CentER
for
Economic Research

Discussion paper

[redacted]

Tilburg University

Center
for
Economic Research

go

No. 9660

A CLASSICAL PROBLEM IN LINEAR REGRESSION OR HOW TO ESTIMATE THE MEAN OF A UNIVARIATE NORMAL DISTRIBUTION WITH KNOWN VARIANCE

By Jan R. Magnus and J. Durbin

June 1996

t regression analysis
v linear regression
t statistical distribution
t estimation
t variance

ISSN 0924-7815

A classical problem in linear regression or how to estimate the mean of a univariate normal distribution with known variance *

By

Jan R. Magnus

CentER for Economic Research, Tilburg University, The Netherlands

and

London School of Economics and Political Science, London, UK

and

J. Durbin

London School of Economics and Political Science, London, UK.

April 1996

*Preliminary versions of this paper were presented at Tilburg University, Humboldt University, LSE, Yale, the University of Amsterdam and the New Economic School of Moscow. We are grateful to the participants at these seminars. In particular we thank Jim Berger, Bernd Droge, Ed Leamer, Peter Phillips, Hans Schumacher and Aart de Vos for their useful comments, Anurag Banerjee for his help in producing the graphs and Inez Hoondert for expert and cheerful typing.

Title: A classical problem in linear regression or how to estimate the mean of a univariate normal distribution with known variance.

Author: Jan R. Magnus and J. Durbin.

Affiliation: CentER for Economic Research, Tilburg University, The Netherlands and London School of Economics and Political Science, London, UK.

Corresponding author:
Jan R. Magnus
CentER for Economic Research
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands

phone: +31-13-466-3092
fax: +31-13-466-3066
e-mail: magnus@kub.nl

Keywords: regression analysis, model selection, biased estimation, univariate normal mean, pretesting, mean squared error criterion, minimax regret, admissibility, Bayes estimation, Laplace distribution.

Abstract: The problem considered is the estimation of k coefficients of interest in a linear regression model when the $(k+1)st$ coefficient is of no interest. It is shown that this problem is equivalent to the problem of estimating the unknown mean of a univariate normal distribution with variance one given a single observation. Questions of admissibility, risk and regret are studied for this problem. The traditional pretest estimator of the mean is shown to have undesirable properties. A shrinkage estimator with better performance than the pretest estimator is considered. Further improvements in performance are achieved by estimators derived from the Burr family of distributions. However, these estimators are found to be still open to several objections. Generalizations of the Burr estimators are developed which are free from some or all of these objections. The final estimator considered is a Bayes estimator chosen from this class with a Laplace distribution as its prior. Of all the estimators studied in the paper, this has the best all-round performance. The optimal Burr estimator, however, is a close competitor.

1 Introduction

The purpose of this paper is to develop an optimal method of estimation under model uncertainty. Our starting point is the linear regression model

$$y = X\beta + Z\gamma + u, \quad u \sim N(0, \sigma^2 I_n),$$

where X is an $n \times k$ matrix of explanatory variables that are required to be in the model on theoretical or other grounds, while Z is an $n \times l$ matrix of additional explanatory variables about which there is doubt as to whether they should be in the model or not.[1] We are interested in estimating β (or specified linear combinations of its elements), while the value of γ is of no interest to us. In standard parlance, β is a vector of parameters of interest and γ is a vector of nuisance parameters. The only reason for including Z in the model is that by doing so we expect to obtain a 'better' estimate of β. In this context we assess the relative performance of estimators by the mean squared error (MSE) criterion.

The most common approach to this problem is to test the hypothesis that $\gamma = 0$ and to include Z if the hypothesis is rejected and exclude it otherwise. Inference on parameter estimates is then carried out as if the resulting model were correct. Inferences made in this way will, however, be invalid since the model selection procedure influences the properties of the estimator. Two classical problems can be distinguished. The first problem is that of selecting the set of regressors to be included in the regression. The second problem, more relevant to applied workers, is that of determining the implications of data mining (or model selection) on the estimates of the parameters of interest. This paper is concerned with the second problem and offers a practical solution to it. Both problems were already heavily discussed following Tinbergen's (1939) monumental study for the League of Nations. Both Keynes (1939) and Friedman (1940), in their respective critiques of Tinbergen's work, focused on the method of model selection when the estimation procedure repeatedly uses the same data to discriminate among plausible competing theories. The same point was made in Haavelmo's (1944, Section 17) seminal paper. Koopmans (1949) suggested that a completely new theory of inference was required to solve the dilemmas implied by the model selection problem. The theory of pretest estimation is an attempt to address this problem, but, as we shall see, not

[1] In Leamer's terminology, X contains the 'free' variables and Z the 'doubtful' variables. There is considerable confusion about these names and about Leamer's extreme bounds analysis. See Leamer (1978, 1983), McAleer, Pagan and Volker (1985), Leamer (1985), McAleer and Veall (1989), Breusch (1990) and Granger and Uhlig (1990).

a succesful one.[2] Leamer's (1978) book provides many important new insights, Sawa (1978) and Amemiya (1980) discuss the problem of selection of regressors, Lovell (1983) investigates certain consequences of data mining, Pötscher (1991) and Zhang (1992) consider asymptotic properties of estimators after model selection. See Chatfield (1994) for a recent survey and Draper (1995) for a Bayesian perspective.

For simplicity we assume that Z contains a single explanatory variable ($l = 1$) and that the disturbance variance σ^2 is known. The problem under consideration is therefore to estimate k coefficients of interest (or linear combinations of these) in a regression model containing $k + 1$ regressors using a mean squared error criterion in a situation where the value of the $(k + 1)st$ coefficient is of no interest. We call this *the regression problem*. Remarkably, it turns out that this problem is equivalent to another problem which appears on the face of it to be quite different, namely to find the 'best' estimator for θ when we have one observation x from a $N(\theta, 1)$ distribution. As we shall show, this problem is not as easy as it appears at first sight. We call this *the* $N(\theta, 1)$ *problem*. We shall show that the $N(\theta, 1)$ problem is the nucleus of the regression problem. A full solution to both problems is provided. This solution, we believe, should be acceptable to both frequentists and Bayesians.

Section 2 considers the relation between the regression problem and the $N(\theta, 1)$ problem. Let b_r be the maximum likelihood estimator (MLE) of β assuming that $\gamma = 0$, let b_u be the MLE of β assuming that γ is unknown and let θ be the value of γ divided by the standard error of its MLE. In the traditional approach to the regression problem a choice is made between b_r and b_u depending on the outcome of a test of the hypothesis $\theta = 0$. A smoother and more appealing procedure is to estimate β as a weighted average of b_u and b_r, that is, $b = \lambda(\hat{\theta})b_u + (1 - \lambda(\hat{\theta}))b_r$, where $\hat{\theta}$ is the MLE of θ and $\lambda(\cdot)$ is a function satisfying appropriate conditions. Let $\omega = \lambda(\hat{\theta})\hat{\theta} - \theta$. Remarkably, Theorem 2.2 shows that the MSE of b depends only on the MSE of ω. Since $\hat{\theta} \sim N(\theta, 1)$, it follows that the regression problem and the $N(\theta, 1)$ problem are equivalent. This explains the title of the paper and it also explains why the next five sections are concerned solely with the $N(\theta, 1)$ problem, even though our original target was the regression problem.

In section 3 we consider estimators of θ of the form $t = \lambda(x)x$ where $x \sim N(\theta, 1)$. Taking first the case $\lambda(x) = \lambda$ (constant), we show that $\inf_{\lambda}$ MSE $(\lambda x) = \theta^2/(1 + \theta^2)$.

[2]See Judge and Bock (1978) and Judge and Yancey (1986) for a survey of pretest estimation.

Interestingly, Theorem 3.7 demonstrates that the same infimum holds for the estimator t when $\lambda(x)$ is allowed to vary over a wide class of functions. Starting from this infimum we define the *regret* of the estimator for a particular value of θ as $\mathrm{MSE}(t) - \theta^2/(1+\theta^2)$. Based on this definition we consider minimax regret and minimum average regret estimators.

The traditional pretest estimator has the form $b = b_r$ if $|\hat{\gamma}| \leq c\ (\mathrm{SE}(\hat{\gamma}))$ and $b = b_u$ if $|\hat{\gamma}| > c\ (\mathrm{SE}(\hat{\gamma}))$, where $\hat{\gamma}$ is the MLE of γ, $\mathrm{SE}(\hat{\gamma})$ is its estimated standard error and $c > 0$. The equivalent pretest estimator for the $N(\theta, 1)$ problem is $t = 0$ if $|x| \leq c$ and $t = x$ if $|x| > c$. Section 4 demonstrates a number of undesirable properties of this estimator. First, it is inadmissible. Secondly, there is a range of values of θ for which the MSE of t is greater than the MSE of both the 'usual' estimator $t = x$ and the 'silly' estimator $t = 0$. Thirdly, in the neighbourhood of the value $\theta = 1$, which we show to be of crucial significance at various points of the paper, the MSE is maximized when $c = 1.91$. This means that when the traditional pretest is carried out at the usual 5% level, the resulting estimator is close to having worst possible performance.

In section 5 we consider an estimator, named by us the HTF estimator, which has the form $t = x^3/(c^2 + x^2)$. This is a generalization of estimators suggested by previous writers and it has strong intuitive appeal based on Theorem 3.7. We show that, while still inadmissible, the HTF estimator has better properties than the pretest estimator.

Section 6 introduces a more general class of estimators of which the 'usual', pretest and HTF estimators are special or limiting cases. Estimators in this class have the form $t = \lambda(x)x$ with $\lambda(-x) = \lambda(x)$, where, for $x \geq 0$, $\lambda(x)$ is a distribution function belonging to the Burr family defined by (6.1). This class has lower minimax regret and lower minimum average regret than the HTF class. The 'optimal Burr estimator' (the minimax regret estimator in the Burr class) has a simple form (with $\lambda(x)$ defined in (6.7) with $c = 0.545$) as well as a strong performance and would therefore appear to be a strong candidate for 'the best' estimator for θ. There are, however, several objections which can be raised against the optimal Burr estimator: it is inadmissible, it is not smooth at $x = \pm c$, it does not depend on x when $|x| \leq c$ and it is not a Bayes estimator.

Section 7 considers generalizations of the Burr class of estimators. One of these leads to an estimator which meets all four objections to the Burr family and is optimal or nearly optimal in other respects. We begin by asking whether there is a Bayes estimator of form (7.2) and we then show that this is so when the prior is the Laplace (or double exponential) density. We take as our estimator the posterior mean, which has the form $t = x - h(x)c$, where $h(x)$ is given by (7.8), and we show that this estimator is free from the objections to the Burr class. Since the estimator depends on c we need to select an appropriate value. We find that the minimax regret value is $c = 0.66$. However, the desirability of choosing a neutral prior for θ (that is, a prior density which is symmetric around zero and has the property that the median of θ^2 is one) leads to the value $c = \log 2 = 0.69$. Since we attach greater priority to the neutrality of the prior than to minimax regret, and since the two values of c are very close, we choose as our final estimator the estimator from the Laplace prior with $c = \log 2$. We call this the 'ideal' Laplace estimator. It can be seen from Figure 7.1 that the performances of the two Laplace estimators are very close for all θ (1 = minimax regret, 2 = 'ideal' Laplace). Various properties of the estimator are discussed.

Section 8 relates the $N(\theta, 1)$ results to the regression problem and discusses the relative merits of the optimal Burr estimator and the 'ideal' Laplace estimator. It presents the main conclusions of the paper and discusses possible extensions for future work.

2 Equivalence of the regression problem and the $N(\theta, 1)$ problem

In the regression problem we are concerned with the estimation of (linear combinations of the elements of) β in the linear regression model

$$y = X\beta + \gamma z + u, \tag{2.1}$$

where $y(n \times 1)$ is the vector of observations, $X(n \times k)$ and $z(n \times 1)$, both non-random, represent the values of the regressors, $u(n \times 1)$ is a random vector of unobservable disturbances, and $\beta(k \times 1)$ and the scalar γ are unknown non-random parameters. We assume that the design matrix $(X : z)$ has full column-rank and that the disturbances $u_1, \ldots, u_n$ are i.i.d. $N(0, \sigma^2)$, that is,

$$u \sim N(0, \sigma^2 I_n), \quad \sigma^2 > 0. \tag{2.2}$$

We assume also that σ^2 is known. This is, of course, unrealistic but it simplifies the analysis without affecting the main results. We shall come back to this assumption in our concluding remarks. We introduce the following notation:

$$M = I_n - X(X'X)^{-1}X', \tag{2.3}$$

$$q = \frac{\sigma}{\sqrt{z'Mz}}(X'X)^{-1}X'z, \tag{2.4}$$

$$\theta = \frac{\gamma}{\sigma/\sqrt{z'Mz}}. \tag{2.5}$$

(Notice that the rank condition implies that $z'Mz > 0$.) The idempotent matrix M and the vector q are known non-random quantities, while θ is unknown (since γ is unknown). The parameter θ plays an important role in our analysis and we shall call it the *theoretical t-ratio*.

The maximum likelihood (OLS) estimators of β and γ are

$$b_u = b_r - \hat{\theta} q \tag{2.6}$$

and

$$\hat{\gamma} = z'My/z'Mz, \tag{2.7}$$

where

$$b_r = (X'X)^{-1}X'y \tag{2.8}$$

and

$$\hat{\theta} = \frac{\hat{\gamma}}{\sigma/\sqrt{z'Mz}}. \tag{2.9}$$

The subscripts 'u' and 'r' stand for 'unrestricted' and 'restricted' (with $\gamma = 0$), respectively. It is clear that

$$\hat{\theta} \sim N(\theta, 1). \tag{2.10}$$

We emphasize again that our interest lies in the estimation of (linear combinations of the elements of) β, the parameters of interest, while γ is essentially a nuisance parameter.

The traditional approach, used by the large majority of applied econometricians and statisticians, is to consider only two estimators of β: the restricted estimator b_r defined in (2.8) (where $\gamma = 0$) and the unrestricted estimator b_u defined in (2.6). The choice between them is then based on the t-ratio $\hat{\theta}$ and the estimator b for β can thus be written as

$$b = \begin{cases} b_r, & \text{if} \quad |\hat{\theta}| \leq c, \\ b_u, & \text{if} \quad |\hat{\theta}| > c, \end{cases} \tag{2.11}$$

for some $c \geq 0$. For example, $c = 1.96$ and $c = 2.58$ correspond to the 5% and the 1% significance levels respectively. Hence, the estimator employed in the traditional approach is (2.11), even though the large majority of investigators acts as if the estimator is either b_r or b_u.

Given that we are not interested in γ but only in the best possible estimation of β, this procedure makes very little sense. We are testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ (or equivalently, $\gamma = 0$ against $\gamma \neq 0$) and this gives us an answer to the question: Is it true that $\theta = 0$? But this is the wrong question in this context. The right question is: Is b_r a better estimator for β than b_u? If we agree to judge an estimator by its mean squared error,

$$\mathrm{MSE}(b) \equiv \mathrm{E}(b-\beta)(b-\beta)' = \mathrm{var}(b) + \mathrm{E}(b - \mathrm{E}b)(b - \mathrm{E}b)', \tag{2.12}$$

then the following theorem will help us to answer this question.

Theorem 2.1.[3] We have

$$\mathrm{MSE}(b_r) - \mathrm{MSE}(b_u) = (\theta^2 - 1)qq'$$

and hence

$$\begin{array}{lclll} \mathrm{MSE}(b_r) & \leq & \mathrm{MSE}(b_u) & \text{if} & \theta^2 < 1 \\ \mathrm{MSE}(b_r) & = & \mathrm{MSE}(b_u) & \text{if} & \theta^2 = 1 \\ \mathrm{MSE}(b_r) & \geq & \mathrm{MSE}(b_u) & \text{if} & \theta^2 > 1. \end{array}$$

Theorem 2.1 shows, still assuming that b_r and b_u are the only two estimators to choose between, that the choice should be based on the null hypothesis $H_0 : \theta^2 \leq 1$ against the alternative $H_1 : \theta^2 > 1$. Toro-Vizcarrondo and Wallace (1968) made this point and they showed that a uniformly most powerful test is obtained from the probability

$$\Pr\left(|\hat{\theta}| \leq c | \theta^2 = 1\right) = 1 - \alpha, \tag{2.13}$$

[3] This result is 'well-known' in the sense that, in some form, it has been around for a long time, but not in the sense that many econometricians know it. The earliest reference is possibly exercise 12 in Wold (1953, p. 246) which he attributes to J. Durbin. See Wallace (1964) and Leamer (1983, p. 308).

where α denotes the level of significance.[4] For example, a 5% test and a 1% test now correspond to $c = 2.65$ and $c = 3.33$, respectively. Application of the second test obviously leads to more frequent use of the restricted estimator.

We notice that both the first and the second test lead to an estimator b for β of the form (2.11). There are two problems in applying either of these two tests. The first is that the choice of significance level is largely arbitrary. The second problem is that, after the preliminary test in which we decide whether to use b_r or b_u, neither of these two is the correct estimator for β. The correct estimator is given in (2.11) and is known as the *traditional pretest estimator*. Let us rewrite (2.11) as

$$b = \lambda(\hat{\theta})b_u + (1 - \lambda(\hat{\theta}))b_r, \tag{2.14}$$

where

$$\lambda(\hat{\theta}) = \begin{cases} 0 & \text{if} \quad |\hat{\theta}| \leq c, \\ 1 & \text{if} \quad |\hat{\theta}| > c. \end{cases} \tag{2.15}$$

This formulation shows b as a weighted average of b_u and b_r. Any estimator of β of the form (2.14), where λ is a real-valued function of $\hat{\theta}$ satisfying certain regularity conditions will be called a *weighted-average least squares* (WALS) estimator. The pretest estimator is a very simple (and not a very good) example of such an estimator. It is intuitively appealing to think of a WALS estimator as a continuous function of $\hat{\theta}$ such that the larger is $|\hat{\theta}|$ the more weight is given to b_u relative to b_r. WALS estimation can thus be viewed as a stepwise regression procedure, where we first obtain $\hat{\theta}$ and then b as a function of $\hat{\theta}$. As such it relates to Mallow's (1973) C_p criterion, the Akaike (1974) Information Criterion, Sawa's (1978) BIC criterion and Amemiya's (1980) PC criterion. Judge et al. (1985) show that all of these are essentially stepwise regression procedures based on $\hat{\theta}$. The mean, variance and mean squared error of the WALS estimator b, defined in (2.14), are given in Theorem 2.2.

<u>Theorem 2.2</u>. Let $\omega = \lambda(\hat{\theta})\hat{\theta} - \theta$. Then the conditional distribution of b given $\hat{\theta}$ is

---

[4] See also Bancroft (1944, 1964), Wallace (1964), Wallace and Toro-Vizcarrondo (1969), Wallace (1972), Goodnight and Wallace (1972), Wallace and Ashtar (1972) and Feldstein (1973).

$$b|\hat{\theta} \sim N(\beta - \omega q, \sigma^2(X'X)^{-1})$$

and the (unconditional) mean, variance and MSE of b are

$$\begin{aligned}
\mathrm{E}b &= \beta - (\mathrm{E}\omega)q \\
\mathrm{var}(b) &= \sigma^2(X'X)^{-1} + (\mathrm{var}(\omega))qq' \\
\mathrm{MSE}(b) &= \sigma^2(X'X)^{-1} + (\mathrm{E}\omega^2)qq'.
\end{aligned}$$

Our task is to find an optimal weighting function λ such that the WALS estimator b is best in the sense of having lowest mean squared error. One glance at the last line of Theorem 2.2 reveals that MSE(b) is minimized if and only if $\mathrm{E}\omega^2$ is minimized, that is, if and only if $\mathrm{MSE}(\lambda(\hat{\theta})\hat{\theta})$ is minimized. The striking consequence of Theorem 2.2 is therefore that finding the best WALS estimator for β is equivalent to finding the best estimator $\lambda(\hat{\theta})\hat{\theta}$ for θ. If we estimate a linear combination of the elements of β, say $R\beta$, by Rb, where b is a WALS estimator for β, then MSE (Rb) is minimized when MSE(b) is minimized. Hence, even if we are interested in estimating a particular linear combination of the β's, we need to find the best WALS estimator for β first. It is therefore permitted to think of the regression problem as one where we wish to estimate the complete β-vector.

