Econometric Theory, **24**, 2008, 493–527. Printed in the United States of America.
DOI: 10.1017/S0266466608080201

MINIMIZING AVERAGE RISK IN REGRESSION MODELS

GERDA CLAESKENS
Katholieke Universiteit Leuven

NILS LID HJORT
University of Oslo

Most model selection mechanisms work in an "overall" modus, providing models without specific concern for how the selected model is going to be used afterward. The focused information criterion (FIC), on the other hand, is geared toward optimum model selection when inference is required for a given estimand. In this paper the FIC method is extended to weighted versions. This allows one to rank and select candidate models for the purpose of handling a range of similar tasks well, as opposed to being forced to focus on each task separately. Applications include selecting regression models that perform well for specified regions of covariate values. We derive these weighted focused information criteria (wFIC), give asymptotic results, and apply the methods to real data. Formulas for easy implementation are provided for the class of generalized linear models.

1. INTRODUCTION

Model selection is most often the starting point of any data analysis. It is therefore of importance to carefully examine this initial step. Most model selection tools attach to each potential model a number, or a score value, and then proceed by picking the model with the best (usually either smallest or largest) value of this number. Traditional model selection techniques work this way, such as the Mallows (1973) C_p for linear regression or the more generally applicable Akaike information criteria (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978). Once a model is selected, the actual estimation takes place inside the selected model.

The focused information criterion (FIC) (Claeskens and Hjort, 2003) is also in this spirit, though it distinguishes itself from the other information criteria in that it can be directed—focused—toward a specific purpose. A model selected for estimating the mean return of an investment taking place in one month,

We express our sincere thanks to all reviewers of this paper, including the special issue guest editors and editor Professor Phillips, whose comments and questions have contributed to significant improvements. We also thank Dr. Ronald Klein for kindly giving permission to use the WESDR data. The work of Claeskens has been supported in part by the Fund for Scientific Research Flanders (G.0542.06). Address correspondence to Gerda Claeskens, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium; e-mail: Gerda.Claeskens@econ.kuleuven.be.

© 2008 Cambridge University Press 0266-4666/08 $15.00

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

e.g., should not necessarily be the same as the model to be used for estimating the mean return one year further in time. Similarly, a good statistical model for predicting salary for men is not necessarily good for predicting salary for women, and in medical studies, the model that is best for estimating the survival probability should not be expected to be the same as the best model for estimating the median survival time. In the construction of the FIC we therefore start by specifying the focus parameter, the quantity we wish to estimate, and then use this information to obtain the actual FIC value via an estimate of the mean squared error of the estimator of this focus parameter. Thus different focus parameters lead to different FIC values and to possibly different ranking lists of the best models.

The main issue addressed in this paper is that in many situations, the FIC demands too much focus of its user. For example, for regression models the FIC can easily be used to select a model for the mean response value for a given single covariate position. This is sometimes of relevance, and one may take a "median" or "average" or "90% extreme" covariate position, but in other situations one wishes to construct a model that does well across many covariate positions.

To address this problem, we derive a weighted focused information criterion (wFIC), where we attach to each potential model a single wFIC value, valid over all or part of the covariate space. To explain matters clearly, we mainly concentrate on the class of generalized linear models, though our arguments and methods extend without serious difficulties to more general regression models and indeed to other model selection problems, such as that of selecting among different covariance structure models for multivariate data vectors. More information on generalized linear models can be found in the books by McCullagh and Nelder (1989) and Dobson (2002).

Section 2 briefly reviews the FIC method, in a setting of generalized linear models. Working with weighted versions of mean squared errors, across all or a subset of covariates, leads in Section 3 to the wFIC method. In Section 4 the scope is broadened by allowing more general focus parameters in rather general parametric models. A further extension is given in Section 5, where we start from the linex loss function, often used in econometric contexts where losses are nonsymmetric, instead of trying to minimize averages of squared error loss.

The wFIC method relies on weights that are user- and context-specific. Applying a specific weighting scheme, with what we term generalized linear model (GLM) weights, gives a procedure that turns out to be large-sample equivalent to the AIC. This is discussed in Section 6, along with some other weighting schemes of interest. Three practical data examples, showing the applicability of the method, appear in Section 7. These applications also serve to illustrate that the wFIC's aim of achieving low average risk, with weights determined by the actual application context, may differ from the aim of optimal classification, or optimal equal-weight prediction. An explicit discussion of the performance of the criterion wFIC compared to the classical AIC is in Section 8.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

Section 9 offers some concluding comments, indicating the further scope of our work.

2. THE FIC FOR GLMs

Let the observed data be denoted $Y_1, \ldots, Y_n$, together with observed covariate information $u_1, \ldots, u_n$. In a model selection situation the u_i's are vector valued, and we wish to build methods that somehow manage to select the components of most relevance. Often, we are sure about including some of these covariates in the model (an intercept is an example of this), so that the actual selection should take place over the other variables. Notationally we split the covariate vector u_i into two parts, the "protected" x_i with, say, p covariates that are deemed necessary a priori and the "open" z_i containing the, say, q remaining covariates, among which we intend to perform the selection.

2.1. The GLM Class

For a GLM, there is a monotone differentiable link function $g(\cdot)$ such that

$$g(\mathrm{E}(Y_i|x_i, z_i)) = x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma \quad \text{for } i = 1, \ldots, n,$$

mapping the mean response to a linear predictor defined in terms of regression coefficients $\beta_1, \ldots, \beta_p$ and $\gamma_1, \ldots, \gamma_q$. Thus selecting which covariate components $z_{i,j}$ to include amounts to determining which γ_j's to keep in the model; excluding component $z_{i,j}$ corresponds to using $\gamma_j = 0$. For (classical) linear models, the link function is the identity function. For other examples, see Section 2.3.

We assume that $Y_1, \ldots, Y_n$ are independent with density function belonging to an exponential family of the form

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \quad \text{for } y \in \mathcal{Y},$$

where the sample space $\mathcal{Y}$ is the same in each case and does not depend on the unknown parameters θ_i and ϕ. The functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$ are fully specified. The $b(\cdot)$ function plays a crucial role because its derivatives yield the mean and variance function, whereas $a(\phi)$ is a scale parameter. From the two first Bartlett identities, about moment properties of the first and second log-derivatives of the density, comes the following expressions:

$$\xi_i = \mathrm{E}(Y_i|x_i, z_i) = b'(\theta_i) \quad \text{and} \quad \mathrm{Var}(Y_i|x_i, z_i) = a(\phi)b''(\theta_i).$$

This expresses θ_i as a function of the mean response, given the covariates. When the so-called canonical link function is used, i.e., $g(\cdot) = (b')^{-1}(\cdot)$, then $\theta_i = g(\xi_i) = x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma$.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

2.2. The Pointwise FIC for GLM

Here we derive explicit expressions for the FIC when applied to the linear predictor $\mu = x^{\mathrm{t}}\beta + z^{\mathrm{t}}\gamma = g(\mathrm{E}(Y|x,z))$ associated with a given position (x,z) in the covariate space. The (pointwise) FIC is large-sample equivariant under smooth transformations, and so the FIC for $\mu(x,z)$ will essentially yield the same model ranking as the FIC for $\xi(x,z) = \mathrm{E}(Y|x,z)$.

One of the main ingredients for computing the FIC is the (normalized) Fisher information matrix $J_{n,\mathrm{wide}}$, computed via the second-order partial derivatives of the log-likelihood function with respect to (β,γ). We partition the matrix $J_{n,\mathrm{wide}}$ as

$$J_{n,\mathrm{wide}} = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix},$$

such that the block $J_{n,11}$ in the lower right corner has dimension $q \times q$. In our calculations we frequently need the lower right submatrix of dimension $q \times q$ of $J_{n,\mathrm{wide}}^{-1}$, which we name K_n. This matrix may be found as $K_n = (J_{n,11} - J_{n,10}J_{n,00}^{-1}J_{n,01})^{-1}$.

For several members of the generalized linear model family, the scale parameter ϕ is either known or completely specified. Examples are the Poisson and binomial distributions, where $a(\phi) = 1$, or the normal distribution with known variance. In the case where ϕ is known, the information matrix takes a simple and general form, making it easy to compute in the full generality of GLMs. We first decompose the $n \times (p+q)$ design matrix into $X = (x_1^{\mathrm{t}}, \ldots, x_n^{\mathrm{t}})^{\mathrm{t}}$ of dimension $n \times p$ and $Z = (z_1^{\mathrm{t}}, \ldots, z_n^{\mathrm{t}})^{\mathrm{t}}$ of dimension $n \times q$. Then

$$J_{n,\mathrm{wide}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{a(\phi)} v_i \begin{pmatrix} x_i \\ z_i \end{pmatrix}\begin{pmatrix} x_i \\ z_i \end{pmatrix}^{\mathrm{t}} = \frac{1}{n}\frac{1}{a(\phi)}\begin{pmatrix} X^{\mathrm{t}}VX & X^{\mathrm{t}}VZ \\ Z^{\mathrm{t}}VX & Z^{\mathrm{t}}VZ \end{pmatrix}, \tag{1}$$

where V is the diagonal weight matrix $\mathrm{diag}(v_1, \ldots, v_n)$, with different formulas available for $v_i = v(x_i, z_i)$. One has

$$v_i = [b''(\theta_i)\{g'(\xi_i)\}^2]^{-1} = b''(\theta_i)(\partial\theta_i/\partial\eta_i)^2, \tag{2}$$

where $\eta_i = g^{-1}(\xi_i) = x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma$ and $\xi_i = \mathrm{E}(Y_i|x_i,z_i) = b'(\theta_i)$. For the situation with ϕ known, therefore, the K_n matrix takes the form

$$K_n = a(\phi)\{n^{-1}Z^{\mathrm{t}}V(I_n - X(X^{\mathrm{t}}VX)^{-1}X^{\mathrm{t}}V)Z\}^{-1}. \tag{3}$$

For canonical link functions, where $\theta_i = \eta_i$, matters simplify to $v_i = b''(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)$.

When the scale parameter ϕ is not known, such as with the normal distribution with unknown mean and variance, or the gamma distribution, we shall argue that formula (3) is still valid for K_n. This is because of an orthogonality property, namely, that the mixed second derivatives of the log-likelihood function,

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

with respect to ϕ and β or γ, are seen to have mean zero. Thus the full information matrix takes the form

$$J_{n,\text{wide}} = \begin{pmatrix} J_{n,\text{scale}} & 0 & 0 \\ 0 & n^{-1}a(\phi)^{-1}X^{\text{t}}VX & n^{-1}a(\phi)^{-1}X^{\text{t}}VZ \\ 0 & n^{-1}a(\phi)^{-1}Z^{\text{t}}VX & n^{-1}a(\phi)^{-1}Z^{\text{t}}VZ \end{pmatrix},$$

where $J_{n,\text{scale}}$ is the required $-n^{-1}\sum_{i=1}^{n}\text{E}\partial^2 \log f(Y_i;\theta_i,\phi)/\partial\phi^2$. This implies that the $q \times q$ lower right-hand corner of the inverse information matrix remains as in (3). For the normal distribution with $\phi = \sigma$ and $a(\sigma) = \sigma^2$, e.g., one finds $J_{n,\text{scale}} = 2/\sigma^2$.

We may now describe the FIC procedure for selecting the tentatively best model for estimating the focus parameter $\mu = \mu(x,z) = x^{\text{t}}\beta + z^{\text{t}}\gamma$. The criterion works specifically for the given (x,z) position in covariate space; if required, the procedure can be repeated for several positions. Different submodels are indexed by the various subsets S of covariate components $1,\ldots,q$, ranging from the empty set to the full list. For each submodel S one can evaluate the estimate

$$\hat{\mu}_S(x,z) = x^{\text{t}}\hat{\beta}_S + (\pi_S z)^{\text{t}}\hat{\gamma}_S = x^{\text{t}}\hat{\beta}_S + z_S^{\text{t}}\hat{\gamma}_S \qquad \textbf{(4)}$$

that uses maximum likelihood estimates $(\hat{\beta}_S,\hat{\gamma}_S)$ in the model that employs all of $\beta_1,\ldots,\beta_p$ but only those γ_j for which $j \in S$. Here π_S is the $|S| \times q$ projection matrix that sends z to $\pi_S z = z_S$, the vector of only those z_j for which $j \in S$, and $|S|$ is the number of components in S. The essence of the FIC is to estimate the mean squared error for each candidate estimator and then to pick the one with lowest mean squared error estimate.

We need the vector $\omega = J_{n,10}J_{n,00}^{-1}(\partial\mu/\partial\beta) - (\partial\mu/\partial\gamma)$, which here reads

$$\omega = Z^{\text{t}}VX(X^{\text{t}}VX)^{-1}x - z,$$

whether or not ϕ is known. Note the dependence of ω on (x,z). Note furthermore that premultiplying $v_S = \pi_S v$ with π_S^{t} produces a vector $\pi_S^{\text{t}}\pi_S v$ of full length q, with zeros inserted for those components j with $j \notin S$. Define next $K_{n,S} = (\pi_S K_n^{-1}\pi_S^{\text{t}})^{-1}$, which is the $|S| \times |S|$ lower right block of the inverse of $J_{n,S}$, the information matrix for the S model, and let finally $G_{n,S} = \pi_S^{\text{t}}K_{n,S}\pi_S K_n^{-1}$. The value of the FIC is now obtained as

$$\text{FIC}(S;x,z) = n\omega^{\text{t}}(I_q - G_{n,S})\hat{\gamma}_{\text{wide}}\hat{\gamma}_{\text{wide}}^{\text{t}}(I_q - G_{n,S})^{\text{t}}\omega + 2\omega^{\text{t}}\pi_S^{\text{t}}K_{n,S}\pi_S\omega, \qquad \textbf{(5)}$$

with $\hat{\gamma}_{\text{wide}}$ being the maximum likelihood estimator of γ in the largest of the considered models, i.e., the one containing all of $\gamma_1,\ldots,\gamma_q$. For given values of (x,z), the best model according to the FIC is that model, indexed by S, for which $\text{FIC}(S;x,z)$ is the smallest.

