Econometrics Journal (2002), volume **5**, pp. 225–236.

Estimation of the mean of a univariate normal distribution with known variance

JAN R. MAGNUS

CentER, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands
E-mail: magnus@kub.nl

Received: February 2001

Summary We consider the estimation of the unknown mean η of a univariate normal distribution $\mathrm{N}(\eta, 1)$ given a single observation x. We wish to obtain an estimator which is admissible and has good risk (and regret) properties. We first argue that the 'usual' estimator $t(x) = x$ is not necessarily suitable. Next, we show that the traditional pretest estimator of the mean has many undesirable properties. Thus motivated, we suggest the Laplace estimator, based on a 'neutral' prior for η, and obtain its properties. Finally, we compare the Laplace estimator with a large class of (inadmissible) estimators and show that the risk properties of the Laplace estimator are close to those of the minimax regret estimator from this large class. Thus, the Laplace estimator has good risk (regret) properties as well. Questions of admissibility, risk and regret are reviewed in the appendix.

Keywords: *Pretest, Laplace distribution, Regret, Neutral prior.*

1. INTRODUCTION AND MOTIVATION

Let x be a single observation from a univariate normal distribution with unknown mean η and variance 1, that is, $x \sim \mathrm{N}(\eta, 1)$, $-\infty < \eta < \infty$. In this paper we address the problem of how to estimate η. There are two reasons for raising this seemingly trivial problem. The first reason is to show that this problem is not trivial at all, that the 'usual' estimator $t(x) = x$ is not necessarily a good estimator of η, and to present an improved, possibly optimal, estimator. The second reason is based on the fact, discussed in Magnus and Durbin (1999), that this statistical problem is *equivalent* to an important econometric problem, namely the problem of estimating $\boldsymbol{\beta}$ in the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{z} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n), \tag{1}$$

where $\boldsymbol{X}(n \times k)$ is a matrix of explanatory variables that are required to be in the model on theoretical or other grounds, while there is doubt whether the additional explanatory variable $\boldsymbol{z}(n \times 1)$ should be in the model or not.[1] We call the first problem *the* $\mathrm{N}(\eta, 1)$ *problem* and the second problem *the regression problem.*

The equivalence between the two problems needs to be made a little more precise. Assume that $(\boldsymbol{X} : \boldsymbol{z})$ has full column-rank and that σ^2 is known. (The latter assumption is, of course, unrealistic, but it simplifies the analysis without affecting the main results.) Let $\boldsymbol{b}_u$ and $\hat{\gamma}$ denote

[1] Of course we follow the notation proposed in Abadir and Magnus (2002).

the OLS estimators of $\boldsymbol{\beta}$ and γ in the 'unrestricted' model (1), and let $\boldsymbol{b}_r$ denote the 'restricted' estimator when $\gamma = 0$. Let $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, and define

$$\eta = \frac{\gamma}{\sigma/\sqrt{\boldsymbol{z}'\boldsymbol{M}\boldsymbol{z}}}, \qquad \hat{\eta} = \frac{\hat{\gamma}}{\sigma/\sqrt{\boldsymbol{z}'\boldsymbol{M}\boldsymbol{z}}}. \tag{2}$$

The parameter η is the usual non-centrality parameter associated with the t-statistic for testing $\gamma = 0$. Since σ^2 is assumed known, $\hat{\eta} \sim \mathrm{N}(\eta, 1)$. One can show that $\boldsymbol{b}_r$ dominates $\boldsymbol{b}_u$ (in the mean squared error sense) if and only if $\eta^2 < 1$ (Magnus and Durbin, 1999, Theorem 1). Since both $\boldsymbol{b}_r$ and $\boldsymbol{b}_u$ appear to be too extreme in this context, we define a new estimator, called a weighted-average least squares (WALS) estimator,

$$\boldsymbol{b} = \lambda\boldsymbol{b}_u + (1-\lambda)\boldsymbol{b}_r, \tag{3}$$

where $\lambda = \lambda(\hat{\eta})$ is a scalar function of $\hat{\eta}$ such that $0 \leq \lambda \leq 1$. We should think of λ as a non-decreasing function of $|\hat{\eta}|$, so that the larger is $|\hat{\eta}|$ the larger will be λ and hence the more weight will be put on $\boldsymbol{b}_u$ relative to $\boldsymbol{b}_r$. Theorem 2 in Magnus and Durbin (1999) shows that $\mathrm{MSE}(\boldsymbol{b})$ is minimized if and only if $\mathrm{MSE}(\lambda(\hat{\eta})\hat{\eta})$ is minimized. Hence, finding the best WALS estimator of $\boldsymbol{\beta}$ is equivalent to finding the best estimator of η, and the regression problem is solved if and only if we can solve the $\mathrm{N}(\eta, 1)$ problem.