Although we have emphasized that we are interested in the parameters β and not in θ (that is, γ) which is a nuisance parameter, we now see, ironically, that optimal WALS estimation for β depends completely on finding the optimal estimator for θ. The only information we have on θ is that $\hat{\theta} \sim N(\theta, 1)$, see (2.10). Hence, the regression problem is solved if and only if we can solve the $N(\theta, 1)$ problem: given one observation $x \sim N(\theta, 1)$, find the best estimator for θ (of the form $\lambda(x)x$, but this is of no consequence). To the solution of this problem we now turn.

3 Risk, admissibility and regret

Let x be a single observation from a univariate normal distribution with unknown mean θ and variance 1, that is, $x \sim N(\theta, 1), -\infty < \theta < \infty$. We wish to estimate θ using estimators of the form

$$t(x, \lambda) = \lambda(x)x, \quad \lambda \in \mathcal{L}, \tag{3.1}$$

where $\mathcal{L}$ is some class of real-valued functions to be specified below. The class of estimators will be denoted $\mathcal{T}(\mathcal{L})$. Thus,

$$\mathcal{T}(\mathcal{L}) = \{t : t = t(x, \lambda), \lambda \in \mathcal{L}\}. \tag{3.2}$$

Throughout we assume squared error loss $(t(x, \lambda)-\theta)^2$. The *risk* of an estimator $t \in \mathcal{T}(\mathcal{L})$ is then defined as its mean squared error,

$$R(\theta, \lambda) = \mathrm{E}_\theta(t(x, \lambda) - \theta)^2, \quad \lambda \in \mathcal{L}, \tag{3.3}$$

where E_θ denotes expectation with respect to the $N(\theta, 1)$ distribution. For any $\lambda_1 \in \mathcal{L}, \lambda_2 \in \mathcal{L}$ we say that $t_1 = t(x, \lambda_1)$ *dominates* $t_2 = t(x, \lambda_2)$ if

$$R(\theta, \lambda_1) \leq R(\theta, \lambda_2) \quad \text{for all } \theta \tag{3.4}$$

with strict inequality for some θ. An estimator $t \in \mathcal{T}(\mathcal{L})$ is said to be $\mathcal{L}$-*admissible* if no estimator in $\mathcal{T}(\mathcal{L})$ dominates t. If t is $\mathcal{L}$-admissible for every $\mathcal{L}$, then t is *admissible*. If t is dominated by some estimator t^*, not necessarily in $\mathcal{T}(\mathcal{L})$, then t is *inadmissible*. If t is dominated by $t^* \in \mathcal{T}(\mathcal{L})$, then t is $\mathcal{L}$-*inadmissible*. Thus an estimator can be (and often will be) inadmissible, but still $\mathcal{L}$-admissible for some $\mathcal{L}$. (See Theorem 4.2 for an example.)

An estimator $t(x, \lambda^*)$ is said to be $\mathcal{L}$-*minimax* if

$$\sup_\theta R(\theta, \lambda^*) = \inf_{\lambda \in \mathcal{L}} \sup_\theta R(\theta, \lambda) \tag{3.5}$$

for some $\lambda^* \in \mathcal{L}$. If $t(x, \lambda^*)$ is $\mathcal{L}$-minimax for every $\mathcal{L}$ which contains λ^*, then $t(x, \lambda^*)$ is said to be *minimax*.

Let us first consider the 'usual' estimator for θ, $t(x, \lambda) = x$, obtained by choosing $\lambda \equiv 1$.

Theorem 3.1 ($\lambda \equiv 1$). The 'usual' estimator for θ, $t(x, \lambda) = x$, is unbiased, admissible, has constant risk equal to 1, and is the unique minimax estimator.

These are strong properties in favour of x as an estimator of θ.[5] Let us see why we might want to choose an estimator different from x. Define $\lambda_c^{(1)} = 1/(1+c)$ for all x (c constant and $\neq -1$), so that in particular $\lambda_0^{(1)}(x) = 1$ and $\lambda_\infty^{(1)}(x) = 0$.[6] Now consider

$$t^{(1)}(x; c) \equiv t(x, \lambda_c^{(1)}) = x/(1+c) \tag{3.6}$$

as an estimator of θ. The risk is

$$R^{(1)}(\theta, c) \equiv R(\theta, \lambda_c^{(1)}) = \mathrm{E}_\theta(x/(1+c) - \theta)^2 = \frac{1 + c^2\theta^2}{(1+c)^2} \tag{3.7}$$

and this is minimized at $c^* = 1/\theta^2$ with minimum risk

$$R^{(1)}(\theta, c^*) = \theta^2/(1+\theta^2). \tag{3.8}$$

It is easy to see that for $-1 < c < 0$, $t^{(1)}(x; c)$ is dominated by $t^{(1)}(x; 0)$ and similarly, for $c < -1$, that $t^{(1)}(x; c)$ is dominated by $t^{(1)}(x; \infty)$. Hence $t^{(1)}(x; c)$ is inadmissible for $c < 0$. For $0 \leq c \leq \infty$ we have

[5]The admissibility of the 'usual' estimator is not trivial. If we wish to estimate θ given one observation x from a p-dimensional normal distribution $N(\theta, I_p)$, then the 'usual' estimator $t(x) = x$ is admissible if $p = 1$ and $p = 2$, but inadmissible for $p \geq 3$ (Stein (1955)). This remarkable result gave rise to a huge literature on 'improved' (or Stein-rule) estimators for the mean of a (multivariate) normal distribution. We mention James and Stein (1961), Cohen (1965), Brown (1966, 1971), Sclove (1968), Baranchik (1970, 1973), Efron and Morris (1971, 1972, 1973), Bock (1975), Judge and Bock (1976), Stein (1981) and Berger (1982). See Judge and Bock (1978, Chapters 8 and 10) for a survey.

[6]The reason for not defining $\lambda_c^{(1)} = c$ is to facilitate comparison with $\lambda_c^{(2)} - \lambda_c^{(6)}$ to be defined later.

Theorem 3.2 (λ constant). For any c, $0 \le c \le \infty$, the estimator $t^{(1)}(x;c) = x/(1+c)$ is admissible.

The simple estimator (3.6) for $c \ge 0$ will be called the *normal Bayes estimator* of θ, because it is the Bayes estimator induced by a normal prior (see the proof of Theorem 3.2). In Figure 3.1 we graph the risk $R^{(1)}(\theta, c)$ as a function of θ for three values of c, labeled 1-3, namely $c = 0, 1$ and ∞. The dotted line gives the lower bound (3.8). The figure

FIGURE 3.1

confirms that no estimator dominates another. For example, at $c = 1$, the risk of the estimator $t^{(1)}(x;1)$ lies between the risks of $t^{(1)}(x;0)$ and $t^{(1)}(x;\infty)$, except when $\frac{1}{3}\sqrt{3} < \theta < \sqrt{3}$ in which case $t^{(1)}(x;1)$ is better (has lower risk) than both. This does not mean that $t^{(1)}(x;1)$ or any of the normal Bayes estimators is particularly desirable. But it does show that the 'usual' estimator $t^{(1)}(x;0)$ is not necessarily the best. In Figure 3.2 we have drawn three important relations between θ and c : $R^{(1)}(\theta, c) = 1$,

FIGURE 3.2

$R^{(1)}(\theta, c) = \theta^2$ and $R_c^{(1)\prime}(\theta, c) = 0$ (where $R^{(1)}$ reaches a minimum with respect to c). For each c, the interval between the curves $R^{(1)} = \theta^2$ and $R^{(1)} = 1$ determines the θ's where the normal Bayes estimator has lower risk than both $t^{(1)}(x;0)$ and $t^{(1)}(x;\infty)$.

Since for any $c \ge 0$ the estimator $t^{(1)}(x;c)$ is a weighted average of $t^{(1)}(x;0)$ and $t^{(1)}(x;\infty)$, let us compare these two estimators: the 'usual' one, $t^{(1)}(x;0) = x$, and the 'silly' one, $t^{(1)}(x;\infty) = 0$. Clearly,

$$R^{(1)}(\theta, 0) = 1, \quad R^{(1)}(\theta, \infty) = \theta^2 \tag{3.9}$$

and hence

$$R^{(1)}(\theta, \infty) \le R^{(1)}(\theta, 0) \quad \text{if and only if } |\theta| \le 1. \tag{3.10}$$

This is, of course, the essence of Theorem 2.1. It suggests that the 'usual' estimator $t^{(1)}(x;0) = x$ is good when $|\theta|$ is large, but not so good when $|\theta|$ is small. In fact, the

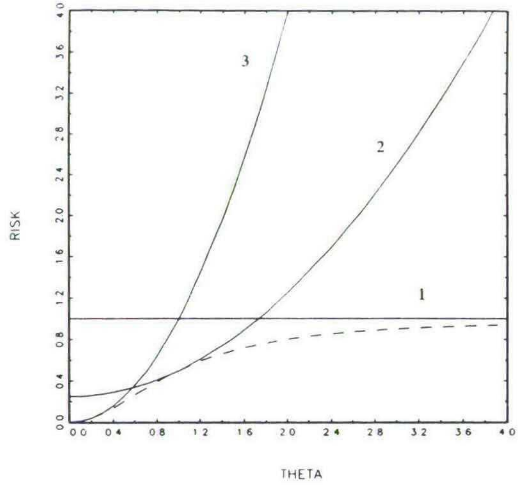

Figure 3.1. - Risk $R^{(1)}(\theta,c)$ of the normal Bayes estimator for three values of c.

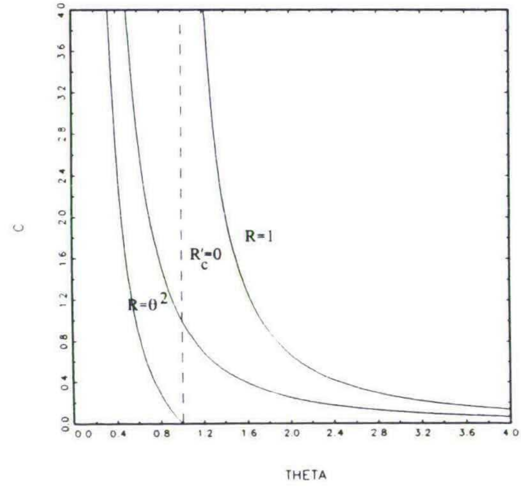

Figure 3.2. - Three relations between θ and c for the normal Bayes estimator $t^{(1)}$.

result in (3.8), that $R^{(1)}(\theta, c)$ is minimized at $c^* = 1/\theta^2$, tells us the same, namely that $\lambda = 0 (c = \infty)$ performs well when $|\theta|$ is small, $\lambda = 1 (c = 0)$ performs well when $|\theta|$ is large and that the larger is $|\theta|$ the larger should be λ. Since $\lambda_c^{(1)} = 1/(1+c)$ and the optimal c is given by $c^* = 1/\theta^2$, we find the optimal λ to be $\lambda^* = \lambda_{c^*}^{(1)} = \theta^2/(1+\theta^2)$. The optimal λ^*, as a function of θ, thus satisfies $0 \le \lambda^*(\theta) \le 1, \lambda^*(-\theta) = \lambda^*(\theta)$, and λ^* is increasing on $(0, \infty)$. Now, θ is not known. But if we think of x as a *preliminary estimator* of θ, then these ideas lead quite naturally and intuitively to the following minimal regularity conditions for λ.

Regularity Conditions R1: λ is a real-valued function defined on $\mathbf{R}$ and satisfies the following conditions:

(a) $0 \le \lambda(x) \le 1$ for all x,

(b) $\lambda(-x) = \lambda(x)$ for all x,

(c) λ is nondecreasing on $[0, \infty)$,

(d) λ is continuous except possibly on a set of measure zero.

Condition (a) defines $t(x, \lambda)$ as a shrinkage estimator (towards 0). This makes good sense, since, if $\lambda(x)$ were such that $\lambda(x) \ge \lambda > 1$ for all $|x| > M$, then $t(x) = \lambda(x)x$ would be inadmissible (Strawderman and Cohen (1971), Theorem 5.5.1). Condition (b) has several rationales. The simplest, perhaps, is the following. Let $\pi(\theta)$ be a prior density of θ. Since we are ignorant about θ, let us assume that $\pi(\theta)$ is symmetric around 0. Then the mean $t(x)$ of the posterior distribution of $\theta|x$, that is, the Bayes estimator for θ, satisfies $t(x) = -t(-x)$ and hence $\lambda(x) = \lambda(-x)$. Condition (c) makes sense too if we think of $t(x)$ as a weighted average of x and 0: $t(x) = \lambda(x) + (1 - \lambda(x))0$. The larger is $|x|$, the better is x as an estimator for θ. Hence, when $|x|$ increases we wish to put more weight on x and less on 0, that is, we wish to make $\lambda(x)$ larger. Condition (d), finally, is a minimal smoothness condition.

The class of functions satisfying regularity conditions R1 is denoted $\mathcal{L}^0$. Subclasses of $\mathcal{L}^0$ will be denoted $\mathcal{L}^{(1)}, \mathcal{L}^{(2)}, \ldots$. Thus, the class of normal Bayes estimators is denoted $\mathcal{L}^{(1)}$. In many cases we shall have $\lambda(0) = 0, \lambda(\infty) = 1$, so that λ can be interpreted as a distribution function on $[0, \infty)$. Condition (b) immediately leads to the following two results.

Theorem 3.3 (antisymmetry of the bias). Let $\mathrm{BIAS}(\theta, \lambda) = \mathrm{E}_\theta(t(x, \lambda) - \theta)$ denote the bias of $t(x, \lambda), \lambda \in \mathcal{L}^0$. Then,

(a) $\mathrm{BIAS}(\theta, \lambda) = -\mathrm{BIAS}(-\theta, \lambda)$,

(b) The only unbiased estimator of θ is $t(x, \lambda) = x$ (obtained for $\lambda = \lambda_0^{(1)} \equiv 1$),

(c) If $\lambda \neq \lambda_0^{(1)}$, then

$$\mathrm{BIAS}(\theta, \lambda) \begin{cases} > 0 & \text{if} \quad \theta < 0 \\ = 0 & \text{if} \quad \theta = 0 \\ < 0 & \text{if} \quad \theta > 0. \end{cases}$$

Theorem 3.4 (symmetry of the risk). For any estimator $t(x, \lambda), \lambda \in \mathcal{L}^0$, we have $R(\theta, \lambda) = R(-\theta, \lambda)$.

We know from Theorem 3.1 that the 'usual' estimator $t^{(1)}(x; 0)$ is unbiased. Theorem 3.3 shows, inter alia, that it is the only unbiased estimator. In view of Theorems 3.3 and 3.4 we shall report results for $\theta \geq 0$ only.

Let us reconsider the class $\mathcal{L}^{(1)}$ where $t^{(1)}(x; c) = x/(1 + c)$, $c \geq 0$. These estimators satisfy R1 and they are admissible. However, it is clear from (3.7) that their risk is unbounded unless $c = 0$. This property is unattractive, especially to non-Bayesians. Thus we shall often impose the following condition.

Regularity Condition R2 (bounded risk): Let $\varepsilon(x) = (1 - \lambda(x))x$. Then there exists a $K < \infty$ such that $|\varepsilon(x)| \leq K$ for all x.

Theorem 3.5 (Brown). Let $t(x, \lambda), \lambda \in \mathcal{L}^0$, be an estimator of θ. $R(\theta, \lambda)$ is bounded if and only if R2 holds.

Condition R2 requires that $\lambda(x)$ approaches 1 sufficiently fast as $x \to \infty$. In particular, estimators which, for large $|x|$, behave like

$$t(x, \lambda) = \left(1 - \frac{c}{|x|^p}\right) x \tag{3.11}$$

have bounded risk when $p \geq 1$.

We next address the question of admissibility. There are estimators, such as $t(x, \lambda) = 0$, which are admissible but nevertheless unappealing. There are other estimators, such as

$$t(x, \lambda) = \frac{x^2}{1 + x^2} x \quad \text{or} \quad t(x, \lambda) = \left(1 - e^{-x^2}\right) x \tag{3.12}$$

which have attractive properties, but can, surprisingly perhaps, be shown to be inadmissible (Strawderman and Cohen (1971)). The fact that we can prove an estimator to be inadmissible does not imply that we can find a better one. Our approach to the admissibility problem is therefore a compromise.[7] Admissibility is important (and eventually we shall arrive at an admissible estimator), but inadmissible estimators will be considered if they appear to be attractive otherwise. In order to prove (in)admissibility the following condition is required.

Regularity Condition R3: Let $\varepsilon(x) = (1 - \lambda(x))x$. Then

(a) $\varepsilon(x)$ is continuously differentiable,

(b) there exists a measure $G(\theta)$ such that

$$\exp\left[-\int_0^x \varepsilon(y)dy\right] = \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(x - \theta)^2\right] dG(\theta).$$

Actually, condition R3(a) is implied by R3(b). It is stated separately to emphasize that estimators which are not differentiable (or worse still, discontinuous) can not satisfy R3(b). We now have

Theorem 3.6 (admissibility). Let $t(x, \lambda)$ be an estimator for θ and assume that R1 holds. Then,

[7] A lot has been written about the importance of admissibility, see e.g. Berger (1985, section 4.8). Any 'kinked' estimator (that is an estimator $t(x)$ which is continuous but not everywhere differentiable, for example $t^{(4)}$ defined in (6.10)), is not differentiable and therefore not admissible (Theorem 3.6), but the difference in risk between the 'kinked' estimator and an admissible improvement can be miniscule. Some authors distinguish between inadmissible and 'seriously inadmissible', but we shall not do this.

(a) R3 is a necessary condition for $t(x, \lambda)$ to be admissible,

(b) R2 and R3 together are sufficient for $t(x, \lambda)$ to be admissible.

Theorem 3.6 is a powerful result which gives a complete characterization of the admissibility of bounded risk shrinkage estimators.