The FIC stems from estimating and then adding a squared bias term and a variance term; see Claeskens and Hjort (2003) for further discussion and appli-

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

cations. There is a modification of (5) for the case of the squared bias being estimated with a negative number; in such cases we truncate that term to zero. For further discussion of this point, see also relevant comments in Hjort and Claeskens (2006).

2.3. Examples

Here we present a short list of examples that fit in with the preceding general framework. Each model may be fitted using statistical software packages, and the J_n and K_n matrices are easily computed via the appropriate form of the v_i weights. For each situation one may use (5) to determine the best submodel for estimating $x^{\mathrm{t}}\beta + z^{\mathrm{t}}\gamma$, or for any smooth function thereof, such as $\mathrm{E}(Y|x,z)$ or the median response $\mathrm{median}(Y|x,z)$.

1. Consider a nonlinear normal regression setup where Y_i is normal (θ_i, σ^2), with $\theta_i = r(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)$ for some specified function $r(\eta)$. Then $b(\theta) = \frac{1}{2}\theta^2$, and the GLM weights of (2) become $v_i = r'(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)^2$. The ordinary linear normal model corresponds to $r(\eta) = \eta$ with weights $v_i = 1$. On some occasions it is important to take variance heterogeneity into account, e.g., by modeling the standard deviation σ_i as a function of covariates. This calls for extensions of the standard GLM model; see Section 7.1 for an illustration.
2. Assume that the Y_i's are Poisson with parameters $\xi_i = \exp(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)$. This is Poisson regression with canonical link function. Then $b(\theta) = \exp(\theta)$, and $v_i = \exp(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)$.
3. Let Y_i be binomial (m_i, p_i), with $p_i = H(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)$, for a suitable distribution function H. This is again a GLM with $b(\theta_i) = m_i \log\{1 + \exp(\theta_i)\}$, and one finds

   $$v_i = m_i H'(\eta_i)^2/[H(\eta_i)\{1 - H(\eta_i)\}],$$

   with $\eta_i = x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma$. This can be used for probit regression, e.g., where $p_i = \Phi(\eta_i)$ with Φ the cumulative standard normal distribution. For logistic regression matters simplify to $v_i = m_i p_i(1 - p_i)$.
4. Suppose positive observations Y_i are modeled with gamma distributions (c, d_i), where c is fixed but $d_i = \exp(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)$. We use the parametrization where the mean of Y_i is $\xi_i = c/d_i$. Here one finds $v_i = c$, simply. Thus the $J_{n,\mathrm{wide}}$ matrix is proportional to the sample variance matrix of the covariate vectors.

   Suppose on the other hand that Y_i is taken to be gamma with parameters (c_i, d), this time with d fixed and flexible $c_i = \exp(\eta_i)$, again with $\eta_i = x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma$. This actually corresponds to a GLM in terms of the log Y_i, and one finds $v_i = \exp(2\eta_i)\psi'(\exp(\eta_i))$, where ψ' is the trigamma function, the derivative of $\psi = \Gamma'/\Gamma$.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

3. THE wFIC FOR GLMs

Because of the dependence of the focus on the covariate values (x, z), the FIC also takes different values, and will produce different rankings of candidate models, for different locations in the covariate space. Sometimes we choose an average or median value for the regression variables (x, z), to represent some "average" or "median" subject, sometimes inside a stratum. Often, such a detailed focus point is not wanted; one might rather wish to find a good model that works well over a major part of, or over all of, the covariate space. We wish to select a good model valid for a range of individuals simultaneously, say, for a subgroup of the population.

We continue to use the GLM setting of the previous section. To reach precise results and concise arguments for our weighted-focused model selection schemes we shall employ a local misspecification framework where the true parameter is of the form $(\beta, \gamma) = (\beta_0, \delta/\sqrt{n})$. Here $\delta = (\delta_1, \ldots, \delta_q)^{\mathrm{t}}$ is fixed and unknown. This is as in Hjort and Claeskens (2003a) and Claeskens and Hjort (2003); see also the rejoinder Hjort and Claeskens (2003b) to the discussion of these papers. The idea is to prove results about model selectors and estimators in terms of the local model misspecification parameter δ and to use such results for developing appropriate model information criteria.

3.1. Some Preliminary Results

Suppose in general terms that a parameter of interest $\mu(\beta, \gamma; u)$ depends on some quantity u that varies in the population being studied. Under the framework outlined previously, henceforth assumed to hold, the true parameter value is $\mu(\beta_0, \delta/\sqrt{n}; u)$. We shall now use results developed in Hjort and Claeskens (2003a) and Claeskens and Hjort (2003), with suitable modifications. These results rely on certain mild regularity conditions, detailed in these papers. One such condition that we need to mention here, to adequately identify the appropriate limits subsequently, is the existence of a limit matrix J to which $J_{n,\mathrm{wide}}$ converges, with blocks J_{00}, etc. Similarly this defines limit versions K of K_n and a fortiori J_S and K_S, limit versions of $J_{n,S}$ and $K_{n,S}$. Also, $K_S = (\pi_S K^{-1} \pi_S^{\mathrm{t}})^{-1}$, the lower right block $(J_S^{-1})_{11}$ of J_S.

For each u the theory as developed in Claeskens and Hjort (2003) applies to the subset-model-based maximum likelihood estimators $\hat{\mu}_S(u)$, for which we have the following convergence in distribution:

$$\sqrt{n}\{\hat{\mu}_S(u) - \mu_{\mathrm{true}}(u)\} \xrightarrow{d} \Lambda_S(u)$$
$$= \left(\frac{\partial \mu(\beta, \gamma; u)}{\partial \beta}\right)^{\mathrm{t}} J_{00}^{-1} M + \omega(u)^{\mathrm{t}}(\delta - \pi_S^{\mathrm{t}} K_S \pi_S K^{-1} D).$$

Here $\omega(u) = J_{10} J_{00}^{-1} \partial \mu(\beta, \gamma; u)/\partial \beta - \partial \mu(\beta, \gamma; u)/\partial \gamma$, with partial derivatives evaluated at the null point $(\beta_0, 0)$. Furthermore, $M \sim \mathrm{N}_p(0, J_{00})$ and

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

$D \sim \mathrm{N}_q(\delta, K)$, and these random vectors are independent. The (M, D) variables are needed in Section 4, and furthermore the variables (C_S, D_S) that appear now are linear functions of (M, D).

For a fixed set S and a fixed focus point, Lemma 3.2 of Hjort and Claeskens (2003a) applies and yields

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_S - \beta_0) \\ \sqrt{n}\,\hat{\gamma}_S \end{pmatrix} \xrightarrow{d} \begin{pmatrix} C_S \\ D_S \end{pmatrix} \sim \mathrm{N}_{p+|S|}(\xi_S, J_S^{-1}),$$

where

$$\xi_S = \begin{pmatrix} J_{00}^{-1} J_{01}(I_q - \pi_S^{\mathrm{t}} K_S \pi_S K^{-1})\delta \\ K_S \pi_S K^{-1} \delta \end{pmatrix}.$$

A special case of importance is that of S being the full set $\{1, \ldots, q\}$, for which

$$\hat{\delta}_{\mathrm{wide}} = \sqrt{n}\,\hat{\gamma}_{\mathrm{wide}} \xrightarrow{d} D \sim \mathrm{N}_q(\delta, K). \tag{6}$$

For convenience of notation we adopt the notion that estimators without subset-subscript correspond to the full model, with all $\gamma_1, \ldots, \gamma_q$ parameters; thus $\hat{\delta} = \hat{\delta}_{\mathrm{wide}}$, etc.

The S-indexed model works with estimates $\hat{\gamma}_S$ for γ_j with $j \in S$ but uses 0 for $j \notin S$. To deal efficiently with different functions of these it will be convenient to introduce the extended projection matrix of dimension $(p + |S|) \times (p + q)$,

$$\tilde{\pi}_S = \begin{pmatrix} I_p & 0_{p,q} \\ 0_{|S|,p} & \pi_S \end{pmatrix}.$$

With $G_S = \pi_S^{\mathrm{t}} K_S \pi_S K^{-1}$, the limit version of $G_{n,S}$ encountered in (5), we have

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_S - \beta_0 \\ \pi_S^{\mathrm{t}} \hat{\gamma}_S - \delta/\sqrt{n} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} C_S \\ \pi_S^{\mathrm{t}} D_S - \delta \end{pmatrix},$$

which is seen to have mean vector and variance matrix

$$F_S \delta = \begin{pmatrix} J_{00}^{-1} J_{01}(I_q - G_S) \\ -(I_q - G_S) \end{pmatrix} \delta \quad \text{and} \quad \Gamma_S = \tilde{\pi}_S^{\mathrm{t}} J_S^{-1} \tilde{\pi}_S.$$

3.2. The Case of No Common Parameters

In the previous section we assumed there to be a parameter vector θ common to all models. Often this is the case, e.g., in regression models when always including an intercept term. However, there are situations where one does not

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

wish to include such a parameter in every model. Using the notation introduced earlier, this is the situation of only having a γ parameter (θ is not present). In fact, this simplifies the calculations quite a bit. Let γ_0 be the vector of null values for γ, where $\gamma_j = \gamma_{0,j}$ corresponds to the jth modeling factor not being present. Most often $\gamma_0 = 0_q$, e.g., when modeling the mean in a GLM where $\gamma_{0,j} = 0$ indicates absence of covariate z_j, though other values are possible also. Consider, e.g., a skewed logistic regression model where the probability of success is

$$P(Y = 1|z) = \left\{ \frac{\exp(z^{\mathrm{t}}\beta)}{1 + \exp(z^{\mathrm{t}}\beta)} \right\}^{\eta}.$$

Here $\gamma = (\beta, \eta)$ and $\gamma_0 = (0,1)$. The computation of the Fisher information matrix is simpler because we now do not have to take derivatives with respect to θ. One finds that the full Fisher information matrix $\tilde{J}_{n,\mathrm{wide}}$ for the situation without common parameter θ is equal to the lower right block $J_{n,11}$ of the partitioned matrix in the (θ, γ) situation. Equivalently, $\tilde{J}_{n,\mathrm{wide}} = \tilde{J} = K_n^{-1}$. To avoid confusion, we place a tilde above matrices to denote the situation of the case of no common parameters. Define the score function $\overline{V}_{n,S} = (1/n)\sum_{i=1}^{n}(\partial/\partial\gamma_S)\log f(Y_i, \gamma_0)$. By a Taylor expansion around $\gamma_{\mathrm{true}} = \gamma_0 + \delta/\sqrt{n}$, under the local misspecification setting, it follows that $\sqrt{n}\,\overline{V}_{n,S} \xrightarrow{d} \pi_S\tilde{J}\delta + \pi_S N$, where $N \sim \mathrm{N}_q(0_q, \tilde{J})$. This implies that $\pi_S N \sim \mathrm{N}_{|S|}(0_{|S|}, \tilde{J}_S)$. From this it follows that

$$\sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) = \tilde{J}_S^{-1}\sqrt{n}\,\overline{V}_{n,S} + o_P(1) \xrightarrow{d} \tilde{J}_S^{-1}(\pi_S\tilde{J}\delta + \pi_S N) = \tilde{D}_S.$$

Defining $\widetilde{W} = \tilde{J}^{-1}N$ and $\tilde{D} = \delta + \widetilde{W}$, we have $\tilde{D}_S = \tilde{J}_S^{-1}\pi_S\tilde{J}(\delta + \widetilde{W}) = \tilde{J}_S^{-1}\pi_S\tilde{J}\tilde{D}$, where $\tilde{D}$ corresponds to $\tilde{D}_{\mathrm{wide}}$, the limiting distribution of $\sqrt{n}(\hat{\gamma}_{\mathrm{wide}} - \gamma_0)$. It now immediately follows that

$$\begin{aligned}\mu(\hat{\gamma}_S;u) - \mu(\gamma_{\mathrm{true}};u) &= \{\mu(\hat{\gamma}_S;u) - \mu(\gamma_0;u)\} + \{\mu(\gamma_0;u) - \mu(\gamma_{\mathrm{true}};u)\} \\ &= \left(\frac{\partial\mu(\beta,\gamma;u)}{\partial\gamma_S}\right)^{\mathrm{t}}(\hat{\gamma}_S - \gamma_0) \\ &\quad + \left(\frac{\partial\mu(\beta,\gamma;u)}{\partial\gamma}\right)^{\mathrm{t}}(-\delta/n^{1/2}) + o_P(1/\sqrt{n}).\end{aligned}$$

Hence, with $\tilde{\omega} = -(\partial\mu/\partial\gamma)(\gamma_0;u)$, $\sqrt{n}\{\mu(\hat{\gamma}_S;u) - \mu(\gamma_{\mathrm{true}};u)\} \xrightarrow{d} \Lambda_S(u) = \tilde{\omega}^{\mathrm{t}}\delta - \tilde{\omega}^{\mathrm{t}}\tilde{G}_S\tilde{D}$. This limiting variable Λ_S has mean $\tilde{\omega}^{\mathrm{t}}(I_q - \tilde{G}_S)\delta$ and variance $\tilde{\omega}^{\mathrm{t}}\pi_S^{\mathrm{t}}\tilde{J}_S^{-1}\pi_S\tilde{\omega}$. Except for the simplified forms of the matrix $\tilde{J}$ and the vector $\tilde{\omega}$, this distribution differs from that in the (θ, γ) situation by the absence of a variance term for the model only containing θ. The risk function or mean squared error corresponds to $\tilde{\omega}^{\mathrm{t}}(I_q - \tilde{G}_S)\delta\delta^{\mathrm{t}}(I_q - \tilde{G}_S^{\mathrm{t}})\tilde{\omega} + \tilde{\omega}^{\mathrm{t}}\pi_S^{\mathrm{t}}\tilde{J}_S^{-1}\pi_S\tilde{\omega}$. We now proceed by inserting the unbiased estimator $\hat{\delta}\hat{\delta}^{\mathrm{t}} - \tilde{J}^{-1}$ for $\delta\delta^{\mathrm{t}}$ and replac-

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

ing unknown parameters by their maximum likelihood estimators in the full model. Substituting $\tilde{\omega}$, $\tilde{G}_S$, and $\tilde{J}$ for ω, G_S, and J, the formula (5) for the pointwise FIC remains valid for the case of no common parameters.