This equivalence implies that associated to any estimator of η in the $\mathrm{N}(\eta, 1)$ problem there exists a unique estimator of $\boldsymbol{\beta}$ in the regression problem. For example, the unrestricted estimator $\boldsymbol{b}_u$ of $\boldsymbol{\beta}$ corresponds to the estimator $t(x) = x$ of η (which we call the 'usual' estimator) and the restricted estimator $\boldsymbol{b}_r$ corresponds to the estimator $t(x) = 0$ (which we call the 'silly' estimator). Now, the 'usual' estimator may make a lot of sense in the $\mathrm{N}(\eta, 1)$ context, but the unrestricted estimator $\boldsymbol{b}_u$ makes less sense in the regression context, because it implies choosing $\boldsymbol{b}_u$ whatever the value of the t-statistic for γ. The equivalence theorem thus shows that we have to reconsider the usefulness of the 'usual' estimator also in the $\mathrm{N}(\eta, 1)$ context, and try and find an alternative to it.

This paper attempts to find an 'optimal' estimator of η in the $\mathrm{N}(\eta, 1)$ problem. Let $t(x) = \lambda(x)x$ be this 'optimal' estimator. Then the equivalence theorem guarantees that *the same* λ-function will provide the 'optimal' WALS estimator (3), irrespective of the values of the regressors.

In Section 2 we briefly review the normal Bayes and the pretest estimator. The first is admissible but has unbounded risk; the second has bounded risk but is inadmissible. In Section 3 we introduce the Laplace estimator based on a neutral prior and argue that this is the estimator we require. Section 4 concludes. An appendix reviews some properties of risk, regret, and admissibility.

2. THE 'USUAL' AND THE PRETEST ESTIMATOR

So, let $x \sim \mathrm{N}(\eta, 1)$ and let $t(x, \lambda) = \lambda(x)x$ be an estimator of η. The 'usual' estimator $t(x, \lambda) = x$ (with $\lambda \equiv 1$) is unbiased, admissible and minimax (Theorem A.1). These are strong properties in favor of x as an estimator of η. 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 $\lambda_0^{(1)}(x) = 1$

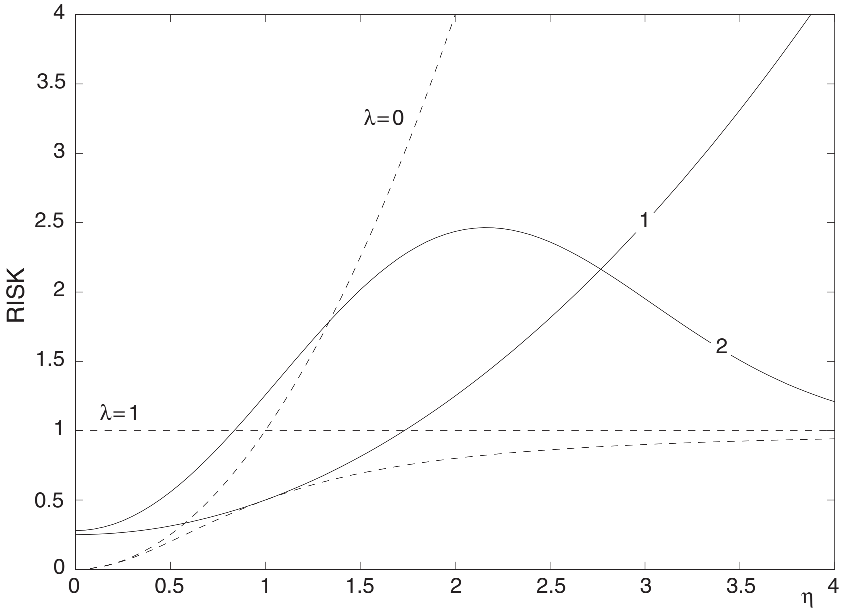

Figure 1. Risk of the normal Bayes (1) and the pretest (2) estimators.

and $\lambda_\infty^{(1)}(x) = 0$. (The reason for not defining $\lambda_c^{(1)} = c$ is to facilitate comparison with $\lambda_c^{(2)}$–$\lambda_c^{(4)}$ to be defined later.) Now consider

$$t^{(1)}(x; c) \equiv t(x, \lambda_c^{(1)}) = x/(1+c) \tag{4}$$

as an estimator of η. The risk is

$$R^{(1)}(\eta, c) \equiv R(\eta, \lambda_c^{(1)}) = \mathrm{E}_\eta(x/(1+c) - \eta)^2 = \frac{1 + c^2\eta^2}{(1+c)^2} \tag{5}$$