Finally, we need to discuss how we shall judge an estimator's performance. We shall do this by looking at the risk function $R(\theta, \lambda)$. But when two estimators are both $\mathcal{L}$-admissible in some class $\mathcal{L} \in \mathcal{L}^0$, then neither estimator dominates the other and some further criterion is needed. The minimax approach sometimes leads to unreasonable or trivial results (Hodges and Lehmann (1950)). Indeed, in our case, minimizing the maximum risk always leads to the 'usual' estimator $t(x, \lambda) = x$, see Theorem 3.1. On both theoretical and practical grounds (Savage (1951), Chernoff and Moses (1959), Sawa and Hiromatsu (1973)) we shall adopt the minimax regret approach where we minimize the maximum regret instead of the maximum risk. The *$\mathcal{L}$-regret* of an estimator $t \in \mathcal{T}(\mathcal{L})$ is defined as

$$r_{\mathcal{L}}(\theta, \lambda) = R(\theta, \lambda) - \inf_{\lambda \in \mathcal{L}} R(\theta, \lambda) \tag{3.13}$$

and an estimator $t(x, \lambda^*) \in \mathcal{T}(\mathcal{L}^*)$ is *$\mathcal{L}$-minimax regret* with respect to $\mathcal{L}^* \subset \mathcal{L}$ if

$$\sup_{\theta} r_{\mathcal{L}}(\theta, \lambda^*) = \inf_{\lambda \in \mathcal{L}^*} \sup_{\theta} r_{\mathcal{L}}(\theta, \lambda). \tag{3.14}$$

In order to implement the minimax regret approach we require, for each θ, the lower bound of the risk $R(\theta, \lambda)$ over all estimators $t(x, \lambda), \lambda \in \mathcal{L}^0$. In (3.8) we showed that the class of estimators $t(x, \lambda_c^{(1)})$ has lower bound $\theta^2/(1+\theta^2)$. This, in fact, is the lower bound in $\mathcal{L}^0$ as well.

<u>Theorem 3.7.</u>

$$\inf_{\lambda \in \mathcal{L}^0} R(\theta, \lambda) = \frac{\theta^2}{1+\theta^2}.$$

We note that in the class $\mathcal{L}^{(1)}$ of normal Bayes estimators no minimax regret solution exists since the risk is not bounded. These estimators are therefore not acceptable from our point of view.

As an alternative to finding the minimum (over λ) of the *maximum* (over θ) regret, we could try and obtain the minimum of the *average* regret. This assumes specification of a prior distribution $\pi(\theta)$ of θ. Since the difference between average regret and average risk does not depend on λ, minimizing average regret is the same as minimizing average risk. In Bayesian parlance, average risk is the Bayes risk of an estimator $t(x,\lambda)$ with respect to π. If we minimize the average risk (Bayes risk) over all λ, then we obtain the Bayes estimator (which is the mean of the posterior distribution of θ given x). If we minimize the average risk (or, equivalently, the average regret) over a subset of $\mathcal{L}^0$, then the resulting estimator will, in general, not be the Bayes estimator.

Our ultimate goal in this paper is to find an estimator which performs well under both criteria: minimax regret and minimum average regret. Our approach to this goal is to concentrate on minimax regret first. Of each estimator of interest we report both the maximum regret and the average regret. In calculating average regret we must specify a prior density for θ. The choice of prior is, of course, difficult.[8] In view of Theorem 2.1 and (3.10), we are particularly interested in priors where the distribution of θ is located at 0 and the distribution of θ^2 is located at 1. Such a prior will be called *neutral*. (More on this at the end of section 4.) The simplest such distribution is $N(0,1)$.

Theorem 3.8 (average risk). Assume a $N(0,1)$ prior density for θ, denoted $\pi(\theta)$. Then the average risk of an estimator $t(x,\lambda), \lambda \in \mathcal{L}^0$, is given by

$$\mathrm{E}_\pi R(\theta,\lambda) = \frac{1}{2} + 4\int_0^\infty (\lambda(u\sqrt{2}) - \frac{1}{2})^2 u^2 \phi(u)du,$$

where E_π denotes expectation with respect to $\pi(\theta)$.

Theorem 3.8 allows us to calculate the average risk (regret) with respect to a $N(0,1)$ prior in a simple manner for any estimator $t(x,\lambda)$. We emphasize, however, that our purpose is not to find the minimum average regret estimator with respect to a $N(0,1)$ prior for θ. We know that the answer to this is $\lambda \equiv 1/2$, that is, $t(x,\lambda) = x/2$. (This is confirmed by Theorem 3.8.) The risk of this estimator is not bounded, but its expected

[8]This is the 'elicitation' problem. See Kadane et. al. (1980), Kadane and Winkler (1988), and Leamer (1992) for possible solutions to this problem and further references.

risk (with respect to the chosen prior) is 1/2 and the expected risk of any other estimator thus exceeds 1/2. We shall use average regret results with a $N(0,1)$ prior as a benchmark, but our primary purpose remains to obtain the best possible maximum regret estimator. We have already considered one subclass of $\mathcal{L}^0$, namely $\mathcal{L}^{(1)}$. In sections 4-7 we shall find $\mathcal{L}^0$-minimax regret estimators with respect to various other subclasses of $\mathcal{L}^0$. In section 4 we consider the class of traditional pretest estimators, which, in spite of their sad properties, are used in almost every applied econometrics paper.

4 The traditional pretest estimator

On the tiny remote island of I, the I-landers lived mainly on fish. Since the wind around the island varied from day to day, the I-landers had built two boats for their fishermen, the R-boat and the U-boat. The R-boat (R for 'rest') was ideal when there was no wind, the U-boat (U for 'unrest') in a heavy storm. Each evening after dinner the King announced his weather forecast for the next day upon which one of the two boats was prepared. An incorrect forecast of the weather and hence a wrong choice of boat could have serious consequences. All I-landers remembered the recent hurricane, which the King had failed to forecast and where the R-boat capsized, resulting in the death of all fishermen on board. One day a young adventurer A found himself stranded at I. A inspected the two fishing boats and found them well-built for their purpose. He noticed, however, that the extreme weather conditions for which the boats were designed rarely occurred. Most days at the island saw a moderate breeze. A decided to build a boat himself. After several months, his work completed, he proudly presented his new boat to the assembled I-landers. 'How does your boat perform when there is no wind?' asked one of the fishermen. 'Well,' said A, 'you can't expect my boat to do quite so well as your R-boat, which was built for the purpose, but it does definitely better than the U-boat when there is little or no wind.' 'And how does your boat perform in stormy weather?' asked a second fisherman. 'Again,' answered A, 'your U-boat does better in a storm, but my boat performs better than the R-boat. In particular, my boat will not capsize in a storm.' Then the King said: 'If the weather is fair, with a gentle breeze, what performance has your boat?' Somewhat embarrassed A replied: 'I must admit that under such conditions my boat performs worse that both the R-boat and the U-boat.' 'Throw this man into the ocean!' cried the King, and A was never heard of again.

We shall see in this section that the pretest estimator performs like the boat designed by our hero in the above story. The traditional pretest estimator for the regression problem is defined in (2.11).[9] The equivalent estimator for the $N(\theta, 1)$ problem is

$$t^{(2)}(x;c) \equiv t(x, \lambda_c^{(2)}) = \begin{cases} 0, & \text{if} \quad |x| \leq c, \\ x, & \text{if} \quad |x| > c, \end{cases} \tag{4.1}$$

[9] Early work on the traditional pretest estimator includes Bancroft (1944, 1964), Huntsberger (1955), Larson and Bancroft (1963), Cohen (1965), Wallace and Ashtar (1972), Sclove, Morris and Radhakrishnan (1972), Bock, Yancey and Judge (1973) and Bock, Judge and Yancey (1973). See the surveys by Judge and Bock (1978, 1983). For a Bayesian perspective see Zellner and Vandaele (1974), Leamer and Chamberlain (1976) and Giles and Rayner (1979).

where

$$\lambda_c^{(2)}(x) = \begin{cases} 0, & \text{if} \quad |x| \leq c, \\ 1, & \text{if} \quad |x| > c. \end{cases} \tag{4.2}$$

We shall refer to the estimator (4.1) as the pretest estimator as well, if there is no possible ambiguity. The class of λ-functions of type (4.2) is denoted $\mathcal{L}^{(2)}$, that is,

$$\mathcal{L}^{(2)} = \{\lambda : \lambda = \lambda_c^{(2)},\ 0 \leq c \leq \infty\}. \tag{4.3}$$

We notice that $\lambda_c^{(2)}$ is well defined at $c = 0$ and at $c = \infty$. At $c = 0$ we have $\lambda_0^{(2)} \equiv 1$ and $t^{(2)}(x; 0) = x$, the 'usual' estimator. At $c = \infty$ we have $\lambda_\infty^{(2)} \equiv 0$ and $t^{(2)}(x; \infty) = 0$, the 'silly' estimator.

The pretest estimator (4.1) has a discontinuity at $x = \pm c$ and we would therefore expect the estimator to behave badly. We shall see that this is indeed the case. Judge and Bock (1978) were the first to provide a thorough discussion of pretest estimators in a regression context, but they did not notice the essential equivalence mentioned above. We provide a full and simplified treatment of this important case.

We first obtain an expression for the risk function $R(\theta, \lambda_c^{(2)})$, which we shall write as $R^{(2)}(\theta, c)$.

Theorem 4.1. The risk of the pretest estimator $t^{(2)}(x; c)$ is given by

$$R^{(2)}(\theta, c) = 1 + (c + \theta)\phi(c + \theta) + (c - \theta)\phi(c - \theta) + (\theta^2 - 1)P(\theta, c),$$

where ϕ denotes the standard normal density and $P(\theta, c) = \int_{-\theta-c}^{-\theta+c} \phi(u)du$.

An alternative expression can be obtained from Bock (1975, Theorems A and B); see also Judge and Bock (1978).

There appears to be some confusion about the admissibility of the pretest estimator. This confusion arises because in the class of pretest estimators no estimator dominates any other. But outside this class there are estimators which dominate the pretest estimator, because of its discontinuity; see Cohen (1965). We have

Theorem 4.2. The estimator $t^{(2)}(x;c)$, $0 \leq c \leq \infty$, is

(a) admissible if $c = 0$ or $c = \infty$, inadmissible otherwise;

(b) $\mathcal{L}^{(2)}$-admissible.

Some further properties of the pretest estimator are given in Theorem 4.3.

Theorem 4.3. The risk $R^{(2)}(\theta, c)$ of the pretest estimator satisfies:

(a) $R^{(2)}(\theta, c)$ is symmetric in θ,

(b) $R^{(2)}(\theta, 0) = 1$, $R^{(2)}(\theta, \infty) = \theta^2$,

(c) $R^{(2)}(\theta, c)$ is bounded for every $c < \infty$,

(d) for every $c < \infty, R^{(2)}(\theta, c) \to 1$ as $|\theta| \to \infty$.

These properties are clarified in Figure 4.1, where $R^{(2)}(\theta, c)$ is given as a function of θ for seven different values of c, labeled 1-7. These values are, respectively: 0.0, 1.0, 1.2007, 1.3692, 1.96, 2.576, ∞. The figure confirms that no estimator in this class dominates any other (Theorem 4.2(b)), and that the risk is bounded and converges to 1 as $\theta \to \infty$ (unless $c = \infty$).

FIGURE 4.1

Closer inspection of Figure 4.1 reveals a particularly damaging property of the pretest estimator: for θ close to 0 we see, as expected, that the pretest estimator is better than the 'usual' estimator $t^{(2)}(x;0) = x$, but worse that the 'silly' estimator $t^{(2)}(x;\infty) = 0$. When θ is large, the situation is reversed. This, again, is what we would expect. However, for moderate values of θ, in particular around $\theta = 1$, we would like an improved estimator (such as the pretest estimator) to perform better (have lower risk) than both the 'usual' *and* the 'silly' estimator. Figure 4.1 shows that the opposite is the case![10] We

[10] This result is not new - it even appears in some of the textbooks, see Judge et al. (1985, p. 75) - but its harmful consequences don't seem to have been fully appreciated.

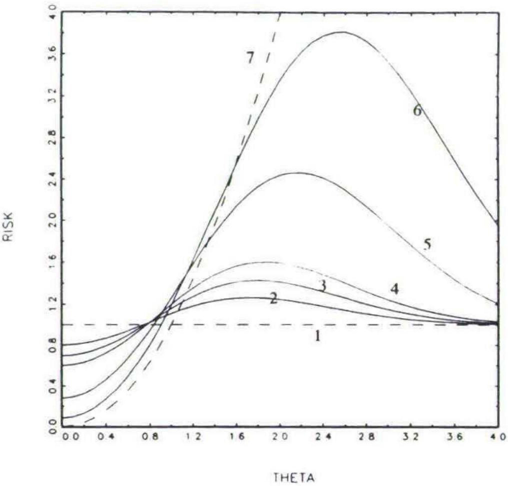

Figure 4.1. - Risk $R^{(2)}(\theta,c)$ of the pretest estimator for seven values of c.

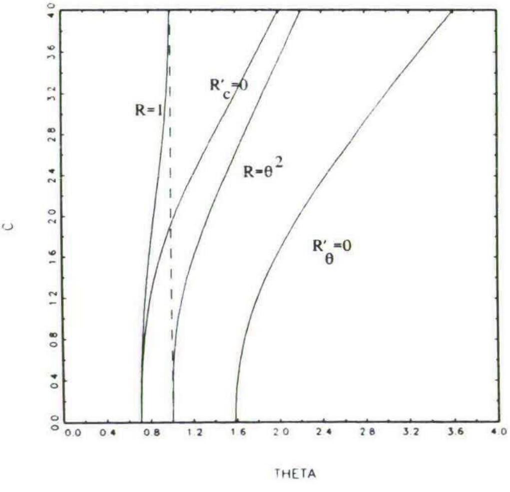

Figure 4.2. - Four relations between θ and c for the pretest estimator $t^{(2)}$.

shall return to this property shortly. We have seen that $t^{(1)}$ does not have this bad property and we shall see that other improved estimators also behave as we would expect. These other estimators thus either outperform both the 'usual' and the 'silly' estimator or they are better than one and worse than the other, but never are they worse than both. In the language of our story, it is possible to design a boat which performs well in all weather conditions, but this is not the boat designed by our hero.

The values for c in Figure 4.1 are chosen with care. We have $c = 0$ and $c = \infty$ as our extremes (in dotted lines). In between we have the important cases $c = 1, c = 1.96$ and $c = 2.576$. These last two values correspond of course to the usual 5% and 1% levels of significance. There are two further graphs, $c = 1.2007$ and $c = 1.3692$, corresponding to minimax regret solutions to be discussed later.

To obtain further insight we study the behaviour of the risk function for c close to 0 and for θ close to 0.

<u>Theorem 4.4</u>.

(a) For c close to 0,

$$R^{(2)}(\theta, c) = 1 + \frac{4}{3}\phi(\theta)(\theta^2 - \frac{1}{2})c^3 + \mathcal{O}(c^5);$$

(b) For θ close to 0,

$$R^{(2)}(\theta, c) = h_0(c) + h_1(c)\theta^2 + \mathcal{O}(\theta^4),$$

where

$$h_0(c) = 1 + 2c\phi(c) - \int_{-c}^{c} \phi(u)du, \quad h_1(c) = c(c^2 - 2)\phi(c) + \int_{-c}^{c} \phi(u)du;$$

(c) $0 \leq h_0(c) \leq 1,\ h_0'(c) \leq 0,\ h_1(c) \geq 0,\ h_1(c) = 1$ for $c = 1.6149$ and $c = \infty$ only.

There are four relations between θ and c of particular interest. These are graphed in Figure 4.2.

FIGURE 4.2

The first relation is $R^{(2)}(\theta, c) = 1$, giving all points where the risk of a pretest estimator equals the risk of the 'usual' estimator ($c = 0$). The relation can be solved explicitly for θ in terms of c. That is, there exists a unique function θ_0, defined on $[0, \infty)$, such that $R^{(2)}(\theta_0(c), c) = 1$. The function θ_0 is monotonically increasing and

$$\frac{1}{2}\sqrt{2} \leq \theta_0(c) \leq 1 \quad \text{for all } c \in (0, \infty], \tag{4.4}$$

using Theorem 4.4(a). The second relation is $R^{(2)}(\theta, c) = \theta^2$, which gives all points where the risk of a pretest estimator is equal to the risk of the 'silly' estimator ($c = \infty$). There exists a unique function θ_∞, defined on $[0, \infty)$, such that $R^{(2)}(\theta_\infty(c), c) = \theta_\infty^2$. The function θ_∞ is monotonically increasing and $\theta_\infty(c) \geq 1$, for all $c \geq 0$. Next we consider the relation $R_\theta^{(2)'}(\theta, c) \equiv \partial R^{(2)}(\theta, c)/\partial\theta = 0$, providing all points where $R^{(2)}(\theta, c)$, considered as a function of θ, attains a maximum (or a minimum, but this does not occur). It follows from Theorem 4.4(a) that, for c close to 0,

$$R_\theta^{(2)'}(\theta, c) = -\frac{4}{3}\theta(\theta^2 - \frac{5}{2})\phi(\theta)c^3 + \mathcal{O}(c^5). \tag{4.5}$$

Thus, there exists a unique function $\theta_{\max}$, defined on $(0, \infty)$ such that $R^{(2)}(\theta_{\max}(c), c) = \sup_\theta R^{(2)}(\theta, c)$. The function $\theta_{\max}$ is monotonically increasing and $\theta_{\max}(c) > \frac{1}{2}\sqrt{10}$ for $c > 0$. Furthermore, $R^{(2)}(\theta_{\max}(c), c)$ is monotonically increasing in c. The fourth relation in Figure 4.2 is $R_c^{(2)'}(\theta, c) \equiv \partial R^{(2)}(\theta, c)/\partial c = 0$. At these points, $R^{(2)}(\theta, c)$, considered as a function of c, attains a maximum! Again applying Theorem 4.4(a) we have, for c close to 0,

$$R_c^{(2)'}(\theta, c) = 4(\theta^2 - \frac{1}{2})\phi(\theta)c^2 + \mathcal{O}(c^4). \tag{4.6}$$

Thus, there exists a unique function $c_{\max}$, defined on $[\frac{1}{2}\sqrt{2}, \infty)$, such that $R^{(2)}(\theta, c_{\max}(\theta)) = \sup_c R^{(2)}(\theta, c)$. The functions $c_{\max}(\theta)$ and $R^{(2)}(\theta, c_{\max}(\theta))$ are both monotonically increasing in θ.

With θ_0 and θ_∞ as defined above we have, for every $c > 0$,

$$\begin{array}{lll} \theta^2 < R^{(2)}(\theta,c) \leq 1 & \text{if} & |\theta| \leq \theta_0(c), \\ R^{(2)}(\theta,c) > 1 \geq \theta^2 & \text{if} & \theta_0(c) < |\theta| \leq 1, \\ R^{(2)}(\theta,c) \geq \theta^2 > 1 & \text{if} & 1 < |\theta| \leq \theta_\infty(c), \\ 1 < R^{(2)}(\theta,c) < \theta^2 & \text{if} & |\theta| > \theta_\infty(c). \end{array} \tag{4.7}$$

This shows again that there exists a range of θ's where the pretest estimator has higher risk than both the 'usual' and the 'silly' estimator.

An immediate consequence of (4.7) is Theorem 4.5.

Theorem 4.5. In the class of pretest estimators,

$$\inf_c R^{(2)}(\theta,c) = \min(1,\theta^2).$$

In order to understand the importance of Theorem 4.5, let us consider once again the 'usual' estimator ($c = 0, R^{(2)}(\theta,c) = 1$) and the 'silly' estimator ($c = \infty, R^{(2)}(\theta,c) = \theta^2$). If we had to choose between these two we would choose the 'usual' estimator if $|\theta| > 1$ and the 'silly' one if $|\theta| \leq 1$. (Of course, θ is unknown.) If we now consider the whole class of pretest estimators, then Theorem 4.5 tells us that of all pretest estimators we would still choose the 'usual' estimator x if $|\theta| > 1$ and the 'silly' estimator 0 if $|\theta| \leq 1$. Thus, in the context of pretest estimation, the only question which appears to be important is whether $|\theta| \leq 1$ or $|\theta| > 1$. (We can go even one step further, see the discussion following Theorem 4.7.)