3.3. The wFIC for GLM

Consider again the linear predictor $\mu(x,z) = x^{\mathrm{t}}\beta + z^{\mathrm{t}}\gamma$ and the collection of submodel-based estimators (4). The vector $\hat{\beta}_S$ is always of length p but takes on different values for different index sets S. In contrast, the estimator $\hat{\gamma}_S$ has length $|S|$. The construction of the wFIC proceeds as follows. As opposed to using the pointwise loss function, we start with the *weighted* average quadratic loss function on the scale of the linear predictor, of the form

$$L_n(S) = \sum_{i=1}^{n} w(x_i,z_i)\{\hat{\mu}_S(x_i,z_i) - \mu_{\mathrm{true}}(x_i,z_i)\}^2$$
$$= \sum_{i=1}^{n} w(x_i,z_i)(x_i^{\mathrm{t}}\hat{\beta}_S + z_{i,S}^{\mathrm{t}}\hat{\gamma}_S - x_i^{\mathrm{t}}\beta_0 - z_i^{\mathrm{t}}\delta/\sqrt{n})^2. \quad \textbf{(7)}$$

The weights $w(x_i,z_i)$ are user-specified and in general different from the weights $v(x_i,z_i)$ in the GLM weight matrix V. We shall show that this random loss has a limit distribution, under mild regularity conditions. Let

$$\Omega_{n,w} = \frac{1}{n}\sum_{i=1}^{n} w(x_i,z_i)\begin{pmatrix} x_i \\ z_i \end{pmatrix}\begin{pmatrix} x_i \\ z_i \end{pmatrix}^{\mathrm{t}} \quad \textbf{(8)}$$

and assume this matrix converges in probability to a nonnegative definite Ω_w, depending of course on the weight function $w(x,z)$.

To properly study the random loss function it is convenient to express it as

$$L_n(S) = n\begin{pmatrix} \hat{\beta}_S - \beta_{\mathrm{true}} \\ \pi_S^{\mathrm{t}}\hat{\gamma}_S - \gamma_{\mathrm{true}} \end{pmatrix}^{\mathrm{t}}\Omega_{n,w}\begin{pmatrix} \hat{\beta}_S - \beta_{\mathrm{true}} \\ \pi_S^{\mathrm{t}}\hat{\gamma}_S - \gamma_{\mathrm{true}} \end{pmatrix}.$$

From conditions and results noted earlier,

$$L_n(S) \xrightarrow{d} L(S) = \begin{pmatrix} C_S \\ \pi_S^{\mathrm{t}}D_S - \delta \end{pmatrix}^{\mathrm{t}}\Omega_w\begin{pmatrix} C_S \\ \pi_S^{\mathrm{t}}D_S - \delta \end{pmatrix}.$$

Under mild conditions, the expected loss $w\text{-risk}_n(S) = \mathrm{E}\,L_n(S)$ will converge to $w\text{-risk}(S,\delta) = \mathrm{E}\,L(S)$. The limit loss $L(S)$ is a quadratic form in normal variables and has mean value

$$w\text{-risk}(S,\delta) = \mathrm{E}\,L(S) = \delta^{\mathrm{t}}F_S^{\mathrm{t}}\Omega_w F_S\delta + \mathrm{Tr}(\Omega_w\Gamma_S)$$
$$= \mathrm{Tr}(\Omega_w F_S\delta\delta^{\mathrm{t}}F_S^{\mathrm{t}}) + \mathrm{Tr}(\Omega_w\Gamma_S) = \mathrm{I}(S) + \mathrm{II}(S),$$

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

involving matrices F_S and Γ_S defined previously. We see that the $\mathrm{I}(S)$ term corresponds to weighted squared bias, whereas the second term $\mathrm{II}(S)$ is related to weighted variance. By earlier efforts, this second term can be expressed as $\mathrm{II}(S) = \mathrm{Tr}(\Omega_w \bar{\pi}_S^{\mathrm{t}} J_S^{-1} \bar{\pi}_S)$.

This leads upon estimating unknown quantities to a wFIC. The second term is not problematic, and we use

$$\widehat{\mathrm{II}}(S) = \mathrm{Tr}(\Omega_{n,w} \bar{\pi}_S^{\mathrm{t}} \hat{J}_{n,S}^{-1} \bar{\pi}_S),$$

where $\hat{J}_{n,S}$ is the appropriate submatrix of

$$\hat{J}_n = \frac{1}{n} \frac{1}{a(\hat{\phi})} \sum_{i=1}^{n} \hat{v}_i \begin{pmatrix} x_i \\ z_i \end{pmatrix} \begin{pmatrix} x_i \\ z_i \end{pmatrix}^{\mathrm{t}} = \frac{1}{n} \frac{1}{a(\hat{\phi})} \begin{pmatrix} X^{\mathrm{t}} \hat{V} X & X^{\mathrm{t}} \hat{V} Z \\ Z^{\mathrm{t}} \hat{V} X & Z^{\mathrm{t}} \hat{V} Z \end{pmatrix}$$

and $\hat{v}_i$ is the estimated version of (2), inserting $x_i^{\mathrm{t}} \hat{\beta}_{\mathrm{wide}} + z_i^{\mathrm{t}} \hat{\gamma}_{\mathrm{wide}}$ for the linear predictor $x_i^{\mathrm{t}} \beta + z_i^{\mathrm{t}} \gamma$. See the examples in Section 2.3. For the first term, we note that

$$\hat{\delta}\hat{\delta}^{\mathrm{t}} = n\hat{\gamma}\hat{\gamma}^{\mathrm{t}} \xrightarrow{d} DD^{\mathrm{t}},$$

a variable with mean $\delta\delta^{\mathrm{t}} + K$; recall the convention noted around (6) that $\hat{\gamma}$ means $\hat{\gamma}_{\mathrm{wide}}$, etc. Hence $\hat{\mathrm{I}}^0(S) = \mathrm{Tr}\{\Omega_{n,w} \hat{F}_S (\hat{\delta}\hat{\delta}^{\mathrm{t}} - \hat{K}_n) \hat{F}_S^{\mathrm{t}}\}$ is an asymptotically unbiased estimator of $\mathrm{I}(S)$, and $\hat{\mathrm{I}}^0(S) + \widehat{\mathrm{II}}(S)$ an asymptotically unbiased estimator of w-risk(S, δ). To avoid estimating a nonnegative quantity with a negative number, we advocate truncation to zero, and we use $\hat{\mathrm{I}}(S) = \max\{\hat{\mathrm{I}}^0(S), 0\}$. The wFIC consists in evaluating

$$\mathrm{wFIC}(S) = \hat{\mathrm{I}}(S) + \widehat{\mathrm{II}}(S) = \max\{\hat{\mathrm{I}}^0(S), 0\} + \widehat{\mathrm{II}}(S) \qquad \textbf{(9)}$$

for each candidate model S and in the end selecting the model with smallest value of this estimated average risk. All components of the preceding expressions are easily obtained via standard output of any software fitting GLMs.

Remark 1. When more components are put into S, the $\mathrm{I}(S)$ term becomes smaller; for S equal to the full $\{1, \ldots, q\}$ we find $F_S = 0$, making $I(\mathrm{wide}) = 0$. On the other hand, the more components in S, the bigger is the variance $\mathrm{II}(S)$. Thus the wFIC method reflects the squared modeling bias against variance balance.

Remark 2. The $\mathrm{wFIC}(S) = \hat{\mathrm{I}}(S) + \widehat{\mathrm{II}}(S)$ is not algebraically equivalent to the simpler one of averaging individual $\mathrm{FIC}(S)$ scores. The two methods are the same only in cases where there are no modifications of setting negative estimates of squared bias to zero. The wFIC method, as outlined in Sections 3 and 4, performs the squared bias modification only once, at the end, as opposed to performing this operation for each individual application.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

Remark 3. The methods stem from precise limit distribution results that involve quantities such as J and K, along with further relatives such as K_S and G_S. Our wFIC(S) formulas involve estimates of these quantities, $\hat{J}_{n,S}$, $\hat{K}_{n,S}$, etc. The theory behind the methods ensures that they work well as long as estimates are used that are consistent, under the local misspecification framework $(\beta,\gamma) = (\beta_0, \delta/\sqrt{n})$. Among various possibilities we have chosen to use the "wide model perspective" in our implementations, starting from estimators $\hat{J}_n$ that use estimates $(\hat{\beta}_{\text{wide}}, \hat{\gamma}_{\text{wide}})$.

3.4. Examples Revisited

Here we make the formula for wFIC much more explicit for the class of GLM examples, with particular attention to the classical linear model. Again we use $V = \text{diag}\{v_i\}$ and $W = \text{diag}\{w(x_i, z_i)\}$. Examples of v_i for the case of the linear, Poisson, binomial, and gamma models are given in Section 2.3. We turn to the choice of weights $w(x_i, z_i)$ in Section 6.

First we work with $\hat{\text{I}}(S)$ of (9). With $J_{n,\text{wide}}$ as in (1), $G_S = \pi_S^{\text{t}} K_S \pi_S K_n^{-1}$, and K_n as in (3), one has

$$\hat{F}_S = \begin{pmatrix} (X^{\text{t}}\hat{V}X)^{-1}X^{\text{t}}\hat{V}Z(I_q - G_S) \\ -(I_q - G_S) \end{pmatrix} \quad \text{and} \quad \Omega_{n,w} = \frac{1}{n}\begin{pmatrix} X^{\text{t}}WX & X^{\text{t}}WZ \\ Z^{\text{t}}WX & Z^{\text{t}}WZ \end{pmatrix}.$$

Next we simplify the matrix multiplications needed to compute $\hat{\text{I}}^0(S) = \text{Tr}\{\hat{F}_S^{\text{t}}\Omega_{n,w}\hat{F}_S(\hat{\delta}\hat{\delta}^{\text{t}} - \hat{K}_n)\}$. After rearranging the resulting terms and introducing the generalized "hat matrix" for GLM, $H_V = X(X^{\text{t}}\hat{V}X)^{-1}X^{\text{t}}\hat{V}$, we arrive at $\hat{\text{I}}^0(S) = n^{-1}\text{Tr}\{(I_q - G_S)^{\text{t}}Z^{\text{t}}(I_n - H_V)^{\text{t}}W(I_n - H_V)Z(I_q - G_S)(\hat{\delta}\hat{\delta}^{\text{t}} - \hat{K}_n)\}$, and hence $\hat{\text{I}}(S) = \max\{\hat{\text{I}}^0(S), 0\}$. For $\widehat{\text{II}}(S)$ we need the matrix $J_{n,S}$, which is that submatrix of $J_{n,\text{wide}}$ for which the row and column indices belong to the index set S. For example, if $S = \{3,5\}$ this gives the parameter $\mu_S(x,z) = x^{\text{t}}\beta + z_3\gamma_3 + z_5\gamma_5$. In this example, of the matrix $J_{n,S}$ we keep of the last q rows only rows 3 and 5, and of the last q columns we keep only columns 3 and 5. In general, for any subset S of $\{1,\ldots,q\}$, we denote the reduced design matrix $Z_S = Z\pi_S^{\text{t}}$ and

$$\hat{J}_{n,S} = \frac{1}{na(\phi)}\begin{pmatrix} X^{\text{t}}\hat{V}X & X^{\text{t}}\hat{V}Z_S \\ Z_S^{\text{t}}\hat{V}X & Z_S^{\text{t}}\hat{V}Z_S \end{pmatrix} \quad \text{and} \quad \hat{J}_{n,S}^{-1} = na(\phi)\begin{pmatrix} B_S^{00} & B_S^{01} \\ B_S^{10} & B_S^{11} \end{pmatrix}.$$

Using formulas for the inverse of a partitioned matrix, $B_S^{11} = \{Z_S^{\text{t}}\hat{V}(I_n - H_V)Z_S\}^{-1}$, and

$$B_S^{00} = (X^{\text{t}}\hat{V}X)^{-1} + (X^{\text{t}}\hat{V}X)^{-1}X^{\text{t}}\hat{V}Z_S B_S^{11} Z_S^{\text{t}}\hat{V}X(X^{\text{t}}\hat{V}X)^{-1},$$

$$B_S^{01} = -(X^{\text{t}}\hat{V}X)^{-1}X^{\text{t}}\hat{V}X_S B_S^{11}.$$

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

This leads to

$$\widehat{\mathrm{II}}(S) = a(\phi)\mathrm{Tr}[X^{\mathrm{t}}WX\{(X^{\mathrm{t}}\hat{V}X)^{-1} + H_V Z_S B_S^{11} Z_S^{\mathrm{t}} H_V^{\mathrm{t}} - 2X^{\mathrm{t}}WZ_S B_S^{11} H_V^{\mathrm{t}} + Z^{\mathrm{t}}WZ_S B_S^{11}\}],$$

and finally to wFIC(S) as in (9). These expressions are straightforward to program.

