerable variance heterogeneity in the original metric which is corrected by strong

transformations in the neighborhood of the reciprocal. Finally $p(\lambda|AHN)$ is even more concentrated because transformations of this type also remove possible non-additivity. The analysis suggests among other things that for this data the choices of transformations yielding approximate additivity, homogeneity of variance and normality are not in conflict. From Figure 14 (taken from [22]) we see that for this example appropriate adjustment of the discrepancy parameter $\beta = \lambda$ affects not only the location of the poison main effects, it also has a profound effect on their precision*. Indeed the effect of including λ in estimating main effects is equivalent to increasing the sample size by a factor of almost three.

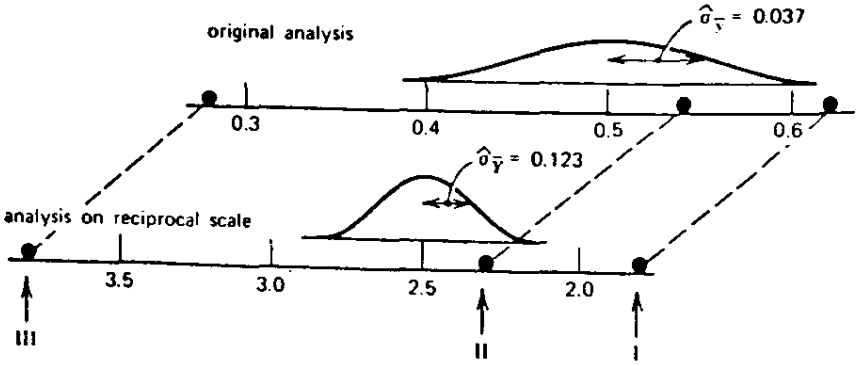


Figure 14. Posterior distributions for individual means (poisson main effects) on original and reciprocal scale. Note greatly increased precision induced by appropriate transformation.

PSYCHING OUT THE ROBUSTIFIERS

To apply Bayesian analysis we must choose a $p(\underset{\sim}{\beta})$ which roughly describes the world we believe applies in the problem context.

There are a number of ways we can be assisted in this choice.

(1) We can look at extensive sets of data and build up suitable distributions $p(\underset{\sim}{\beta})$ from experience.

* Similar results are obtained for the treatment effects and as noted before the transformation eliminates the need for interaction parameters.

(ii) We can consult experts who have handled a lot of data of the kind being considered.
(iii) We can consider the nature of the robust estimates proposed and what they reveal about the proposer's prior beliefs.

Consider, for example, the heavy tailed error problem. Gina Chen in a recent Ph.D. thesis [23] has found prior distributions $p(\underset{\sim}{\beta})$ yielding posterior means which approximate robust estimates of location already proposed on other grounds. In one part of her study she considers a model in which data come from an exponential power distribution. $p(y|\theta,\sigma,\beta)$ of the form of (11) with probability $1 - \alpha$, and, with probability α to have come from a similar distribution but with a standard deviation k times as large. Thus

$$p(y|\theta,\sigma,\alpha,\beta) = (1 - \alpha)p(y|\theta,\sigma,\beta) + \alpha \cdot p(y|\theta,k\sigma,\beta) \ . \qquad (22)$$

It turns out in fact that priors which put all the mass at individual points in the $p(\alpha,\beta)$ plane can very closely approximate suggested M estimators as well as trimmed means, Winsorized means, and other L estimators.

Three objectives of her study were
(i) To make it possible to examine more closely and hence to criticize the assumptions about the real world which would lead to the various robust estimates.
(ii) To compare these revealed assumptions with the properties of actual data.
(iii) To allow conclusions obtained from simple problems to be applied more generally. Once we agree on what $p(\underset{\sim}{\beta})$ should be for a location parameter then the same $p(\underset{\sim}{\beta})$ can be used for more complicated problems occurring in the same scientific context. Direct application of Bayes theorem can then, for example, indicate the appropriate analysis for all linear and nonlinear models formerly analyzed by least squares.