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

$$R^{(1)}(\eta, c^*) = \eta^2/(1+\eta^2). \tag{6}$$

For $c \geq 0$, the estimator $t^{(1)}(x; c)$ is admissible (Theorem A.2), and will be called the *normal Bayes estimator* of η, because it is the Bayes estimator induced by a normal prior with mean 0 and variance $1/c$. In Figure 1 we graph the risk $R^{(1)}(\eta, c)$ as a function of η for three values of c, namely $c = 0$, 1 and ∞. The broken line (denoted $\lambda = 1$) corresponds to $c = 0$, the broken line (denoted $\lambda = 0$) corresponds to $c = \infty$, and the solid line (labeled 1) corresponds to $c = 1$. The third broken line gives the lower bound (6). The figure 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 $\frac{1}{3}\sqrt{3} < \eta < \sqrt{3}$ 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. The class of normal Bayes estimators is admissible, but—as Figure 1 shows—has unbounded risk (unless $c = 0$). Since for any $c \geq 0$ 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, $R^{(1)}(\eta, 0) = 1$ and $R^{(1)}(\eta, \infty) = \eta^2$, and hence

$$R^{(1)}(\eta, \infty) \leq R^{(1)}(\eta, 0) \qquad \text{if and only if} \quad |\eta| \leq 1. \tag{7}$$

This is, of course, the essence of Theorem 1 in Magnus and Durbin (1999). It suggests that the 'usual' estimator $t^{(1)}(x; 0) = x$ is good when $|\eta|$ is large, but not so good when $|\eta|$ is small.

The second estimator in common use for the $\mathrm{N}(\eta, 1)$ problem is the traditional pretest estimator[2] defined as

$$t^{(2)}(x; c) \equiv t(x, \lambda_c^{(2)}) = \begin{cases} 0 & \text{if } |x| \leq c, \\ x & \text{if } |x| > c, \end{cases} \tag{8}$$

where

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

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 (8) has a discontinuity at $x = \pm c$ and we would therefore expect the estimator to behave badly. This is indeed the case.

First, the pretest estimator is inadmissible.[3] When we graph the risk $R^{(2)}(\eta, c)$ as a function of η for $c = 1.96$ (Figure 1, solid line, labeled 2), we see that no estimator in this class dominates any other, and that the risk is bounded and converges to 1 as $\eta \to \infty$ (unless $c = \infty$).

Closer inspection of Figure 1 reveals the second bad property of the pretest estimator: for η close to 0 we see, as expected, that the pretest estimator is better than the 'usual' estimator $t^{(2)}(x; 0) = x$, but worse than the 'silly' estimator $t^{(2)}(x; \infty) = 0$. When η is large, the situation is reversed. This, again, is what we would expect. However, for moderate values of η, in particular around $\eta = 1$, 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 1 shows that the opposite is the case! This result is not new—it even appears in some of the textbooks, see Judge *et al.* (1985, p. 75)—but its harmful consequences do not seem to have been fully appreciated.

In summary, the class of pretest estimators—in addition to other bad properties—is inadmissible, but has bounded risk (unless $c = \infty$).

3. THE LAPLACE ESTIMATOR

Motivated by the previous section, we are looking for an estimator of η which is admissible, has bounded risk, has good properties around $\eta = 1$, and is optimal or near-optimal in terms of minimax regret (see appendix).

[2] See Magnus (1999) for a recent survey of the traditional pretest estimator.

[3] 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 appendix). Thus, the estimator $t^{(2)}(x; c)$ is inadmissible (unless $c = 0$ or $c = \infty$), but $\mathcal{L}^{(2)}$-admissible.

In order to assure admissibility we assume a prior density for η, say $\pi(\eta)$. Since $x|\eta \sim \mathrm{N}(\eta, 1)$, we obtain

$$t(x) = \lambda(x)x = \frac{\int \eta\pi(\eta)\exp(-(x-\eta)^2/2)\,r\,d\eta}{\int \pi(\eta)\exp(-(x-\eta)^2/2)\,r\,d\eta}.$$

We shall say that a prior for η is *neutral* if the distribution of η is located at 0 and the distribution of η^2 is located at 1; see the discussion in the introduction and following (7). More precisely, we demand that

$$\Pr(\eta < 0) = \Pr(\eta > 0) = \tfrac{1}{2}, \qquad \Pr(|\eta| < 1) = \Pr(|\eta| > 1) = \tfrac{1}{2},$$

that is, $\mathrm{median}(\eta) = 0$ and $\mathrm{median}(\eta^2) = 1$.

