Journal of Econometrics 40 (1989) 33–44. North-Holland

PREDICTIVE EFFICIENCY FOR SIMPLE NON-LINEAR MODELS*

Thomas F. COOLEY

University of Rochester, Rochester, NY 14627, USA

William R. PARKE

University of Rochester, Rochester, NY 14627 USA
University of California, Santa Barbara, CA 93106, USA

Siddhartha CHIB

University of Missouri, Columbia, MO 65211, USA

This paper demonstrates the use of exact predictive likelihood functions for simple non-linear models. A measure of predictive efficiency based on the concept of expected information loss is introduced as a way of comparing alternative prediction functions. It is shown that the predictive likelihood function minimizes expected information loss over a wide class of potential prediction functions. Some Monte Carlo experiments illustrate the performance of alternative prediction functions in settings where prediction is difficult.

1. Introduction

Problems of prediction are distinguished from classical parameter estimation by the fact that the object of interest is an unknown probability distribution rather than an unknown, but non-stochastic parameter. Prediction functions based on mean squared analysis, Monte Carlo simulation or Bayesian procedures are commonly used devices for approximating the unknown future distribution. In Cooley and Parke (1987a, b) we have described approximate prediction functions based on a definition of predictive likelihood. In this paper we propose a way of comparing alternative prediction functions and illustrate its use in simple non-linear models.

Practical as well as theoretical considerations suggest that successful prediction functions should (i) be free of unknown parameters, (ii) reflect parameter estimation uncertainty and (iii) converge in probability to the true density of

*This is a substantially abridged version of the paper 'Prediction Functions' presented at the Conference on Forecasting at Arizona State University in March 1987. We are grateful to Adrian Pagan, Peter Schmidt and an anonymous referee for helpful comments. Responsibility for errors remains ours. The first author acknowledges financial support from the John M. Olin Foundation and the Center for Research in Government Policy and Business.

the future observations as the sample size goes to infinity. A prediction function with these characteristics can be summarized by a measure of location such as the mean or median or via a predictive confidence interval that incorporates the shape and dispersion. In an earlier paper [Cooley and Parke (1987a)] we compared the performance of several alternative prediction functions in the context of a simple dynamic model. The basis for comparison was largely how well they captured the mean and median and whether they had appropriate length confidence intervals. The results of that paper suggest that, although there are important theoretical differences among them, the techniques seem to perform about equally well when judged in terms of their moments and confidence intervals whether or not they account for parameter uncertainty or capture the correct functional form. If that were generally the case, there would not seem to be much reason to worry about exact or approximate predictive densities.

Our goal in this paper is to shed more light on this issue in two respects. First, we introduce a measure of expected information loss as a way of comparing predictive densities. This provides an informative summary of the relative predictive efficiencies of alternative prediction functions by accounting for the entire shape of the distribution and by taking expectations over realizations of both past and future data. The second notable feature of this paper is that we illustrate the information loss associated with different prediction functions for some difficult prediction problems. The examples themselves are simple non-linear models: a log-linear function and a logistic function. The prediction problems are difficult, however, in that they arise from either policy interventions or exogenous circumstances of low probability. The results demonstrate significant differences in the information loss associated with alternative prediction functions.

2. Exact predictive likelihood functions

Before introducing the measure of information loss, we define here the exact predictive likelihood functions for the linear regression model and two non-linear extensions. Although the concept of predictive likelihood is discussed in Cooley and Parke (1987a, b), those papers develop approximate versions based on asymptotic expansions and asymptotic distributions. The cases considered here are logically prior in that they lend themselves to exact small sample derivations. Interesting analytic results can be obtained for these models that can only be illustrated numerically for other models.

