Post-model Selection Inference and Model Averaging

Georges Nguefack-Tsague
Biostatistics Unit, Department of Public Health
University of Yaounde I, P. O. Box 1364
Yaoundé, Cameroon
gnguefack@yahoo.fr

Walter Zucchini
Institute for Statistics and Econometrics
Georg-August-Universität
Platz der Göttinger Sieben 5
37073 Göttingen, Germany
walter.zucchini@wi-wiss.uni-goettingen.de

Abstract

Although model selection is routinely used in practice, little is known about its precise effects on any subsequent inference that is carried out. The same goes for the effects induced by the closely related technique of model averaging. This paper is concerned with the use of the same data first to select a model and then to carry out inference, in particular point estimation and point prediction. The properties of the resulting estimator, called a *post-model-selection estimator* (PMSE), are difficult to derive. Using selection criteria such as hypothesis testing, AIC, BIC, HQ and C_P, we illustrate that, in terms of risk function, no single PMSE dominates the others. The same conclusion holds more generally for any penalized likelihood information criterion. We also compare various model averaging schemes and show that no single one dominates the others in terms of risk function. Since PMSEs can be regarded as a special case of model averaging, with 0-1 random weights, we propose a connection between the two theories, in the frequentist approach, by taking account of the selection procedure when performing model averaging. We illustrate the point by simulating a simple linear regression model.

Keywords: Model averaging, Model selection, Inference after model selection, Post-selection.

1. Introduction

In most statistical modeling applications, several models are plausible a priori, and so some model selection procedure is applied to choose the (single) model that will be used in the subsequent analysis, e.g. to estimate quantity or quantities of interest. Overviews, explanations, discussions and examples of model selection procedures can be found in the books by Linhart and Zucchini (1986), McQuarrie and Tsai (1998), Burnham and Anderson (2002) and Claeskens and Hjort (2008).

An alternative to selecting a single model for estimation purposes is to use a weighted average of the estimates resulting from each of the models under consideration. This leads to the class of model averaging estimators. Several options are available for specifying the weights; e.g. they can be based on the Akaike's information criterion, AIC (Akaike, 1973) or, in the Bayesian paradigm, on the Bayesian information criterion, BIC (Schwarz, 1978). It is not the *construction* of the estimator that causes difficulties; the problem is to determine its *properties*.

The same problem arises for estimators obtained after model selection. This is not surprising, since model selection corresponds to the special case of model averaging in which the weight one is assigned to the selected model, and the weight zero to all other models. We refer to these estimators as *post-model selection estimators* (PMSE, Leeb and Pötscher, 2005). If both model selection and estimation are based on the same set of data, then ignoring this fact in the subsequent analysis leads to invalid inferences. Literature on this issue includes, *inter alia*, Bancroft (1944) for pretest estimators, Breiman (1992), Hjorth (1994), Chatfield (1995), Draper (1995), Buckland et al. (1997), Zucchini (2000), Candolo et al. (2003), Hjort and Claeskens (2003), Efron (2004), Leeb and Pötscher(2005), Longford (2005), Nguefack-Tsague and Zucchini (2005), Claeskens and Hjort (2008) and Zucchini et al. (2011). A bibliography on the topic is provided in Nguefack-Tsague (2006).

Section 2 gives brief accounts of model averaging estimators and PMSEs. In Section 3 we propose a new approach for computing the weights for the competing models, one that takes into account the selection probability of each model. Given a selection procedure, we suggest to weight the likelihood of each model by the estimated probability that the model is selected. Section 4 illustrates the point that, even in a very simple example (simple linear regression) no single PMSE is best over the entire spectrum of possibilities, where ``best" here is defined as the mean squared error (risk). Nevertheless, on the whole, the proposed averaging scheme compares favourably with the more established estimators. Following a brief summary (Section 5) the appendix gives proofs of some of the statements in the paper.

2. Model averaging and post-model selection estimators

2.1 Model averaging estimators

Let $M=\{M_1,\ldots,M_K\}$ be a set of K plausible models to estimate μ, the quantity of interest. Denote by $\hat{\mu}_k$ the estimator of μ obtained when using model M_k. Model averaging involves finding non-negative weights, $w_1,\ldots,w_K$, that sum to one, and then estimating μ by

$$\hat{\mu}=\sum_{k=1}^{K} w_k\,\hat{\mu}_k\,. \tag{1}$$

Some classical model averaging weights base the weights on penalized likelihood values. Let IC_k denote an `information criterion' of the form

$$IC_k=-2\log L_k+s_k, \tag{2}$$

where s_k is a penalty term, and L_k is the maximized likelihood value for the model M_k. The Akaike information criterion (AIC) (Akaike, 1973) is the special case with $s_k = 2q_k$, where q_k is the number of parameters of model M_k. Buckland et al. (1997) proposed using weights of the form:

$$w_k = \frac{\exp(-s_k/2)L_k}{\sum_{l=1}^{K}\exp(-s_l/2)L_l} = \frac{\exp(-IC_k/2)}{\sum_{l=1}^{K}\exp(-IC_l/2)}. \tag{3}$$

“Akaike weights" (denoted by $w_{aic,k}$) refer to the case with $IC_k = AIC_k$. Numerous applications of Akaike weights are given in Burnham and Anderson (2002). Candolo et al. (2003) applied them to a linear regression example.

The analogous Bayesian model averaging is described in Hoeting (1999) and Wasserman (2000). In the context of regression and classification, LeBlanc and Tibshirani (1996) proposed using no penalty term, that is setting $w_k = L_k / \sum_{l=1}^{K} L_l$. Hjort and Claeskens (2003) proposed the smooth focused information criterion (FIC) and other model averaging schemes to study model averaged, or compromise estimators, together with their limiting distributions and risk properties. Model averaging in semi-parametric regression with AIC-based or BIC-based weights was studied by Claeskens and Carroll (2007).