There are various continuous distributions which satisfy these two requirements: uniform on $[-2, 2]$, triangular on $[-(2+\sqrt{2}), 2+\sqrt{2}]$, normal $\mathrm{N}(0, 2.1981)$, logistic(log 3), Cauchy, and Laplace(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. The Laplace prior is the one whose maximum regret is lowest (0.5127 vs. 0.6332 for Cauchy and 1.2150 for logistic), and hence is the one we prefer.

The Laplace (or *double exponential*) density is given by

$$\pi(\eta; c) = \frac{c}{2}\exp(-c\,|\eta|), \quad -\infty < \eta < \infty, \qquad c > 0. \tag{9}$$

Laplace, in his fundamental memoir on inverse probability, deduced this distribution from the principle of 'insufficient reason'; see Stigler (1986, pp. 109–113) for a summary of Laplace's own argument. The density (9) is unimodal and symmetric around 0. Hence the median of η is 0 and the median of η^2 is $(\log 2)^2/c^2$.

Assuming a Laplace prior density $\pi(\eta; c)$ for η, the mean of the posterior distribution of $\eta|x$ can be expressed as

$$t^{(3)}(x; c) = \frac{1+h(x)}{2}(x-c) + \frac{1-h(x)}{2}(x+c) = x - h(x)c, \tag{10}$$

where

$$h(x) = \frac{1-\mathrm{e}^{2cx}\psi(x)}{1+\mathrm{e}^{2cx}\psi(x)}, \qquad \psi(x) = \frac{\Phi(-x-c)}{\Phi(x-c)},$$

and $\Phi(\cdot)$ denotes the standard-normal c.d.f.[4] 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 $t^{(3)}$ with $c = \log 2$ will be called the 'neutral' Laplace estimator, since it is based on a neutral prior. The estimator is easy to compute, has a strong justification based on the neutral prior, is admissible, and has bounded risk. Its risk function (labeled 3) is graphed in Figure 2.

We notice that the risk $R^{(3)}(\eta, c)$ increases monotonically with $|\eta|$ and that $R^{(3)}(\eta, c) \to 1 + c^2$ as $|\eta| \to \infty$. The risk of the 'neutral' Laplace estimator is bounded by $0.4446 \leq R^{(4)}(\eta, c) \leq$

[4]See Pericchi and Smith (1992). Our expression (10) is easier for computational purposes than their formula (6), because h is monotonic.

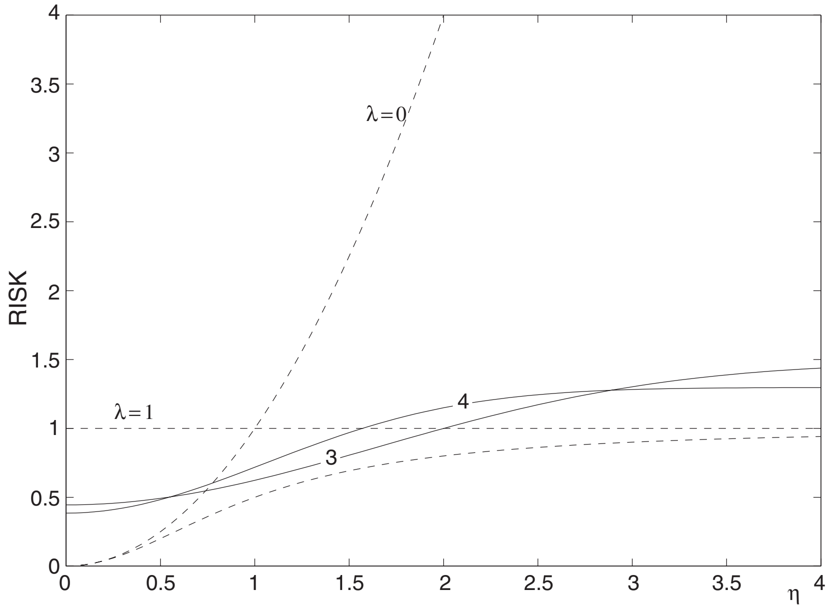

Figure 2. Risk of the 'neutral' Laplace (3) and the optimal Burr (4) estimators.

1.4805, and its regret by $0.1006 \leq R^{(4)}(\eta, c) - \eta^2/(1+\eta^2) \leq 0.5127$. The minimum regret is obtained for $\eta = 1.27$ and the maximum regret for $\eta = 4.93$. For η close to 0, the 'neutral' Laplace estimator is better (has lower risk) than the 'usual' estimator $t(x) = x$, but worse than the 'silly' estimator $t(x) = 0$. This is what we would expect. For $|\eta|$ large, the situation is reversed. Again, this is what we would expect. But for quite a large and important interval, $0.739 < |\eta| < 2.001$, the 'neutral' Laplace estimator is better than both the 'usual' and the 'silly' estimator.