Given that $|\theta| = 1$ appears as the natural pivot in the class of pretest estimators, we would like to choose c such that, at the very least, the estimator performs well (has low risk) around $|\theta| = 1$. One glance at Figure 4.2 tells us that the worst estimator in this respect is obtained around $c = 1.9$ where the curve $R'_c = 0$ crosses the line $\theta = 1$. (The exact value is $c = 1.9150$.) This corresponds to a significance level of 5,55%. Thus, ironically, the usual 5% estimator ($c = 1.96$) is very close to being the worst choice of c in the sense that it gives the highest risk in the interval around $\theta = 1$!

Let us now consider the minimax regret solutions. We have two options here, since the minimum risk in $\mathcal{L}^0$ is $\theta^2/(1+\theta^2)$ (Theorem 3.7), while the minimum risk in $\mathcal{L}^{(2)}$ is higher, namely $\min(1,\theta^2)$ (Theorem 4.5). Thus, we have two regret functions:

$$r_0(\theta,c) = R^{(2)}(\theta,c) - \frac{\theta^2}{1+\theta^2} \tag{4.8}$$

and

$$r_2(\theta,c) = R^{(2)}(\theta,c) - \min(1,\theta^2). \tag{4.9}$$

Sawa and Hiromatsu (1973) obtained the $\mathcal{L}^{(2)}$-minimax regret solution, that is, they found c which minimizes $\sup_\theta r_2(\theta,c)$.[11] We provide both the $\mathcal{L}^{(2)}$ and the $\mathcal{L}^0$-minimax regret solutions. To prove that these solutions are unique, we notice that, for θ close to 0, both regret functions can be written as

$$r(\theta,c) = h_0(c) + (h_1(c)-1)\theta^2 + \mathcal{O}(\theta^4), \tag{4.10}$$

see Theorem 4.4(b), and hence both functions attain a local maximum at $\theta = 0$ if and only if $0 < c < 1.6149$ (Theorem 4.4(c)).

Theorem 4.6 (Minimax regret). With respect to $\mathcal{L}^{(2)}$, the class of pretest estimators,

(a) the $\mathcal{L}^0$-minimax regret estimator is obtained for $c = 1.2007$ with $\mathcal{L}^0$-minimax regret $r_0^* = 0.6958$;

(b) the $\mathcal{L}^{(2)}$-minimax regret estimator is obtained for $c = 1.3692$ with $\mathcal{L}^{(2)}$-minimax regret $r_2^* = 0.5988$.

The risk functions associated with $c = 1.2007$ and $c = 1.3692$ are graphed in Figure 4.1.

Finally, let us discuss the minimum average regret approach, discussed in section 3. In order to implement this approach we need to specify a prior distribution for θ. Which

[11] See also Farebrother (1975), Brook (1976) and Droge (1993).

prior should we choose? Toyoda and Wallace (1976) considered this question and chose a diffuse prior for θ. They then concluded that the 'usual' estimator x is the minimum average regret estimator. Given our previous results, this is not surprising. A diffuse prior on θ implies that $\Pr(|\theta| > 1) = 1$ and we know that x has lowest risk when $|\theta| > 1$. In our view, a diffuse prior is not appropriate. For example, it puts equal weight on the intervals (100, 101) and $(0, 1)$ and this is counterintuitive. In section 3 we suggested a $N(0,1)$ prior for θ. This prior is neutral between the 'usual' and the 'silly' estimator (the extremes).

For a normal prior, not necessarily neutral, we have

Theorem 4.7 (Minimum average risk). Let $\pi(\theta)$ be a $N(\mu, \tau^2)$ prior density of θ. Then,

(a) the average risk (with respect to π) of the pretest estimator is given by

$$\mathrm{E}_\pi R^{(2)}(\theta, c) = 1 - (\alpha + \beta\tau^2)\phi(\alpha) + (\beta + \alpha\tau^2)\phi(\beta) + (\mu^2 + \tau^2 - 1)\int_\alpha^\beta \phi(u)du,$$

where

$$\alpha = \frac{-c-\mu}{\sqrt{1+\tau^2}}, \quad \beta = \frac{c-\mu}{\sqrt{1+\tau^2}};$$

(b) the minimum average risk in $\mathcal{L}^{(2)}$ is

$$\inf_c \mathrm{E}_\pi R^{(2)}(\theta, c) = \min(1, \mu^2 + \tau^2)$$

and the minimum is obtained for

$$c_\pi = \begin{cases} \infty, & \text{if } \mu^2 + \tau^2 \leq 1, \\ 0, & \text{if } \mu^2 + \tau^2 > 1. \end{cases}$$

The surprising consequence of Theorem 4.7 is that, with a normal prior for θ, the minimum average regret solution is the 'usual' estimator x when $E\theta^2 > 1$ and the 'silly' estimator 0 when $E\theta^2 < 1$. This emphasizes once again that, for the traditional pretest estimator, the key issue is whether (the expected value of) θ^2 is larger or smaller than 1.

If our neutral prior for θ is $N(0,1)$, than Theorem 4.7(a) shows that the expected risk equals 1 *for every value of* c. (We used this result in the proof of Theorem 4.2.) This prior does therefore not yield a minimum average regret estimator.

c	maximum regret	average regret
0.0	1.0	0.6557
1.0	0.8013	0.6557
1.2007	0.6958*	0.6557
1.3692	0.8412	0.6557
1.96	1.6450	0.6557
2.576	2.9522	0.6557
∞	∞	0.6557

Table 4.1. - Maximum and average $\mathcal{L}^0$-regret for seven pretest estimators.

In Table 4.1 we present the maximum and average $\mathcal{L}^0$-regret of seven pretest estimators. In each case the regret function is r_0, given in (4.8). The maximum regret is minimized at $c = 1.2007$ in accordance with Theorem 4.6(a). (In this and subsequent tables a star * denotes the minimum.) The average regret is taken with a $N(0,1)$ prior. We know that the risk is always 1 in this case and since

$$\int_{-\infty}^{\infty} \frac{\theta^2}{1+\theta^2}\phi(\theta)d\theta = 0.3443 \tag{4.11}$$

we find a constant average regret of 0.6557.

We have discussed the traditional pretest estimator in some detail because it is an estimator which is routinely used in econometrics and other areas of applied statistics. This does not mean that we believe it is a good estimator. On the contrary, we have emphasized its poor properties throughout this section. We have seen in section 3 that the normal Bayes estimator $t^{(1)}$ is admissible, but has unbounded risk (unless $c = 0$) and therefore permits no minimax solution. In the current section we have seen that the pretest estimator $t^{(2)}$ is inadmissible (it is not even continuous) and has pathological behaviour around $|\theta| = 1$, but that its risk is bounded. We next turn to an estimator which, while still inadmissible, has considerable intuitive appeal and combines bounded risk with good behaviour around $|\theta| = 1$.

5 The Huntsberger-Thompson-Feldstein estimator

We know from Theorem 3.7 that $R(\theta, \lambda) \geq \theta^2/(1+\theta^2)$ for every θ and every $\lambda \in \mathcal{L}^0$. The lower bound is attained when $\lambda = \theta^2/(1+\theta^2)$. An obvious and suggestive choice for λ is therefore $\lambda(x) = x^2/(1+x^2)$. In this section we shall study a slight generalization of this, namely

$$\lambda_c^{(3)}(x) = \frac{x^2}{c^2 + x^2}, \quad 0 \leq c \leq \infty. \tag{5.1}$$

The class of λ-functions of this type is denoted $\mathcal{L}^{(3)}$. The estimator takes the simple form

$$t^{(3)}(x; c) \equiv t(x, \lambda_c^{(3)}) = \frac{x^3}{c^2 + x^2}. \tag{5.2}$$

For $c = 0$ and $c = \infty$ we again find the 'usual' and the 'silly' estimators as special cases.

The estimator $t^{(3)}$ was first considered by Huntsberger (1955) in the context of pooling two estimators based on a preliminary test of significance. Thompson (1968) proposed (a generalization of) this estimator as a shrinkage estimator, while Feldstein (1973) considered it in the context of multicollinearity. We shall refer to the estimator (5.2) as the HTF estimator.

In spite of its intuitive appeal the HTF estimator is inadmissible, although *within* the class of HTF estimators no estimator dominates another.

Theorem 5.1. The HTF estimator $t^{(3)}(x; c)$, $0 \leq c \leq \infty$, is

(a) admissible if $c = 0$ or $c = \infty$, inadmissible otherwise;

(b) $\mathcal{L}^{(3)}$-admissible.

Denoting the risk of the HTF estimator by $R^{(3)}(\theta, c)$ we find, again, that $R^{(3)}(\theta, c)$ is symmetric in θ, is bounded for every $c < \infty$ (but not for $c = \infty$) and approaches 1 as $|\theta| \to \infty$ for every $c < \infty$. In Table 5.1 we present the maximum and average $\mathcal{L}^0$-regret (that is, $R^{(3)}(\theta, c) - \theta^2/(1+\theta^2)$) for five values of c. Average regret is taken with respect to a $N(0,1)$ prior for θ.

c	maximum regret	average regret
0.0	1.0	0.6557
1.0	0.4670	0.3606
1.0920	0.4251*	0.3379
2.1647	1.1591	0.2327*
∞	∞	0.6557

Table 5.1. - Maximum and average $\mathcal{L}^0$-regret for five HTF estimators.

The minimax regret solution is this class of estimators is obtained for $c = 1.0920$ and the minimum average regret solution for $c = 2.1647$. The risk functions of these five estimators, labeled 1-5, are graphed in Figure 5.1.

FIGURE 5.1

The minimax regret estimator ($c = 1.0920$) in this class, in spite of being inadmissible, looks good. Its average regret is also quite acceptable. One particularly pleasing aspect of the HTF estimator (in contrast to the pretest estimator) is that, there now exists, for every c, an interval around $\theta = 1$, where the HTF estimator is better (has lower risk) than the 'usual' estimator ***and*** the 'silly' estimator. Also, there is no value of θ where the HTF estimator is worse than both the 'usual' and the 'silly' estimator.

In Figure 5.2 further insight is provided through the four relations $R = 1, R = \theta^2, R'_\theta = 0, R'_c = 0$. The most important difference with Figure 4.2 is that the curve $R = 1$ now lies to the right of $R = \theta^2$. The area between $R = \theta^2$ and $R = 1$ is the area where the HTF estimator performs better than both the 'usual' and the 'silly' estimators. We comment briefly on each of the four graphs.

FIGURE 5.2

(i) $R^{(3)}(\theta, c) = 1$: This relation implies a unique function θ_0, defined on $(0, \infty]$, such that $R^{(3)}(\theta_0(c), c) = 1$. The function θ_0 approaches 1.31 as $c \to 0$, reaches a maximum at $c = 0.90$ where $\theta_0(c) = 1.45$ (point A in the figure) and $1 \leq \theta_0(c) \leq 1.45$ for all $c \in (0, \infty]$.

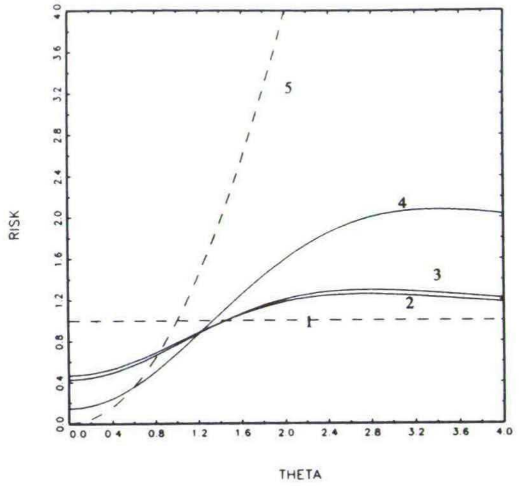

Figure 5.1. - Risk $R^{(3)}(\theta,c)$ of the HTF estimator for five values of c.

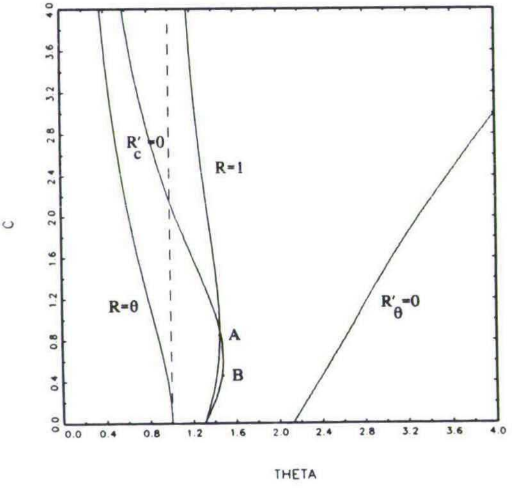

Figure 5.2. - Four relations between θ and c for the HTF estimator $t^{(3)}$.

(ii) $R^{(3)}(\theta, c) = \theta^2$: This implies a unique function θ_∞, defined on $[0, \infty)$, such that $R^{(3)}(\theta_\infty(c), c) = \theta_\infty^2$. The function θ_∞ is monotonically decreasing and $0 < \theta_\infty(c) \leq 1$ for all $c \geq 0$.

(iii) $R_\theta^{(3)'}(\theta, c) = 0$: A unique function $\theta_{\max}$ exists, defined on $(0, \infty)$, such that $R^{(3)}(\theta_{\max}(c), c) = \sup_\theta R^{(3)}(\theta, c)$. The function $\theta_{\max}$ is monotonically increasing and $\theta_{\max}(c) > 2.13$ for $c > 0$. Furthermore, $R^{(3)}(\theta_{\max}(c), c)$ is monotonically increasing in c.

(iv) $R_c^{(3)'}(\theta, c) = 0$: There exist two functions $c_{\min}$ and $c_{\max}$. The function $c_{\min}$, defined on $(0, 1.48]$, satisfies $R^{(3)}(\theta, c_{\min}(\theta)) = \inf_c R^{(3)}(\theta, c)$ and decreases monotonically. At $\theta = 1.48, c_{\min}(\theta) = 0.58$ (point B). The function $R^{(3)}(\theta, c_{\min}(\theta))$ defines the lower bound in $\mathcal{L}^{(3)}$. The function $c_{\max}$, defined on [1.31, 1.48], satisfies $R^{(3)}(\theta, c_{\max}(\theta)) = \sup_c R^{(3)}(\theta, c)$ and increases monotonically. $R^{(3)}(\theta, c_{\max}(\theta))$ increases monotonically as well.

At this point we conclude that the minimax regret estimator in the HTF class (with $c = 1.0920$) is our 'best' estimator so far. There are two reasons why we believe a better estimator can be found. First, the estimator is inadmissible and therefore a better estimator must exist. Secondly, the class of HTF-estimators is small. In a larger class we expect to find an estimator with lower maximum regret. In the next section we define a much larger class of estimators of which all three classes $\mathcal{L}^{(1)}, \mathcal{L}^{(2)}$ and $\mathcal{L}^{(3)}$ are special or limiting cases.

6 The Burr family

In search of a more general class of λ-functions we begin by noticing that any λ-function satisfying R1 for which $\lambda(0) = 0$ and $\lambda(\infty) = 1$ is a distribution function on $[0, \infty)$. So our objective is to select an appropriate class of distribution functions. It would seem desirable that this class of distribution functions has as special or limiting cases the three classes of λ-functions discussed so far ($\lambda = \lambda_c^{(1)}$ (normal Bayes), $\lambda = \lambda_c^{(2)}$ (pretest), $\lambda = \lambda_c^{(3)}$ (HTF)) and that $\lambda(x)$ is given explicitly in terms of x and not as an integral. The class of distribution functions we shall use is

$$\lambda_c^B(x; \alpha, \beta) = 1 - [1 + (x^2/c^2)^\alpha]^{-\beta}, \tag{6.1}$$

where $\alpha > 0, \beta > 0$ and c is again a scale parameter. This distribution function was first proposed by Burr (1942). Burr and Cislak (1968) showed that the Burr family covers important regions of many well-known distribution functions.

For $\alpha \to 0, \lambda_c^B$ approaches $1 - 2^{-\beta}$, a constant $\in (0,1)$. For $\alpha \to \infty, \lambda_c^B \to \lambda_c^{(2)}$. For $\alpha = 1, \beta = 1, \lambda_c^B = \lambda_c^{(3)}$. So the Burr family (6.1) contains all three previously discussed estimators as special or limiting cases.

Since estimators based on $\lambda_c^{(3)}$ are inadmissible, Burr estimators cannot, in general, be admissible, although certain limiting cases (like $\alpha \to 0$, but not $\alpha \to \infty$) will be admissible.

Defining, as usual, $\varepsilon(x) = (1 - \lambda(x))x$, we obtain

$$\varepsilon(x) = \frac{c^{2\alpha\beta} x}{(x^{2\alpha} + c^{2\alpha})^\beta}, \tag{6.2}$$

and hence, for large x,

$$\varepsilon(x) \approx c^{2\alpha\beta} x^{-(2\alpha\beta - 1)}. \tag{6.3}$$

This leads to

Theorem 6.1. The risk $R(\theta, c; \alpha, \beta)$ of an estimator $t(x, \lambda_c^B)$ in the Burr class is bounded in θ if and only if $2\alpha\beta \geq 1$. Furthermore, for $|\theta| \to \infty$,

$$R(\theta, c; \alpha, \beta) \to \begin{cases} 1 + c^2 & , \text{ if } \; 2\alpha\beta = 1 \\ 1 & , \text{ if } \; 2\alpha\beta > 1. \end{cases}$$

Since minimax regret solutions cannot exist when the risk is unbounded, we must impose $2\alpha\beta \geq 1$. For specific values of β we now find the optimal values α^*, c^*, θ^* for which the maximum (over θ) of the regret function is minimized (over α and c). The minimax regret, which is a function of β, is denoted r^*.

β	α^*	$2\,\alpha^*\beta$	c^*	θ^*	r^*
10.0	0.40	8.00	24.24	3.28	0.414
5.0	0.43	4.30	8.43	3.24	0.411
2.0	0.52	2.08	2.24	3.18	0.405
1.0	0.69	1.38	1.01	3.09	0.397
0.5	1.08	1.08	0.64	3.00	0.390
0.2	2.50	1.00	0.55	2.81	0.385
0.1	5.00	1.00	0.55	2.74	0.385
0.0	∞	1.00	0.54	2.73	0.385

Table 6.1. - Minimax regret results for the Burr family.

Table 6.1 (and many more calculations not reported here) suggests very strongly that the optimal estimator from the minimax regret point of view is obtained by letting $2\alpha\beta = 1$ and $\alpha \to \infty, \beta \to 0$. In order to find this limit, we let $h(t, \alpha) = (1 + t^\alpha)^{1/\alpha}$, for $t \geq 0$. A simple application of l'Hôspital's rule then yields

$$\lim_{\alpha \to \infty} h(t, \alpha) = \begin{cases} 1 & , \text{ if } \; 0 \leq t \leq 1, \\ t & , \text{ if } \; t > 1. \end{cases} \tag{6.4}$$

Now writing

$$\lambda_c^B(x; \alpha, \beta) = 1 - [h(x^2/c^2, \alpha)]^{-\alpha\beta}, \tag{6.5}$$

we find that, along the path $2\alpha\beta = 1$,

$$\lim_{\alpha\to\infty} \lambda_c^B(x;\alpha,\beta) = \lambda_c^{(4)}(x), \tag{6.6}$$

where

$$\lambda_c^{(4)}(x) = \begin{cases} 0 & , \text{ if } |x| \le c, \\ 1 - \dfrac{c}{|x|} & , \text{ if } |x| > c. \end{cases} \tag{6.7}$$

The estimator $t^4(x;c) = t(x, \lambda_c^{(4)})$ will be called the *optimal Burr estimator* and the class of λ-functions defined by (6.7) is denoted $\mathcal{L}^{(4)}$. It is remarkable that the optimal Burr estimator, obtained as the minimax regret solution for the very large Burr class, should take such a simple form. Its risk is given in Theorem 6.2.