We note that, if one were able to find closed-form expressions for the model selection probabilities $\Pr(M_k \textit{ is selected})$ for each model, then an obvious weighting scheme would be to use an estimator of these probabilities, but this is not often recommended, as we show below.

2.2 Post-model selection estimators

A post-model selection estimator (PMSE) can be regarded as a special case of a model averaging estimator in which one of the weights is equal to one and the remaining $K-1$ weights are equal to zero. The model selection criterion determines which model is assigned the weight one and hence used to estimate

μ. The index of the selected model, $\hat{k}$, is a random variable. We denote the selected model by $M_{\hat{k}}$, and the PMSE of the quantity of interest by $\hat{\mu}_{\hat{k}}$. Let $I(.)$ denote the indicator function that has the value 1 if the argument is true, and 0 if it is false. Then

$$M_{\hat{k}} = \sum_{k=1}^{K} I(\textit{model } k \textit{ is selected}) M_k, \qquad \hat{\mu}_{\hat{k}} = \sum_{k=1}^{K} I(\textit{model } k \textit{ is selected}) \hat{\mu}_k.$$

Clearly, the properties of $\hat{\mu}_{\hat{k}}$ depend on (among other things) the set of candidate models, M, and on the selection procedure, which we denote by S.

3. Combining model averaging and model selection

As pointed out, PMSEs are special cases of model averaging estimators whose properties depend on the selection procedure S. We therefore suggest estimating μ by a weighted average of the $\hat{\mu}_k$ in which the weights take account of S, specifically where they depend on estimators $p(M_k|S) = \hat{\Pr}(M_k \ is\ selected|S),\ k = 1,\ldots K$. This suggestion is new because classical model averaging does not take the selection probabilities into account. We propose the following *adjusted likelihood weights*

$$w_{al,k} = \frac{p(M_k \mid S)L_k}{\sum_{i=1}^{K} p(M_i \mid S)L_i}, \quad k = 1,2,\ldots,K. \qquad (4)$$

The likelihoods are taken into account because they quantify the relative plausibility of the data under each competing model; the estimated selection probability $p(M_k \mid S)$ adjusts the weights for the selection procedure. Both of these components are required. If one were to use only the likelihoods to determine the weights then complex models (i.e. models having many parameters) would automatically be assigned larger weights. The weights $w_{al,k}$ are similar to the weights w_k defined in (3) but they differ in the way the likelihood is adjusted. With the proposed method a ``bad" model will be penalized by any reasonable selection procedure through the probability $p(M_k \mid S)$, even if it is complex in terms of the number of parameters. We let the selection procedure determine in how far a model is penalized.

The problem that needs to be solved is that of constructing estimators, $p(M_k|S)$, of the model selection probabilities. Hjort and Claeskens (2003) showed that a naive bootstrap estimator of the selection probability of model M_k (namely the proportion of resamples in which M_k is selected) does not work. If the selection probabilities depend on some parameter for which a closed form expression exists, and if one can find an estimator of the parameter, then it is possible to obtain estimators for these probabilities. For the case where there is no close form, Miller (2002) suggested using a Monte Carlo method based on projection.

For finite samples one can use the variance formulae proposed in Buckland et al. (1997); the first when estimates are perfectly correlated and the second when they are independent:

$$\begin{aligned} Var(\hat{\mu}) &= \{\sum_{k=1}^{K} w_{al,k}\sqrt{Var(\hat{\mu}_k) + [\hat{\mu}_k - \hat{\mu}]^2}\}^2 \quad && if\ estimators\ are\ perfectly\ correlated, \\ Var(\hat{\mu}) &= \sum_{k=1}^{K} w_{al,k}^2\{Var(\hat{\mu}_k) + [\hat{\mu}_k - \hat{\mu}]^2\} \quad && if\ estimators\ are\ independent, \end{aligned} \qquad (5)$$

where $\hat{\mu}$ is the weighted estimator and $w_{al,k}$ the weight for model M_k. Although neither of these two extreme scenarios are likely to occur in practice, these expressions provide an indication of the range of values in which the variance falls. For large samples, one can use the limiting risk properties and limiting distributions of general model weights as given in Hjort and Claeskens (2003).

4. Illustration with simple linear regression

4.1 The set-up

Consider the familiar simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,\ldots,n, \tag{6}$$

where the $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$. For simplicity we consider the case in which σ is known, but similar results are obtained when σ is unknown. Without loss of generality we assume that $\bar{x} = 0$, since the model can be reparameterized as $Y_i = \lambda_0 + \lambda_1 z_i + \varepsilon_i$, where $z_i = x_i - \bar{x}$. The OLS estimators of the parameters of this model are given by

$$\hat{\beta}_0 = \bar{y} \text{ and } \hat{\beta}_1 = \sum\nolimits_{i=1}^n x_i (y_i - \bar{y}) / \sum\nolimits_{i=1}^n x_i^2 .$$

As is well-known, these estimators are normally distributed, unbiased, such that

$$v_0 = \text{Var}(\hat{\beta}_0) = \sigma^2 / n , \quad v_1 = \text{Var}(\hat{\beta}_1) = \sigma^2 / \sum\nolimits_{i=1}^n x_i^2 , \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0 ,$$

and therefore independently distributed.

We consider the problem of estimating $\mu = E(Y \mid x_+)$ for a given (e.g. future) value, x_+, of the covariate, but (as often occurs in practice) suppose that we are uncertain whether or not it is advantageous to take the covariate into account when doing so. In other words, estimation can be based on either of the following two models:

$$\textit{Under model } M_1 : \mu = \beta_0, \qquad \textit{under model } M_2 : \mu = \beta_0 + \beta_1 x_+ . \tag{7}$$

Suppose now that we first apply a model selection procedure to determine which model to use and then estimate μ using the selected model. Our objective in the next section is to derive the properties of the resulting estimator.