In order to study the minimax regret properties of the Laplace estimator, we consider briefly some (inadmissible) competitors. We know from Theorem A.7 that $R(\eta, \lambda) \geq \eta^2/(1+\eta^2)$ for every η and every $\lambda \in \mathcal{L}^0$. The lower bound is attained when $\lambda = \eta^2/(1+\eta^2)$. An obvious and suggestive choice for λ is therefore $\lambda(x) = x^2/(c^2+x^2)$, where $0 \leq c \leq \infty$. The estimator then takes the simple form $t(x) = x^3/(c^2+x^2)$. For $c = 0$ and $c = \infty$ we again find the 'usual' and the 'silly' estimators as special cases. This estimator is inadmissible (Strawderman and Cohen (1971, p. 278)). In spite of this, the estimator has two big advantages over the pretest estimator. First, there now exists, for every c, an interval around $\eta = 1$, where the estimator is better (has lower risk) than the 'usual' estimator *and* the 'silly' estimator, while there is no value of η where the estimator is worse than both the 'usual' and the 'silly' estimator. Secondly, the minimax $\mathcal{L}^0$-regret solution takes the value 0.4251, which is substantially lower than for the pretest estimator (0.6958). Hence a simple generalization of the pretest estimator already produces an estimator with more attractive properties.

We now search for the 'optimal' λ-function in a much larger class. 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. The class of distribution functions we shall use is $\lambda(x; \alpha, \beta) = 1 - (1 + (x^2/c^2)^\alpha)^{-\beta}$, 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.

Extensive calculations (not reported here) show that the minimax regret estimator is obtained along the path $2\alpha\beta = 1$ when $\alpha \to \infty$. This gives

$$\lambda_c^{(4)}(x) = \begin{cases} 0 & \text{if } |x| \leq c, \\ 1 - \dfrac{c}{|x|} & \text{if } |x| > c. \end{cases} \tag{11}$$

The estimator $t^{(4)}(x; c)$ obtained from (11) is called the Burr estimator. It is 'kinked', hence not differentiable and thus inadmissible by Theorem A.6. The minimax $\mathcal{L}^0$-regret solution in the class of Burr estimators (and hence in the whole Burr class) is obtained for $c = 0.545$ with maximum regret 0.3850. The estimator $t^{(4)}(x; c)$ with $c = 0.545$ is the optimal Burr estimator and its risk function is graphed in Figure 2 (labeled 4). We see that the difference in risk profile between the 'neutral' Laplace estimator and the optimal Burr estimator is very small. Hence, even though risk and regret were not primary considerations in choosing the Laplace estimator, it behaves very well from the minimax regret point of view as well.

4. CONCLUDING REMARKS

We briefly consider the application of these results to the regression problem. For every estimator $t(x) = \lambda(x)x$ for the $N(\eta, 1)$ problem we have a corresponding WALS estimator $\boldsymbol{b} = \lambda(\hat{\eta})\boldsymbol{b}_u + (1 - \lambda(\hat{\eta}))\boldsymbol{b}_r$. 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 $N(\eta, 1)$ problem imply that a fixed λ (for example $\lambda = 1/2$) is unattractive, because it corresponds to the normal Bayes estimator $t^{(1)}$. Also, the traditional pretest estimator (choose the restricted estimator $\boldsymbol{b}_r$ when the t-statistic for γ is small, choose the unrestricted estimator $\boldsymbol{b}_u$ otherwise) is unattractive, 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. It is not surprising then that we do not recommend the traditional pretest estimator. The 'neutral' Laplace estimator does not suffer from these problems. It is admissible, has a strong and plausible Bayesian interpretation, has good minimax regret properties, and is easy to compute.

The current paper has tried to concentrate on the main issues by making several simplifying assumptions. First, the assumption that σ^2 is known is clearly unrealistic. Magnus and Durbin (1999) show that the assumption that σ^2 is known is unnecessary. Preliminary investigations show 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, see also Droge and Georg (1995). Nevertheless, further work in this direction is necessary. Secondly, we have assumed that there is only one nuisance parameter γ. Again, the basic set-up is still valid when there are more than one nuisance parameters; see Magnus and Durbin (1999). Application of this basic framework to finding optimal estimators in the case of more than one nuisance parameter is, however, more complicated.

ACKNOWLEDGEMENTS

I am grateful to: Richard Smith, two referees, Jim Berger, Bernd Droge, Ed Leamer, Peter Phillips, Hans Schumacher, Aart de Vos and Arnold Zellner for constructive comments; to Dmitri Danilov for help in producing the graphs; and to Jim Durbin for continuous encouragement and positive criticism.

REFERENCES

Abadir, K. M. and J. R. Magnus (2002). Notation in econometrics: a proposal for a standard. *Econometrics Journal 5*, 76–90.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.