In the special case of the classical linear model, the "hat matrix" is $H = X(X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}$, and

$$\hat{\mathrm{I}}(S) = \max[0, n^{-1}\mathrm{Tr}\{(I_q - G_{n,S})^{\mathrm{t}}Z^{\mathrm{t}}(I_n - H)^{\mathrm{t}}W(I_n - H)Z(I_q - G_{n,S}) \times (\hat{\delta}\hat{\delta}^{\mathrm{t}} - \hat{K}_n)\}],$$

$$\widehat{\mathrm{II}}(S) = \hat{\sigma}^2\mathrm{Tr}[X^{\mathrm{t}}WX\{(X^{\mathrm{t}}X)^{-1} + HZ_S B_S^{11} Z_S^{\mathrm{t}} H^{\mathrm{t}} - 2X^{\mathrm{t}}WZ_S B_S^{11} H^{\mathrm{t}} + Z^{\mathrm{t}}WZ_S B_S^{11}\}],$$

with $B_S^{11} = \{Z_S^{\mathrm{t}}(I_n - H)Z_S\}^{-1}$ and again with wFIC(S) being the sum of these two terms.

Remark 4. In linear models the expression for the pointwise (unweighted) FIC has been shown to be exact; see Claeskens and Hjort (2003, Sect. 5.5). There it is shown that, without assuming normality and when using least squares estimators, the exact expression for the mean squared error matches that obtained by FIC. This is a favorable property that helps in appreciating the obtained expression for the FIC. More precisely, in a linear model with $Y_i = x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma + \varepsilon_i$, where $\mu = x^{\mathrm{t}}\beta + z^{\mathrm{t}}\gamma$ is to be estimated, we use least squares estimators in a submodel indexed by S, leading to the estimator $\hat{\mu}_S = x^{\mathrm{t}}\hat{\beta}_S + (\pi_S z)^{\mathrm{t}}\hat{\gamma}_S$. Computing the *exact bias and variance* of this estimator leads to the following expression for the mean squared error of $\hat{\mu}_S$:

$$n^{-1}(x^{\mathrm{t}}J_{n,00}^{-1}x + \omega^{\mathrm{t}}\pi_S^{\mathrm{t}}K_{n,S}\pi_S\omega) + \omega^{\mathrm{t}}(I_q - \pi_S^{\mathrm{t}}K_{n,S}\pi_S K_n^{-1})\gamma\gamma^{\mathrm{t}}(I_q - K_n^{-1}\pi_S^{\mathrm{t}}K_{n,S}\pi_S)\omega. \quad \textbf{(10)}$$

This is, up to a constant not depending on the subset S, identical to the limiting expression on which the FIC is based; compare with (5). The same property holds when a (nonrandom) weight function w is included in the loss function L_n in (7). For linear models, computing the exact mean squared error of $L_n(S)$, with least squares estimators inserted for $\hat{\beta}_S$ and $\hat{\gamma}_S$, leads to an expression that is equivalent to that upon which wFIC as in (9) is based. Note that the error variance σ^2 appears in the denominator of the matrix J_n as the scaling factor $a(\sigma)$. There is accordingly a σ^2 factor implicitly featuring in the first two terms of (10) but not in the final term.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

4. THE wFIC FOR GENERAL RISK AVERAGES

Earlier we developed a wFIC method for estimating weighted averaged risks in GLMs. Sometimes different risk averages are called for, however, and this section extends the wFIC to handle such cases. We choose to stay inside the GLM framework exposited at the start of Section 3, although more general situations can be considered. A brief motivating example is as follows: Suppose Y_i data are exponentially distributed with parameters $\theta_i = \exp(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)$ and that one wishes to estimate the quantile distribution

$$\mu(u) = \mu(u|x_i, z_i) = \{-\log(1-u)\}/\exp(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma) \quad \text{for } u \in (0,1).$$

How can we select a model that provides good estimates $\hat{\mu}_S(u)$ across all deciles $u = 0.1, \ldots, 0.9$, say, or does the context of the application perhaps dictate that the upper quantiles between 0.80 and 0.99 are the most important? One might also wish to have a model that works well for, e.g., all upper quantiles for a specific stratum of z covariates.

Assume in general terms that a parameter $\mu(u)$ is to be estimated, defined in terms of the parameters (β, γ) of a GLM, and depending on some parameter u that may or may not depend on the covariates. As in previous sections limit distributions will be established in the framework where $(\theta, \gamma) = (\theta_0, \gamma_0 + \delta/\sqrt{n})$, aiming at providing adequate finite-sample approximations for risks and averaged risks. We use θ instead of β here, and γ_0 rather than simply 0, to indicate a somewhat more general model framework than merely models indexed by regression coefficients; θ could for some models encompass such regression coefficients along with some scale and skewness parameters, or it could be associated also with, say, correlation structure in multidimensional data. We have

$$\sqrt{n}\{\hat{\mu}_S(u) - \mu_{\mathrm{true}}(u)\} \xrightarrow{d} \Lambda_S(u) = \left(\frac{\partial \mu(u)}{\partial \theta}\right)^{\mathrm{t}} J_{00}^{-1} M + \omega(u)^{\mathrm{t}}(\delta - G_S D), \quad \textbf{(11)}$$

where $\omega(u) = J_{10}J_{00}^{-1}\partial\mu/\partial\theta - \partial\mu/\partial\gamma$ and $G_S = \pi_S^{\mathrm{t}} K_S \pi_S K^{-1}$; also, M and D are independent, and $M \sim \mathrm{N}_p(0, J_{00}^{-1})$ and $D \sim \mathrm{N}_q(\delta, K)$. Consider the random average loss incurred by using estimators associated with submodel S, say,

$$L_n(S, \delta) = n \int \{\hat{\mu}_S(u) - \mu_{\mathrm{true}}(u)\}^2 \, \mathrm{d}W_n(u). \quad \textbf{(12)}$$

Here W_n represents some relevant distribution of u values, such as the deciles in the preceding quantile example, or some empirical distribution over some interesting subset of the covariate space. This random loss would also depend on θ_0, but we concentrate here on what happens in terms of the local $\delta = \sqrt{n}(\gamma - \gamma_0)$ scale around γ_0. Assuming W_n converging to a suitable weight distribution W (or that it simply stays fixed, independent of sample size), we have $L_n(S, \delta)$ converging in distribution to $L(S, \delta) = \int \Lambda_S(u)^2 \, \mathrm{d}W(u)$ under mild conditions. We measure the total averaged risk via the expected value of L_n,

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

which converges to

$$w\text{-risk}(S,\delta) = \mathrm{E}\, L(S,\delta) = \int \mathrm{E}\Lambda_S(u)^2\, \mathrm{d}W(u),$$

again under weak conditions. Here

$$\begin{aligned}\mathrm{E}\Lambda_S(u)^2 &= \tau_0(u)^2 + \omega(u)^{\mathrm{t}}\mathrm{E}(\delta - G_S D)(\delta - G_S D)^{\mathrm{t}}\omega(u)\\ &= \tau_0(u)^2 + \omega(u)^{\mathrm{t}}\{(I_q - G_S)\delta\delta^{\mathrm{t}}(I_q - G_S)^{\mathrm{t}} + G_S K G_S^{\mathrm{t}}\}\omega(u),\end{aligned}$$

in terms of $\tau_0(u)^2 = (\partial\mu(u)/\partial\theta)^{\mathrm{t}} J_{00}^{-1}(\partial\mu(u)/\partial\theta)$ and the variance matrix

$$\mathrm{Var}\, G_S D = G_S K G_S^{\mathrm{t}} = \pi_S^{\mathrm{t}} K_S \pi_S K^{-1} \pi_S^{\mathrm{t}} K_S \pi_S = \pi_S^{\mathrm{t}} K_S \pi_S.$$

This leads after subtracting out the immaterial term $\int \tau_0(u)^2\, \mathrm{d}W(u)$ to the expression

$$w\text{-risk}(S,\delta) = \mathrm{Tr}\{A(I_q - G_S)\delta\delta^{\mathrm{t}}(I_q - G_S)^{\mathrm{t}}\} + \mathrm{Tr}(A\pi_S^{\mathrm{t}} K_S \pi_S) \qquad \textbf{(13)}$$

for the limit risk, where $A = \int \omega(u)\omega(u)^{\mathrm{t}}\, \mathrm{d}W(u)$. Hence the generalized wFIC becomes

$$\begin{aligned}\mathrm{wFIC}(S) &= \hat{\mathrm{I}}(S) + \widehat{\mathrm{II}}(S)\\ &= \max[\mathrm{Tr}\{\hat{A}(I_q - \hat{G}_{n,S})(\hat{\delta}\hat{\delta}^{\mathrm{t}} - \hat{K}_n)(I_q - \hat{G}_{n,S})^{\mathrm{t}}\}, 0]\\ &\quad + \mathrm{Tr}(\hat{A}\pi_S^{\mathrm{t}}\hat{K}_{n,S}\pi_S).\end{aligned}$$

Here $\hat{A}$ is a sample-based estimate of the A matrix. The candidate model with lowest value of $\mathrm{wFIC}(S)$ is selected. This more general wFIC can be applied when one wishes to consider, e.g., average risk across both covariates and quantiles.

5. MODEL SELECTION WITH WEIGHTED LINEX LOSS

There are many areas of application where precision of estimators and predictiors is best assessed using nonsymmetric loss functions, including situations in econometrics and social sciences. Among the more important alternatives to traditional squared error loss is the linex loss, defined here as $L_0(\sqrt{n}(\hat{\mu} - \mu))$, where $L_0(v) = \exp(cv) - cv - 1$. The parameter c determines the degree of asymmetry and skewness and can be both positive and negative, and $c \to 0$ corresponds to the usual squared error loss again. See Varian (1975) and Zellner (1986) for discussion and applications. Suppose a family of estimands $\mu(u)$ needs to be estimated, with a combined weighted loss of the form $\int L_0(\sqrt{n}(\hat{\mu}(u) - \mu(u)))\, \mathrm{d}W_n(u)$. Here we develop a model selection scheme that aims at minimizing the resulting risk function.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

When model S is used, the limit loss becomes

$$\text{loss}(S,\delta) = \int L_0(\Lambda_0(u) + \omega(u)^{\mathrm{t}}(\delta - G_S D))\,\mathrm{d}W(u).$$

The inner variable here is normal with mean $b_S(u)^{\mathrm{t}}\delta = \omega(u)^{\mathrm{t}}(I_q - G_S)\delta$ and variance $\tau_S(u)^2 = \tau_0(u)^2 + \omega(u)^{\mathrm{t}}G_S K G_S^{\mathrm{t}}\omega(u)$. Hence the limit risk associated with selecting model S is

$$\begin{aligned}\text{risk}(S,\delta) &= \mathrm{E}\,\text{loss}(S,\delta)\\ &= \int [\exp\{cb_S(u)^{\mathrm{t}}\delta + \tfrac{1}{2}c^2\tau_S(u)^2\} - cb_S(u)^{\mathrm{t}}\delta - 1]\,\mathrm{d}W(u).\end{aligned}$$

But this may be estimated unbiasedly in the limit experiment, starting from the observation that

$$\mathrm{E}\exp\{cb_S(u)^{\mathrm{t}}D\} = \exp\{cb_S(u)^{\mathrm{t}}\delta + \tfrac{1}{2}c^2 b_S(u)^{\mathrm{t}}Kb_S(u)\}.$$

Hence

$$\begin{aligned}r^*(S) = \int (&\exp[cb_S(u)^{\mathrm{t}}D + \tfrac{1}{2}c^2\{\tau_S(u)^2 - b_S(u)^{\mathrm{t}}Kb_S(u)\}]\\ &- cb_S(u)^{\mathrm{t}}D - 1)\,\mathrm{d}W(u)\end{aligned}$$

has the required property. This leads to the following selection scheme: for each candidate model, compute the weighted linex FIC score

$$\begin{aligned}\hat{r}^*(S) = \int (&\exp[c\hat{b}_S(u)^{\mathrm{t}}\hat{\delta} + \tfrac{1}{2}c^2\{\hat{\tau}_S(u)^2 - \hat{b}_S(u)^{\mathrm{t}}\hat{K}\hat{b}_S(u)\}]\\ &- c\hat{b}_S(u)^{\mathrm{t}}\hat{\delta} - 1)\,\mathrm{d}\hat{W}_n(u)\end{aligned}$$

and choose in the end the model with the lowest value of $\hat{r}^*(S)$. Here

$$\hat{b}_S(u) = \hat{\omega}(u)^{\mathrm{t}}(I_q - G_S) \quad \text{and} \quad \hat{\tau}_S(u)^2 = \hat{\tau}_0(u)^2 + \hat{\omega}(u)^{\mathrm{t}}\hat{G}_S\hat{K}\hat{G}_S^{\mathrm{t}}\hat{\omega}(u).$$

One may show that the limit version of this criterion, for $c \to 0$, is identical to our earlier wFIC.