4.2 Post-model selection estimators

We denote the standardized intercept and standardized slope by $b_0 = \beta_0 / \sqrt{v_0}$ and $b_1 = \beta_1 / \sqrt{v_1}$, respectively. It follows that the OLS estimators of β_0 and β_1 can be represented as

$$\hat{\beta}_0 = \sqrt{v_0}(Z_0 + b_0) \text{ and } \hat{\beta}_1 = \sqrt{v_1}(Z_1 + b_1) ,$$

where $Z_0, Z_1 \overset{iid}{\sim} N(0,1)$. For the two models in (7), M_1 has $q_1 = 1$ and M_2 has $q_2 = 2$ parameters. Let IC_1 and IC_2 be information criteria of the form (2) for M_1 and M_2, respectively. The following propositions are proved in the appendix.

Proposition 1: For a criterion of the form (2) with $s_k = hq_k$, $h > 0$,

$$IC_2 - IC_1 = -(Z_1 + b_1)^2 + h. \qquad (8)$$

In terms of the criterion model M_2 is chosen if $| Z_1 + b_1 | \geq \sqrt{h}$; otherwise M_1 is chosen. Denoting the PMSE of μ by $\hat{\mu}_{\hat{k},h}$ we have that

$$\begin{aligned}\hat{\mu}_{\hat{k},h} &= \hat{\beta}_0 I(| Z_1 + b_1 | < \sqrt{h}) + (\hat{\beta}_0 + \hat{\beta}_1 x_+) I(| Z_1 + b_1 | \geq \sqrt{h}) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_+ I(| Z_1 + b_1 | \geq \sqrt{h}) \\ &= \hat{\beta}_0 + x_+ \tilde{\beta}_{1h},\ say. \qquad (9)\end{aligned}$$

Thus the behaviour of $\hat{\mu}_{\hat{k},h}$ is largely determined by that of $\tilde{\beta}_{1h}$.

Proposition 2: The expectation, bias and variance of the PMSE in (9) are given by

$$\begin{aligned} E(\hat{\mu}_{\hat{k},h}) &= \beta_0 + x_+ \sqrt{v_1}\left(b_1 d + e\right), \\ Bias(\hat{\mu}_{\hat{k},h}) &= x_+ \sqrt{v_1}\left\{b_1(d-1) + e\right\}, \qquad (10) \\ Var(\hat{\mu}_{\hat{k},h}) &= \sigma^2 / n + x_+^2 v_1 \{b_1^2 d(1-d) + 2b_1 e(1-d) + d - e^2 + f\}, \end{aligned}$$

where $d = \Phi(r) + 1 - \Phi(q)$, $e = \phi(q) - \phi(r)$, $f = q\phi(q) - r\phi(r)$, $r = -\sqrt{h} - b_1$, $q = \sqrt{h} - b_1$; ϕ is the density function of the standard normal and Φ the distribution function.

Note that these moments of the PMSEs do not depend on the values of the dependent variable, and they depend on the covariate x *only* through $\sum_{i=1}^n x_i^2$ via v_1=Var$(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^n x_i^2$. (This explains why the properties of PMSEs in this example are not sensitive to other aspects of the data set.) In multivariate regression, the important quantity is $(X'X)^{-1}$. We used simulated data to investigate the properties of different PMSEs, namely 10^6 samples of size $n = 20$, with $\sigma^2 = 1$, $x_i \overset{iid}{\sim} U(1,10)$, $i = 1,2,\ldots,n$, $\beta_0 = 0$, $x_+ = 6.5$. The reported results are not sensitive to the choice of these selected values; in particular, increasing the sample size, has minimal impact on the results. All expectations here were taken with respect to the full model, M_2. As the risk functions are symmetric around zero we only display the graphs for $b_1 > 0$. All computations are performed with the software R (R Development Core Team, 2010).

Suppose we apply a test of hypothesis to select one of the two models in (7), i.e. we test the null hypothesis, H$_0$: $\beta_1 = 0$ (i.e. that model M_1 holds) at the level α and select model M_1 if the hypothesis can't be rejected; otherwise we select M_2. (Unlike the literature, instead of using the term "*pretest*" for the estimator obtained after a test of hypothesis, we prefer "*post-testing*" to be consistent with the PMSE terminology.) In particular M_2 is chosen if $|\hat{\beta}_1|/\sqrt{v_1} \geq z_{1-\frac{\alpha}{2}}$, i.e if $|Z_1 + b_1| \geq z_{1-\frac{\alpha}{2}}$ where $z_{1-\frac{\alpha}{2}}$ is the quantile of the N(0,1). Thus hypothesis testing can be regarded as an information criterion of the form (2) with $h = z^2_{1-\frac{\alpha}{2}}$. (Equivalently the significant level is $\alpha = 2[1-\Phi(\sqrt{h})]$, so that a test of hypothesis with $\alpha = 0.16$ is equivalent to the AIC.)

The values of h for some classical selection procedures are given by

$$h = \begin{cases} z^2_{1-\frac{\alpha}{2}} & \textit{for hypothesis testing} \\ 2 & \textit{for AIC (Akaike information criterion)} \\ \ln n & \textit{for BIC (Bayesian information criterion)} \\ \ln(\ln n) & \textit{for HQ (Hannan and Quinn, 1979).} \end{cases}$$

The Mallows Cp (Mallows, 1973) is an example of a selection criterion that is not of the form of (2). As is shown in the appendix, and for the two models defined in (7), this criterion leads to model M_2 being selected if $F(1, n-2, b_1^2) > 2$, where $F(1, n-2, b_1^2)$ is the non-central F distribution with 1 and n-2 degrees of freedom and non-central parameter b_1^2.