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

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

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*, 215–32.

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*, 627–43.

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.

Giles, D. E. A. (2002). Preliminary-test and Bayes estimation of a location parameter under 'reflected normal' loss. In A. Ullah, A. Wan and A. Chaturvedi (eds), *Handbook of Applied Econometric and Statistical Inference*. New York: Marcel Dekker.

Giles, J. A. and D. E. A. Giles (1996). Risk of a homoscedasticity pre-test estimator of the regression scale under LINEX loss. *Journal of Statistical Planning and Inference 50*, 21–35.

Hodges, J. L. and E. L. Lehmann (1950). Some problems in minimax point estimation. *Annals of Mathematical Statistics 21*, 182–97.

Huntsberger, D. V. (1955). A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics 26*, 734–43.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lütkepohl and T. C. Lee (1985). *The Theory and Practice of Econometrics*, 2nd edn. New York: Wiley.

Magnus, J. R. (1999). The traditional pretest estimator. *Theory of Probability and Its Applications 44*, 293–308.

Magnus, J. R. and J. Durbin (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica 67*, 639–43.

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.

Stigler, S. M. (1986). *The History of Statistics*. Cambridge, MA: Harvard University Press.

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*, 270–96.

APPENDIX: A REVIEW OF RISK, ADMISSIBILITY AND REGRET

This appendix contains some results about risk, admissibility, and regret. Most of these results are known, although not necessarily easily accessible. Only Theorem A.7 is new.

Let x be a single observation from a univariate normal distribution with unknown mean η and variance 1, that is, $x \sim \mathrm{N}(\eta, 1)$, $-\infty < \eta < \infty$. We wish to estimate η using estimators of the form $t(x, \lambda) = \lambda(x)x$, $\lambda \in \mathcal{L}$, where $\mathcal{L}$ is some class of real-valued functions to be specified later. The class of estimators will be denoted $\mathcal{T}(\mathcal{L})$. Thus, $\mathcal{T}(\mathcal{L}) = \{t : t = t(x, \lambda),\ \lambda \in \mathcal{L}\}$. Throughout we assume squared error loss $(t(x, \lambda) - \eta)^2$, and the relative merits of the estimators considered depend on this choice.[5] The *risk* of an estimator $t \in \mathcal{T}(\mathcal{L})$ is then defined as its mean squared error,

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

where E_η denotes expectation with respect to the $\mathrm{N}(\eta, 1)$ 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 $R(\eta, \lambda_1) \leq R(\eta, \lambda_2)$ for all η, with strict inequality for some η. 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 $\mathcal{L}$.

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

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

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*.

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

Proof. 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)). □

Theorem A.2 (λ constant). *For any c, $0 \leq c \leq \infty$, the estimator $t^{(1)}(x; c) = x/(1 + c)$ is admissible.*

Proof. Since $R(0, \lambda) \geq 0$ with equality if and only if $\lambda \equiv 0$, we see that $t^{(1)}(x; \infty)$ dominates every other estimator at $\eta = 0$ and hence is admissible. Also, $t^{(1)}(x; 0)$ is admissible by Theorem A.1. For $0 < c < \infty$, the result follows from Berger (1985, pp. 127–128, 161 and 254). □

Since $\lambda_c^{(1)} = 1/(1 + c)$ and the optimal c is given by $c^* = 1/\eta^2$, we find the optimal λ to be $\lambda^* = \lambda_{c^*}^{(1)} = \eta^2/(1 + \eta^2)$. The optimal λ^*, as a function of η, thus satisfies $0 \leq \lambda^*(\eta) \leq 1$, $\lambda^*(-\eta) = \lambda^*(\eta)$, and λ^* is increasing on $(0, \infty)$. Now, η is not known. But if we think of x as a *preliminary estimator* of η, then these ideas lead quite naturally and intuitively to the following minimal regularity conditions for λ.

Regularity Condition R1. λ is a real-valued function defined on $\mathbb{R}$ and satisfies the following conditions: (a) $0 \leq \lambda(x) \leq 1$ for all x, (b) $\lambda(-x) = \lambda(x)$ for all x, (c) λ is non-decreasing on $[0, \infty)$, (d) λ is continuous except possibly on a set of measure zero.

[5] See Giles and Giles (1996) for examples of the use of an asymmetric loss function. In addition, boundedness of the loss function may be important; see Giles (2002).