Theorem 6.2. The risk of the optimal Burr estimator $t^{(4)}(x;c)$ is given by

$$R^{(4)}(\theta,c) = 1 + c^2 - (c-\theta)\phi(c+\theta) - (c+\theta)\phi(c-\theta) + (\theta^2 - 1 - c^2)P(\theta,c)$$

where $P(\theta,c) = \int_{-\theta-c}^{-\theta+c} \phi(u)du$.

Unfortunately the optimal Burr estimator is 'kinked', hence not differentiable and thus inadmissible by Theorem 3.6.

Theorem 6.3. The optimal Burr estimator $t^{(4)}(x;c), 0 \le c \le \infty$, is

(a) admissible if $c = 0$ or $c = \infty$, inadmissible otherwise;

(b) $\mathcal{L}^{(4)}$-admissible.

The risk $R^{(4)}(\theta,c)$ is bounded, but it does not approach 1 as $|\theta| \to \infty$. In view of Theorem 6.1 we have

$$R^{(4)}(\theta,c) \to 1 + c^2 \quad \text{as } |\theta| \to \infty. \tag{6.8}$$

The maximum and average $\mathcal{L}^0$-regret for five selected values of c is presented in Table 6.2.

c	maximum regret	average regret
0.0	1.0	0.6557
0.545	0.3850*	0.2927
0.866	0.8164	0.2508*
1.0	1.0603	0.2564
∞	∞	0.6557

Table 6.2. - Maximum and average $\mathcal{L}^0$-regret for five optimal Burr estimators.

The minimax regret solution in the class of optimal Burr estimators (and hence in the whole Burr class) is obtained for $c = 0.545$ and the minimum average regret in $\mathcal{L}^{(4)}$ is at $c = 0.866$. Compared with the minimax regret estimators in the HTF class we see that the best estimator in the optimal Burr class has not only lower maximum regret (0.3850 versus 0.4251) but also lower average regret (0.2927 versus 0.3379). The risk functions of the five estimators in Table 6.2 are labeled 1-5 and graphed in Figure 6.1.

FIGURE 6.1

Before we graph various relations between c and θ we prove the following theorem which relates to the behavior of the risk when c or θ is close to 0 (compare Theorem 4.4).

Theorem 6.4.

(a) For c close to 0,

$$R^{(4)}(\theta, c) = 1 - 4\phi(\theta)c + c^2 - \frac{2}{3}(\theta^2 + 1)c^3 + \mathcal{O}(c^5);$$

(b) For θ close to 0,

$$R^{(4)}(\theta, c) = h_0(c) + h_1(c)\theta^2 + \mathcal{O}(\theta^4),$$

where

$$h_0(c) = (c^2 + 1)(1 - \int_{-c}^{c} \phi(u)du) - 2c\phi(c), \quad h_1(c) = \int_{-c}^{c} \phi(u)du;$$

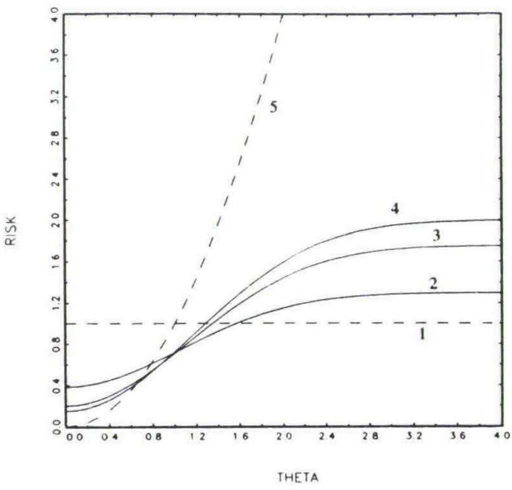

Figure 6.1. - Risk $R^{(4)}(\theta,c)$ of the optimal Burr estimator for five values of c.

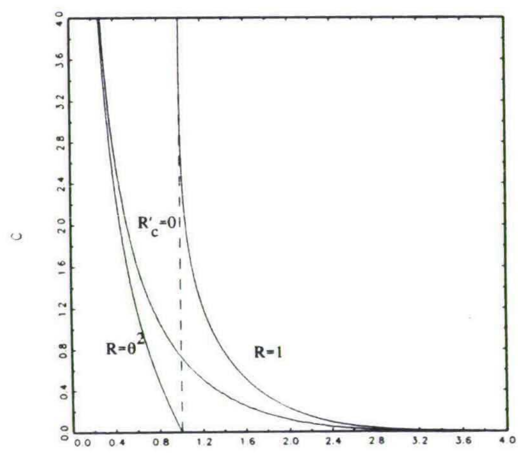

Figure 6.2. - Three relations between θ and c for the optimal Burr estimator $t^{(4)}$.

(c) $0 \leq h_0(c) \leq 1, \;\; h_0'(c) < 0, \;\; 0 \leq h_1(c) \leq 1, \;\; h_1'(c) > 0.$

Theorem 6.4 shows, inter alia, that the regret function can, for θ close to 0, be written as

$$r(\theta, c) = h_0(c) - (1 - h_1(c))\theta^2 + \mathcal{O}(\theta^4). \tag{6.9}$$

Since $0 \leq h_1(c) \leq 1$, the regret function attains a local maximum at $\theta = 0$ for every value of c.

In Figure 6.2 we graph the relations $R = 1, R = \theta^2$ and $R_c' = 0$. (R_θ' is always > 0 in this case.) The relationship $R_c' = 0$ defines points where $R^{(4)}(\theta, c)$ is minimized with respect to c. The area between $R = \theta^2$ and $R = 1$ is the area where the optimal Burr estimator has lower risk than both the 'usual' and the 'silly' estimator.

FIGURE 6.2

To summarize, we have identified, within a very large class of estimators (the Burr class), one estimator which minimizes the maximum regret (defined as $R^{(4)}(\theta, c) - \theta^2/(1 + \theta^2)$). This estimator, the optimal Burr estimator, can be written as

$$t^{(4)}(x; c) = \begin{cases} x + c & , \text{ if } \;\; x < -c, \\ 0 & , \text{ if } \;\; -c \leq x \leq c, \\ x - c & , \text{ if } \;\; x > c, \end{cases} \tag{6.10}$$

with $c = 0.545$. The estimator also has good average regret (with respect to a $N(0, 1)$ prior for θ) and would therefore appear to be a strong candidate for 'the best' estimator for θ.

7 Some generalizations and a proper Bayes solution

There are, however, several objections which can be raised against the optimal Burr estimator (6.10). The estimator is inadmissible, is not smooth at $x = \pm c$, does not depend on x when $|x| \leq c$, and it is not a Bayes estimator. In this section we shall consider some generalizations of (6.10) which deal with some of these objections. Eventually this leads to an estimator, which is very close to the optimal Burr estimator, but has none of the objections just raised. This will be our 'ideal' estimator.

To try and remedy these four objections, consider the following generalization of (6.10):

$$t^*(x;c) = \begin{cases} x + (1-a(c))c, & \text{if } \ x < -c, \\ a(x)x, & \text{if } \ -c \leq x \leq c, \\ x - (1-a(c))c, & \text{if } \ x > c, \end{cases} \tag{7.1}$$

where the function $a(x)$ satisfies: (a) $0 \leq a(x) \leq 1$, (b) $a(-x) = a(x)$, (c) $a(x)$ is nondecreasing and continuous for all $|x| \leq c$. Any estimator t^* satisfies conditions R1 and R2 and is continuous at $x = \pm c$. We consider four special cases:

(i) $a(x) = 0$. This is the optimal Burr estimator.

(ii) $a(x) = a$ (constant). This is the *limited translation estimator* proposed by Efron and Morris (1971). The estimator was developed in order to control the unbounded risk associated with normal priors. Efron and Morris were not aware of the fact that their estimator has near-optimal minimax regret properties. Searching over a and c we find the minimax regret solution at $a = 0.15$ and $c = 0.65$. The minimax regret at this point is 0.3843, slightly lower than the minimax regret in the optimal Burr class (0.3850). The limited translation estimator depends on x when $|x|$ is small, but it still suffers from three of the four objections raised above. In particular, it is not differentiable at $x = \pm c$. The next two estimators are continuously differentiable.

(iii) $a(x) = x^2/(c^2+x^2)$. This estimator is the HTF estimator for $|x| \leq c$ and approaches the optimal Burr estimator for $|x| > c$. The estimator is continuously differentiable and slightly curved around $x = 0$, but it is neither admissible nor Bayes. The minimax regret solution is found for $c = 1.102$ and $\theta = 2.861$ and takes the value 0.3850. This is exactly the same value as the minimax regret for the optimal Burr estimator.

(iv) $a(x) = |x|/(2c)$. This estimator is inspired by Huber's favourite choice of function for M-estimation; see Huber (1977, p. 13) in the context of robust procedures. The estimator, like the previous one, is continuously differentiable and suitably curved at $x = 0$, but neither admissible nor Bayes. The minimax regret is obtained for $c = 1.107$ and $\theta = 2.91$ and it takes the value 0.3857, slightly higher than for the optimal Burr estimator.

These estimators are slight generalizations of the optimal Burr estimator and they meet two of the four objections raised at the beginning of this section. We now develop an estimator which meets all four objections. To this end we rewrite (6.10) as

$$t^{(4)}(x; c) = w(x)(x - c) + (1 - w(x))(x + c), \tag{7.2}$$

where

$$w(x) = \begin{cases} 0, & \text{if } \; x < -c, \\ \frac{1}{2}(1 + x/c), & \text{if } \; -c \leq x \leq c, \\ 1, & \text{if } \; x > c. \end{cases} \tag{7.3}$$

Equation (7.2) shows the optimal Burr estimator as a data-based weighted average of $x - c$ and $x + c$. In particular, for all $c \geq 0$,

$$x - c \leq t^{(4)}(x; c) \leq x + c. \tag{7.4}$$

We now ask ourselves the following questions: Does there exist a Bayes estimator of the form (7.2)? (The answer is yes.) If so, does the prior for θ, which underlies this estimator have an appealing intuition? (Yes again.) The prior we are looking for turns out to be the *Laplace* (or *double exponential*) density given by

$$\pi(\theta; c) = \frac{c}{2} \exp\left(-c|\theta|\right), \quad -\infty < \theta < \infty, \; c > 0. \tag{7.5}$$

Laplace (1774), in his fundamental memoir on inverse probability, deduced this distribution from the principle of 'insufficient reason'.[12] The density (7.5) is unimodal and symmetric around 0. Hence the mean and median of θ are both 0. The mean and median of θ^2 are

$$\mathrm{E}(\theta^2) = \frac{2}{c^2}, \quad \mathrm{median}(\theta^2) = \frac{(\log 2)^2}{c^2}. \tag{7.6}$$

Since $x|\theta \sim N(\theta, 1)$ and assuming a Laplace prior density $\pi(\theta; c)$ for θ, the mean of the posterior distribution of $\theta|x$ can be expressed as

$$t^{(5)}(x;c) = \frac{1+h(x)}{2}(x-c) + \frac{1-h(x)}{2}(x+c) = x - h(x)c, \tag{7.7}$$

where

$$h(x) = \frac{1-e^{2cx}d(x)}{1+e^{2cx}d(x)}, \quad d(x) = \frac{\Phi(-x-c)}{\Phi(x-c)} \tag{7.8}$$

and Φ denotes the standard normal distribution function.[13] We notice that h is monotonically increasing on $(-\infty, \infty)$ with $h(-\infty) = -1$, $h(0) = 0$, $h(\infty) = 1$, and that $h(-x) = -h(x)$.

The estimator $t^{(5)}$ is of the form (7.2). It is a Bayes estimator and hence admissible. It is smooth at $x = \pm c$ and suitably curved when x is close to 0. Hence the estimator does not suffer from any of the objections against the optimal Burr estimator. We shall refer to the estimator (7.7) as the *Laplace estimator* and we shall denote the class of λ-functions induced by (7.7) as $\mathcal{L}^{(5)}$. The maximum and minimum $\mathcal{L}^0$-regret and $\mathcal{L}^0$-risk for three selected values of c are presented in Table 7.1.

	regret		risk	
c	maximum	minimum	maximum	minimum
0.6567	0.4645*	0.1051	1.4313	0.4645
0.6931	0.5127	0.1006	1.4805	0.4446
1.4142	2.0215	0.0345	3.0000	0.1858

[12] See Stigler (1986, p. 111) on the wonderful historical details. One of Laplace's problems was that he lacked a symbol for absolute value. The current symbol $|x|$ was introduced by Weierstrass in 1841, see Cajori (1993, vol. II, 123-124).

[13] See Pericchi and Smith (1992). Our expression (7.7) is easier for computational purposes than their formula (6), because h is monotonic.

Table 7.1. - Maximum and minimum $\mathcal{L}^0$-regret and risk for three Laplace estimators.

The minimax regret solution for the Laplace estimator is obtained for $c = 0.6567$ with minimax regret 0.4645, somewhat higher than the optimal Burr estimator (0.3850). The other two selected values for c are $c = \log 2 (= 0.6931)$ and $c = \sqrt{2} (= 1.4142)$. To understand the rationale behind these values, we recall from section 3 that a neutral prior is one where the distribution of θ is located at 0 and the distribution of θ^2 is located at 1. For the Laplace prior this implies $\mathrm{E}\theta = 0$ and either $\mathrm{E}(\theta^2) = 1$ or, more appropriately, median $(\theta^2) = 1$. From (3.6) we see that $\mathrm{E}(\theta^2) = 1$ when $c = \sqrt{2}$ and that median $(\theta^2) = 1$ when $c = \log 2$. The risk functions of the three estimators, labeled 1-3, are graphed in Figure 7.1.

FIGURE 7.1

It is evident from the graph that the risks for estimators 1 and 2 are very close for all θ. We note that the risk $R^{(5)}(\theta, c)$ increases monotonically with $|\theta|$ and $R^{(5)}(\theta, c) \to 1 + c^2$ as $|\theta| \to \infty$. Since $c = \log 2$ is close to the minimax solution $c = 0.6567$ and has the neutrality properties

$$\Pr(\theta < 0) = \Pr(\theta > 0) = \frac{1}{2}, \quad \Pr(|\theta| < 1) = \Pr(|\theta| > 1) = \frac{1}{2}, \tag{7.9}$$

we choose the estimator $t^{(5)}$ given by (7.7) with $c = \log 2$ as the 'ideal' Laplace estimator. For its risk we have

$$0.4446 \leq R^{(5)}(\theta, c) \leq 1.4805, \tag{7.10}$$

and for its regret

$$0.1006 \leq R^{(5)}(\theta, c) - \frac{\theta^2}{1 + \theta^2} \leq 0.5127. \tag{7.11}$$

The minimum regret is obtained for $\theta = 1.27$ and the maximum regret for $\theta = 4.93$. We see from Figure 7.2 that, for θ close to 0, the 'ideal' Laplace estimator is better (has lower risk) than the 'usual' estimator $t(x) = x$, but worse than the 'silly' estimator $t(x) = 0$.

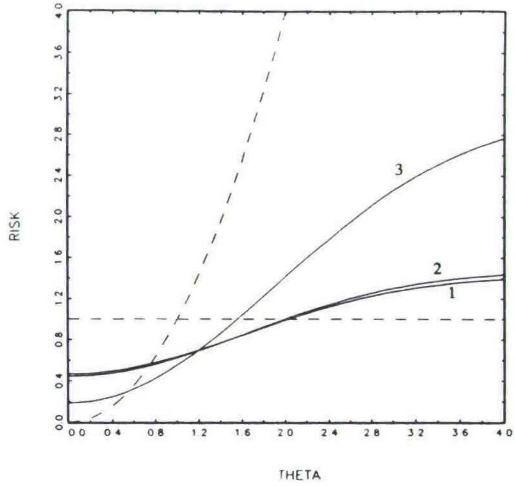

Figure 7.1. - Risk $R^{(5)}(\theta,c)$ of the Laplace estimator for three values of c.

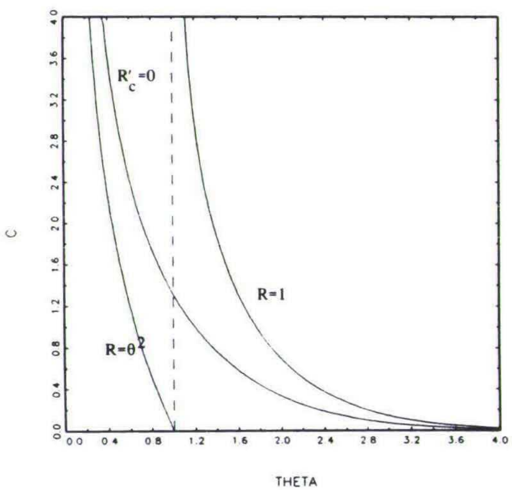

Figure 7.2. - Three relations between θ and c for the Laplace estimator $t^{(5)}$.

This is what we would expect. For $|\theta|$ large, the situation is reversed. Again, this is what we would expect. But for quite a large and important interval, $0.739 < |\theta| < 2.001$, the Laplace estimator is better than both the 'usual' and the 'silly' estimator. The maximum regret is obviously small compared to other estimators, since the Laplace estimator is close to the minimax regret estimator. In fact, both risk and regret compare very favourably with the other estimators.

In Figure 7.2 we present again the relations $R = 1$, $R = \theta^2$ and $R'_c = 0$. (R'_θ is always > 0 as in the optimal Burr class.) The relationship $R'_c = 0$ defines points where the risk is minimized with respect to c.

FIGURE 7.2

The 'ideal' estimator in the Laplace class should appeal to both Bayesians and non-Bayesians. To non-Bayesians because it is near minimax regret. To Bayesians (a) because it is a proper Bayes estimator, (b) because the prior is neutral (in the sense of section 2 and (7.9)) with respect to the 'usual' and the 'silly' estimators, and (c) because the Laplace prior is a particularly suitable one from the viewpoint of maximum entropy.[14] This aspect of the Laplace distribution was recently emphasized by Zellner (1994). The argument is based on two well-known facts:

(i) Given two random variables θ and σ^2, the conditional density for $\theta|\sigma^2$ which maximizes the entropy subject to the conditions $\mathrm{E}(\theta|\sigma^2) = 0$ and $\mathrm{var}(\theta|\sigma^2) = \sigma^2$ is the normal distribution $N(0, \sigma^2)$; and

(ii) Given one positive random variable σ^2, the density for σ^2 which maximizes the entropy subject to $\mathrm{E}\sigma^2 = s^2$ is the exponential density $(1/s^2)\exp[-\sigma^2/s^2]$.

From (i) and (ii) we can obtain the joint density for θ and σ^2 and, integrating over σ^2, the marginal density for θ. Zellner (1994) showed that this density for θ is Laplace.

But there is more to be said in favour of Laplace. Suppose we wish to determine a neutral prior. There are six continuous densities $\pi(\theta)$ in common use which are symmetric around zero: uniform on $[-a, a]$, triangular on $[-b, b]$, normal $N(0, \tau^2)$, logistic (λ), student $t(\nu)$ and Laplace (c).[15] Imposing the second condition for neutrality (median

[14] See Rao (1973, p. 162 and 172-173) for a definition and brief discussion.

[15] The Laplace density is given in (7.5). The logistic density is defined as $\pi(\theta; \lambda) = \lambda e^{-\lambda\theta}/(1 + e^{-\lambda\theta})^2$.