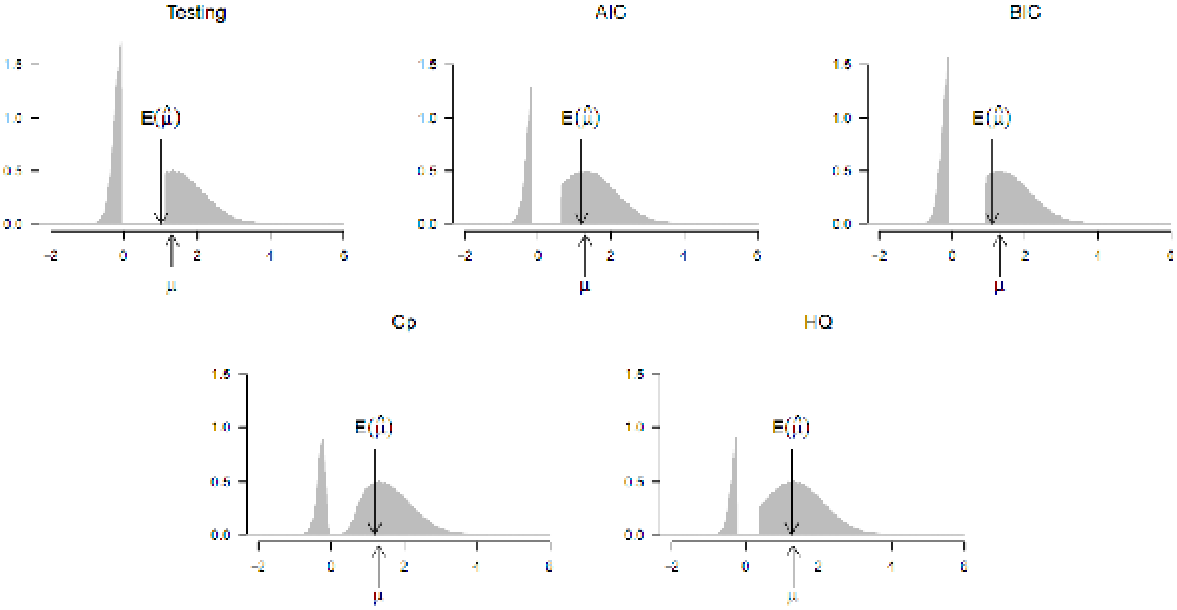

Figure 1: Densities of five PMSEs for $\mu = E(y \mid x = 6.5)$ following model selection using a test of hypothesis ($\alpha = 5\%$), AIC, BIC, CP and HQ when $b_1 = 0.2$. The expected value of each PMSE is shown.

Figure 1 displays the densities of post-testing (with $\alpha = 0.05$), post-AIC, post-BIC, post-C$_P$ and post-HQ estimators of μ. The densities are all bimodal, reflecting the mixture nature of the PMSEs. Note also the gap in their support. None of them are even approximately normally distributed. The figure also displays μ and $E(\hat{\mu})$, from which one can ascertain the bias.

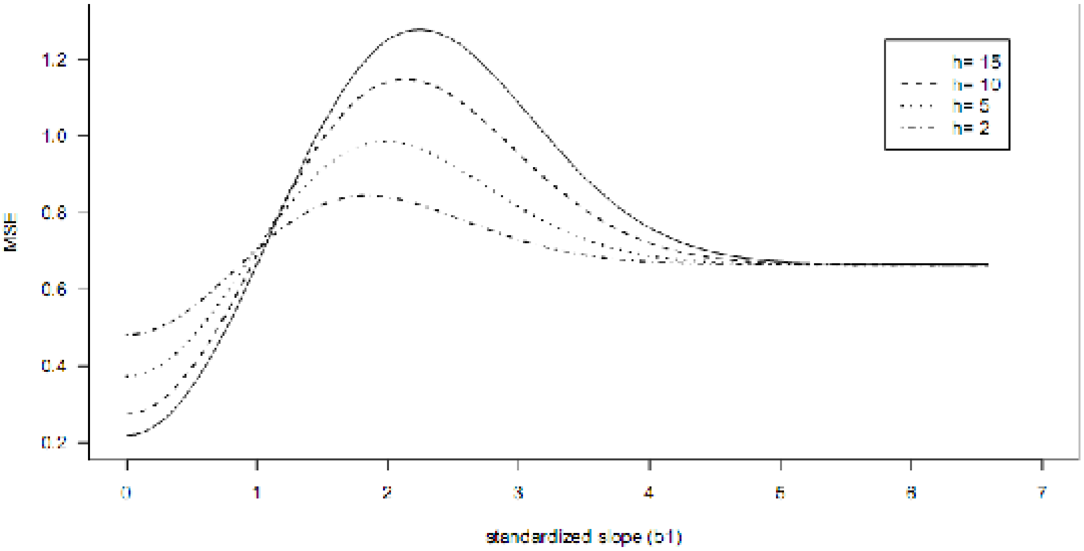

Figure 2: Risk function of the PMSE as a function of the standardized slope for different levels, of the penalty factor h.

Figure 2 displays the MSE (risk) of PMSEs, for selection with criteria of the type (2), as a function of b_1 (or, strictly speaking, as a function of $|b_1|$) for different values of the penalty factor, h. For small values of b_1 the risk is reduced by increasing h, but the opposite is true for large values of b_1. All the lines in the figure cross; no single value of the penalty factor h is optimal over the entire range of b_1.