SUMMARY AND CONCLUSIONS

A major activity of statisticians should be to help the scientist in his iterative search for useful but necessarily inexact parsimonious models. While inexact models may

mislead, attempting to allow for every contingency a priori is impractical. Thus models must be built by an iterative feedback process in which an initial parsimonious model may be modified when diagnostic checks applied to residuals indicate the need.

When discrepancies may occur which are unlikely to be detected by diagnostic checks, this feedback process could fail and therefore procedures must be robustified with respect to these particular kinds of discrepancies. This writer believes that this may best be done by suitably modifying the model rather than by modifying the method of inference.

In particular a Bayes approach offers many advantages. Suppose the scientist wishes to protect inferences about primary parameters $\underset{\sim}{\theta}$ from effects of discrepancy parameters $\underset{\sim}{\beta}$. Bayes analysis automatically brings into the open a number of important elements.

(i) The prior distribution $p(\underset{\sim}{\beta})$ reveals the nature of the supposed universe of discrepancies from which the procedure is being protected.

(ii) The distribution $p_u(\underset{\sim}{\beta}|\underset{\sim}{y}) = p(\underset{\sim}{\beta}|\underset{\sim}{y})/p(\underset{\sim}{\beta})$ represents information about $\underset{\sim}{\beta}$ coming from the data itself. This distribution may be inspected for concordance with $p(\underset{\sim}{\beta})$.

(iii) The conditional posterior distribution $p(\underset{\sim}{\theta}|\underset{\sim}{\beta},\underset{\sim}{y})$ shows the sensitivity of inferences to choice of $\underset{\sim}{\beta}$.

(iv) From the marginal posterior distribution $p(\underset{\sim}{\theta}|\underset{\sim}{y})$ appropriate inferences which are robust with respect to $\underset{\sim}{\beta}$ may be made.

(v) Implications of inspired empiricism can lead to useful models. For example, we can ask "What kind of $p(\underset{\sim}{\beta})$ will make some empirical robust measure of location a Bayesian estimator?" This $p(\underset{\sim}{\beta})$ may then be examined, criticized and perhaps compared with distributions of $\underset{\sim}{\beta}$ encountered in the real world.

(vi) Once $p(\underset{\sim}{\beta})$ is agreed on then that same $p(\underset{\sim}{\beta})$ can be applied to other problems. For example, we do not need to give special consideration to robust regression, robust

analysis of variance, robust non-linear estimation. We simply carry through the Bayesian analysis with the agreed $p(\underset{\sim}{\beta})$.

(vii) In the past the available capacity and speed of computers might have limited this approach but this is no longer true. It will be necessary however, to make a major effort to produce suitable programs which can readily perform analyses and display results of the kind exemplified in this paper.

APPENDIX 1

Suppose that, in model (1), n observations are available and standard assumptions of independence and homoscedasticity are made about the errors $\{\varepsilon_u\}$. Suppose finally that the object is to estimate $E(y)$ over a region in the space of $\underset{\sim}{\xi}$ "covered" by the experiments $\{\underset{\sim}{\xi}_u\}$. Then the number of parameters p employed in the expectation function is a natural measure of prodigality and its reciprocal $1/p$ of parsimony.

Now denote by $\hat{y}_u^{(p)} = f^{(p)}(\underset{\sim}{\xi}_u, \underset{\sim}{\hat{\theta}}_p)$ a fitted value with estimates $\underset{\sim}{\hat{\theta}}_p$ obtained by least squares and by

$$\bar{V}^{(p)} = \sum_u V\{\hat{y}_u^{(p)}\}/n \qquad (A.1)$$

the average variance of the n fitted values.

It is well known that (exactly if the expectation function is linear in $\underset{\sim}{\theta}$, and in favorable circumstances, approximately otherwise) no matter what experimental design $\{\underset{\sim}{\xi}_u\}$ is used

$$\bar{V}^{(p)} = p\sigma^2/n \ . \qquad (A.2)$$

Now if the $\{\eta_u\}$ can be regarded as a sampling of the function over the region of interest, then $\bar{V}^{(p)}$ provides a measure of average variance of estimate of the function over the experimental region.

