Predictive Model Selection

West and Harrison 1989)) and nonlinear models see

For the model selection problem in general, one can replace 1.2) and the accompan ing descriptions of distributions for various quantities b

densit to inferential use, adapting the philosoph advocated in Geisser 19 1). The imagined replication makes y and comparable; in fact, exchangeable a priori. oreover, the parameters in the model pla a minimal role under replication. It seems clear that good models, among those under consideration, should make predictions close to what has been observed for an identical experiment. The criteria below are defined with this motivation.

For a given model m, consider

$$L_m^2 = E[\quad - y)' \quad - y)] \ ,$$

where the expectation is taken with respect to the PDRE f_m. The measure L_m^2 has the decomposition

$$L_m^2 = \sum_{i=1}^{n} \left\{ [E \quad _i) - y_i]^2 + V \ r \quad _i) \right\} \ ,$$

as a sum of two components, one involving the means of the predictive distribution, and the other involving the variances. Thus a model's performance is measured b a combination of how close its predictions are to the observed data and the variabilit of the predictions. Good models will have small values of L_m^2. It is often more convenient to use the measure

$$L_m = \sqrt{L_m^2}$$

since it is a distance on the response axis, measured in the same units as the response variable. We refer to L_m as the *L criterion*.

To define the second criterion, consider

$$M_m^* = f_m \ y) \ .$$

This is the PDRE under model m, evaluated at the observed response y. good model will have a large value of M_m^*. ratio of M_m^*'s for two different models is an instance of what itkin 1991) calls the *osterior Ba es factor*. gain, to facilitate interpretation, let

$$M_m = \ M_m^*)^{-1/n}$$

which is in the units of the response variable, and small values of it indicate good models. We refer to M_m as the *M criterion*.

The third criterion we introduce for model selection is the Kullback-Leibler KL) divergence between two predictive densities. Suppose f_1 and f_2 are two densities with respect to Lebesgue measure. Then, the KL divergence between f_1 and f_2 is defined b

$$f_1, f_2) = \int \ log \left[f_1 \ x)/f_2 \ x) \right] \ f_1 \ x) \ dx \ .$$

In general, $f_1, f_2) \neq f_2, f_1)$, and $, b) \geq 0$ with equalit occurring onl if $= b$. The KL divergence has been used in the literature for a wide variet of statistical problems, and in connection with the Ba esian predictive distribution. For example, with its use itchison 19 5) shows that the predictive distribution best approximates the sampling distribution, Johnson and Geisser 1983) detect influential observations in linear regression, and cCulloch 1989) assesses the influence of model assumptions. Bhattacharjee and Dunsmore 1991) use the KL directed divergence to select variables in logistic regression.

For our purposes, suppose m_0 is a fixed model in $\mathcal{M}$ from which we measure other models. In variable selection, for instance, a natural choice for m_0 might be the full model 1.1) with all of the k predictors. Using PDRE's of m_0 and m, we define

$$_m = m_0, m) + m, m_{130000310}$$

2 Prior Distributions

with $c \geq 0$ quantif ing, in multiples of the present experiment, the importance one wishes to attach to the prior guess η_0. Thus under model m, T_m is a scalar multiple of the Fisher information matrix for $\beta^{(m)}$. Zellner's g-priors Zellner, 1986) also have this structure for the precision matrix. It has the advantage of leading to anal ticall tractable and computationall feasible solutions.

Now, we take $\beta^{(m)}|\tau$ to be normall distributed, i.e.,

$$\beta^{(m)}|\tau \sim No_{k_m}\ \mu^{(m)}, \tau\ T_m)\ . \qquad 2.5)$$

s a result of focusing on the observables, onl a few easil interpreted quantities are needed to specif the prior. In particular, the prediction η_0 is turned into a prior for $\beta^{(m)}|\tau$ for each m in an automated fashion.

Finall , the prior distribution for τ is taken to be a gamma distribution with parameters $_0/2, \gamma_0/2)$, i.e., with densit

$$\pi\ \tau)\ d\tau \propto\ \tau^{\delta_0/2-1}\ exp\{-\gamma_0\tau/2\}\ d\tau\ . \qquad 2.6)$$

For a fixed model m, 2.5) and 2.6) result in the conjugate normal-gamma prior.

With this prior and the likelihood implied b 1.2) for each m, a straightforward derivation ields

$$\sim S_n\left(n+\ _0, \eta_m, s_m^2\ I+\ 1-\gamma)P_m)\right)\ , \qquad 2.\ )$$

where $\gamma\ =\ c/\ 1+c)$, $\eta_m\ =\ P_m\ \gamma\eta_0+\ 1-\gamma)y)$, $s_m^2\ =\ \ n+\ _0)^{-1}\ q_m+\gamma p_m+\gamma_0)$, $q_m\ =\ y'\ I-P_m)y$, and $p_m=\ y-\eta_0)'P_m\ y-\eta_0)$. The PDRE in 2.2) for noninformative priors can be obtained from 2.) b formall setting $\gamma=0$, $_0=-k_m$, and $\gamma_0=0$. oreover if X_m has rank $r_m<k_m$, replace k_m b r_m in 2.) above. For brevit , an relevant expressions are given onl for the case of conjugate priors in the remainder of this article.