Condition (a) defines $t(x, \lambda)$ as a shrinkage estimator (towards 0). This makes good sense, since, if $\lambda(x)$ were such that $\lambda(x) \geq \lambda > 1$ for all $|x| > M$, then $t(x) = \lambda(x)x$ would be inadmissible Strawderman and Cohen (1971, Theorem 5.5.1). Also, $\lambda = 0$ and $\lambda = 1$ are natural endpoints because they correspond—in view of the equivalence between the $\mathrm{N}(\eta, 1)$ problem and the regression problem—to the restricted and unrestricted estimator, respectively. Condition (b) has several rationales. The simplest, perhaps, is the following. Let $\pi(\eta)$ be a prior density of η. Since we are ignorant about η, let us assume that $\pi(\eta)$ is symmetric around 0. This seems reasonable, because ignorance about η in the $\mathrm{N}(\eta, 1)$ problem means ignorance about γ in the equivalent regression problem, and in particular ignorance about the *sign* of γ. The assumption $\Pr(\gamma > 0) = \Pr(\gamma < 0)$ leads to a prior which is symmetric about $\eta = 0$. Then the mean $t(x)$ of the posterior distribution of $\eta|x$, that is, the Bayes estimator of η, 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: $t(x) = \lambda(x)x + (1 - \lambda(x))0$. The larger is $|x|$, the better is x as an estimator for η. 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 A.3 (antisymmetry of the bias). *Let* $\mathrm{bias}(\eta, \lambda) = \mathrm{E}_\eta(t(x, \lambda) - \eta)$ *denote the bias of* $t(x, \lambda)$, $\lambda \in \mathcal{L}^0$. *Then, (a)* $\mathrm{bias}(\eta, \lambda) = -\mathrm{bias}(-\eta, \lambda)$, *(b) the only unbiased estimator of* η *is* $t(x, \lambda) = x$ *(obtained for* $\lambda \equiv 1$*), and (c) if* $\lambda(x) \neq 1$ *for some* x, *then* $\mathrm{bias}(\eta, \lambda) > 0$ *if* $\eta < 0$, $\mathrm{bias}(\eta, \lambda) = 0$ *if* $\eta = 0$, *and* $\mathrm{bias}(\eta, \lambda) < 0$ *if* $\eta > 0$.

Proof. Let $x = \eta + \varepsilon$, where $\varepsilon \sim \mathrm{N}(0, 1)$, and define $h(\varepsilon, \eta) = \lambda(x)x - \eta$. Then $h(-\varepsilon, \eta) = -h(\varepsilon, -\eta)$ and $\mathrm{bias}(\eta, \lambda) = -\mathrm{bias}(-\eta, \lambda)$. This proves (a). To prove (b) and (c), let $\zeta(x) = (1 - \lambda(x))x$. Since $\lambda(-x) = \lambda(x)$, we have $\zeta(-x) = -\zeta(x)$. Hence, following Huntsberger (1955) and letting $\phi(\cdot)$ denote the standard-normal p.d.f., $\mathrm{bias}(\eta, \lambda) = \int_0^\infty \zeta(x)\phi(x + \eta)(1 - \mathrm{e}^{2\eta x})\,r\,dx$. For $\eta \neq 0$, the integral is zero if and only if $\zeta(x) = 0$ 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 $1 - \mathrm{e}^{2\eta x}$, which completes the proof. □

Theorem A.4 (symmetry of the risk). *For any estimator* $t(x, \lambda)$, $\lambda \in \mathcal{L}^0$, *we have* $R(\eta, \lambda) = R(-\eta, \lambda)$.

Proof. Let $x = \eta + \varepsilon$, where $\varepsilon \sim \mathrm{N}(0, 1)$ and define h as in the proof of Theorem A.3. Then, $R(\eta, \lambda) = \mathrm{E}_\eta h^2(\varepsilon, \eta) = \mathrm{E}_\eta h^2(-\varepsilon, \eta) = \mathrm{E}_\eta h^2(\varepsilon, -\eta) = R(-\eta, \lambda)$. □

We know from Theorem A.1 that the 'usual' estimator $t^{(1)}(x; 0)$ is unbiased. Theorem A.3 shows, inter alia, that it is the only unbiased estimator.

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 (5) that their risk is unbounded unless $c = 0$. Thus we impose the following condition.

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

Theorem A.5 (Brown). *Let* $t(x, \lambda)$, $\lambda \in \mathcal{L}^0$, *be an estimator of* η. *Then* $R(\eta, \lambda)$ *is bounded if and only if R2 holds.*

Proof. To prove that R2 is sufficient, let again $x = \eta + \varepsilon$, $\varepsilon \sim \mathrm{N}(0, 1)$. Then, $R(\eta, \lambda) = \mathrm{E}(\varepsilon - \zeta(\varepsilon + \eta))^2 \leq 2\mathrm{E}(\varepsilon^2 + \zeta^2(\varepsilon + \eta)) \leq 2(1 + K^2)$. Necessity follows from Brown (1971, Theorem 3.3.1). □

Condition R2 requires that $\lambda(x)$ approaches 1 sufficiently fast as $x \to \infty$. In particular, estimators which, for large $|x|$ and $p \geq 1$ behave like $t(x, \lambda) = (1 - c|x|^{-p})x$ have bounded risk.