Lauritzen (1974) and Hinkley (1979) define a predictive likelihood function based on minimal sufficient statistics S_d for the m data period observations $Y_d = (Y_1, \ldots, Y_m)$, S_f for the n future period observations $Y_f = (Y_{m+1}, \ldots, Y_{m+n})$, and S_{d+f} for the combined data and forecast period observations Y_{d+f}. Sufficiency ensures that the conditional probability densities

$f(S_f, S_d|S_{d+f})$ and $f(Y_f|S_f)$ do not involve the true parameters. Loosely stated, the predictive likelihood function is intended to reflect the degree to which Y_f and Y_d are compatible with a common sufficient reduction S_{d+f}. We state this formally as:

Definition 1 [Lauritzen (1974), Hinkley (1979)]

$$\text{plik}(Y_f|Y_d) = f(Y_f|S_f) \cdot f(S_f, S_d|S_{d+f}). \tag{1}$$

We can illustrate this idea for the linear regression model,

$$Y_i = X_i\beta + \varepsilon_i, \qquad i = 1, \ldots, m+n, \tag{2}$$

where X_i is $1 \times k$ and $\varepsilon_{d+f} \sim \text{N}(0, \sigma^2 I_{m+n})$. The minimal sufficient statistics S_{d+f} are the independent quantities $X'_{d+f}Y_{d+f}$ and $SSR_{d+f} = Y'_{d+f}[I - M]Y_{d+f}$, where $M = X_{d+f}(X'_{d+f}X_{d+f})^{-1}X'_{d+f}$ [Cox and Hinkley (1974, p. 14)].[1] (The sufficient statistics for the data period are X'_dY_d and SSR_d.) Applying Definition 1, we obtain:

Proposition 1. For (2), let $\hat{\beta}_d = (X'_dX_d)^{-1}X'_dY_d$ *and let*

$$A = \left(Y_f - X_f\hat{\beta}_d\right)'\left[I_n + X_f(X'_dX_d)^{-1}X'_f\right]^{-1}\left(Y_f - X_f\hat{\beta}_d\right).$$

For β *and* σ^2 *unknown,*

$$\text{plik}(Y_f|Y_d) \propto \{1 + A/SSR_d\}^{-(m+n-k)/2}.$$

For β *unknown and* σ^2 *known,*

$$\text{plik}(Y_f|Y_d) \propto e^{-\frac{1}{2}A/\sigma^2}.$$

Proof. Appendix.

The predictive likelihood function (plik) thus takes on a familiar form for the linear regression model. If both β and σ^2 are unknown, it has the form of a multivariate t with $m-k$ degrees of freedom, mean $X_f\hat{\beta}_d$, and covariance matrix $s^2I_n + s^2X_f(X'_dX_d)^{-1}X'_f$, where $s^2 = SSR_d/(m-k)$ is the usual estimate of σ^2. The term s^2I_n in the covariance matrix is due to error term uncertainty, and the term $s^2X_f(X'_dX_d)^{-1}X'_f$ is due to parameter uncertainty. If σ^2 is known, the predictive likelihood function is the corresponding multivariate normal density. For simplicity, we will use the latter density and the case $n = 1$ (so that f denotes $m+1$) in the remainder of this paper.

[1]Any one-to-one transformation of a minimal sufficient statistic is also minimal sufficient, and we choose $\log(SSR)$ because the latter is consistent with an invariant uniform prior [Jeffreys (1983)].

Extending Proposition 1 to simple non-linearities is straightforward. The minimal sufficient statistics for the non-linear model

$$h(Z_i) = Y_i, \qquad Y_i = X_i\beta + \varepsilon_i, \tag{3}$$

are again $X'Y$ and SSR. Letting J denote the Jacobian dY_i/dZ_i,

$$\text{plik}(Z_f|Z_d) = |J| \cdot \text{plik}(Y_f|Y_d). \tag{4}$$

This functional form parallels the true density

$$f(Z_f; \beta, \sigma^2) = |J| \cdot f(Y_f; \beta, \sigma^2). \tag{5}$$

In particular, under the usual assumption that $(X_d'X_d)^{-1} = O(m^{-1})$, we have $\hat{\beta}_d = \beta + O_p(m^{-1/2})$ and $\text{plik}(Z_f|Z_d)$ converges to the true density $f(Z_f; \beta, \sigma^2)$.