The L criterion under model m is now given b

$$L_m=\{\ 1+\lambda_m)q_m+\gamma\ \gamma+\lambda_m)p_m+\lambda_m\gamma_0\}^{1/2}\ , \qquad 2.8)$$

where $\lambda_m=\frac{n+(1-\gamma)k_m}{n+\delta_0-2}$. We see that L_m^2 above is a linear function of q_m and p_m. The quantit q_m is the squared length of the projection of the data onto the error space of model m, i.e., the error sum of squares for model m. The quantit p_m represents a penalt for a bad prior guess at Y. It is the squared length of the projection of the "guessing error" onto the model's column space. Under reference priors, 2.8) reduces to $L_m=(2\ n-1)\ n-k_m-2)^{-1}q_m)^{1/2}$. In this case, L_m is similar to the root mean square criterion.

To calculate the calibration number S , one can sample from the marginal distribution

$$Y\sim S_n\left(\ _0, \eta_{m^*}, \gamma_0\ _0{}^{-1}\ I+\gamma^{-1}\ 1-\gamma)P_{m^*})\right)\ ,$$

and calculate L_{m^*} with each sample. Here m^* is the model that minimizes L_m.) The standard deviations of these values provides a onte Carlo approximation to S . If one is using the reference priors in 2.1), however, it is well known that the marginal distribution of Y is improper. In this case, one could sample from the conditional distribution $Y|\tau \sim No_n\ 0, \tau\ I - P_{m^*}))$ with τ replaced b $\tilde{\tau}$, the mode of the posterior distribution of τ using m^*. The standard deviation of the resulting samples of L_{m^*} can be viewed as an approximation to $= [V\ r\ L_{m^*}|\tau = \tilde{\tau})]^{1/2}$. For large n one can obtain the anal tic approximation

$$\approx \tilde{\tau}^{-1}$$

where

$$= \frac{2}{}$$

number S . Under the improper reference priors, however, one is not guaranteed such automatic protection and hence must be careful to not include in $\mathcal{M}$ an model that can ield a

3.2 Transformation Selection

In linear regression, transformations of the predictor variables can often lead to more accurate predictions and a model that better fits the data. Box and Cox 1964) discuss transformations with an emphasis on transforming the response variable. The also mention briefl a possible Ba esian approach. It appears, however, that the literature on Ba esian transformation methods is sparse at best.

Here, we show how two of the predictive criteria can be used to select a specific member of a suitabl chosen parametric transformation famil . The criterion as defined in this article is not applicable to this problem. Consider equation 1.2) where a single model $m \in \mathcal{M}$

an intercept and the Box-Cox transformation on x. gain, to denote the dependence of the criteria on α, we write $M_m \equiv M\ \alpha)$ and $L_m \equiv L\ \alpha)$. Under the noninformative prior 2.1), $M\ \alpha)$ and $L\ \alpha)$ are equivalent and we get the minimizer $\hat{\alpha} = -1.325$, with $L\ \hat{\alpha}) = 0.160$ and $M\ \hat{\alpha}) = 0.099$. Results for the coefficient of determination R^2), residual sums of squares RSS), and the criterion functions are given for three regression models in Table 2.

Table 2 - Comparison of odels, Vapor Pressure Data $\mathcal{M} \in$

odel	R^2	RSS	$L\ \alpha)$

4 Discussion

The minimizations of the criterion functions L and M for the transformation problems were carried out numericall since anal tic methods are not readil available. The computations were greatl facilitated b LISP-ST T Tierne , 1990), which made it possible to carr out the calculations with relativel few lines of code. The functions NEWTON X and NEL E D X were used with good success. For the examples of this paper, the calculations proceeded quite fast on a SUN SP RC station. Starting values of $\alpha = 1, \ldots, 1)'$ worked well. Other starting values were also used.

n important issue in an model selection procedure is that of model assumptions. It is well known that violations of the same can result in the addition or omission of variables in a variable selection procedure. IC and BIC, for instance, are not robust to outliers or influential points. The criteria proposed in this article likel suffer from the same problems. Simultaneousl checking and selecting models is difficult, and there are no definitive solutions to this problem. Ho

given b

$$M_m = \pi^{1/2} \left(\frac{\ , \ \frac{n+\delta_0}{2})}{\ , \ n + \ _0/2)} \ \ 2-\gamma)^{k_m/2} \right)^{1/n} \ \ _m^{1/2} \left(1 + \frac{b_m}{\ _m}\right)^{1+\frac{\delta_0}{2n}} \ ,$$

where $_m = q_m + \gamma p_m + \gamma_0$ and $b_m = q_m + \frac{\gamma^2}{2-\gamma} p_m$. gain, both $_m$ and b_m are linear combinations of the residual sum of squares and the "guessing error".

n exact expression for the criterion is not available since the necessar integral is not tractable. However, for large n, we can approximate the distribution in 2.) b a $No_n \left(\eta_m, \left(\frac{n+\delta_0}{n+\delta_0-2}\right)^{-1} s_m^{-2} \ I + \ 1-\gamma) P_m)^{-1} \right)$ distribution. Taking m_0 to be the full model, define

$$v = \frac{n + \ _m) \ n + \ _{m_0} - 2)}{n + \ _{m_0}) \ n + \ _m - 2)} \ ,$$

where $_m = \ _{m_0} = \ _0$ for the normal-gamma priors, and $_m = -k_m$, $_{m_0} = -k_{m_0}$ under nonio 2))Tjfl013.9995TDfl m)Tjfl/T131T42 c5Dfl and)[e995TD0Tjfl/T1.fine36499813.9With16)Tjfl/T81Tffl3

[2] itchison, J., and Dunsmore, I. R. 19 5), *Statistical Prediction Anal sis*, New York : Cambridge Universit Press.