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

$$t(x, \lambda) = \frac{x^2}{1 + x^2} x \qquad \text{or} \qquad t(x, \lambda) = (1 - \mathrm{e}^{-x^2})x$$

which have attractive properties, but are inadmissible (Strawderman and Cohen (1971)). In order to prove (in)admissibility the following condition is required.

Regularity Condition R3. Let $\zeta(x) = (1 - \lambda(x))x$. Then there exists a measure $G(\eta)$ such that

$$\exp\left(-\int_0^x \zeta(y)\, r\, dy\right) = \int_{-\infty}^{\infty} \exp\left(-\tfrac{1}{2}(x - \eta)^2\right) r\, dG(\eta).$$

This condition implies that $\zeta(x)$ is continuously differentiable. Hence, any estimator which is not differentiable (or worse still, discontinuous) cannot satisfy R3.

Theorem A.6 (admissibility). *Let $t(x, \lambda)$ be an estimator of η and assume that R1 holds. Then, (a) R3 is a necessary condition for $t(x, \lambda)$ to be admissible, and (b) R2 and R3 together are sufficient for $t(x, \lambda)$ to be admissible.*

Proof. Assume that R3 does not hold. Then $t(x, \lambda)$ is not Generalized Bayes (Strawderman and Cohen (1971) and Berger and Srinivasan (1978)) and therefore not admissible (Berger (1985, pp. 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, pp. 552, 553) for further references.) □

Theorem A.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 $R(\eta, \lambda)$. 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 leads to the 'usual' estimator $t(x, \lambda) = x$, see Theorem A.1. On both theoretical and practical grounds, we shall adopt the much more appealing 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

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

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

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

In order to implement the minimax regret approach we require, for each η, the lower bound of the risk $R(\eta, \lambda)$ over all estimators $t(x, \lambda)$, $\lambda \in \mathcal{L}^0$. In (6) we showed that the class of estimators $t(x, \lambda_c^{(1)})$ has lower bound $\eta^2/(1 + \eta^2)$. This, in fact, is the lower bound in $\mathcal{L}^0$ as well.

Theorem A.7. *The lower bound of the risk in $\mathcal{L}^0$ is given by*

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

Proof. Using the symmetry condition $\lambda(-x) = \lambda(x)$, we have

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

where $\phi(\cdot)$ denotes the standard-normal p.d.f., and hence

$$\begin{aligned} R(\eta, \lambda) &= \int_{-\infty}^{\infty} (x\lambda(x) - \eta)^2 \phi(x - \eta)\, r\, dx \\ &= \int_{0}^{\infty} ((x\lambda(x) - \eta)^2 \phi(x - \eta) + (x\lambda(x) + \eta)^2 \phi(x + \eta))\, r\, dx. \end{aligned}$$

Now, let

$$\omega(x) = \mathrm{e}^{-2x} \qquad \text{and} \qquad h(x) = \frac{1 - \omega(x)}{x(1 + \omega(x))}.$$

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

$$\begin{aligned} R(\eta, \lambda) = \int_{0}^{\infty} & x^2 (1 + \omega(\eta x))(\lambda(x) - \eta^2 h(\eta x))^2 \phi(x - \eta)\, r\, dx \\ & + 4\eta^2 \int_{0}^{\infty} \frac{\omega(\eta x)}{1 + \omega(\eta x)} \phi(x - \eta)\, r\, dx. \end{aligned}$$

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 non-decreasing on $(0, \infty)$, there exists a unique x_0 such that $x_0 = 0$ if $\lambda(0) \geq \eta^2$ and $\lambda(x_0) = \eta^2 h(\eta x_0)$ if $\lambda(0) < \eta^2$. This can easily be seen by considering graphs of the functions $\lambda(x)$ and $\eta^2 h(\eta x)$ for both cases. (If $\lambda(x)$ is not continuous, the second condition is replaced by $\lambda(x_0 - \epsilon) \leq \eta^2 h(\eta x_0) \leq \lambda(x_0 + \epsilon)$ for all $\epsilon > 0$ sufficiently small.) With x_0 so defined we have $|\lambda(x) - \eta^2 h(\eta x)| \geq |\lambda(x_0) - \eta^2 h(\eta x)|$ 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 $R(\eta, \lambda) \geq R(\eta, \lambda_0)$. We also know from (6) that $R(\eta, \lambda_0) \geq \eta^2/(1 + \eta^2)$. This completes the proof. □