Convergence to the true density helps in constructing minimum-length predictive confidence intervals with prespecified probability levels. Consider, for example, a log-linear model

$$\log(Z_i) = Y_i, \qquad Y_i = X_i\beta + \varepsilon_i. \tag{6}$$

In this case,

$$\text{plik}(Z_f|Z_d) \propto Z_f^{-1} \cdot \exp\left\{-\tfrac{1}{2}\left(\log(Z_f) - X_f\hat{\beta}_d\right)^2/(\sigma^2 + \tau^2)\right\}, \tag{7}$$

where

$$\hat{\beta}_d = (X_d'X_d)^{-1}X_d'Y_d \quad \text{and} \quad \tau^2 = \sigma^2 X_f(X_d'X_d)^{-1}X_f'.$$

$\text{Plik}(Z_f|Z_d)$ thus has the form of a log-linear density with log-mean $X_f\hat{\beta}_d$ and log-variance $\sigma^2 + \tau^2$. While one might construct a confidence interval for Z_f by simply transforming a confidence interval for $Y_f = \log(Z_f)$, a Neyman–Pearson type construction based on (7) will yield shorter confidence intervals for a given probability level.

The distinction between region forecasts for Z_f and transformed region forecasts for Y_f is even more striking for the logistic model

$$\log(Z_i/(1 - Z_i)) = Y_i, \qquad Y_i = X_i\beta + \varepsilon_i, \qquad 0 < Z_i < 1. \tag{8}$$

If σ^2 is large enough, the true density for Z_f will be bimodal with a region of highest probability consisting of two disjoint intervals. While the region of highest predictive likelihood will also be composed of two intervals, transforming a confidence interval for Y_f will yield a single interval that totally misses the bimodal nature of the true density.

3. Predictive efficiency

These simple examples illustrate that the entire shape of a predictive density may well be relevant for non-linear prediction problems. As a quantitative measure of how well a candidate normalized prediction function $f^*(Z_f|Z_d)$ based on estimated parameters resembles the unknown true density $f(Z_f; \beta, \sigma^2)$, we adopt the Kullback–Leibler information measure of the difference between $f^*(Z_f|Z_d)$ and $f(Z_f; \beta, \sigma^2)$.[2]

Definition 2 [Predictive Efficiency]. The K–L information measure is

$$I(f^\circ, f^*) = \int \left[\log\left(f\left(Z_f; \beta, \sigma^2\right)\right) - \log\left(f^*\left(Z_f|Z_d\right)\right)\right] \times f\left(Z_f; \beta, \sigma^2\right) \mathrm{d}Z_f. \tag{9}$$

The expected value of this measure over realizations of Z_d,

$$\bar{I}(f^\circ, f^*) = \int I(f^\circ, f^*) f\left(Z_d; \beta, \sigma^2\right) \mathrm{d}Z_d, \tag{10}$$

measures the information loss associated with f^*. We say that f^* is predictive efficient relative to f^{**} if $\bar{I}(f^\circ, f^*) < \bar{I}(f^\circ, f^{**})$.

Two advantages of this approach should be noted. First, unlike other efficiency measures such as mean squared prediction error, $\bar{I}(f^\circ, f^*)$ is naturally tailored to the functional form of the true future density. Furthermore, it is invariant to common non-linear transformations $Y_i = h(Z_i)$ in the following sense.

Proposition 2. For any monotone, differentiable transformation $Y_i = h(Z_i)$, *let* $f^*(Y_f|Y_d) = |J|^{-1} \cdot f^*(Z_f|Z_d)$. *Then*

$$I\left(f^\circ\left(Z_f; \beta, \sigma^2\right), f^*\left(Z_f|Z_d\right)\right) = I\left(f^\circ\left(Y_f; \beta, \sigma^2\right), f^*\left(Y_f|Y_d\right)\right), \tag{11}$$

and

$$\bar{I}\left(f^\circ\left(Z_f; \beta, \sigma^2\right), f^*\left(Z_f|Z_d\right)\right) = \bar{I}\left(f^\circ\left(Y_f; \beta, \sigma^2\right), f^*\left(Y_f|Y_d\right)\right). \tag{12}$$

[2]Aitchison (1975) and Larimore (1983) also advocate (10) as an information measure of goodness of prediction fit. Akaike's (1973) information criterion (AIC) for model selection is based on a sample variant of (10).