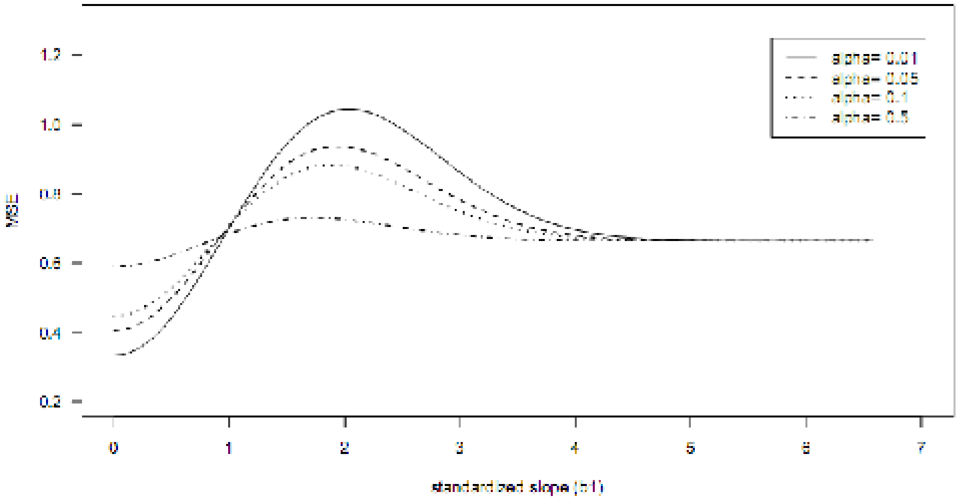

Figure 3: Risk function of the post-testing estimator as a function of the standardized slope for different significance levels, α.

Figure 3 displays the risk of the PMSE as a function of b_1 when selection is based on a test of hypothesis using different values of the significance level, α. For small values of b_1 the risk is reduced by using small values of α, but the opposite holds for large values of b_1. Here too all the lines in the figure cross and no single value of α is optimal over the entire range of b_1. This fact is not surprising, since the hypothesis testing is equivalent to an information criterion of the form (2) with a penalty factor $h = z^2_{1-\frac{\alpha}{2}}$.

The PMSE for the Mallows C P is given by

$$\hat{\mu}_{\hat{k},CP} = \sqrt{v_0}(Z_0 + b_0) + x_+\sqrt{v_1}(Z_1 + b_1)I(F(1, n-2, b_1^2) > 2). \tag{11}$$

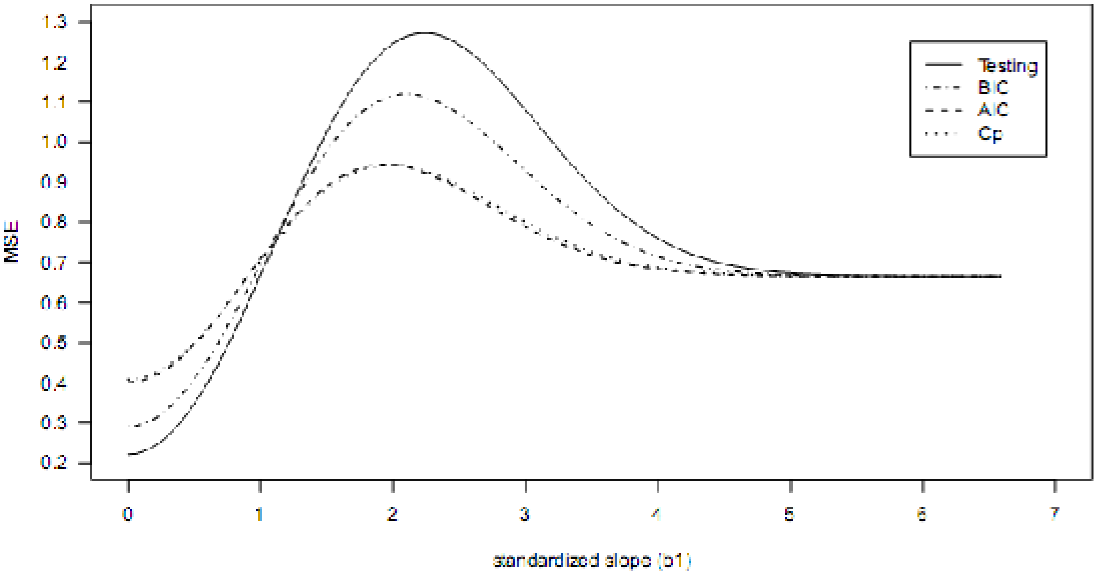

Figure 4: The risk of the post-C P, post-BIC, post-AIC and post-testing estimators as functions of the standardized slope b_1.

Figure 4 compares the risk for PMSEs based on 4 standard model selection criteria. The post-testing leads to the smallest risk for small values of b_1, followed by the BIC, AIC, and finally C P. The reverse is true when b_1 is large. No selection criterion dominates the others; none of them is best over the entire range of the slope parameter.

4.3 Model averaging

Consider a model averaging estimator with weights $w_1 = w$ and $w_2 = 1-w$ for the models M_1 and M_2 in (7):

$$\hat{\mu} = w\hat{\mu}_1 + (1-w)\hat{\mu}_2 = w\hat{\beta}_0 + (1-w)(\hat{\beta}_0 + x_+\hat{\beta}_1) = \hat{\beta}_0 + (1-w)\hat{\beta}_1 x_+. \tag{12}$$

Using an information criterion of the form (2) with $s_k = hq_k$, M_2 is selected if $|Z_1 + b_1| \geq \sqrt{h}$. Thus the probability of selecting model M_2 is given by $\Pr(M_2 \mid S) = \Phi(Z_1 \leq -\sqrt{h} - b_1) + 1 - \Phi(Z_1 \geq \sqrt{h} - b_1)$, which can be estimated using $p_2 = \Phi(\sqrt{h} - \hat{b}_1) + 1 - \Phi(-\sqrt{h} - \hat{b}_1) = \Phi(\sqrt{h} - (Z_1 + b_1)) + 1 - \Phi(-\sqrt{h} - (Z_1 + b_1)).$