6. CHOICE OF WEIGHTS

The wFIC method developed earlier provides a quite general and versatile model selection scheme, in that the weights $w_i = w(x_i, z_i)$ are fully user-specified, meant to reflect what aspects are deemed more important than others for the use of the finally selected model. The only caveat is that the weights should not be seriously unstable; the mathematical requirement for our asymptotics to go through is that the $\Omega_{n,w}$ matrix converges in probability with increasing sample size. This section discusses various types of weights.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

6.1. GLM Weights

The log-likelihood structure of a GLM itself suggests a type of weights for use in the wFIC method. If one chooses $w(x_i, z_i) = v_i/a(\phi)$, then $\Omega_{n,w} = J_{n,\text{wide}}$. This leads to simplifications in the w-risk(S,δ) and wFIC(S) expressions, as we shall see now. For simplicity of notation and presentation we work directly in the limit experiment, where J and K, etc., replace $\hat{J}_n$ and $\hat{K}_n$, etc.; we also write $G_S = \pi_S^{\mathrm{t}} K_S \pi_S K^{-1}$. First look at $\mathrm{I}(S) = \delta^{\mathrm{t}} F_S^{\mathrm{t}} J F_S \delta$, where some manipulations give

$$JF_S = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \begin{pmatrix} J_{00}^{-1} J_{01}(I_q - G_S) \\ -I_q + G_S \end{pmatrix} = \begin{pmatrix} 0 \\ -K^{-1}(I_q - G_S) \end{pmatrix}.$$

This leads with some further efforts to

$$\mathrm{I}(S) = \delta^{\mathrm{t}}(I_q - G_S)^{\mathrm{t}} K^{-1}(I_q - G_S)\delta = \delta^{\mathrm{t}}(K^{-1} - K^{-1}\pi_S^{\mathrm{t}} K_S \pi_S K^{-1})\delta,$$

which is estimated by

$$\begin{aligned} \hat{\mathrm{I}}(S) &= \mathrm{Tr}\{(K^{-1} - K^{-1}\pi_S^{\mathrm{t}} K_S \pi_S K^{-1})(DD^{\mathrm{t}} - K)\} \\ &= D^{\mathrm{t}} K^{-1} D - D^{\mathrm{t}} K^{-1} \pi_S^{\mathrm{t}} K_S \pi_S K^{-1} D - q + |S|, \end{aligned}$$

as long as this expression is positive; it is otherwise truncated to zero. Next,

$$\mathrm{II}(S) = \mathrm{Tr}(J\tilde{\pi}_S^{\mathrm{t}} J_S^{-1} \tilde{\pi}_S) = p + |S|.$$

This is easily obtained because the effect of pre- and postmultiplication by $\tilde{\pi}_S$ is that only that submatrix of J^{-1} is kept for which the row (and column) numbers belong to the index set S, entries on all other rows and columns being replaced by zero. This implies that J is multiplied by part of its inverse matrix, leading to the simplification $p + |S|$. This is also true for the data-based version $\widehat{\mathrm{II}}(S)$.

To summarize this, with GLM weights the limit risk takes the form

$$w\text{-risk}(S,\delta) = \delta^{\mathrm{t}}(K^{-1} - K^{-1}\pi_S^{\mathrm{t}} K_S \pi_S K^{-1})\delta + p + |S|,$$

and the canonical risk estimate (for the limit experiment) is

$$\hat{\mathrm{I}}(S) + \widehat{\mathrm{II}}(S) = \begin{cases} D^{\mathrm{t}}(K^{-1} - K^{-1}\pi_S^{\mathrm{t}} K_S \pi_S K^{-1})D + 2|S| + p - q & \\ & \text{if } N(S) \text{ takes place,} \\ p + |S| & \text{otherwise.} \end{cases} \quad \textbf{(14)}$$

Here $N(S)$ is the event that the trace in $\hat{\mathrm{I}}(S)$ is positive, i.e., $D^{\mathrm{t}}(K^{-1} - K^{-1}\pi_S^{\mathrm{t}} K_S \pi_S K^{-1})D > q - |S|$. The $N(S)$ takes place with high probability if

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

δ is some distance away from zero, but in situations where the underlying γ vector is close to zero, i.e., the narrow model is close to being correct, the probability that $N(S)$ does not take place is significant. The finite-sample version of the risk estimate uses $\hat{\delta}$ for D and $\hat{K}_n$ and $\hat{K}_{n,S}$ for K and K_S.

6.2. The wFIC and the AIC

Consider the AIC, which in the present circumstances takes the form

$$\mathrm{AIC}(S) = 2\sum_{i=1}^{n} \log f(y_i; \hat{\theta}_i, \hat{\phi}) - 2(p + |S|),$$

with $\hat{\theta}_i$ being the appropriate maximum likelihood estimate of $\theta_i = \theta(x_i, z_i)$, which again depends on $\hat{\beta}_S$ and $\hat{\gamma}_S$. The AIC scores may be computed for each submodel S, down to that of the most narrow model, which corresponds to $S = \emptyset$ and which uses only $\beta_1, \ldots, \beta_p, \phi$ as model parameters. When subtracting the smallest model's AIC value from $\mathrm{AIC}(S)$, and performing a one-step Taylor expansion, we find that, for the limiting situation where $n \to \infty$,

$$\mathrm{aic}(S) = \mathrm{AIC}(S) - \mathrm{AIC}(\emptyset) \xrightarrow{d} D^{\mathrm{t}} K^{-1} \pi_S^{\mathrm{t}} K_S \pi_S K^{-1} D - 2|S|. \qquad \textbf{(15)}$$

See Claeskens and Hjort (2003, eqn. (2.5)). The best models have the highest AIC scores.

We see from this and (14) that the wFIC method, when using GLM weights, is essentially large-sample equivalent to the AIC method. The word *essentially* relates to the modification for truncating an estimate of a squared bias to zero, when relevant, spelled out in (14). Thus the wFIC provides a fresh perspective on the AIC, and our arguments even suggest a correction to the AIC scores in cases where the event $N(S)$ does not take place. The result just described generalizes further, to the setting of Section 4. When the A matrix of (13) is proportional to K^{-1}, then the wFIC and the AIC are large-sample equivalent, inside the framework of local neighborhood models.

It is interesting to go through the list of examples in Section 2.3 to see the appropriate random loss functions that correspond to the AIC. For the logistic regression setup of Example 3 in that section, e.g., model selection by estimating the mean of

$$L_n(S) = \sum_{i=1}^{n} m_i p(x_i, z_i)\{1 - p(z_i, z_i)\}\left(\log \frac{\hat{p}_{i,S}}{1 - \hat{p}_{i,S}} - \log \frac{\hat{p}_i}{1 - \hat{p}_i}\right)^2$$

is essentially the same as AIC. Here $\hat{p}_{i,S}$ is the estimated probability under model S. Similarly, for Poisson regression, basing model selection on estimating

$$w\text{-risk}_n(S) = \mathrm{E}\sum_{i=1}^{n} \exp(x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma)(\log \hat{\lambda}_{i,S} - \log \lambda_i)^2$$

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

will be large-sample equivalent to the AIC scheme, where $\hat{\lambda}_{i,S}$ is the estimate of Poisson rate i inside the S model. Of course other weights and other transformations can be worked with, for logistic and Poisson regression, and such alternatives can in the perspective developed here be seen as cousins to the AIC method.

6.3. Other Types of Weights

In addition to the perhaps canonical GLM weights choice discussed previously, the following types of weights may be considered. We emphasize again that choosing an appropriate weight function is a context- and application-dependent issue.

Uniform weights. A simple choice of weights could just assign mass $1/n$ to each contribution in the summand, with the Ω_n matrix simply equal to the empirical variance matrix of the $p + q$-sized covariate vectors. Another choice is to let $w_i = 1$ precisely for individuals i belonging to a stratum of interest and $w_i = 0$ for the others. This is exemplified in Section 7.

Gliding covariate window. A version of the preceding situation is to use $w_i = w(x_i, z_i; x_0)$ equal to 1 for individuals inside a certain neighborhood of some fixed x_0. This gives a ranking of candidate models around each fixed x_0. Smoothed versions can also be contemplated, also in a context of model assessment where the x_0 is moved in its covariate space. A good final model, then, should rank highly in all these local competitions.

Averaging over z for fixed x. Another version with some appeal is to assess models in terms of how well they perform for a given covariate x, averaged across all likely z values. To indicate in one particular fashion how this may be handled, let

$$w_i \propto f(z_i - \xi_x, \Sigma_x),$$

in terms of the density f of the multinormal density $\mathrm{N}_q(0, \Sigma_x)$, where ξ_x and Σ_x are the estimated mean and variance matrix for z given x.

Dimension reduction for discriminant analysis. Assume that a classifier is to be constructed for discriminating between two populations 0 and 1 with prior frequencies π_0 and π_1. For given models for the two populations, say, $f_0(x, \theta, \gamma)$ and $f_1(x, \theta, \gamma)$ in terms of the feature vector x, where θ signifies protected and γ open parameters, the optimal classifier says "1" if $\pi_1 f_1(x) > \pi_0 f_0(x)$ and "0" otherwise. Using ordinary direct likelihood methods to estimate parameters and deciding on a good submodel, e.g., via the AIC, is often not adequate, because effort is spent to estimate densities well where it does not really matter. What counts for good classification is to estimate log-densities well where $\log f_1(x) - \log f_0(x)$ is close to the threshold $c = \log(\pi_0/\pi_1)$. The wFIC may be used for such purposes, taking as a departure point loss functions of the type

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

$$\sum_{i=1}^{n_0} w_0(x_{0,i})\{\log \hat{f}_0(x_{0,i}) - \log f_0(x_{0,i})\}^2$$

$$+ \sum_{i=1}^{n_1} w_1(x_{1,i})\{\log \hat{f}_1(x_{1,i}) - \log f_0(x_{1,i})\}^2,$$

with weight functions w_0 and w_1 chosen in a suitable fashion such that weights are large in the (estimated) area where the log-density difference is close to the threshold c. Here $x_{0,i}$ and $x_{1,i}$ denote training vectors from populations 0 and 1.

Robust weights. There is a whole area of research addressing issues related to robust model choice. Robustness is concerned with the downweighting (and sometimes identification) of data vectors that are "outliers" or that may exert too strong an influence on maximum likelihood estimators. The methodology developed in this paper sticks to the maximum likelihood estimators, as such, but the weight function of the wFIC allows downweighting schemes that make model selection less dependent on extreme data vectors. One version of this would be

$$w(x_i, z_i) = h(\|(x_i, z_i) - (x_0, z_0)\|) \quad \text{for } i = 1, \ldots, n,$$

where the norm in question could measure a suitable distance from covariate vectors to some robustly identified center location (x_0, z_0) and where $h(u)$ could be taken as 1 over a broad interval but made to go to zero beyond that interval. Another choice is $h(u) = \exp(-c|u|)$ for a perhaps small value of c. Such a scheme assures robustness with respect to extreme or overly influential covariate vectors.

Note before we come to the next point that the wFIC methodology also works with weights more general than $w_i = w(x_i, z_i)$, as long as the $\Omega_{n,w}$ matrix of (8) converges in probability. Thus weights of the form $w_i = h(\text{res}_i)$, functions of suitably defined GLM residuals, are allowed. One such version, among several, is

$$w_i = h(\text{res}_i) = \begin{cases} 1 & \text{if } |\text{res}_i| \le c, \\ \text{c}/|\text{res}_i| & \text{if } |\text{res}_i| > c, \end{cases}$$

where $\text{res}_i = (Y_i - \hat{\xi}_i)/\hat{\sigma}_i$, featuring (perhaps robustified) estimates of mean and standard deviation for Y_i. Weights of this type are discussed, in a rather different context of robust linear regression, by Ronchetti and Staudte (1994), who also find that $c = 1.345$ is a reasonable default value for this weight function.

Weights in general regression models. In the more general framework of Section 4, selecting weights appropriate for the occasion amounts to specifying first the relevant class of focus estimands $\mu(u)$ and then a weight function W for the u index. This is tantamount to constructing the weight matrix A of (13); see the illustration in Section 7.1.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

7. ILLUSTRATIONS

7.1. Quantiles for Wage Distributions

For the following illustration of our methodology we shall consider the data set cps85.data, available from the Internet, pertaining to a random sample of 528 individuals from the 1985 Current Population survey by the U.S. Census Bureau; for a brief discussion, see Hansen (2007, p. 31). Eight covariates are recorded for each individual, along with the person's hourly wage (in dollars). We order these covariates as x_1, length of education, in years; x_2, gender (1 if female, 0 if male); x_3, potential labor market experience, in years; x_4, union status (1 if in union job, 0 if not); z_1, region of residence (1 if south, 0 if not); z_2, indicator for nonwhite and non-Hispanic; z_3, indicator for Hispanic; and z_4, indicator for being married. The statistical task is to model and then analyze, how these influence the salary level. Our notation reflects the modeling decision to keep the four x_j's protected (viewing it as prior knowledge that these variables have a bearing on salary) whereas the z_j's may or may not enter the finally selected model. Because salary distributions tend to be log-normal, and the data exhibit variance heterogeneity, our widest model takes log-salary Y_i for individual i to be normally distributed with mean ξ_i a linear function in the eight covariates and with standard deviation σ_i, also allowed to depend on covariates.

Write $\xi_i = x_i^{\mathrm{t}}\beta + z_i^{\mathrm{t}}\gamma$, where in our application $x_i = (1, x_{i,1}, \ldots, x_{i,4})^{\mathrm{t}}$ and $z_i = (1, z_{i,1}, \ldots, z_{i,4})^{\mathrm{t}}$, and $\sigma_i = \sigma \exp(v_i^{\mathrm{t}}\kappa)$, where v_i depends on the covariates. Here we normalize the v_i's so as to have mean zero, for simplicity of both computation and interpretation; thus σ is the standard deviation for Y_i's for which v_i is equal to the average value zero. We shall in fact take $\sigma_i = \sigma \exp(\kappa_1 x'_{i,1} + \kappa_3 x'_{i,3})$, with $x'_{i,1} = (x_{i,1} - \bar{x}_1)/\mathrm{sd}(x_1)$ and $x'_{i,3} = (x_{i,3} - \bar{x}_3)/\mathrm{sd}(x_3)$. The narrow model has $p = 1 + 5$ parameters (σ and four β_j's) whereas the wide model has an additional $q = 4 + 2$ parameters (four γ_j's and two κ_j's). A total of $2^6 = 64$ candidate models is considered, corresponding to setting subsets of the six extra parameters to zero.