Proof. Appendix.

We can thus frame predictive efficiency questions for many non-linear models in terms of the underlying linear model. For the linear model, Levy and Perng (1986) show that plik($Y_f|\, Y_d$) minimizes $\bar{I}(f^\circ, f^*)$ over a wide class of potential prediction functions of the form $g(Y_f - X_f\hat{\beta}_d)$ for some function g. Proposition 2 effectively extends Levy and Perng's optimality result to a general class of non-linear models.

The resulting bound on predictive efficiency, which is attained by the predictive likelihood function, follows directly from Kullback (1959, p. 189):

$$\bar{I}(f^\circ, \text{plik}) = \frac{n}{2}\log(1 + \tau^2/\sigma^2). \tag{13}$$

We can compare this expected information loss to that obtained by other prediction functions. One example is the naive plug-in function that simply substitutes parameter estimates for the unknown parameters in the true density: $\text{CEQ}(Y_f|\, Y_d) = f(Y_f; \hat{\beta}_d, s^2)$. The expected information loss for this function is

$$\bar{I}(f^\circ, \text{CEQ}) = \frac{n}{2}\tau^2/\sigma^2, \tag{14}$$

where $\tau^2 = \sigma^2 X_f(X_d'X_d)^{-1}X_f'$. Correcting for parameter uncertainty is thus important to the extent that $\log(1 + \tau^2/\sigma^2) < \tau^2/\sigma^2$.

4. Monte Carlo results[3]

We conclude this paper with some Monte Carlo experiments that compare predictive efficiencies for alternative prediction functions. The experiments are deliberately chosen to illustrate situations where non-linearities and parameter uncertainty make an important difference in the prediction problem. We report results for three techniques and for each technique we have a version that corrects for parameter uncertainty and one that ignores it. The predictive likelihood function, plik, is as discussed in section 2. The certainty equivalence function, CEQ, is essentially the same as the plik (it has the correct functional form) but it does not correct for parameter uncertainty. Rather, it treats estimated parameters as known.

The other two approaches considered yield symmetric prediction functions with the form of a normal density. The mean squared error prediction functions, *MSE/MSE**, are based on a linear approximation to a non-linear estimated model. The asterisk denotes the version that corrects for parameter

[3]The results reported here are a brief summary of a more complete Monte Carlo study described in Cooley, Parke and Chib (1987).

uncertainty. A direct linear regression of Z_i on the explanatory variables yields the misspecified prediction functions *REG/REG**. We use truncated normal densities in calculating the information loss for the normal prediction functions to take into account the restriction on the permissible range of Z_i inherent in the true specification. This makes the results for these normal prediction functions more favorable than would otherwise be the case.

The first two examples involve prediction from a log-linear model (6) with three explanatory variables exhibiting moderate collinearity. For the first experiment, the sample period exogenous data are drawn from a mixing process designed to permit the possible rare occurrence of extreme observations. Future period observations are drawn from the density that can generate extreme values. The second example, also a log-linear model, represents the sort of prediction problem encountered when the X's are altered by a policy intervention. To capture this, we add a constant to the draw for one of the X's for the future observations.

Figs. 1a and 1b illustrate the expected information losses for the two experiments just described for sample sizes of 50, 100, 200, 400 and 800. These