Equation (A.2) says that this average variance of estimate of the function is proportional to the prodigality p. Alternatively it is reasonable to regard $I^{(p)} = \{\bar{V}^{(p)}\}^{-1}$ as a measure of information supplied by the experiment about the function and

$$I^{(p)} = n/p\sigma^2 \ . \qquad (A.3)$$

Thus this measure of information is proportional to the parsimony $1/p$. For example, if the expectation function needed as many parameters as there were observations so that $p = n$ then $\hat{y}_t = y_t$ and

$$\bar{V}^{(n)} = V(y_t) = \sigma^2 \qquad I^{(n)} = 1/\sigma^2 \ . \tag{A.4}$$

In this case the model does not summarize information and does not help in reducing the variance of estimate of the function.

At the other extreme if the model needed to contain only a single parameter, for example,

$$y_t = \theta + \varepsilon_t \tag{A.5}$$

then $\hat{\theta} = \hat{y}_t = \bar{y}$ and

$$\bar{V}^{(1)} = V(\bar{y}) = \sigma^2/n \qquad I^{(1)} = n/\sigma^2 \ . \tag{A.6}$$

In this case the use of the model results in considerable summarizing of information and reduces the variance of estimate of the function n times or equivalently increases the information measure n-fold.

Considerations of this sort weigh heavily against unnecessarily complicated models.

As an example of unnecessary complication consider an experimenter who wished to model the deviation $\tilde{y}_t$ from its mean of the output from a stirred mixing tank in terms of the deviation $\tilde{\xi}_t$ from its mean of input feed concentration. If data were available to equal intervals of time, he might use a model

$$\tilde{y}_t = \{\theta_o\tilde{\xi}_t + \theta_1\tilde{\xi}_{t-1} + \theta_2\tilde{\xi}_{t-2} + \cdots + \theta_k\tilde{\xi}_{t-k}\} + \varepsilon_t \tag{A.7}$$

in which k was taken sufficiently large so that deviations in input ξ_{t-k-q} for $q > 0$ were assumed to have negligible effect on the output at time t. This model contains $k + 1$ parameters $\underset{\sim}{\theta}$ which need to be estimated. Alternatively if he knew something about the theory of mixing he might instead tentatively entertain the model

$$\tilde{y}_t = \theta_o\{\tilde{\xi}_t + \theta_1\tilde{\xi}_{t-1} + \theta_1^2\tilde{\xi}_{t-2} + \cdots\} + \varepsilon_t \tag{A.8}$$

or equivalently

$$\tilde{y}_t = \theta_1 \tilde{y}_{t-1} + \theta_o \tilde{\xi}_t + \varepsilon_t - \theta_1 \varepsilon_{t-1} \qquad (A.9)$$

which contains only two parameters $\underset{\sim}{\theta}$.

Thus if the simpler model provided a fair approximation, it could result in greatly increased precision as well as understanding.

APPENDIX 2

The practical importance of worrying about the right things is illustrated, for example, by the entries in the following table taken from [7], [22]. This shows the results of a sampling experiment designed to compare the robustness to non-normality and to serial correlation of the significance level of the t test and the non-parametric Mann-Whitney test. One thousand pairs of samples of ten of independent random variables u_t were drawn from a rectangular distribution, a normal distribution and a highly skewed distribution (a χ^2 with 4 degrees of freedom) all adjusted to have mean zero. In the first row of the table the errors $\varepsilon_t = u_t$ were independently distributed, in the second and third rows a moving average model $\varepsilon_t = u_t - \theta u_{t-1}$ was used to generate errors with serial correlation -0.4 and $+0.4$ respectively. The numbers on the right show the corresponding results when the pairs of samples were randomized.

In this example the performance characteristic studied is the numbers of samples showing significance at the 5% level when the null hypotheses of equality of means was in fact true. Under ideal assumptions the number observed would, of course, vary about the expected value of 50 with a sampling standard deviation of about 7. It is not intended to suggest by this example that the performance of significance tests when the null hypothesis is true is the most important thing to be concerned about. But rightly or wrongly, designers of non-parametric tests have been concerned about it, and demonstrations of this kind suggest that their labors are to some extent misdirected. In this example it is evident that it is the physical act of randomization and much less so the choice of criterion that protects the significance level.