After a careful scrutiny of the data we removed four outlying individuals (salary of nos. 247, 270, and 348 is too high and salary of no. 502 is too low) to avoid parameter estimates and hence model selection mechanisms being overly influenced by highly unusual data values (outside the 0.999 probability range under the assumed regression structure). We first computed AIC values for all models, and the five best are 12456, 123456, 1456, 13456, and 145. In this notation, "145," e.g., means including $\gamma_1, \gamma_4, \kappa_1$ in the model while setting $\gamma_2, \gamma_3, \kappa_3$ to zero, and "0" and "123456" correspond to the two extreme models of the narrow and the wide. That variance heteroskedasticity is called for is clear from the fact that the best variance homogeneous model (124 actually) is only ranked no. 25 on the AIC preference list.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

To compute FIC and wFIC values we start with the log-densities of the model,

$$-\frac{1}{2\sigma^2\exp(2v_i^{\mathrm{t}}\kappa)}(y_i-x_i^{\mathrm{t}}\beta-z_i^{\mathrm{t}}\gamma)^2-\log\sigma-v_i^{\mathrm{t}}\kappa-\frac{1}{2}\log(2\pi).$$

Computing derivatives with respect to $\sigma,\beta,\gamma,\kappa$ at the null model where $\gamma=0$ and $\kappa=0$ yields $(\varepsilon_i^2-1)/\sigma,\varepsilon_i x_i/\sigma,\varepsilon_i z_i/\sigma,(\varepsilon_i^2-1)v_i$, in terms of $\varepsilon_i=(Y_i-\xi_i)/\sigma_i$. Because these are independently and identically distributed (i.i.d.) standard normal, one finds the normalized information matrix

$$J_n=n^{-1}\sum_{i=1}^n \mathrm{Var}\begin{pmatrix}(\varepsilon_i^2-1)/\sigma\\ \varepsilon_i x_i/\sigma\\ \varepsilon_i z_i/\sigma\\ (\varepsilon_i^2-1)v_i\end{pmatrix}=\begin{pmatrix}2/\sigma^2 & 0 & 0 & 0\\ 0 & \Sigma_{n,00}/\sigma^2 & \Sigma_{n,01}/\sigma^2 & 0\\ 0 & \Sigma_{n,10}/\sigma^2 & \Sigma_{n,11}/\sigma^2 & 0\\ 0 & 0 & 0 & 2L_n\end{pmatrix},$$

where

$$\Sigma_n=\frac{1}{n}\sum_{i=1}^n\begin{pmatrix}x_i\\ z_i\end{pmatrix}\begin{pmatrix}x_i\\ z_i\end{pmatrix}^{\mathrm{t}}\quad\text{and}\quad L_n=\frac{1}{n}\sum_{i=1}^n v_i v_i^{\mathrm{t}}.$$

These are of sizes 9×9 and 2×2 in the application. For a given focus estimand $\mu=\mu(\sigma,\beta,\gamma,\kappa)$, the general method of Section 4 leads to

$$\omega=\begin{pmatrix}0 & \Sigma_{n,10}\\ 0 & 0\end{pmatrix}\begin{pmatrix}2 & 0\\ 0 & \Sigma_{n,00}\end{pmatrix}^{-1}\begin{pmatrix}\partial\mu/\partial\sigma\\ \partial\mu/\partial\beta\end{pmatrix}-\begin{pmatrix}\partial\mu/\partial\gamma\\ \partial\mu/\partial\kappa\end{pmatrix}$$
$$=\begin{pmatrix}\Sigma_{n,10}\Sigma_{n,00}^{-1}\partial\mu/\partial\beta-\partial\mu/\partial\gamma\\ -\partial\mu/\partial\kappa\end{pmatrix}.$$

Also, $K_n=\mathrm{diag}(\sigma^2\Sigma_n^{11},\frac{1}{2}L_n^{-1})$, with Σ_n^{11} the lower right 4×4 block of the inverse of Σ_n. From this we may compute all 64 $\hat K_{n,S}=(\pi_S\hat K_n^{-1}\pi_S^{\mathrm{t}})^{-1}$ matrices, involving the maximum likelihood estimate $\hat\sigma$ computed from the widest model, and all 64 $G_{n,S}=\pi_S^{\mathrm{t}}\hat K_{n,S}\pi_S\hat K_n^{-1}$ matrices. Actually, the scale factors cancel each other for $G_{n,S}$, and so these matrices are observed without estimation noise.

The apparatus of Section 4 may now be applied for each given class of relevant estimands, such as probabilities of low wages as a function of covariates within certain strata of workers, the ratio of men's wages to women's wages, etc. Here we choose to consider quantiles of the wages per hour distribution for given covariates, i.e.,

$$\mu(x,z)=\exp\{x^{\mathrm{t}}\beta+z^{\mathrm{t}}\gamma+c\sigma\exp(v^{\mathrm{t}}\kappa)\},$$

where v is determined by x, as detailed previously, and c is the appropriate normal quantile. We shall focus on married women in unionized work, living

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

TABLE 1. Variable selection for quantiles of wage distributions. The table presents the six best models according to AIC, wFIC with $x_{3,0} = 3$, and wFIC with $x_{3,0} = 20$. See the text for more explanation.

		wFIC, $x_{3,0} = 3$			wFIC, $x_{3,0} = 20$		
	AIC	0.95	0.50	0.05	0.95	0.50	0.05
1	12456	56	0	0	5	0	0
2	123456	356	5	5	0	6	6
3	1456	256	6	6	35	5	5
4	13456	6	56	56	56	56	56
5	145	5	36	3	6	3	3
6	1245	125	356	35	356	35	36

in the south region, with both race indicators $z_2 = z_3 = 0$, and we consider the salary quantile as a function of the length x_1 of the woman's education. In other words, our focus is the function $\mu(x_1) = \mu(x_1, \sigma, \beta, \gamma, \kappa)$ emerging from the preceding formula for covariate vectors $(1, x_1, 1, x_{3,0}, 1, 1, 0, 0, 1)^t$, when x_1 traverses the sampled values (from 6 to 18 years), for some fixed length of work experience $x_{3,0}$. There are 2^6 different curve estimates $\hat{\mu}_S(x_1)$, one for each submodel.

Table 1 gives the best six AIC models, along with the best six wFIC models, for the 0.95, 0.50, 0.05 quantile functions and for $x_{3,0}$ set equal to, respectively, 3 years and 20 years, for the stratum of married women described previously. Figure 1 gives quantile curves for the case of $x_{3,0}$ being 3 years, with estimates stemming both from the AIC-best model and for those found by the wFIC mechanism. Similar analyses for other functions of other covariates and for other strata of workers can easily be carried out, in each case leading to different lists of top models. That such versatility and flexibility of method are of importance and relevance in econometric model selection situations is also discussed in Hansen (2005). Note that all $n = 524$ data vectors are used to estimate the wage quantiles for each stratum.

7.2. Diabetic Retinopathy Data

The Wisconsin Epidemiologic Study of Diabetic Retinopathy (see Klein, Klein, Moss, Davis, and DeMets, 1984) provides information for studying how diabetic retinopathy may be influenced by several other measurements. The data set consists of patient information for 348 men and 343 women. The binary outcome variable $Y = 0$ indicates that there is no or only mild nonproliferate retinopathy for both eyes. A value $Y = 1$ corresponds to moderate to severe nonproliferate retinopathy, or proliferate retinopathy, for one or both eyes. Pre-

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

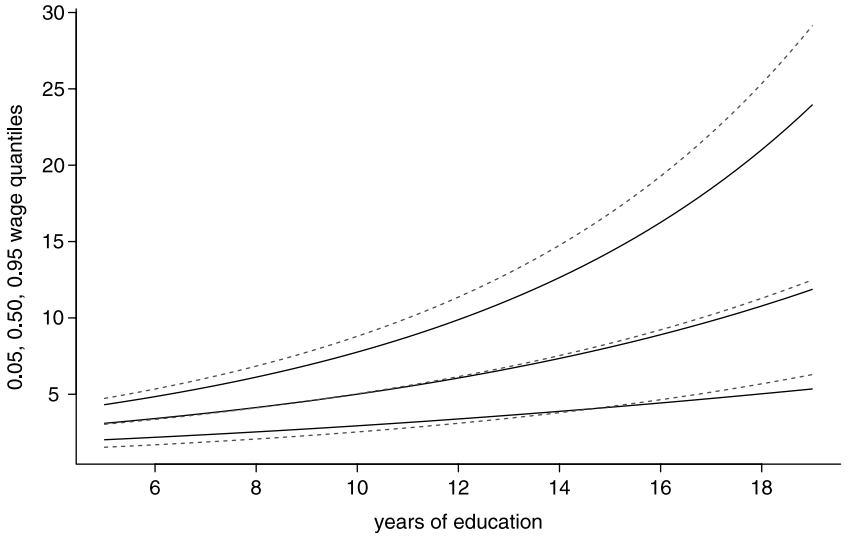


FIGURE 1. Quantiles for wages: estimates of the 0.95, 0.50, 0.05 quantile curves for hourly wages for the stratum of women described in the text, as a function of the length of their education, when the work market experience $x_{3,0}$ is fixed equal to 3 years. The AIC-model-based estimates are given with full lines, whereas the dotted lines refer to estimates based on the best wFIC model, with better estimated precision.

vious analysis of similar data, in addition to prior knowledge among ophthalmologists, suggests that duration of diabetes in years (x_1) and presence of macular edema in at least one eye (x_2) are of direct relevance for the outcome of Y, and so these are protected covariates in each of the models that we shall consider. Further covariates, seen as candidates for inclusion in a final model, are percentage of glycosylated hemoglobin (z_1), body mass index (z_2), pulse rate in beats per 30 seconds (z_3), sex $(z_4$, with 1 for male and 0 for female), presence of urine protein (z_5), and area of residence $(z_6$, 1 or 0 for urban or rural).

A logistic regression model is used for the analysis. This means that the $\hat{V}$ matrix is diagonal with elements $\hat{p}_i(1-\hat{p}_i)$, where $\hat{p}_i = H(\hat{\beta}_0 + x_i^{\mathrm{t}}\hat{\beta} + z_i^{\mathrm{t}}\hat{\gamma})$ with $H(u) = \exp(u)/\{1+\exp(u)\}$ the logistic transform. Here β_0 is the intercept, whereas $x_i = (x_{i,1}, x_{i,2})^{\mathrm{t}}$ and $z_i = (z_{i,1}, \ldots, z_{i,6})^{\mathrm{t}}$ are the covariate parts for individual i, for $i = 1, \ldots, n = 691$. At the full model, we obtain estimates -6.344, 0.084, 2.036, 0.069, 0.031, 0.039, 0.119, 1.035, 0.241 of the nine parameters, with corresponding standard errors 1.148, 0.011, 0.411, 0.054, 0.026, 0.017, 0.231, 0.276, 0.228. Here the intercept and variables x_1, x_2, z_5 are highly significant, and variable z_3 is associated with a p-value of 0.022. The other variables are individually not significant.

We perform model selection among all six z variables and allow for all possible subsets of the full model, leading to 64 candidate models. As model selec-

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

TABLE 2. Values of the wFIC, where weights are indicator values for men (1st column), indicator values for women (4th column), and the GLM weights for logistic regression (7th column), together with the selected covariates, in addition to the estimated error rate using leave-one-out cross-validation for 691 predictions. All indicated logistic regression models contain an intercept term and variables x_1 and x_2.

Men			Women			GLM weights		
wFIC_M	z-variables	CV	wFIC_F	z-variables	CV	wFIC_glm	z-variables	CV
21.201	3, 5	0.185	24.828	3, 5	0.185	5.688	3, 5	0.185
23.270	2, 3, 5	0.184	24.974	3, 5, 6	0.181	6.032	3, 5, 6	0.181
24.230	1, 3, 5	0.184	26.155	1, 3, 5	0.184	6.149	2, 3, 5	0.184
24.393	3, 5, 6	0.181	26.160	2, 3, 5	0.184	6.163	1, 3, 5	0.184
27.590	1, 2, 3, 5	0.177	28.072	1, 3, 5, 6	0.177	7.000	1, 2, 3, 5	0.178
27.767	2, 3, 5, 6	0.178	28.616	2, 3, 5, 6	0.178	7.000	1, 3, 5, 6	0.177
28.680	1, 3, 5, 6	0.177	28.673	1, 2, 3, 5	0.178	7.000	2, 3, 5, 6	0.178
29.982	3, 4, 5	0.184	31.096	3, 4, 5	0.184	7.465	3, 4, 5	0.184
31.186	2, 3, 4, 5	0.185	31.636	3, 4, 5, 6	0.181	7.776	3, 4, 5, 6	0.181
31.345	1, 3, 4, 5	0.184	32.945	2, 3, 4, 5	0.185	7.921	2, 3, 4, 5	0.185

Note: CV = cross-validation.

tion criteria we take first $\text{wFIC}_\text{M}(S)$ with indicator weights $w_i = I(i \text{ is male})$ and next $\text{wFIC}_\text{F}(S)$ with weights $w_i = I(i \text{ is female})$. Note that the values of $\text{wFIC}(S)$ are computed using the complete data set; we are not splitting the data set for model selection. We furthermore consider wFIC_glm with the GLM weights for logistic regression, which translates to taking $w_i = \hat{v}_i = \hat{p}_i(1 - \hat{p}_i)$. Table 2 gives for the three wFIC criteria the top 10 models, identified by their z covariates, along with their associated wFIC scores. The versatility of our method could be further illustrated by choosing other weight functions; thus $w_i = \hat{p}_i$ would give extra emphasis to models that do well for patients with high $Y = 1$ probabilities, whereas $w_i = 1 - \hat{p}_i$ corresponds to finding models that do better for those with low $Y = 1$ probabilities.