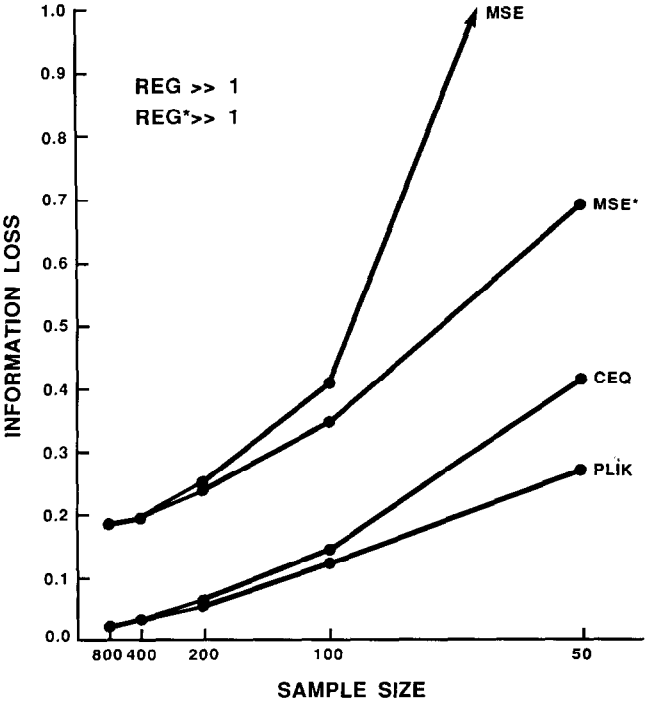

Fig. 1a. Log-linear model mixture experiment.

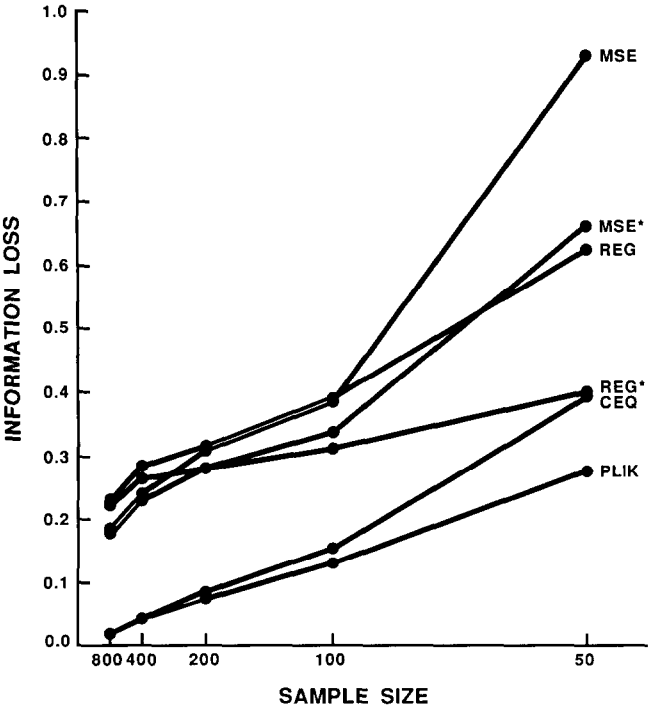

Fig. 1b. Log-linear model policy intervention experiment.

results illustrate the importance of accounting for both parameter uncertainty and functional form. While the latter is apparently the more important of the two, parameter uncertainty is clearly important for the combination of small sample sizes and difficult prediction problems.

The final two examples consider the same prediction problems as above, but in the context of a logistic model (8). The logistic specification increases the extent of the non-linearity: there is both an upper and lower bound on the range of the dependent variable and the function changes from concave to convex. The parameter values are chosen so that the true density is essentially unimodal.[4] Figs. 1c and 1d illustrate the results for these two experiments. These results are basically similar to those reported above although the misspecified *REG/REG** prediction functions perform much worse here. Once again both functional form and parameter uncertainty make a difference. The information loss associated with functions that ignore parameter uncer-

[4] The results for a bimodal true density are not reported because *MSE/MSE** and *REG/REG** fare very poorly in terms of information losses in that case.