TABLE A.1

Tests of Two Samples of Ten Observations Having the Same Mean. Frequency in 1,000 Trials of Significance at the 5 Percent Level Using the t-Test (t) and the Mann-Whitney Test

			With No Randomization			After Randomization		
	ρ_1	Test	Rectangular	Normal	Chi-square*	Rectangular	Normal	Chi-square*
Independent Observations	0.0	t	56	54	47	60	43	49
		MW	43	45	43	58	41	44
Autocorrelated Observations	-0.4	t	5	3	1	48	55	63
		MW	5	1	2	43	49	56
	0.4	t	125	105	114	59	58	54
		MW	110	96	101	46	53	43

*This parent chi-square distribution has four degrees of freedom and is thus highly skewed.

REFERENCES

1. Box, G.E.P. and Jenkins, G.M. (1970). Time Series Analysis: Forecasting and Control, Holden-Day.
2. Box, G.E.P., Hillmer, S.C. and Tiao, G.C. (1976). Analysis and Modelling of Seasonal Time Series. Technical Report No. 465, Department of Statistics, University of Wisconsin, Madison.
3. Watson, James D. (1968). The Double Helix. New York: Atheneum. A personal account of the discovery of the structure of DNA.
4. Box, J.F. (1978). R.A. Fisher, The Life of a Scientist. Wiley, Chapter 7.
5. Box, G.E.P. and Youle, P.V. (1955). "The Exploration and Exploitation of Response Surfaces: An Example of the Link Between the Fitted Surface and the Basic Mechanism of the System", Biometrics 11, 287.
6. Box, G.E.P. and Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley, p. 305, Chapter 5.
7. Box, G.E.P. (1976). "Science and Statistics", JASA 71, 791.
8. Tukey, J.W. (1949). "One Degree of Freedom for Non-additivity", Biometrics 5, 232.
9. Coen, P.G., Gomme, E.D. and Kendall, M.G. (1969). "Lagged Relationships in Economic Forecasting", JRSS, Series A 132, 133.
10. Box, G.E.P. and Newbold, Paul (1971). "Some Comments on a Paper of Coen, Gomme, and Kendall, JRSS, Series A 134, 229.
11. Box, G.E.P. and Cox, D.R. (1964). "An Analysis of Transformations", JRSS, Series B 26, 211.
12. Tiao, G.C. and Box, G.E.P. (1973). "Some Comments on "Bayes" Estimators", The American Statistician 27, No. 1, 12.
13. Box, G.E.P. and Tiao, G.C. (1962). "A Further Look at Robustness via Bayes's Theorem", Biometrika 49, 419.
14. Box, G.E.P. and Tiao, G.C. (1964b). "A Note on Criterion Robustness and Inference Robustness", Biometrika 51, 169.

15. Pallesen, L.C. (1977). "Studies in the Analysis of Serially Dependent Data", Ph.D. thesis, the University of Wisconsin, Madison.
16. Tukey, J.W. (1960). "A survey of sampling from contaminated distributions" in Contributions to Probability and Statistics, Stanford University Press, 448-485.
17. Box, G.E.P. and Tiao, G.C. (1968). "A Bayesian Approach to Some Outlier Problems", Biometrika 55, 119.
18. Abraham, B. and Box, G.E.P. (1978). "Linear Models and Spurious Observations", Appl. Statist. 27, 131-138.
19. Abraham, B. and Box, G.E.P. (1979). "Outliers in Time Series", (to appear) Biometrika, see also Technical Report #440, Department of Statistics, University of Wisconsin, Madison.
20. Anscombe, F.J. (1961). "Examination of Residuals", Proceedings of 4th Berkeley Symposium, Math. Statist. Proc. 1, 1.
21. Anscombe, F.J. and Tukey, W.J. (1963). "The Examination and Analysis of Residuals", Technometrics 5, 141.
22. Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). "Statistics for Experimenters", John Wiley.
23. Chen, Gina (1979). "Studies in Robust Estimation", Ph.D. thesis, the University of Wisconsin, Madison.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

Department of Statistics and
Mathematics Research Center
University of Wisconsin-Madison
Madison, WI 53706