From (3) the special case $h = 2$ leads to the Akaike weight, proposed by Buckland et al. (1997):

$$w_{aic,2} = \frac{e^{(\mathrm{AIC}_1 - \mathrm{AIC}_2)}}{1 + e^{(\mathrm{AIC}_1 - \mathrm{AIC}_2)}} = \frac{e^{\frac{1}{2}(Z_1 + b_1)^2 - 1}}{1 + e^{\frac{1}{2}(Z_1 + b_1)^2 - 1}}. \qquad (13)$$

The analogous weight for the general case of a criterion of the form (2) with $s_k = hq_k$ is

$$w_{\mathrm{h},2} = \frac{e^{\frac{1}{2}(Z_1 + b_1)^2 - \frac{h}{2}}}{1 + e^{\frac{1}{2}(Z_1 + b_1)^2 - \frac{h}{2}}}, \qquad (14)$$

and for the non-penalized information criterion

$$w_2 = \frac{e^{\frac{1}{2}(Z_1 + b_1)^2}}{1 + e^{\frac{1}{2}(Z_1 + b_1)^2}}. \qquad (15)$$

Thus from (4), the adjusted likelihood weight for model M_2 is

$$w_{al,2} = \frac{\delta_2 e^{\frac{1}{2}(Z_1 + b_1)^2}}{1 + \delta_2 e^{\frac{1}{2}(Z_1 + b_1)^2}}, \qquad (16)$$

where $\delta_2 = p_2/(1 - p_2)$. The proposal to base the (model averaging) weight of a model on its likelihood adjusted by its estimated model selection probability is, of course, not restricted to criteria of the form (2).

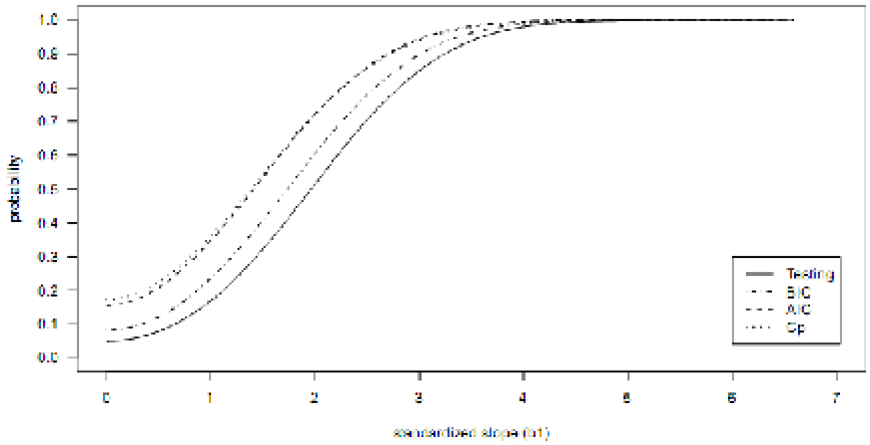

Figure 5: Model selection probabilities as function of b_1

Figure 5 shows the probability of selecting model, M_2, as a function of b_1, using four different criteria. As expected the curves behave similarly: as b_1 increases so $\Pr(M_2 \mid S)$ also increases and approaches one. However, the curves are not identical, and so the properties of the corresponding PMSEs are also different. It is this dependence on the selection criterion that the adjusted likelihood weights are designed to take into account.