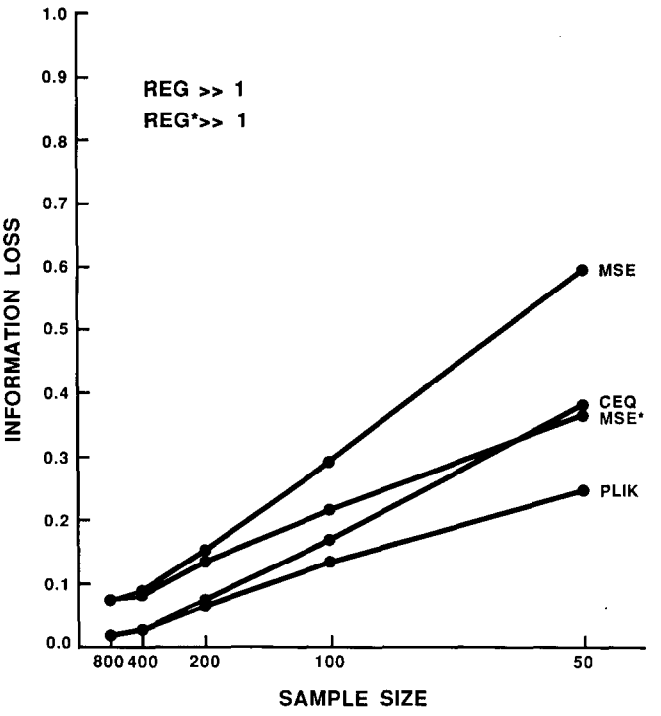

Fig. 1c. Logistic model mixture experiment.

tainty converges relatively quickly to the information loss of the corresponding functions that incorporate it as the sample size increases.

These simple examples do not test the candidate prediction functions on two important points. First, realistic prediction models are often more elaborate than those we have used. In particular, they almost always involve a greater number of parameters and, hence, entail a greater possibility for practically important parameter uncertainty. Second, the non-linearities in more elaborate models may be difficult to analyze in a closed form and, unlike the simple log and logistic transformations, may depend upon estimated parameters. The examples do, however, give some indication of how the importance of parameter uncertainty depends on sample size. They also suggest that there are prediction problems arising in familiar models where it is quite important to have an accurate approximation of the underlying true density. There are many different approaches to obtaining that density. While we have emphasized the predictive likelihood approach it is worth noting that for the problems considered here the plik is equivalent to prediction functions based on Monte Carlo simulation or a Bayesian posterior density.

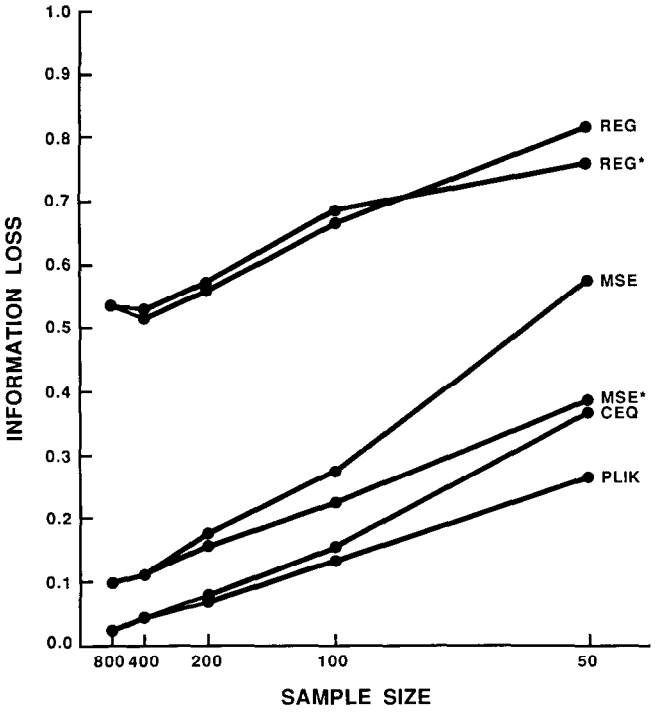

Fig. 1d. Logistic model policy intervention experiment.