$(\theta^2) = 1)$ determines for each of these one-parameter densities the value of the parameter: $a = 2$, $b = 2 + \sqrt{2}$, $\tau^2 = 2.1981$, $\lambda = \log 3$, $\nu = 1$, $c = \log 2$. The uniform and triangular densities are unsuitable because their support is not $(-\infty, \infty)$. The tails of the normal distribution are too thin (leading to unbounded risk). This leaves the logistic, Cauchy and Laplace priors. With a prior density $\pi(\theta)$, the mean of the posterior distribution of $\theta|x$ can be written as

$$t_\pi(x) = x - \frac{\int u\pi(x-u)\phi(u)du}{\int \pi(x-u)\phi(u)du}. \tag{7.12}$$

Hence (7.12) gives the Bayes estimator for θ induced by the prior $\pi(\theta)$. It enables us to calculate the risk and regret functions associated with the logistic and Cauchy estimators. The results are summarized in Table 7.2.

	regret				risk			
prior	maximum		minimum		maximum		minimum	
normal	∞	$(\theta = \infty)$	0.0	$(\theta = 1.48)$	∞	$(\theta = \infty)$	0.4724	$(\theta = 0)$
logistic	1.2150	$(\theta = 10.01)$	0.0377	$(\theta = 1.41)$	2.2069	$(\theta = \infty)$	0.4643	$(\theta = 0)$
Cauchy	0.6332	$(\theta = 3.54)$	0.0	$(\theta = \infty)$	1.5618	$(\theta = 3.68)$	0.3757	$(\theta = 0)$
Laplace	0.5127	$(\theta = 4.93)$	0.1006	$(\theta = 1.27)$	1.4805	$(\theta = \infty)$	0.4446	$(\theta = 0)$

Table 7.2. - Maximum and minimum $\mathcal{L}^0$-regret and risk for four Bayes estimators with neutral priors.

The table compares four Bayes estimators induced by neutral priors: $N(0, 2.1981)$, logistic (with $\lambda = \log 3$), Cauchy and Laplace (with $c = \log 2$). The values for θ at which the maxima and minima are attained are presented as well. The normal prior with variance $\tau^2 = 2.1981$ leads to unbounded risk and regret. The minimum regret is attained at $\theta = \tau = 1.48$. The minimum risk $(\tau^2/(1+\tau^2))^2 = 0.4724$ is attained at $\theta = 0$. For the logistic estimator, both the maximum regret (1.2150) and the maximum risk (2.2069) are higher than for the Laplace estimator. For the Cauchy estimator the maximum regret is 0.6332 (higher than Laplace) and this maximum is attained at $\theta = 3.54$ (lower than Laplace). Also the maximum risk is higher than Laplace and is attained for a lower θ. All these findings favour Laplace over logistic and Cauchy. We may therefore conclude that the Laplace estimator $t^{(5)}$ wins over the logistic and Cauchy estimators on three counts: its maximum regret is lower (which should please the frequentists), its justification as a prior is stronger (which should please the Bayesians), and it is computationally very much easier to work with (which should please both).

This brings us the end of our discussion of the $N(\theta, 1)$ problem. Implications for the regression problem will be discussed in the concluding section.

8 Conclusions

In this paper we have attempted to solve an old and classical problem in applied statistics. The problem is how best to estimate the parameters of interest β in a linear regression model

$$y = X\beta + \gamma z + u. \tag{8.1}$$

The explanatory variables in X are regarded as belonging in the equation according to some theory, and can be thought of as the minimum set of variables required to explain y. The explanatory variable z, however, is only included because the researcher believes it might lead to 'better' estimates of β. The focus of our analysis is the estimation of one or several linear combinations of the parameters β.

We propose to estimate β as a weighted average of the unrestricted estimator b_u and the restricted estimator b_r (with $\gamma = 0$), that is, $b = \lambda b_u + (1 - \lambda)b_r$, where λ is a function of the t-ratio of γ. We call this estimator a WALS (weighted-average least squares) estimator. Judging the estimator's performance by its mean squared error, we see from Theorem 2.2 that if any λ-function is optimal for b as an estimator for β, then it is also optimal for a linear function $\psi' b$ as an estimator for $\psi'\beta$. Thus, β contains the 'parameters of interest' (even when the focus of our analysis is one particular linear combination of the β's) and γ is a 'nuisance parameter'.

This classical problem of estimating β in the presence of a nuisance parameter γ we have called *the regression problem.* In Theorem 2.2 we show that the regression problem is equivalent to a fundamental statistical problem, which we have called *the* $N(\theta, 1)$ *problem*: Given one observation x from a $N(\theta, 1)$ distribution, what is the 'best' estimator for θ. This seemingly trivial problem turns out to be far from trivial. Sections 3-7 of the paper are devoted to it. After a long journey through normal Bayes, pretest, HTF, Burr, Laplace and Cauchy estimators, we finally arrive at an estimator (the 'ideal' Laplace estimator) which is near-optimal from the minimax regret point of view and also has an attractive Bayesian interpretation with the prior median of θ^2 equal to one. (The importance of $\theta^2 = 1$ as a natural pivot is discussed in Theorem 2.1, (3.10) and the discussion following Theorems 4.5 and 4.7.) The Laplace estimator is defined in (7.7) and is 'ideal' for $c = \log 2$.

The six main estimators discussed in sections 3-7 are graphed in Figure 8.1. Each

FIGURE 8.1

graph represents an estimator $t(x)$ as a function of x when $x \sim N(\theta, 1)$. The dotted line gives $t(x) = x$, the 'usual' estimator. It is clear that the normal Bayes estimator $t^{(1)}$ and the pretest estimator $t^{(2)}$ are far from satisfactory. The normal Bayes estimator diverges from x and has therefore unbounded risk (Theorem 3.5). The pretest estimator is inadmissible and discontinuous. Its pathological behaviour is discussed in detail in section 4. The estimators $t^{(3)}, \ldots, t^{(6)}$ are better. However, $t^{(3)}$ and $t^{(4)}$ are inadmissible. The Laplace estimator $t^{(5)}$ and the Cauchy estimator $t^{(6)}$ are both admissible, but $t^{(5)}$ has better minimax regret properties than $t^{(6)}$.

Among these estimators our preferred one is the 'ideal' Laplace estimator, because it has attractive smoothness properties and near-optimal risk performance. Considered as a Bayes estimator it is based on a prior with strong intuitive appeal. Its main competitor is the optimal Burr estimator with $c = 0.545$ (see (6.7)) whose maximum regret is smaller than that of the 'ideal' Laplace estimator and is easier to use in practice. The two estimators are quite different when x is small, but their risk functions are similar as can be seen from Figure 8.2. For $0 \leq |\hat{\theta}| \leq 0.55$ and $|\hat{\theta}| \geq 2.89$ the optimal Burr estimator has slightly smaller risk, while for $0.55 < |\hat{\theta}| < 2.89$ the 'ideal' Laplace estimator has slightly smaller risk.

FIGURE 8.2

We now consider the application of these results to the regression problem. For every estimator $t(x) = \lambda(x)x$ for the $N(\theta, 1)$ problem we have a corresponding WALS estimator $b = \lambda(\hat{\theta})b_u + (1 - \lambda(\hat{\theta}))b_r$. Every WALS estimator has the advantage over a traditional pretest estimator that the completely arbitrary choice of significance level (0.01, 0.05 or something else) is avoided. A second advantage is that the problem that in a large enough sample the classical test will be virtually certain to reject (Berger (1985), p. 20) does not occur here. The results for the $N(\theta, 1)$ problem imply that a fixed λ (for example $\lambda = 1/2$) is unacceptable, because it corresponds to the normal Bayes estimator $t^{(1)}$. Also, the traditional pretest estimator (choose the restricted estimator b_r when the

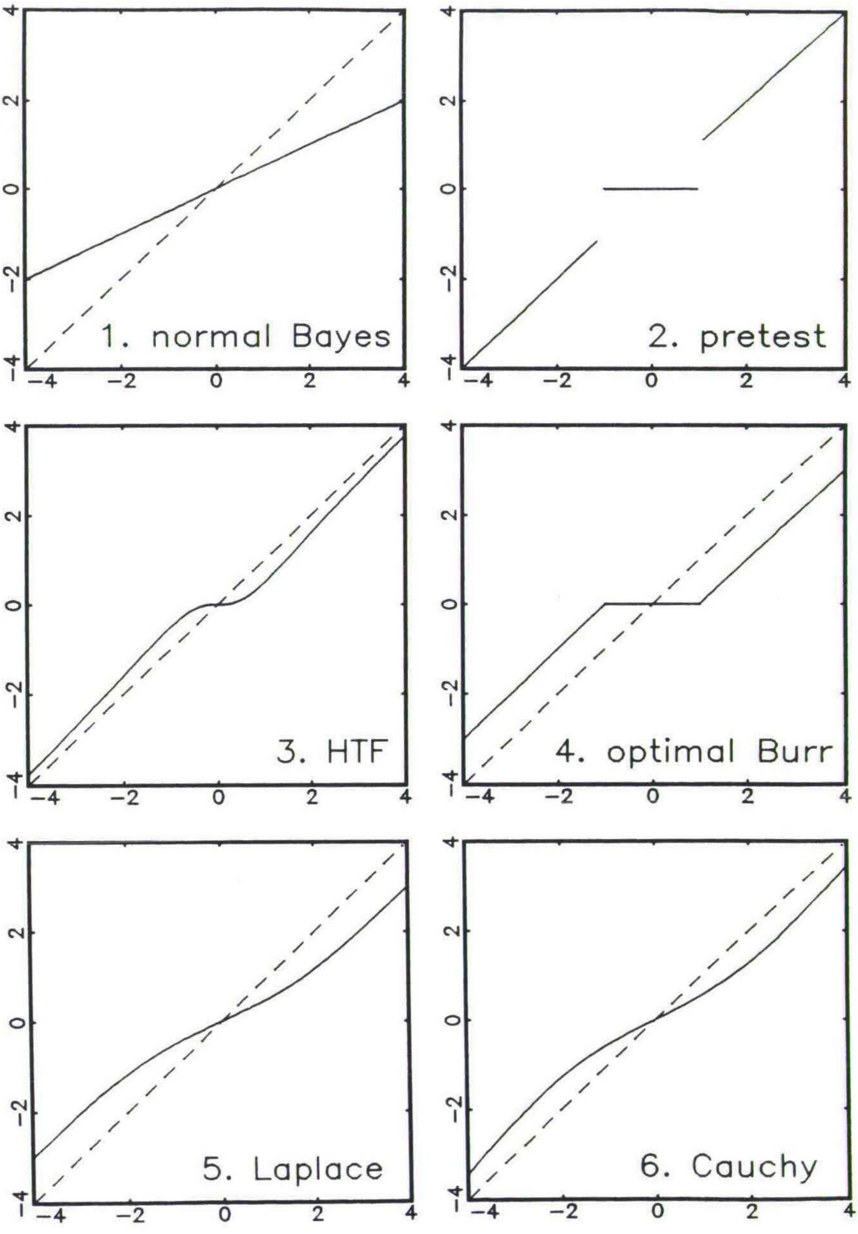

Figure 8.1. - Six estimators t(x) of θ when x ~ N(0,1)

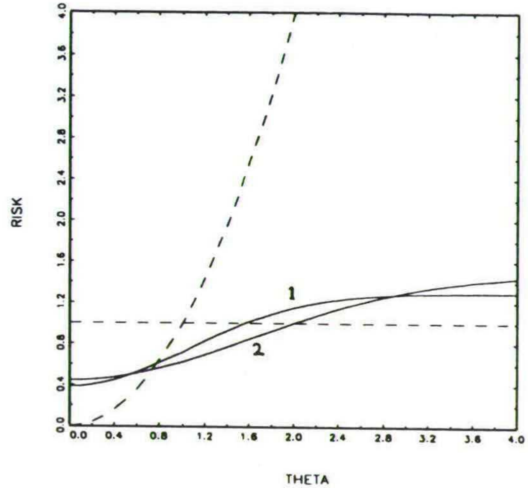

Figure 8.2. - Risk of the optimal Burr estimator (1) and the 'ideal' Laplace estimator (2).

t-ratio for γ is small, choose the unrestricted estimator b_u otherwise) is unacceptable (see section 4), even though it is the estimator routinely used. In fact, the famous 5% pretest estimator is very close to being the worst possible pretest estimator where the whole class of pretest estimators is poor to begin with. (See the discussion between Theorems 4.5 and 4.6.) It is not surprising then that we do not recommend the traditional pretest estimator. The main competitors are the optimal Burr estimator based on $\lambda^{(4)}$ and the 'ideal' Laplace estimator based on $\lambda^{(5)}$. The optimal Burr estimator has lower maximum regret and is easier to compute, while the 'ideal' Laplace estimator is admissible and has a strong and plausible Bayesian interpretation. We have a slight preference for the 'ideal' Laplace estimator.

A few words about the weighting function λ. Condition R1 requires that $0 \leq \lambda(\hat{\theta}) \leq 1$. Bounded risk implies that $\lambda(\hat{\theta}) \to 1$ as $\hat{\theta} \to \infty$ (Theorem 4.5), which shows that the unrestricted estimator b_u is optimal when the t-ratio is very large. This, of course, is plausible. But what would we expect $\lambda(\hat{\theta})$ to be when $\hat{\theta}$ is small, say $\hat{\theta} = 0$? Refering to Figure 8.1, $\lambda(0)$ is given by the slope of $t(x)$ at $x = 0$. For the estimators $t^{(2)}, t^{(3)}$ and $t^{(4)}$ we see that $\lambda(0) = 0$. But for the three Bayes estimators $t^{(1)}, t^{(5)}$ and $t^{(6)}$ we have $\lambda(0) > 0$. In fact, assuming a neutral prior, we find that $\lambda(0) = 0.6873$ in the case of the normal $N(0, 2.1981)$ Bayes estimator, $\lambda(0) = 0.5896$ for the 'ideal' Laplace estimator, and $\lambda(0) = 0.5251$ for the Cauchy estimator. In each case $\lambda(0) > 1/2$. Hence, even when $\hat{\theta} = 0$, a neutral prior will lead to a WALS estimator for β where more than half the weight is put on the unrestricted estimator b_u. For Bayesians this will be perfectly plausible and acceptable and they should be happy with the 'ideal' Laplace estimator. For many non-Bayesians this may also be plausible and they too will be happy with the 'ideal' Laplace estimator. But some non-Bayesians may argue that if $|\hat{\theta}|$ is small, say less than about one half, we should choose the restricted estimator b_r. This view implies that $\lambda(0) = 0$ and leads to the optimal Burr estimator. As we have seen, the risk functions of the two estimators are not very different.

For the application of WALS estimation based on the 'ideal' Laplace estimator we need to compute the relevant $\lambda(\hat{\theta})$. In Table 8.1 we present $\lambda(\hat{\theta})$ for selected values of $\hat{\theta}$. If the table is not sufficiently precise, then $\lambda(\hat{\theta})$ can be computed from

$$\lambda(\hat{\theta}) = 1 - \frac{h(\hat{\theta})\log 2}{\hat{\theta}} \quad (\hat{\theta} \neq 0), \tag{8.2}$$

where $h(\hat{\theta})$ is defined in (7.8) with $c = \log 2$. For $\hat{\theta} \geq 4.5$ the approximation $\lambda(\hat{\theta}) = 1 - (\log 2)/\hat{\theta}$ is accurate up to four decimal places. For $|\hat{\theta}| < 4.5$ we can approximate $h(\hat{\theta})$ from the formulae in Abadir (1995), in particular (26) and (50), or calculate exactly using standard computer software.

$\hat{\theta}$	$\lambda(\hat{\theta})$	$\hat{\theta}$	$\lambda(\hat{\theta})$	$\hat{\theta}$	$\lambda(\hat{\theta})$	$\hat{\theta}$	$\lambda(\hat{\theta})$
0.0	0.5896	2.0	0.6943	4.0	0.8268	6.0	0.8845
0.5	0.5973	2.5	0.7354	4.5	0.8460	6.5	0.8934
1.0	0.6197	3.0	0.7722	5.0	0.8614	7.0	0.9010
1.5	0.6537	3.5	0.8027	5.5	0.8740	∞	1.0000

Table 8.1. Optimal weights for the WALS estimator based on the 'ideal' Laplace prior.

The current paper has tried to concentrate on the main issues by making several simplifying assumptions. Two of these can and should be removed in future work. First, the assumption that σ^2 is known is clearly unrealistic. Preliminary investigations show that the essence of our analysis, for example the analog of Theorem 2.2, still goes through and that the difference for the WALS estimator between the case σ^2 known and σ^2 not known is similar to the difference between a $N(0, 1)$ distribution and a Student distribution.[16] Secondly, we have assumed that there is only one nuisance parameter γ. The basic set-up is still valid when there are more than one nuisance parameters, but the details are more complicated and further work is required.

In addition, we can apply the general idea of the paper to areas other than estimation. For example, instead of asking how to estimate β in the presence of nuisance parameters, we can ask how to predict y. In addition, we need to know more about the distribution of the WALS estimator since this is relevant for inference.[17] Theorem 2.2 is only a beginning in this direction.

[16] See also Droge and Georg (1995).

[17] See Adams (1991) for a comprehensive investigation of the effects of model search on inference in regression.

Appendix: Proofs of Theorems

Proof of Theorem 2.1: It is well-known that

$$b_r \sim N[\beta + \theta q, \sigma^2(X'X)^{-1}]$$

and

$$b_u \sim N[\beta, \sigma^2(X'X)^{-1} + qq'].$$

Hence,

$$\mathrm{MSE}(b_r) - \mathrm{MSE}(b_u) = (\theta^2 - 1)qq'$$

and the results follow. (The theorem can also be proved as a special case of Theorem 2.2.)

Proof of Theorem 2.2: A straightforward but somewhat tedious application of standard results on the multivariate normal distribution (Rao (1973, p. 522)) shows that the conditional distribution of (b_r, b_u) given $\hat{\theta}$ is

$$\begin{pmatrix} b_r \\ b_u \end{pmatrix} \Big| \hat{\theta} \sim N\left[\begin{pmatrix} \beta + \theta q \\ \beta - (\hat{\theta} - \theta)q \end{pmatrix}, \sigma^2 \begin{pmatrix} (X'X)^{-1} & (X'X)^{-1} \\ (X'X)^{-1} & (X'X)^{-1} \end{pmatrix}\right]$$

and the results follow.

Proof of Theorem 3.1: Clearly, $t(x,\lambda) = x$ is unbiased and has risk (variance) equal to 1. Blyth (1951) showed that x is admissible (see also Berger (1985, p. 548)). Since x is admissible and has constant risk it must be unique minimax (Berger (1985, p. 382, exercise 32)).

Proof of Theorem 3.2: For any estimator $t(x,\lambda)$ we have, at $\theta = 0$, $R(0,\lambda) = \mathrm{E}(\lambda^2(u)u^2)$, where $u \sim N(0,1)$. Thus, $R(0,\lambda) \geq 0$ with equality if and only if $\lambda \equiv 0$. Since $t^{(1)}(x;\infty) = \lambda(x)x$ with $\lambda(x) \equiv 0$, we see that $t^{(1)}(x;\infty)$ dominates every other estimator at $\theta = 0$ and hence is admissible. Also, $t^{(1)}(x;0)$ is admissible by Theorem 3.1. Assume next that $0 < c < \infty$ and let $\pi(\theta)$ be a prior density of θ. In particular, let $\pi(\theta)$

be $N(0, 1/c)$. It is well-known (Berger (1985, p 127-128 and p. 161)) that the mean of the posterior distribution of θ given x is given by $E(\theta|x) = x/(1+c) = t^{(1)}(x; c)$ and that this is the Bayes estimator of θ. Since the risk function $R^{(1)}(\theta, c)$ is continuous in θ for every c and the prior π gives positive probability to any interval in $\mathbf{R}$, it follows from Berger (1985, p. 254) that $t^{(1)}(x, c)$ is admissible for $0 < c < \infty$ and hence for $0 \leq c \leq \infty$.