For this particular illustration, the best wFIC model turns out to be the one with variables z_3 and z_5, for each of the three setups. This happens to correspond precisely to those variables that were found individually significant in the full model at the 5% level. For other weight schemes, different submodels are selected. The three wFIC schemes differ in the ranking of the following top candidate models. From Table 2, there is some indication that the body mass index (variable z_2) is more important for men, whereas the area of residence (z_6) may be more important for women. Thus we learn that, e.g., body mass index influences Y in possibly different ways for men and for women, which may be taken into account for building a final model or for carrying out the most appropriate model averaging for prediction; see the discussion in Sec-

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

tion 9. We note further that the model ranking based on wFIC_{glm} is very similar to that based on the AIC.

It is important to note that the wFIC method aims at obtaining a good approximation to the full probability surface function $p(x, z) = P(Y = 1|x, z)$, across values of (x, z), with appropriate and user-defined weights $w(x, z)$. This is different from the aim of finding a model that produces the lowest possible error rate when used for discriminant analysis purposes. Such an aim relates to the more crude task of sorting individuals into "Y is probably 1" and "Y is probably 0" classes, based on (x, z). Thus it is not surprising that some submodels have slightly better error rate properties than those found best from the wFIC perspective. Examining estimated error rates for the 64 different logistic regression–based classifiers (via leave-one-out cross-validation), the primary feature is that they are all rather similar, spanning a tight range from 0.175 (121 errors of 691) to 0.190 (131 errors of 691). The tentatively best models, from this classification accuracy point of view, are 1256 (error rate 0.175), then 1356, 13456, 125, 12456, 12356 (each with error rate 0.177); here "1256" identifies the model with variables z_1, z_2, z_5, z_6, etc. Entirely similar results are found using, e.g., ten- or fivefold cross-validation. For some further discussion of predictive properties of different models we refer to Claeskens, Croux, and Van Kerckhoven (2006).

7.3. Highway Data

To illustrate robust downweighting, we use Hoffstedt's highway data; see also Weisberg (2005, Sect. 7.2). This data set is used to explain the 1973 accident rate per million vehicle miles as a function of several variables. There are 39 observations made. In every model we include an intercept term and x_1, the length of the highway segment in miles. Additional variables to choose from, when searching for an overall good model, are average daily traffic count in thousands (z_1), truck volume as a percentage of the total volume (z_2), total number of lanes of traffic (z_3), number of access points per mile (z_4), number of signalized interchanges per mile (z_5), number of freeway-type interchanges per mile (z_6), speed limit in 1973 (z_7), lane width, in feet (z_8), width of the outer shoulder on the roadway (in feet) (z_9), and finally an indicator of the type of roadway or the source of funding for the road (z_{10}).

Based on robust C_p model selection, Ronchetti and Staudte (1994) support the model that includes, in addition to x_1, the variables z_5, z_6, z_7, and z_{10}, and also the model with additional variables z_2, z_3, z_4, and z_9. Here we construct weights $w_1, \ldots, w_{39}$ based on the robustification method outlined in Section 6.3; these rely on initially used robust estimators for the regression coefficients obtained in the full model. Five of the observations receive a weight that is smaller than one; all other observations get weight one. The weights that differ from one are 0.903, 0.568, 0.436, 0.577, 0.811. These weights are then used in the wFIC(S) construction and lead to preferring the model with the two vari-

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

TABLE 3. Values of the wFIC for the highway data, together with the selected variables and also the error rate in 39 predictions obtained using leave-one-out cross-validation. Also presented are the 10 best models according to leave-one-out cross-validation.

wFIC			Error rate CV	
z-variables	wFIC	CV	z-variables	CV
10	29.04	1.51	2, 4, 7, 10	0.965
4	29.23	1.03	2, 3, 4, 7, 10	0.980
8	29.51	1.54	2, 4, 7	0.988
7	29.72	1.10	1, 2, 4, 7, 10	0.989
8, 10	44.00	1.58	2, 3, 4, 7	0.991
3, 10	44.14	1.56	3, 4, 7	0.994
4, 10	44.16	1.05	1, 2, 4, 7	0.995
8, 9	44.20	1.40	4, 7, 10	0.997
9, 10	44.20	1.35	2, 4, 5, 7, 10	0.998
7, 10	44.23	1.13	1, 2, 3, 4, 7, 10	1.000

Note: CV = cross-validation.

ables x_1 and z_{10}. For this particular example, the model selected by wFIC is more parsimonious than the model suggested by the robust C_p.

Table 3 presents, in a manner similar to Table 2, the 10 best wFIC values, the corresponding models, and the corresponding leave-one-out prediction error $(1/39)\sum_{i=1}^{39}|y_i - x_i^{\mathrm{t}}\hat{\beta}_{-i}|$ computed in that model. Each time, the values (y_i, x_i) are left out when computing $\hat{\beta}_{-i}$, where we write $x_i = (1, x_{i,1}, z_{i,1}, \ldots, z_{i,10})^{\mathrm{t}}$. The 10 best models according to the leave-one-out prediction error are also presented. The wFIC with robustified weights has larger estimated prediction error than the best models obtained with leave-one-out cross-validation. The model selected by the wFIC method tends for this example to be more parsimonious than the model selected by cross-validation. One should keep in mind that minimizing prediction error was not the goal for the computed values of the FIC.

8. PERFORMANCE OF THE wFIC

We have developed a model selection criterion that aims at doing well for given loss functions of the averaging type (12). In particular, different focus type parameters, and also different weighting strategies W_n, lead to different selection mechanisms, unlike the AIC and most other strategies that work in "overall modus." It is also important to study *the consequences* of given model selection strategies in terms of final performance; point estimates and confidence intervals, etc., are really the second step of a two-step strategy.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

8.1. The Real Risk Function

Consider estimators of the type $\mu^*(u) = \hat{\mu}_{\hat{S}}(u)$, where $\hat{S}$ is the submodel selected by, say, AIC or the wFIC. A more general class is that of

$$\mu^*(u) = \sum_S c_n(S)\hat{\mu}_S(u), \tag{16}$$

where the model weights are allowed to depend upon data and where the sum is over all candidate models (e.g., all inside a nested sequence). The AIC method corresponds to $c_n(S) = 1$ for precisely the set S selected by the AIC, with $c_n(S) = 0$ for the other models, etc. Suppose that the collection of model weights $c_n(S)$ is such that they converge jointly to $c(S|D)$, in the limit experiment associated with variables (M, D) in (11). For this broad class of estimators-after-selection, and more generally model average estimators,

$$\sqrt{n}\{\mu^*(u) - \mu_{\text{true}}(u)\} \xrightarrow{d} \Lambda_0(u) + \omega(u)^{\mathrm{t}}\{\delta - \delta^*(D)\}, \tag{17}$$

where $\delta^*(D) = \sum_S c(S|D)G_S D$. This follows essentially via simultaneous convergence of all variables involved in (11) and the random weights, in combination with the continuity theorem of weak convergence. The important consequence is that we may describe what happens to *the real random loss* $L_n^* = n\int\{\mu^*(u) - \mu_{\text{true}}(u)\}^2\,\mathrm{d}W_n(u)$ incurred by the statistician's strategy, as opposed to only the random loss for strategies using predetermined models. For the real loss we have

$$L_n^* \xrightarrow{d} \int [\Lambda_0(u) + \omega(u)^{\mathrm{t}}\{\delta - \delta^*(D)\}]^2\,\mathrm{d}W(u).$$

From this we also infer the large-sample limit of *the real risk function* (again after subtracting the irrelevant $\int \tau_0^2\,\mathrm{d}W$ term),

$$\begin{aligned}\text{risk}(\delta) &= \mathrm{E}\int [\omega(u)^{\mathrm{t}}\{\delta - \delta^*(D)\}]^2\,\mathrm{d}W(u)\\ &= \int \omega(u)^{\mathrm{t}}\{\text{bias}(\delta)\text{bias}(\delta)^{\mathrm{t}} + \text{var}(\delta)\}\omega(u)\,\mathrm{d}W(u)\\ &= \mathrm{Tr}[A\{\text{bias}(\delta)\text{bias}(\delta)^{\mathrm{t}} + \text{var}(\delta)\}], \end{aligned}\tag{18}$$

in which $\text{bias}(\delta)$ and $\text{var}(\delta)$ are used to denote the q-dimensional bias function and $q \times q$ variance matrix of $\delta^*(D)$. The specifics of the model selection or average strategy (16) determine these bias and variance functions; recall that $D \sim \mathrm{N}_q(\delta, K)$. Also, $A = \int \omega\omega^{\mathrm{t}}\,\mathrm{d}W$, as before. We may note that (13) is a special case of the general result (18).

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

Result (18) covers both the AIC and wFIC strategies, along with smoothed versions thereof (and the original FIC scheme corresponds to the weight measure being concentrated in a single point). That the AIC method is inside the required class essentially follows from (15), which implies that there is joint convergence, with $c_{n,\mathrm{aic}}(S) \to_d c_{\mathrm{aic}}(S|D)$, the latter defined as the indicator for the set S for which $\mathrm{aic}(S) = D^{\mathrm{t}}K^{-1}\pi_S^{\mathrm{t}}K_S\pi_S K^{-1}D - 2|S|$ is larger than all other $\mathrm{aic}(S')$. Similarly, the wFIC method uses random weights $c_{n,\mathrm{fic}}(S)$ that converge jointly to $c_{\mathrm{fic}}(S|D)$, defined as the indicator for the set S for which

$$\mathrm{I}^*(S) + \mathrm{II}^*(S) = \max[\mathrm{Tr}\{A(I_q - G_S)(DD^{\mathrm{t}} - K)(I_q - G_S)^{\mathrm{t}}\}, 0] + \mathrm{Tr}(A\pi_S^{\mathrm{t}}K_S\pi_S)$$

is smaller than all other $\mathrm{I}^*(S') + \mathrm{II}^*(S')$. This follows because $\hat{\delta} \to_d D$ whereas the other quantities are estimated consistently in our local $\gamma = \gamma_0 + \delta/\sqrt{n}$ neighborhood formulation.

8.2. Comparing AIC with wFIC

The performances of different model selection or model averaging strategies may be compared using (18). Space does not allow us any deep investigation into the pros and cons of different methods, but we shall attempt to explain here that there will be many situations where the wFIC method will perform better than the AIC, for large regions of the parameter space.

Let us assume, for simplicity of presentation, that the K matrix is diagonal, with diagonal elements, say, $\kappa_1^2, \ldots, \kappa_q^2$, and that the list of candidate models corresponds to all 2^q subsets of $\{1, \ldots, q\}$. The right-hand side of (15) then becomes $\mathrm{aic}(S) = \sum_{j\in S}(D_j^2/\kappa_j^2 - 2)$, and one finds that (the limit experiment version of) AIC selects the set of j for which $|D_j|/\kappa_j \geq \sqrt{2}$. The AIC therefore acts as a pretest screener for each nonprotected parameter, actually using significance level 0.157 per parameter; the performance of other pretest methods, with harder or softer screening thresholds, can also be analyzed via (17). We can therefore compute the $\mathrm{bias}_{\mathrm{aic}}(\delta)$ and $\mathrm{var}_{\mathrm{aic}}(\delta)$ for the AIC method, via $\delta^*_{\mathrm{aic},j} = D_j I\{|D_j|/\kappa_j \geq \sqrt{2}\}$ for $j = 1, \ldots, q$. These considerations determine the AIC performance as

$$\mathrm{risk}_{\mathrm{aic}}(\delta) = \mathrm{bias}_{\mathrm{aic}}(\delta)^{\mathrm{t}}A\mathrm{bias}_{\mathrm{aic}}(\delta) + \sum_{j=1}^{q} a_{j,j}\mathrm{var}_{\mathrm{aic},j}(\delta).$$

Note that AIC selects these j regardless of the context-driven loss function, i.e., regardless of the matrix $A = \int \omega\omega^{\mathrm{t}}\,\mathrm{d}W$.

The wFIC, on the other hand, selects components with specific regard to A, i.e., to what is deemed important when setting up the loss function in the first place. One may show here, under the diagonal K assumption, that wFIC selects

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

S to give smallest value of $h(S) = \sum_{j\in S} 2a_{j,j}\kappa_j^2 + D_{S^c}^{\mathrm{t}} A_{S^c} D_{S^c}$, with S^c the complement of S. Note, e.g., that if the loss function gives an A matrix with some components j that do not matter, then wFIC does not include them, whereas AIC is programmed to take these on board also, even when this gives an extra yoke of sampling variability. Similarly, if some $a_{j,j}$ components are much larger than others, potentially contributing harshly to the risk in (18), then wFIC tends to avoid including the associated γ_j's in the model.