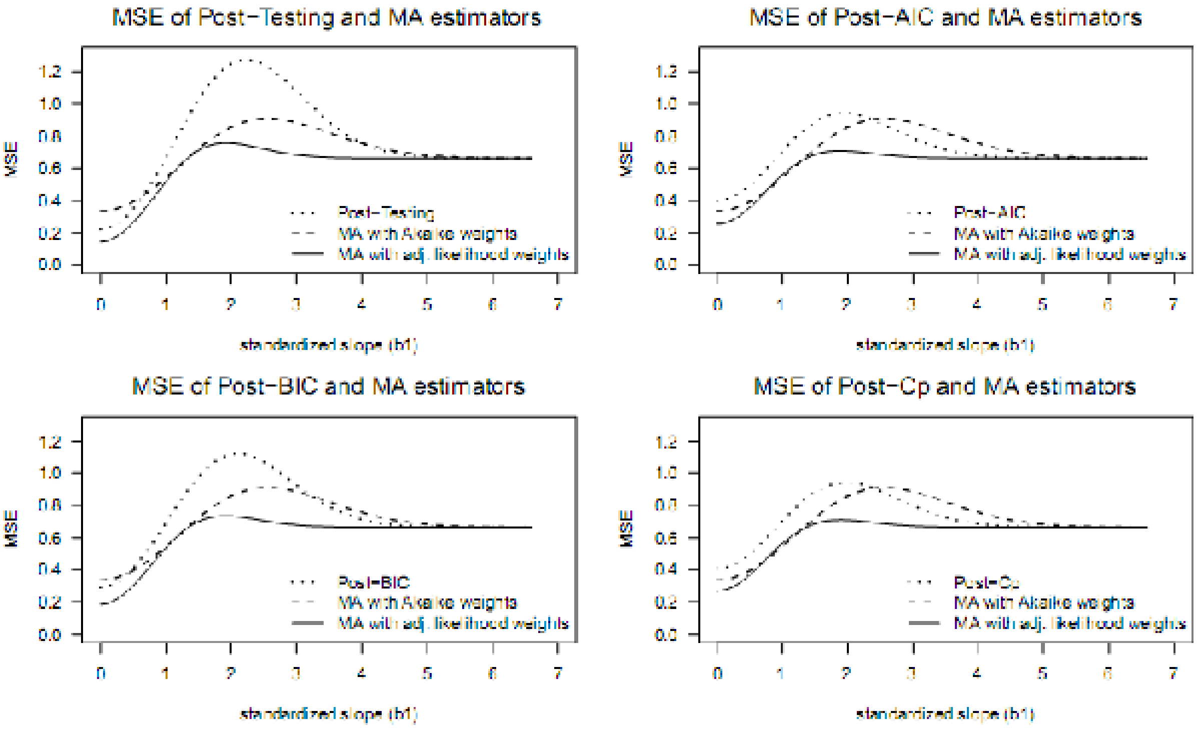

Figure 6: Comparison of risks of PMSEs (broken line) with adjusted weights (solid line) and Akaike weights (dashed line) as function of b_1

The fact that no single PMSE uniformly dominates the others in terms of risk was illustrated in Figure 4. The top right-hand panel of Figure 6 compares the risk of the post-AIC estimator with the risk of two model averaging estimators, one based on Akaike weights and the other on adjusted likelihood weights. The other three panels display similar comparisons for the hypothesis testing, BIC and C $_P$ selection criteria. For each value of b_1 the model averaging estimator based on adjusted likelihood weights either has the smallest risk, or has a risk that is not much larger than the smallest risk attained by the other estimators. Furthermore it is substantially better than the alternatives considered over quite a large interval of b_1 values. This illustrates that, on the whole (at least in this example) the estimator based on the adjusted likelihood weights outperforms that based on Akaike weights, post-testing, post-AIC, post-BIC and post-C$_P$.

5. Concluding Remarks

We have proposed a method of assigning weights for model averaging in a frequentist context. The key idea is to incorporate the estimated probability of selecting each model into the weights. We have illustrated the advantage of the method in the context of a simple regression analysis. We showed that, on the whole, the proposed method is superior (in terms of risk function) to the popular model averaging method based on Akaike weights. In principle the same idea is more generally applicable. However, several problems need to be solved before the proposed method can be applied in typical (more complex) settings. In many applications, e.g. when selecting a model for a density function from a set of candidate models, the competing models are not nested, which they were in our example. It is more difficult to estimate the model selection probabilities if the models are not nested. In general it is not easy to construct estimators for such probabilities if no closed form expression for them is available. As indicated in the text, Monte Carlo methods offer an alternative approach in such cases.

The fact that the proposed method performs very well in the example investigated here suggests that the method is promising and worth further investigation. The main challenges will be to construct estimators for the selection probabilities in more realistic settings, and to investigate the theoretical properties of the resulting model averaging estimators.

6. Appendix

Proof of Proposition 1

From (2) with $s_k = hq_k$ we have that

$$IC_2 - IC_1 = -2(\ln L_2 - \ln L_1) + (s_2 - s_1) \text{ where } s_2 - s_1 = hq_2 - hq_1 = 2h - h = h.$$

Under normality

$$\ln L_2 = -(n/2)\ln 2\pi - (n/2)\ln \sigma^2 - (1/2\sigma^2)\sum_{i=1}^{n}(y_i - \overline{y} - \hat{\beta}_1 x_i)^2,$$
$$\ln L_1 = -(n/2)\ln 2\pi - (n/2)\ln \sigma^2 - (1/2\sigma^2)\sum_{i=1}^{n}(y_i - \overline{y})^2, \text{ and so}$$
$$2\sigma^2(\ln L_2 - \ln L_1) = \sum_{i=1}^{n}\left((y_i - \overline{y})^2 - (y_i - \overline{y} - \hat{\beta}_1 x_i)^2\right) = \sum_{i=1}^{n}\left(2\hat{\beta}_1 x_i (y_i - \overline{y}) - \hat{\beta}_1^2 x_i^2\right).$$

Since $\hat{\beta}_1 = \sum_{i=1}^{n} x_i (y_i - \overline{y}) / \sum_{i=1}^{n} x_i^2$ and $v_1 = \sigma^2 / \sum_{i=1}^{n} x_i^2$ it follows that

$\ln L_2 - \ln L_1 = \hat{\beta}_1^2 / (2v_1)$ and hence that $IC_2 - IC_1 = -\hat{\beta}_1^2 / v1 + h = -(Z_1 + b_1)^2 + h$.

Proof of Proposition 2

From (9), the moments of $\hat{\mu}_{\hat{k},h}$ are given by

$$\begin{aligned}
E(\hat{\mu}_{\hat{k},h}) &= E(\hat{\beta}_0) + x_+ E(\tilde{\beta}_{1h}) &&= \beta_0 + x_+ E(\tilde{\beta}_{1h}),\\
Bias(\hat{\mu}_{\hat{k},h}) &= E(\hat{\mu}_{\hat{k},h}) - \beta_0 - b_1\sqrt{v_0}\,x_+ &&= x_+(E(\tilde{\beta}_{1h}) - \beta_1),\\
Var(\hat{\mu}_{\hat{k},h}) &= Var(\hat{\beta}_0) + x_+^2 Var(\tilde{\beta}_{1h}) &&= \sigma^2/n + x_+^2 Var(\tilde{\beta}_{1h}).
\end{aligned}\tag{17}$$