<u>Proof of Theorem 3.3</u>: Let $x = u + \theta$, where $u \sim N(0, 1)$. Define $h(u, \theta) = \lambda(u + \theta)(u + \theta) - \theta$. Then $h(-u, \theta) = -h(u, -\theta)$ and

$$\text{BIAS}(\theta, \lambda) = E_\theta h(u, \theta) = E_\theta h(-u, \theta) = -E_\theta h(u, -\theta) = -\text{BIAS}(-\theta, \lambda).$$

This proves (a). To prove (b) and (c), let $\varepsilon(x) = (1 - \lambda(x))x$. Since $\lambda(-x) = \lambda(x)$, we have $\varepsilon(-x) = -\varepsilon(x)$. Hence, following Huntsberger (1955),

$$\begin{aligned}
\text{BIAS}(\theta, \lambda) &= -E\varepsilon(x) = -\textstyle\int_{-\infty}^{\infty} \varepsilon(x)\phi(x - \theta)dx \\
&= \textstyle\int_0^{\infty} \varepsilon(x)\phi(x + \theta)dx - \int_0^{\infty} \varepsilon(x)\phi(x - \theta)dx \\
&= \textstyle\int_0^{\infty} \varepsilon(x)\phi(x + \theta)(1 - e^{2\theta x})dx.
\end{aligned}$$

For $\theta \neq 0$, the integral is zero if and only if $\varepsilon(x) = 0$ for all x, that is, $\lambda = \lambda_0^{(1)}$. If $\lambda \neq \lambda_0^{(1)}$, then the sign of the bias depends on the sign of $1 - e^{2\theta x}$, which completes the proof.

<u>Proof of Theorem 3.4</u>: Let $x = u + \theta$, where $u \sim N(0, 1)$ and define h as in the proof of Theorem 3.3. Then,

$$R(\theta, \lambda) = E_\theta h^2(u, \theta) = E_\theta h^2(-u, \theta) = E_\theta h^2(u, -\theta) = R(-\theta, \lambda).$$

<u>Proof of Theorem 3.5</u>: To prove that R2 is sufficient, let $x = u + \theta, u \sim N(0, 1)$. Then,

$$R(\theta, \lambda) = E(u - \varepsilon(u + \theta))^2 \leq 2E(u^2 + \varepsilon^2(u + \theta)) \leq 2(1 + K^2).$$

To prove necessity we write

$$\begin{aligned}
R(\theta, \lambda) &= \textstyle\int_{-\infty}^{\infty} [\lambda(x)x - \theta]^2 \phi(x - \theta)dx \\
&\geq \textstyle\int_{|x-\theta|\leq 1} [\lambda(x)x - \theta]^2 \phi(x - \theta)dx \\
&\geq \phi(1) \textstyle\int_{|x-\theta|\leq 1} [\lambda(x)x - \theta]^2 dx \\
&= \phi(1) \textstyle\int_{-1}^{1} [\varepsilon(\theta + u) - u]^2 du.
\end{aligned}$$

Since $t(x,\lambda) \equiv \lambda(x)x$ is nondecreasing for $x \geq 0$, we have, for $|u| \leq 1$ and $\theta \geq 1$, $t(\theta+u,\lambda) \leq t(\theta+1,\lambda)$ and hence $\varepsilon(\theta+u) - u \geq \varepsilon(\theta+1) - 1$. For every $\theta \geq 1$ satisfying $\varepsilon(\theta+1) \geq 1$ we then find

$$R(\theta,\lambda) \geq 2\phi(1)[\varepsilon(\theta+1)-1]^2.$$

Clearly, if ε is unbounded then so is R. (A more general result requiring a more difficult proof is given by Brown (1971, Theorem 3.3.1).)

Proof of Theorem 3.6: Assume that R3 does not hold. Then $t(x,\lambda)$ is not Generalized Bayes (Strawderman and Cohen (1971), Berger and Srinivasan (1978)) and therefore not admissible (Berger (1985, p. 542-544)). This proves (a). To prove (b) assume that R2 and R3 hold. Then Brown (1971) showed that $t(x,\lambda)$ is admissible. (See also Berger (1985, p. 552-553) for further references.)

Proof of Theorem 3.7: Using the symmetry condition $\lambda(-x) = \lambda(x)$, we have

$$\int_{-\infty}^{0} (x\lambda(x)-\theta)^2\phi(x-\theta)dx = \int_{0}^{\infty} (x\lambda(x)+\theta)^2\phi(x+\theta)dx$$

and hence

$$\begin{aligned} R(\theta,\lambda) &= \textstyle\int_{-\infty}^{\infty}(x\lambda(x)-\theta)^2\phi(x-\theta)dx \\ &= \textstyle\int_0^{\infty}((x\lambda(x)-\theta)^2\phi(x-\theta) + (x\lambda(x)+\theta)^2\phi(x+\theta))dx. \end{aligned}$$

Now, let

$$\eta(x) = e^{-2x} \quad \text{and} \quad h(x) = \frac{1-\eta(x)}{x(1+\eta(x))}.$$

Then we obtain, after some algebra and completing the square,

$$\begin{aligned} R(\theta,\lambda) &= \textstyle\int_0^{\infty} x^2(1+\eta(\theta x))(\lambda(x)-\theta^2 h(\theta x))^2\phi(x-\theta)dx \\ &+ \ 4\theta^2 \textstyle\int_0^{\infty} \dfrac{\eta(\theta x)}{1+\eta(\theta x)}\phi(x-\theta)dx. \end{aligned}$$

The following properties of $h(x)$ should be noted: $h(-x) = h(x), h(x) \to 1$ as $x \to 0$, $h(x) \to 0$ as $x \to \infty, h'(x) < 0$ for all $x > 0$. Since $h(x)$ is strictly decreasing on $(0, \infty)$ and $\lambda(x)$ is nondecreasing on $(0, \infty)$, there exists a unique x_0 such that $x_0 = 0$ if $\lambda(0) \geq \theta^2$ and $\lambda(x_0) = \theta^2 h(\theta x_0)$ if $\lambda(0) < \theta^2$. This can easily be seen by considering graphs of the functions $\lambda(x)$ and $\theta^2 h(\theta x)$ for both cases. (If $\lambda(x)$ is not continuous, the second condition is replaced by

$$\lambda(x_0 - \varepsilon) \leq \theta^2 h(\theta x_0) \leq \lambda(x_0 + \varepsilon)$$

for all $\varepsilon > 0$ sufficiently small.) With x_0 so defined we have

$$|\lambda(x) - \theta^2 h(\theta x)| \geq |\lambda(x_0) - \theta^2 h(\theta x)| \quad \text{for all } x \geq 0.$$

Let λ_0 denote the constant function such that $\lambda_0(x) = \lambda(x_0)$ for all $x \geq 0$. It is then clear that $R(\theta, \lambda) \geq R(\theta, \lambda_0)$. We also know from (3.8) that $R(\theta, \lambda_0) \geq \theta^2/(1 + \theta^2)$. This completes the proof.

<u>Proof of Theorem 3.8</u>: With $\theta \sim N(0, 1)$ and $x|\theta \sim N(\theta, 1)$, we obtain $x \sim N(0, 2)$ and $\theta|x \sim N(x/2, 1/2)$. Then, writing

$$\mathrm{E}_\pi R(\theta, \lambda) = \mathrm{E}_x \mathrm{E}_{\pi|x}(\lambda(x)x - \theta)^2,$$

where $\mathrm{E}_{\pi|x}$ denotes the expectation with respect to the distribution of $\theta|x$ and E_x denotes the expectation with respect to the distribution of x, the result follows.

<u>Proof of Theorem 4.1</u>: With $\lambda_c^{(2)}(x)$ defined in (4.2), we have

$$\begin{aligned}
R^{(2)}(\theta, c) &= \textstyle\int_{-\infty}^{\infty} (\lambda_c^{(2)}(x)x - \theta)^2 \phi(x - \theta) dx \\
&= \theta^2 \textstyle\int_{|x| \leq c} \phi(x - \theta) dx + \int_{|x| > c} (x - \theta)^2 \phi(x - \theta) dx \\
&= 1 + \theta^2 \textstyle\int_S \phi(u) du - \int_S u^2 \phi(u) du \\
&= 1 + (\theta^2 - 1) P(\theta, c) - \textstyle\int_S (u^2 - 1) \phi(u) du,
\end{aligned}$$

where $S = \{u : -\theta - c < u < -\theta + c\}$. Noting that $\phi'(u) = -u\phi(u), \phi''(u) = (u^2 - 1)\phi(u)$, we find

$$\int_S (u^2 - 1)\phi(u)du = \int_S \phi''(u)du = [\phi'(u)]_S = -(c+\theta)\phi(c+\theta) - (c-\theta)\phi(c-\theta),$$

which concludes the proof.

Proof of Theorem 4.2: For $c = 0$ or $c = \infty$, the pretest estimator is admissible by Theorem 3.2. For $0 < c < \infty$, $t^{(2)}(x; c)$ is discontinuous at $x = \pm c$, thus violating condition R3(a). Hence the estimator is inadmissible by Theorem 3.6. This proves (a). To prove (b) we follow Kempthorne (1984); see also Droge (1993). Assume a prior distribution $\pi(\theta)$ for θ. In particular, let $\pi(\theta)$ be $N(0, 1)$. Then we can show either directly or using Theorem 4.7 that $\mathrm{E}_\pi R^{(2)}(\theta, c) = 1$ for every c. This implies that there cannot exist two values $c_1 \neq c_2$ such that $R^{(2)}(\theta, c_1) \leq R^{(2)}(\theta, c_2)$ for all θ with strict inequality for some θ. Hence every pretest estimator is $\mathcal{L}^{(2)}$-admissible. (See Berger (1985, p. 253-254).)

Proof of Theorem 4.3: (a) follows from Theorem 3.4; (b) is easy, see also (3.9); (c) and (d) are clearly *not* true for $c = \infty$. For $c < \infty$, (c) follows from Theorem 3.5 and (d) from Theorem 4.1.

Proof of Theorem 4.4: For small c we have

$$\phi(c+\theta) = \phi(\theta)\left[1 - \theta c + \frac{1}{2}(\theta^2 - 1)c^2 - \frac{1}{6}\theta(\theta^2 - 3)c^3\right] + \mathcal{O}(c^4),$$

from which we obtain

$$\begin{aligned}
&\phi(c+\theta) + \phi(c-\theta) = \phi(\theta)\left[2 + (\theta^2 - 1)c^2\right] + \mathcal{O}(c^4),\\
&\phi(c+\theta) - \phi(c-\theta) = -\phi(\theta)\left[2\theta c + \tfrac{1}{3}\theta(\theta^2 - 3)c^3\right] + \mathcal{O}(c^5),\\
&P(\theta, c) = \phi(\theta)\left[2c + \tfrac{1}{3}(\theta^2 - 1)c^3\right] + \mathcal{O}(c^5).
\end{aligned}$$

Inserting these expansions in $R^{(2)}(\theta, c)$, given in Theorem 4.1, proves (a). (b) is proved similarly. To prove (c) we notice that $2c\phi(c) \leq \int_{-c}^{c}\phi(u)du \leq 1$ and hence $0 \leq 2c\phi(c) \leq h_0(c) \leq 1$ and $h_1(c) \geq c^3\phi(c) \geq 0$. Also, $h_0'(c) = -2c\phi(c) \leq 0$ and $h_1'(c) = c^2(5 - c^2)\phi(c)$. Since $h_1(0) = 0$, $h_1(\infty) = 1$, it follows that there exists a unique $c, 0 < c < \sqrt{5}$, such that $h(c) = 1$. The value of this c can be approximated to any degree of accuracy.

Proof of Theorem 4.5: It follows from (4.7) that, for $c \in (0,\infty)$, $R^{(2)}(\theta,c) > \theta^2$ if $|\theta| \leq 1$ and $R^{(2)}(\theta,c) > 1$ if $|\theta| > 1$. The bounds are obtained for $c = \infty$ and $c = 0$, respectively.

Proof of Theorem 4.6: A formal proof of the uniqueness of the $\mathcal{L}^{(2)}$-minimax regret estimator can be found in Droge (1993). For the case of $\mathcal{L}^0$ the proof is similar, but tedious. Once we know that a unique solution exists, it can be found numerically to any desired degree of precision. (For readers who accept computer output as 'proof', the output shows unambiguously that there is a unique solution.)

Proof of Theorem 4.7: The proof of (a) is a straightforward exercise in integration. To prove (b) let us first consider the case $\mu = 0$. If $\mu = 0$, then $\mathrm{E}_\pi R^{(2)}(\theta,c)$ reduces to

$$\mathrm{E}_\pi R^{(2)}(\theta,c) = 1 + (\tau^2 - 1)\kappa(c),$$

where

$$\kappa(c) = \int_{-\delta}^{\delta} \phi(u)du - 2\delta\phi(\delta), \;\; \delta = \frac{c}{\sqrt{1+\tau^2}}.$$

Now, $\kappa(c)$ is monotonically increasing on $(0,\infty)$ with $\kappa(0) = 0, \kappa(\infty) = 1$. Hence, if $\tau^2 - 1 > 0$ the minimum is obtained by choosing $\kappa(c) = 0$ (as small as possible), that is, $c = 0$. If $\tau^2 - 1 < 0$ we must choose $\kappa(c) = 1$ (as large as possible), that is, $c = \infty$. This proves (b) for the special case $\mu = 0$. The general case is rather tedious and is left to the reader.

Proof of Theorem 5.1: For $c = 0$ or $c = \infty$ we have the 'usual' and the 'silly' estimator respectively and we know that these are admissible (Theorem 3.2). For $0 < c < \infty$ we show that condition R3(b) is not satisfied. Then, by Theorem 3.6(a), the estimator can not be admissible. Now, $\varepsilon(x) = (1 - \lambda(x))x = c^2x/(c^2 + x^2)$. Hence,

$$A(x) \equiv \int_0^x \varepsilon(y)dy = \frac{c^2}{2}\int_0^{x^2} \frac{dt}{c^2+t} = \frac{c^2}{2}\log\frac{c^2+x^2}{c^2}.$$

It follows that

$$\exp[-A(x)] = \exp\left[-\int_0^x \varepsilon(y)dy\right] = \left(\frac{c^2}{c^2+x^2}\right)^{c^2/2}.$$

As noted by Strawderman and Cohen (1971, p. 278) this has a non-removable singularity at $x = \pm ci$. Hence $\exp[-A(x)]$ cannot be extended analytically into the whole complex plane and condition R3(b) cannot hold.

To prove (b) we note that, at $\theta = 0$,

$$R^{(3)}(0,c) = \mathrm{E}\left(\frac{u^3}{c^2+u^2}\right)^2,$$

where $u \sim N(0,1)$. Hence $R^{(3)}(0,c)$ is a *decreasing* function of c. On the other hand, when θ is large, we have

$$R^{(3)}(\theta,c) = 1 + \frac{c^2(c^2+2)}{\theta^2} + \mathcal{O}\left(\frac{1}{\theta^4}\right),$$

which is an *increasing* function of c. Hence no estimator in $\mathcal{L}^{(3)}$ can dominate another.

<u>Proof of Theorem 6.1</u>: This follows from (6.3).

<u>Proof of Theorem 6.2</u>: Similar to the proof of Theorem 4.1.

<u>Proof of Theorem 6.3</u>: Since $t^{(4)}(x;c)$ is not differentiable at $x = \pm c$, it is not admissible (Theorem 3.6(a)). The proof of (b) is similar to the proof of Theorem 5.1. For $\theta = 0, h_0(c) \equiv R^{(4)}(0,c)$ is decreasing in c (see also Theorem 6.4). For θ is large, (6.8) shows that $R^{(4)}(\theta,c) \approx 1 + c^2$, which is increasing in c. Hence no estimator dominates any other in $\mathcal{L}^{(4)}$.

<u>Proof of Theorem 6.4</u>: Similar to the proof of Theorem 4.4.

References

Abadir, K.M. (1995). 'An introduction to hypergeometric functions for economists', Discussion Paper in Economics #95/10, University of Exeter.

Adams, J.L. (1991). 'A computer experiment to evaluate regression strategies', *Proceedings of the Computational Statistics Section*, American Statistical Association, 55-62.

Akaike, H. (1974). 'A new look at the statistical identification model', *I.E.E.E. Transactions on Automatic Control*, 19, 716-723.

Amemiya, T. (1980). 'Selection of regressors', *International Economic Review*, 21, 331-354.

Bancroft, T.A. (1944). 'On biases in estimation due to the use of preliminary tests of significance', *Annals of Mathematical Statistics*, 15, 190-204.

Bancroft, T.A. (1964). 'Analysis and inference for incompletely specified models involving the use of preliminary tests of significance', *Biometrics*, 20, 427-442.

Baranchik, A.J. (1970). 'A family of minimax estimators of the mean of a multivariate normal distribution', *Annals of Mathematical Statistics*, 41, 642-645.

Baranchik, A.J. (1973). 'Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables', *The Annals of Statistics*, 1, 312-321.

Berger, J.O. (1982). 'Bayesian robustness and the Stein effect', *Journal of the American Statistical Association*, 77, 358-368.

Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis, 2nd Edition, Springer-Verlag, New York.

Berger, J.O. and C. Srinivasan (1978). 'Generalized Bayes estimators in multivariate problems', *The Anuals of Statistics*, 6, 783-801.

Blyth, C.R. (1951). 'On minimax statistical decision procedures and their admissibility', *Annals of Mathematical Statistics*, 22, 22-42.

Bock, M.E. (1975). 'Minimax estimators of the mean of a multivariate normal distribution', *The Annals of Statistics*, 3, 209-218.

Bock, M.E., G.G. Judge and T.A. Yancey (1973). 'Some comments on estimation in regression after preliminary tests of significance', *Journal of Econometrics*, 1, 191-200.

Bock, M.E., T.A. Yancey and G.G. Judge (1973). 'The statistical consequences of preliminary test estimators in regression', *Journal of the American Statistical Association*, 68, 109-116.

Breusch, T.C. (1990). 'Simplified extreme bounds', in: Granger, C.W.J. (ed.), Modelling Economic Series, Clarendon Press, Oxford.

Brook, R.J. (1976). 'On the use of a regret function to set significance points in prior tests of estimation', *Journal of the American Statistical Association*, 71, 126-131.

Brown, L.D. (1966). 'On the admissibility of invariant estimators of one or more location parameters', *Annals of Mathematical Statistics*, 37, 1087-1136.

Brown, L.D. (1971). 'Admissible estimators, recurrent diffusions, and insoluble boundary value problems', *Annals of Mathematical Statistics*, 42, 855-903.

Burr, I.W. (1942). 'Cumulative frequency functions', *Annals of Mathematical Statistics*, 13, 215-232.

Burr, I.W. and P.J. Cislak (1968). 'On a general system of distributions. I. Its curve-shape characteristics. II. The sample median', *Journal of the American Statistical Association*, 63, 627-643.

Cajori, F. (1993). A History of Mathematical Notations, 2 vols, Dover Publications, New York.

Chatfield, C. (1995). 'Model uncertainty, data mining and statistical inference' (with discussion), *Journal of the Royal Statistical Society*, Series A, 158, 419-466.