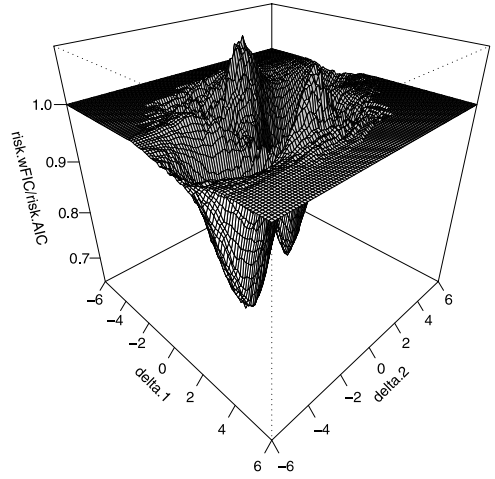
We compared the AIC to the wFIC in the limit experiment situation. We took $q = 2$, which allows us to present the results graphically, and we let δ_1, δ_2 take values in an equispaced grid of 200 points between -6 and 6. For each such combination of "true" values of $\delta = (\delta_1, \delta_2)$, 2,000 values for D were simulated from a bivariate standard normal distribution. We proceeded by performing wFIC and AIC model selection for each simulated value of the vector D to obtain the value of the indicator weight $c(S|D)$, which is either 0 (if the model is not selected) or 1 (when that model is selected) for the four candidate models, which we label "0" (none of γ_1, γ_2 included), "1" (γ_1 included, γ_2 not), "2" (γ_1 excluded, γ_2 included), "12" (both included). This allowed us to compute values of $\mathrm{bias}(\delta) = \mathrm{E}(\delta^*(D) - \delta)$ and of $\mathrm{var}(\delta) = \mathrm{var}(\delta^*(D))$ for wFIC and AIC model selection. The A matrix taken was

$$A = \begin{pmatrix} 10 & 3 \\ 3 & 2 \end{pmatrix}.$$

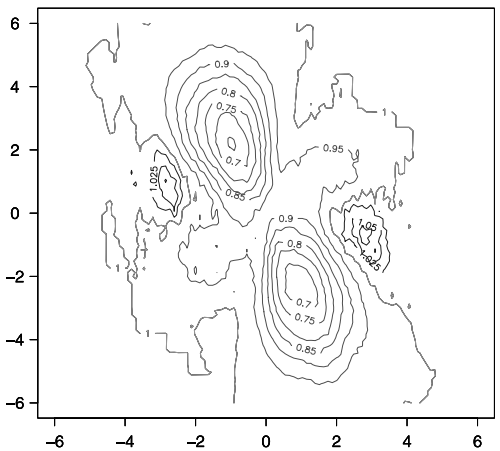
Figure 2 depicts the ratio of the risk values for wFIC by those of AIC. Values below 1 indicate preference for the wFIC, whereas values above 1 are situations where AIC has lower risk. There is no absolute preference of one method above the other. The risk ratio surface converges to 1 when $\|\delta\|$ increases. The figure indicates a significant area in δ-space where wFIC has better performance. In these areas wFIC has a much lower minimal risk than that of AIC. The risk function, computed via equation (18), takes values between 5.216 and 229.914 for wFIC. For AIC the minimal simulated risk is larger, with value 5.704, whereas the upper bound is the same. Computations for the risk ratio of wFIC/AIC lead to a lower bound of 0.643; i.e., the risk for wFIC is only 0.643 times that of the AIC for certain values in δ-space, whereas the maximum of this ratio is equal to 1.090, indicating only a slight advantage for the AIC in those areas where it has a better risk.

Because the limit version of AIC selects those variables with index j for which $|D_j|/\kappa_j \geq \sqrt{2}$, the areas where AIC selects a certain model are easy to depict; see Figure 3a. For the same grid of D_1, D_2 values we can use wFIC model selection and obtain information on the D_1, D_2 values for which wFIC selects a certain model. This leads to the areas of Figure 3b for the A matrix above. To judge the quality of these areas we need to compare those with the actual areas that minimize the risk function for each S. Therefore we employ the w-risk

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press



(a)



(b)

FIGURE 2. (a) Simulated risk ratio wFIC/AIC for model selection in a situation with $q = 2$, K the identity matrix, and the matrix A as given in the text. Values below 1 indicate preference for wFIC; values above 1 indicate preference for AIC. (b) Contour plots of the risk ratio wFIC/AIC.

expression in equation (13) to obtain the areas in δ-space where each model should be selected to have minimal w-risk. These areas are depicted in Figure 3c for the same A matrix as used above. Even though the wFIC does not match these areas perfectly, it comes much closer than, e.g., the AIC.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

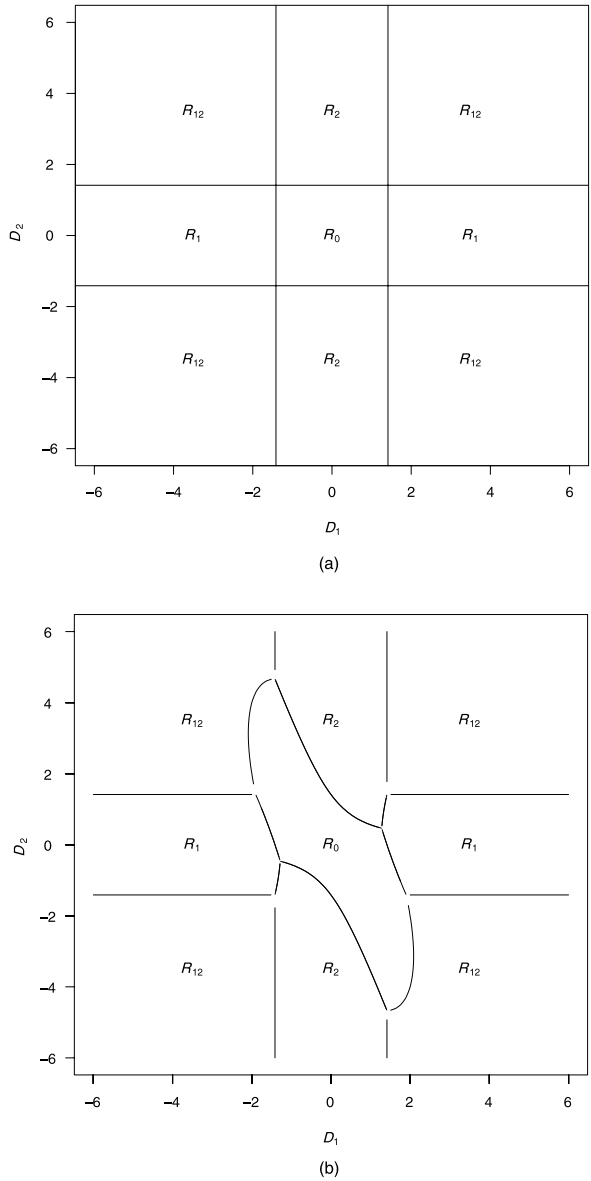


FIGURE 3. Regions for R for (D_1, D_2) where the different models are selected (cases (a) and (b)), and regions S for (δ_1, δ_2) where different models are optimal (case (c)), for the A matrix given in the text. Here R_{12} denotes the full model with both variables, R_1 is the model only including the first variable, R_2 only includes the second variable, and R_0 is the narrow model, not including any variable. (a) Areas where AIC selects a certain model, (b) areas for the wFIC, and (c) optimal areas where the indicated model is best in minimizing the w-risk function. (*Figure continues on facing page*.)

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

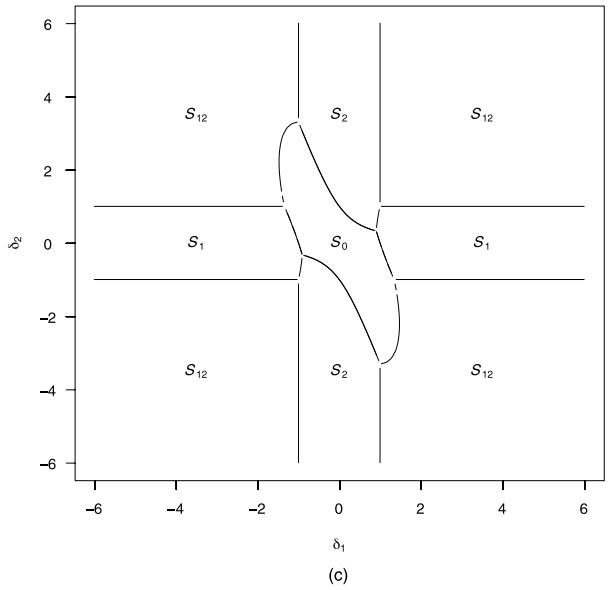


FIGURE 3. *Continued.*

9. CONCLUDING COMMENTS

This paper has provided a fair middle ground between the extremes "blind model selection" (where a model is found via, say, AIC or BIC, without any particular regard to the actual use of the model after selection) and "fully focussed selection" (where a model is selected to perform optimally for a given estimand). We have seen that special versions of the wFIC are essentially equivalent to the AIC method, and so our methods may be seen as suitable generalizations of the AIC for use in situations where the context of one's modeling and analysis dictates more specific weighting schemes than the default ones. Subsequently we give some concluding remarks pertaining to themes related to but outside the main scope of the present paper.

In this paper we chose to concentrate on the generalized linear models framework, where methods have a particularly clear structure, but it is clear that the methods can be generalized to many other regression structures, as per Section 4. Thus parametric regression models for multidimensional data, and for hazard rates with censored data, can be handled with essentially the same methods. Extensions to the semiparametric Cox model are less immediate but can be accomplished via methods of Hjort and Claeskens (2006).

Our paper has developed strategies that for each model rely on maximum likelihood (or asymptotically equivalent) estimators. There is a need for generalizing methods and results to more robust strategies, e.g., involving

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

M-estimators. This is entirely possible, but it requires more work and will result in algebraically and structurally somewhat less elegant methods.

Though this paper has been specifically concerned with model selection, methods of Hjort and Claeskens (2003a) can be applied to provide model average procedures, say, of the type $\hat{\mu} = \sum_S c(\text{wFIC}(S))\hat{\mu}_S$, a data-dependent average across the estimators of the individual models, with

$$c(\text{wFIC}(S)) = \exp\{-\tfrac{1}{2}\lambda\ \text{wFIC}(S)\}\Big/\sum_{S'}\exp\{-\tfrac{1}{2}\lambda\ \text{wFIC}(S')\}.$$

Here λ is an algorithmic parameter, with small λ corresponding to near uniform weighting, whereas larger λ means giving nearly full weight to the model that wins the wFIC competition. It would be natural to smooth across the most promising wFIC quantile curves, e.g., in the wage distribution study of Section 7.1. Methods of Section 8 also make it possible to study performances of such average estimators, compared to the more usual estimators-post-selection. Specifically, the result (18) for the real risk is valid for a broad class of model average estimators, including the smoothed wFIC method.

Our methods have been developed inside a framework of "first-order asymptotics" for general parametric models. It might be important to supplement such methods by suitable second-order corrections to make them work more precisely for moderate or smaller sample sizes. The content of Remark 4 of Section 3 is that the wFIC(S) expression is exactly correct for each finite $n > p + q$, when used in the linear model with nonrandom weights. This is an indication that even the first-order approximations to risks and their estimates are adequate also in other GLMs.

Our "F" in FIC and wFIC is for "focus," making our methods different in spirit and operation from the majority of other information criteria in the literature. Specifically, even though we use Fisher information matrices of different candidate models as a step in the evaluation of our criteria, our methods are not related to the Fisher information criterion developed in Wei (1992) as a continuation and refinement of earlier work on predictive minimum description length methods of Rissanen. The framework of Wei's information criterion is limited to linear regression models. The posterior information criterion (PIC) of Phillips and Ploberger (1996) is related to Wei's criterion, to Rissanen's minimum description length principle, and to the BIC but was developed to work in considerably more general frameworks, encompassing broad classes of stationary time series models. See further remarks also about other selection criteria and their interrelationships in Claeskens and Hjort (2003), Hjort and Claeskens (2003a, 2003b), and the ensuing discussion contributions.

REFERENCES

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

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press

Claeskens, G., C. Croux, & J. Van Kerckhoven (2006) Variable selection for logistic regression using a prediction focussed information criterion. *Biometrics* 62, 972–979.

Claeskens, G. & N.L. Hjort (2003) The focused information criterion (with discussion). *Journal of the American Statistical Association* 98, 900–916.

Dobson, A.J. (2002) *An Introduction to Generalized Linear Models.* Chapman and Hall.

Hansen, B.E. (2005) Challenges for econometric model selection. *Econometric Theory* 21, 60–68.

Hansen, B. (2007) Econometrics. Manuscript, Department of Economics, University of Wisconsin.

Hjort, N.L. & G. Claeskens (2003a) Frequentist model average estimators (with discussion). *Journal of the American Statistical Association* 98, 879–899.

Hjort, N.L. & G. Claeskens (2003b) Rejoinder to "The focused information criterion" and "Frequentist model average estimators." *Journal of the American Statistical Association* 98, 938–945.

Hjort, N.L. & G. Claeskens (2006) Focussed information criteria and model averaging for Cox's hazard regression model. *Journal of the American Statistical Association* 101, 1449–1464.

Klein, R., B.E.K. Klein, S.E. Moss, M.D. Davis, & D.L. DeMets (1984) The Wisconsin epidemiologic study of diabetic retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of Ophthalmology* 102, 520–526.

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

McCullagh, P. & J. Nelder (1989) *Generalized Linear Models*. Chapman and Hall.

Phillips, P.C.B. & W. Ploberger (1996) An asymptotic theory of Bayesian inference for time series. *Econometrica* 64, 381–412.

Ronchetti, E. & R.G. Staudte (1994) A robust version of Mallows' C_p. *Journal of the American Statistical Association* 89, 550–559.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.

Varian, H.R. (1975) Distributive justice, welfare economics, and the theory of fairness. *Philosophy and Public Affairs* 4, 223–247.

Wei, C.Z. (1992) On predictive least squares principles. *Annals of Statistics* 20, 1–42.

Weisberg, S. (2005) *Applied Linear Regression*, 3rd ed. Wiley.

Zellner, A. (1986) Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association* 81, 446–451.

https://doi.org/10.1017/S0266466608080201 Published online by Cambridge University Press