[3] itkin, . 1991), "Posterior Ba es Factors," with discussion) , *Journal of the Ro al Statistical Societ* , Ser. B, 53, 111-142.

[4] kaike, H. 19 3), "Information Theor and an Extension of the aximum likelihood Principle," *International S m osium on Information Theor* , eds. B. N. Petrov and F. Csaki, pp.26 -281. Budapest: kademia Kiado.

[5] Bates, D. ., and Watts, D. G. 1988) *onlinear Regression Anal sis and its A lications*, New York : John Wile .

[6] Bernardo, J. . 1985), Comment on "Outliers and Influential Observations in Linear odels", with discussion), in *Ba esian Statistics 2*, eds. Bernardo, J. ., DeGroot, . H., Lindle , D. V., and Smith, . F. ., msterdam : North-Holland, p.492.

[ ] Bhattacharjee, S. K., and Dunsmore, I. R. 1991) "The Influence of Variables in Logistic Regression", *Biometrika*, 8, 851-856.

[8] Box, G. E. P., and Cox, D. R. 1964), "The nal sis of Transformations" with discussion), *Journal of the Ro al Statistical Societ* , Ser. B, 26, 211-252.

[9] Box, G. E. P., and Jenkins, G. . 19 6). *Time Series Anal sis : Forecasting and Control*, 2nd Ed.), San Francisco: Holden-Da .

[10] Box, G. E. P., and Kanemasu, H. 19 3), "Posterior Probabilities of Candidate odels in odel Discrimination," Technical Report 322, Universit of Wisconsin.

[11] Box, G.E. P., and e er, D. R. 1986), "Dispersion Effects From Fractional Designs," *Technometrics*, 28, 19-2 .

[12] Box, G. E. P., and Tiao, G. C. 19 3), *Ba esian Inference in Statistical Anal sis*. Reading, : ddison-Wesle .

[13] Box, G. E. P. , and Tidwell, P. W. 1962), "Transformation of the Independent Variables," *Technometrics*, 4, 531-550.

[14] Carroll, R. J., and Ruppert, D. 1988), *Transformation and Weighting in Regression*, London : Chapman and Hall.

[15] Christensen, R. 1990), *Log-Linear Models*, New York : Springer-Verlag.

[16] Cla ton, . K., Geisser, S., and Jennings, D. E. 1986), " comparison of Several odel Selection Procedures," in *Studies in Ba esian Econometrics and Statistics*, eds. P. K. Goel and . Zellner, New York : Elsevier.

[1] Cook, R.

[32] itchell, T. J., and Beauchamp, J. J. 1988), "Ba esian Variable Selection in Linear Regression," with discussion), *Journal of the American Statistical Association* , 83, 1023-1036.

[33] Pettit, L. I., and Smith, . F. . 1985), "Outliers and Influential Observations in Linear odels", with discussion), in *Ba esian Statistics 2*, eds. Bernardo, J. ., DeGroot, . H., Lindle , D. V., and Smith, . F. ., msterdam : North-Holland, p.492.

[34] San artini, ., and Spezzaferri, F. 1984), " Predictive odel Selection Criterion," *Journal of the Ro al Societ* , Ser. B, 46, 296-303.

[35] San artini, ., and Spezzaferri, F. 1986) "Selection of Variables in ultiple Regression for Prediction and Control, " *Statistica*, 118-12 .

[36] Schwarz, G. 19 8), "Estimating the Dimension of a odel", *Annals of Statistics*, 6, 461-464.

[3] Smith, . F. ., and Spiegelhalter, D. J. 1980), "Ba es Factors and Choice Criteria for Linear odels," *Journal of the Ro al Statistical Societ* , Ser. B., 42, 213-220.

[38] Spiegelhalter, D. J., and Smith, . F. . 1982), "Ba es Factors for Linear and Log-linear odels with Vague Prior Information", *Journal of the Ro al Statistical Societ* , Ser. B., 44, 3 -38 .

[39] Taguchi, G., and Wu, Y. 1980), *Introduction to ff-Line Qualit Control*, Nago a, Japan : Central Japan Qualit Control ssociation.

[40] Tierne , L. 1990), *Lis -Stat : An bject- riented Environment for Statistical Com uting and D namic Gra hics* , New York: John Wile .

[41] Tuke , J. W. 19), *Ex lorator Data Anal sis,* Reading, : ddison-Wesle .ne 19410TDfl Nago)T