Chernoff, H. and L.E. Moses (1959). Elementary Decision Theory, John Wiley, New York.

Cohen, A. (1965). 'Estimates of linear combinations of the parameters in the mean vector of a multivariate distribution', *Annals of Mathematical Statistics*, 36, 78-87.

Draper, D. (1995). 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society*, Series B, 57, 45-97.

Droge, B. (1993). 'On finite sample properties of adaptive least squares regression estimates', *Statistics*, 24, 181-203.

Droge, B. and T. Georg (1995). 'On selecting the smoothing parameter of least squares regression estimates using the minimax regret approach', *Statistics and Decisions*, 13, 1-20.

Efron, B. and C. Morris (1971). 'Limiting the risk of Bayes and empirical Bayes estimators - Part I: The Bayes case', *Journal of the American Statistical Association*, 66, 807-815.

Efron, B. and C. Morris (1972). 'Limiting the risk of Bayes and empirical Bayes estimators - Part II: The empirical Bayes case', *Journal of the American Statistical Association*, 67, 130-139.

Efron, B. and C. Morris (1973). 'Stein's estimation rule and its competitors - An empirical Bayes approach', *Journal of the American Statistical Association*, 68, 117-130.

Farebrother, R.W. (1975). 'Minimax regret significance points for a preliminary test in regression analysis: comment', *Econometrica*, 43, 1005-1006.

Feldstein, M.S. (1973). 'Multicollinearity and the mean square error of alternative estimators', *Econometrica*, 41, 337-346.

Friedman, M. (1940). 'Review of Jan Tinbergen. Statistical Testing of Business Cycle Theories, II: Business Cycles in the United States of America', *American Economic Review*, 30, 657-661.

Giles, D.E.A. and A.C. Rayner (1979). 'The mean squared errors of the maximum likelihood and natural-conjugate Bayes regression estimators', *Journal of Econometrics*, 11, 319-334.

Goodnight, J. and T.D. Wallace (1972). 'Operational techniques and tables for making weak MSE tests for restrictions in regressions', *Econometrica*, 1972, 699-709.

Granger, C.W.J. and H.F. Uhlig (1990). 'Reasonable extreme-bounds analysis', *Journal of Econometrics*, 44, 159-170.

Haavelmo, T. (1944). 'The probability approach in econometrics', *Econometrica*, 12 (supplement), 1-115.

Hodges, J.L. and E.L. Lehmann (1950). 'Some problems in minimax point estimation', *Annals of Mathematical Statistics*, 21, 182-197.

Huber, P.J. (1977). Robust Statistical Procedures, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 27, Society for Industrial and Applied Mathematics, Philadelphia.

Huntsberger, D.V. (1955). 'A generalization of a preliminary testing procedure for pooling data', *Annals of Mathematical Statistics*, 26, 734-743.

James, W. and C. Stein (1961). 'Estimation with quadratic loss', *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361-379.

Judge, G.G. and M.E. Bock (1976). 'A comparison of traditional and Stein-rule estimators under weighted squared error loss', *International Economic Review*, 17, 234-240.

Judge, G.G. and M.E. Bock (1978). The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics, North-Holland, Amsterdam.

Judge, G.G. and M.E. Bock (1983). 'Biased estimation', in: Griliches, Z. and M.D. Intriligator (eds.), Handbook of Econometrics, Vol.I, Chapter 10, North-Holland, Amsterdam.

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lutkepohl and T.C. Lee (1985). The Theory and Practice of Econometrics, Second Edition, John Wiley, New York.

Judge, G.G. and T.A. Yancey (1986). Improved Methods of Inference in Econometrics, North-Holland, Amsterdam.

Kadane, J.B., J.M. Dickey, R.L. Winkler, W.S. Smith and S.C. Peters (1980). 'Interactive elicitation of opinion for a normal linear model', *Journal of the American Statistical Association*, 75, 845-854.

Kadane, J.B. and R.L. Winkler (1988). 'Separating probability elicitation from utilities', *Journal of the American Statistical Association*, 83, 357-363.

Kempthorne, P.J. (1984). 'Admissible variable-selection procedures when fitting regression models by least squares for prediction', *Biometrika*, 71, 593-597.

Keynes, J.M. (1939). 'Professor Tinbergen's method', *Economic Journal*, 49, 558-568.

Koopmans, T. (1949). 'Identification problems in economic model construction', *Econometrica*, 17, 125-144.

Laplace, P.S. (1774). 'Mémoire sur la probabilité des causes par les événements', *Mémoires de mathématique et de physique présentés à l'Académie royale des sciences*, 6, 621-656. (Translated by S.M. Stigler (1986), 'Memoir on the probability of the causes of events', *Statistical Science*, 1, 364-378.)

Larson, H.J. and T.A. Bancroft (1963). 'Biases in prediction by regression for certain incompletely specified models', *Biometrika*, 50, 391-402.

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

Leamer, E.E. (1983). 'Model choice', in: Griliches, Z. and M.D. Intriligator (eds.), Handbook of Econometrics, Vol. I, Chapter 5, North-Holland, Amsterdam.

Leamer, E.E. (1983). 'Let's take the con out of econometrics', *American Economic Review*, 73, 31-43.

Leamer, E.E. (1985), 'Sensitivity Analyses would help', *American Economic Review*, 75, 308-313.

Leamer, E.E. (1992). 'Bayesian elicitation diagnostics', *Econometrica*, 60, 919-942.

Leamer, E.E. and G. Chamberlain (1976). 'A Bayesian interpretation of pretesting', *Journal of the Royal Statistical Society*, Series B, 38, 85-94.

Lovell, M.C. (1983). 'Data mining', *The Review of Economics and Statistics*, 65, 1-12.

Mallows, C.L. (1973). 'Some comments on C_p', *Technometrics*, 15, 661-675.

McAleer, M., A.R. Pagan and P.A. Volker (1985). 'What will take the con out of econometrics', *American Economic Review*, 75, 293-307.

McAleer, M. and M.R. Veall (1989). 'How fragile are fragile inferences? A re-evaluation of the deterrent effect of capital punishment', *The Review of Economics and Statistics*, 71, 99-106.

Pericchi, L.R. and A.F.M. Smith (1992). 'Exact and approximate posterior moments for a normal location parameter', *Journal of the Royal Statistical Society*, Series B, 54, 793-804.

Pötscher, B.M. (1991). 'Effects of model selection on inference', *Econometric Theory*, 7, 163-185.

Rao, C.R. (1973). Linear Statistical Inference and Its Applications, 2nd Edition, John Wiley, New York.

Savage, L.J. (1951). 'The theory of statistical decision', *Journal of the American Statistical Association*, 46, 55-67.

Sawa, T. (1978). 'Information criteria for discriminating among alternative regression models', *Econometrica*, 46, 1273-1291.

Sawa, T. and T. Hiromatsu (1973). 'Minimax regret significance points for a preliminary test in regression analysis', *Econometrica*, 41, 1093-1101.

Sclove, S.L. (1968). 'Improved estimators for coefficients in linear regression', *Journal of the American Statistical Association*, 63, 596-606.

Sclove, S.L., C. Morris and R. Radhakrishnan (1972). 'Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution', *Annals of Mathematical Statistics*, 43, 1481-1490.

Stein, C. (1955). 'Inadmissibility of the usual estimator for the mean of a multivariate normal distribution', *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, 197-206.

Stein, C. (1981). 'Estimation of the mean of a multivariate normal distribution', *The Annals of Statistics*, 9, 1135-1151.

Stigler, S.M. (1986). The History of Statistics, Harvard University Press, Cambridge, Mass, USA.

Strawderman, W.E. and A. Cohen (1971). 'Admissibility of estimators of the mean vector of a multivariate normal distribution with quadratic loss', *Annals of Mathematical Statistics*, 42, 270-296.

Thompson, J.R. (1968). 'Some shrinkage techniques for estimating the mean', *Journal of the American Statistical Association*, 63, 113-122.

Tinbergen, J. (1939). Statistical Testing of Business Cycle Theories, 2 volumes, League of Nations, Geneva.

Toro-Vizcarrondo, C. and T.D. Wallace (1968). 'A test of the mean square error criterion for restrictions in linear regression', *Journal of the American Statistical Association*, 63, 558-572.

Toyoda, T. and T.D. Wallace (1976). 'Optimal critical values for pre-testing in regression', *Econometrica*, 44, 365-375.

Wallace, T.D. (1964). 'Efficiencies for stepwise regressions', *Journal of the American Statistical Association*, 59, 1179-1182.

Wallace, T.D. (1972). 'Weaker criteria and tests for linear restrictions in regression', *Econometrica*, 40, 689-698.

Wallace, T.D. and V.G. Ashtar (1972). 'Sequential methods in model construction', *The Review of Economics and Statistics*, 54, 172-178.

Wallace, T.D. and C.E. Toro-Vizcarrondo (1969). 'Tables for the mean square error test for exact linear restrictions in regression', *Journal of the American Statistical Association*, 64, 1649-1663.

Wold, H. (1953). Demand Analysis: A Study in Econometrics, John Wiley, New York.

Zellner, A. (1994). 'Bayesian method of moments/instrumental variable (BMOM/IV) analysis of mean and regression models', H.G.B. Alexander Research Foundation, Graduate School of Business, University of Chicago.

Zellner, A. and W. Vandaele (1974). 'Bayes-Stein estimators for k-means, regression and simultaneous equation models, in: S.E. Fienberg and A. Zellner (eds.), Studies in Bayesian Econometrics and Statistics in Honor of Leonard L. Savage, North-Holland, Amsterdam, 627-653.

Zhang, P. (1992). 'On the distributional properties of model selection criteria', *Journal of the American Statistical Association*, 87, 732-737.

No.	Author(s)	Title
9599	H. Bloemen	The Relation between Wealth and Labour Market Transitions: an Empirical Study for the Netherlands
95100	J. Blanc and L. Lenzini	Analysis of Commmunication Systems with Timed Token Protocols using the Power-Series Algorithm
95101	R. Beetsma and L. Bovenberg	The Interaction of Fiscal and Monetary Policy in a Monetary Union: Balancing Credibility and Flexibility
95102	P. de Bijl	Aftermarkets: The Monopoly Case
95103	F. Kumah	Unanticipated Money and the Demand for Foreign Assets - A Rational Expectations Approach
95104	M. Vázquez-Brage, A. van den Nouweland, I. García-Jurado	Owen's Coalitional Value and Aircraft Landing Fees
95105	Y. Kwan and G. Chow	Estimating Economic Effects of the Great Leap Forward and the Cultural Revolution in China
95106	P. Verheyen	The missing Link in Budget Models of Nonprofit Institutions; Two Practical Dutch Applications
95107	J. Miller	Should we Offer the Unemployment Places on Labour Market Programmes With the Intention That They Reject Them?
95108	C. van Raalte and H. Webers	Statial Competition with Intermediated matching
95109	W. Verkooijen, J. Plasmans and H. Daniels	Long-Run Exchange Rate Determination: A Neutral Network Study
95110	E. van der Heijden, J. Nelissen, J. Potters and H. Verbon	Transfers and Reciprocity in Overlapping-Generations Experiments
95111	A. van den Elzen and D. Talman	An Algorithmic approach towards the Tracing Procedure of Harsanyi and Selten
95112	R. Beetsma and L. Bovenberg	Does Monetary Unification Lead to Excessive Debt Accumulation?
95113	H. Keuzenkamp	Keynes and the Logic of Econometric Method
95114	E. Charlier, B. Melenberg and A. van Soest	Estimation of a Censored Regression Panel Data Model Using Conditional Moment Restrictions Efficiently
95115	N. G. Noorderhaven, B. Nooteboom and H. Berger	Exploring Determinants of Perceived Interfirm Dependence in Industrial Supplier Relations

No.	Author(s)	Title
95116	F. Kleibergen and H. Hoek	Bayesian Analysis of ARMA models using Noninformative Priors
95117	J. Lemmen and S. Eijffinger	The Fundamental Determiniants of Financial Integration in the European Union
95118	B. Meijboom and J. Rongen	Clustering, Logistics, and Spatial Economics
95119	A. de Jong, F. de Roon and C. Veld	An Empirical Analysis of the Hedging Effectiveness of Currency Futures
95120	J. Miller	The Effects of Labour Market Policies When There is a Loss of Skill During Unemployment
95121	S. Eijffinger, M. Hoeberichts and E. Schaling	Optimal Conservativeness in the Rogoff (1985) Model: A Graphical and Closed-Form Solution
95122	W. Ploberger and H. Bierens	Asymptotic Power of the Integrated Conditional Moment Test Against Global and Large Local Alternatives
95123	H. Bierens	Nonparametric Cointegration Analysis
95124	H. Bierens and W. Ploberger	Asymptotic Theory of Integrated Conditional Moment Tests
95125	E. van Damme	Equilibrium Selection in Team Games
95126	J. Potters and F. van Winden	Comparative Statics of a Signaling Game: An Experimental Study
9601	U. Gneezy	Probability Judgements in Multi-Stage Problems: Experimental Evidence of Systematic Biases
9602	C. Fernández and M. Steel	On Bayesian Inference under Sampling from Scale Mixtures of Normals
9603	J. Osiewalski and M. Steel	Numerical Tools for the Bayesian Analysis of Stochastic Frontier Models
9604	J. Potters and J. Wit	Bets and Bids: Favorite-Longshot Bias and Winner's Curse
9605	H. Gremmen and J. Potters	Assessing the Efficacy of Gaming in Economics Educating
9606	J. Potters and F. van Winden	The Performance of Professionals and Students in an Experimantal Study of Lobbying
9607	J.Kleijnen, B. Bettonvil and W. van Groenendaal	Validation of Simulation Models: Regression Analysis Revisited
9608	C. Fershtman and N. Gandal	The Effect of the Arab Boycott on Israel: The Automobile Market
9609	H. Uhlig	Bayesian Vector Autoregressions with Stochastic Volatility

No.	Author(s)	Title
9610	G. Hendrikse	Organizational Change and Vested Interests
9611	F. Janssen, R. Heuts and T. de Kok	On the (R,s,Q) Inventory Model when Demand is Modelled as a compound Bernoulli Process
9612	G. Fiestras-Janeiro P. Borm and F. van Megen	Protective Behaviour in Games
9613	F. van Megen, G. Facchini, P. Borm and S. Tijs	Strong Nash Equilibria and the Potential Maximizer
9614	J. Miller	Do Labour Market Programmes Necessarily Crowd out Regular Employment? - A Matching Model Analysis
9615	H. Huizinga	Unemployment Benefits and Redistributive Taxation in the Presence of Labor Quality Externalities
9616	G. Asheim and M. Dufwenberg	Admissibility and Common Knowlegde
9617	C. Fernández, J. Osiewalski and M. Steel	On the Use of Panel Data in Bayesian Stochastic Frontier Models
9618	H. Huizinga	Intrafirm Information Transfers and Wages
9619	B. Melenberg and B. Werker	On the Pricing of Options in Incomplete Markets
9620	F. Kleibergen	Reduced Rank Regression using Generalized Method of Moments Estimators
9621	F. Janssen, R. Heuts and T. de Kok	The Value of Information in an (R,s,Q) Inventory Model
9622	F. Groot, C. Withagen and A. de Zeeuw	Strong Time-Consistency in the Cartel-Versus-Fringe Model
9623	R. Euwals and A. van Soest	Desired and Actual Labour Supply of Unmarried Men and Women in the Netherlands
9624	M. Khanman, M. Perry and P.J. Reny	An Ex-Post Envy-Free and Efficient Allocation Mechanism: Imperfect Information Without Common Priors
9625	C. Eaves, G. van der Laan D. Talman and Z. Yang	Balanced Simplices on Polytopes
9626	H. Bloemen and A. Kalwij	Female Employment and Timing of Births Decisions: a Multiple State Transition Model
9627	P. Bolton	Strategic Experimentation: a Revision
9628	C. Fershtman and Y. Weiss	Social Rewards, Externalities and Stable Preferences

No.	Author(s)	Title
9629	P.M. Kort, G. Feichtinger R.F. Hartl and J.L. Haunschmied	Optimal Enforcement Policies (Crackdowns) on a Drug Market
9630	C. Fershtman and A. de Zeeuw	Tradeable Emission Permits in Oligopoly
9631	A. Cukierman	The Economics of Central Banking
9632	A. Cukierman	Targeting Monetary Aggregates and Inflation in Europe
9633	F. de Roon, C. Veld and J. Wei	A Study on the Efficiency of the Market for Dutch Long Term Call Options
9634	B. van Aarle, L. Bovenberg M. Raith	Money, Fiscal Deficits and Government Debt in a Monetary Union
9635	G. van der Laan, D. Talman and Z. Yang	Existence of and Equilibrium in a Competitive Economy with Indivisibilities and Money
9636	I. Brouwer, J. van der Put C. Veld	Contrarian Investment Strategies in a European Context
9637	M. Berg, F. van der Duyn Schouten and J. Jansen	Optimal Service Policies to Remote Customers with Delay-Limits
9638	F.C. Drost and C.A.J. Klaassen	Efficient Estimation in Semiparametric GARCH Models
9639	A.G. de Kok and F.B.S.L.P. Janssen	Demand Management in Multi-Stage Distribution Chain
9640	D.A. Freund, T.J. Kniesner A.T. LoSasso	How Managed Care Affects Medicaid Utilization A Synthetic Difference-in-Differences Zero-Inflated Count Model
9641	M.J. Lee	Instrumental Variable Estimation For Linear Panel Data Models
9642	W. Härdle, E. Mammen and M. Müller	Testing Parametric versus Semiparametric Modelling in Generalized Linear Models
9643	C. Dustmann, N. Rajah and A. van Soest	Part-Time Work, School Success and School Leaving
9644	S.C.W. Eijffinger and M. Hoeberichts	The Trade Off Between Central Bank Independence and Conservativeness
9645	R. Sarin and P. Wakker	A Single-Stage Approach to Anscombe and Aumann's Expected Utility

No.	Author(s)	Title
9646	J.P. Ziliak and T.J. Kniesner	The Importance of Sample Attrition in Life Cycle Labor Supply Estimation
9647	P.M. Kort	Optimal R&D Investments of the Firm
9648	M.P. Berg	Performance Comparisons for Maintained Items
9649	H. Uhlig and Y. Xu	Effort and the Cycle: Cyclical Implications of Efficiency Wages
9650	M. Slikker and A. van den Nouweland	Communication Situations with a Hierarchical Player Partition
9651	H.L.F. de Groot	The Struggle for Rents in a Schumpeterian Economy
9652	R.M. de Jong and J. Davidson	Consistency of Kernel Estimators of heteroscedastic and Autocorrelated Covariance Matrices
9653	J. Suijs, A. De Waegenaere and P. Borm	Stochastic Cooperative Games in Insurance and Reinsurance
9654	A.N. Banerjee and J.R. Magnus	Testing the Sensitivity of OLS when the Variance Matrix is (Partially) Unknown
9655	A. Kalwij	Estimating the Economic Return to Schooling on the basis of Panel Data
9656	M. Lind and F. van Megen	Order Based Cost Allocation Rules
9657	A. van Soest, P. Fontein and Rob Euwals	Earnings Capacity and Labour Market Participation
9658	C. Fernández and M.F.J. Steel	On Bayesian Modelling of Fat Tails and Skewness
9659	R. Sarin and P. Wakker	Revealed Likelihood and Knightian Uncertainty
9660	J.R. Magnus and J. Durbin	A Classical Problem in Linear Regression or How to Estimate the Mean of a Univariate Normal Distribution with Known Variance