The PMSE of $\tilde{\beta}_{1h}$ can be written as $\tilde{\beta}_{1h} = \sqrt{v_1}(Z_1 + b_1)I(|Z_1 + b_1| \geq \sqrt{h}) = \sqrt{v_1}A_h\ (say)$. If $|Z_0 + b_1| \geq h^{1/2}$ then $Z_0 \geq q$ or $Z_0 < r$, and so

$$\begin{aligned}
E(A_h) &= \int_{-\infty}^{r}(b_1 + z)\phi(z)dz + \int_{q}^{\infty}(b_1 + z)\phi(z)dz\\
&= b_1(\Phi(r) + 1 - \Phi(q)) + \phi(q) - \phi(r)\\
&= b_1 d + e\ (say),\\
E(A_h^2) &= \int_{-\infty}^{r}(b_1 + z)^2\phi(z)dz + \int_{q}^{\infty}(b_1 + z)^2\phi(z)dz\\
&= (b_1^2 + 1)(\Phi(r) + 1 - \Phi(q)) + 2b_1(\phi(q) - \phi(r)) + q\phi(q) - r\phi(r)\\
&= (b_1^2 + 1)d + 2b_1 e + f(say),\\
Var(A_h) &= E(A_h^2) - E^2(A_h) = b_1 d(1-d) + 2b_1(1-d)e + d - e^2 + f.
\end{aligned}$$

Substituting $\mathrm{E}(\tilde{\beta}_{1h}) = \sqrt{v_1}E(A_h)$ and $\mathrm{Var}(\tilde{\beta}_{1h}) = v_1\mathrm{Var}(A_h)$ in (17) yields the required results.

Proof relating to Mallows C P criterion

$C\,P_k = \dfrac{RSS_k}{RSS_2/(n-2)} - n + 2q_k$, where RSS_k is the residual sum of squares for model M_k (k=1,2), $q_1 = 1$ and $q_2 = 2$. Hence $C\,P_1 - C\,P_2 = \dfrac{RSS_1 - RSS_2}{RSS_2/(n-2)} - 2 = F - 2$ (say).

It is well-known (see, e.g., Section 7.2 in Linhart and Zucchini, 1986) that, under the normality assumption, the numerator and denominator of F are independently distributed, and that $RSS_2 = \sigma^2\chi_{n-2}^2$. The numerator, given by $RSS_1 - RSS_2 = \dfrac{\hat{\beta}_1^2\sigma^2}{v_1} = (b_1 + Z_1)^2\sigma^2$, is distributed as $\sigma^2\chi_1^2(b_1^2)$, where b_1^2 is the non-centrality parameter. It follows that F has a non-central Fisher distribution, namely $F(1, n-2, b_1^2)$. Model M_1 is chosen if $C\,P_1 < C\,P_2$, i.e. if $F < 2$.

7. Acknowledgments

We thank participants of the International Workshop on *New trouble for standard regression analysis* held in Regensburg (Germany), November 4-5 2005 for their helpful comments and suggestions; especially to Benedikt M. Pötscher and Hannes Leeb whose comments have led to an improvement on earlier version.

Thanks also to Gerda Claeskens and the two anonymous reviewers for their constructive comments and suggestions.

References

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, eds. B. Petrov and F. Csa'ki, Budapest: Akade'miai Kiado', 267-281.
2. Bancroft, T.A. (1944). On bias in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics,* 15: 190-204.
3. Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-Fixed predictor error. *Journal of the American Statistical Association,* 87: 738-754.
4. Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics,* 53: 603-618.
5. Burnham, P. K. and Anderson, D. R. (2002). *Model selection and multimodel inference, a practical information-theoretic approach*. 2nd Edition. Springer-Verlag: New York.
6. Candolo, C., Davison, A. C. and Deme'trio, C. G. B. (2003). A note on model uncertainty in linear regression. *The Statistician,* 158: 165-177.
7. Chatfied, C. (1995). Model Uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society,* series B 158: 419-466.
8. Claeskens, G. and Carroll, R.J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika,* 94: 249-265.
9. Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association,* 98: 900-916.
10. Claeskens, G. and Hjort, N. L. (2008). Model selection and model averaging, Cambridge University Press: Cambridge.
11. Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society,* series B 57: 45-97.
12. Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association,* 99: 619-642.
13. Hannan, E. J. and Quin, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society,* series B 41: 190-195.
14. Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association,* 98: 879-899.
15. Hjorth, J. (1994). *Computer intensive statistical methods:Validation, model selection, and bootstrap*, Chapman and Hall: London.

16. Hoeting J., Madigan D., Raftery A. and Volinsky C. (1999). Bayesian model averaging: A tutorial. *Statistical Science,* 4: 382-417.

17. Leblanc, M. and Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association,* 91: 1641-1650.

18. Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Fact and fiction. *Econometric Theory,* 21: 21-59.

19. Linhart, H. and Zucchini, W. (1986). *Model selection*. John Wiley and Sons: New York.

20. Longford, N. T. (2005). Editorial: Model selection and efficiency-is 'which model ...?' the right question? *J. R. Statist. Soc. A,* 168, Part 3: 469-472.

21. Mallows, C. L. (1973). Some comments on Cp. *Technometrics,* 15: 661-675.

22. McQuarrie, A. D. R. and Tsai, C.L. (1998) *Regression and time series model selection*, World Scientific: Singapore.

23. Miller, A. J. (2002). *Subset selection in regression*, 2nd Edition, Chapman and Hall: London.

24. Nguefack-Tsague, G. and Zucchini, W. (2005). *Inference after model selection in linear regression*. Internal Report, Institut für Statistik und Ökonometrie, Universität Göttingen.

25. Nguefack-Tsague, G. (2006). *Estimating and correcting the effects of model selection uncertainty*, Cuvillier Verlag: Göttingen.

26. R Development Core Team (2010). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

27. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics,* 6: 461-464.

28. Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology,* 44: 92-107.

29. Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology,* 44: 41-61.

30. Zucchini, W., Claeskens, G. and Nguefack-Tsague, G. (2011). Model selection. In *International Encyclopedia of Statistical Sciences*, Part 13, 830-833, DOI:10.1007/978-3-642-04898-2373, Editor: M. Lovric, Springer.