Appendix

Lemma 1. The sufficient statistics $X_d'Y_d$, $X_{d+f}'Y_{d+f}$, $\log(SSR_d)$ *and* $\log(SSR_{d+f})$ *are related by the relations* $X_{d+f}'Y_{d+f} = X_d'Y_d + Y_f'Y_f$ *and*

$$SSR_{d+f} = SSR_d + A, \tag{A.1}$$

where

$$A = \left(Y_f - X_f\hat{\beta}_d\right)'\left[I_n + X_f\left(X_d'X_d\right)^{-1}X_f'\right]^{-1}\left(Y_f - X_f\hat{\beta}_d\right).$$

Proof of Lemma 1. Let $X = X_{d+f}$, $Y = Y_{d+f}$, $\hat{\beta} = \hat{\beta}_{d+f}$ and $SSR = SSR_{d+f}$. The recursion

$$\hat{\beta} - \hat{\beta}_d = \left(X'X\right)^{-1}X_f'\left(Y_f - X_f\hat{\beta}_d\right) \tag{A.2}$$

follows directly from $X'X(\hat{\beta} - \hat{\beta}_d) = X'Y - Y_d'Y_d - X_f'X_f\hat{\beta}_d$, which can also be

written in the form $X'X(\hat{\beta} - \hat{\beta}_d) = X'(Y - X\hat{\beta}_d)$. We can use this last relation and the device $Y - X\hat{\beta} = Y - X\hat{\beta}_d + X(\hat{\beta}_d - \hat{\beta})$ to write $SSR = (Y - X\hat{\beta})'(Y - X\hat{\beta})$ as

$$SSR = (Y_d - X_d\hat{\beta}_d)'(Y_d - X_d\hat{\beta}_d) + (Y_f - X_f\hat{\beta}_d)'(Y_f - X_f\hat{\beta}_d)$$
$$-(\hat{\beta} - \hat{\beta}_d)'(X'X)(\hat{\beta} - \hat{\beta}_d).$$

We then use the recursion (A.2) to obtain

$$SSR = SSR_d + (Y_f - X_f\hat{\beta}_d)'\left[I_n - X_f(X'X)^{-1}X_f'\right](Y_f - X_f\hat{\beta}_d).$$

Eq. (A.1) follows from $I_n - X_f(X'X)^{-1}X_f' = [I_n + X_f(X_d'X_d)^{-1}X_f']^{-1}$ [Rao (1973, p. 33)]. □

Proof of Proposition 1. Definition 1 can be written as

$$\text{plik}(Y_f|Y_d) = \frac{f(Y_f; \beta, \sigma^2) f(\hat{\beta}_d, \log(SSR_d); \beta, \sigma^2)}{f(\hat{\beta}, \log(SSR); \beta, \sigma^2)}. \tag{A.3}$$

$\hat{\beta}$ and $\log(SSR)$ are independent, with $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$ and $SSR/\sigma^2 \sim \chi^2_{m+n-k}$. Using similar distributions for $\hat{\beta}_d$ and SSR_d, the factors of (A.3) are

$$f(Y_f; \beta, \sigma^2) \propto \exp\{-\tfrac{1}{2}(Y_f - X_f\beta)'(Y_f - X_f\beta)/\sigma^2\}, \tag{A.4}$$

$$f(\hat{\beta}_d, \log(SSR_d); \beta, \sigma^2) \propto \exp\{-\tfrac{1}{2}(\hat{\beta}_d - \beta)'(X_d'X_d)(\hat{\beta}_d - \beta)/\sigma^2\}$$
$$\cdot (SSR_d/\sigma^2)^{(m-k)/2} \exp\{-\tfrac{1}{2}SSR_d/\sigma^2\}, \tag{A.5}$$

$$f(\hat{\beta}, \log(SSR); \beta, \sigma^2) \propto \exp\{-\tfrac{1}{2}(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)/\sigma^2\}$$
$$\cdot (SSR/\sigma^2)^{(m+n-k)/2} \exp\{-\tfrac{1}{2}SSR/\sigma^2\}. \tag{A.6}$$

Using the well-known results

$$(Y - X\beta)'(Y - X\beta) = SSR + (\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta), \tag{A.7}$$

and

$$(Y_d - X_d\beta)'(Y_d - X_d\beta) = SSR_d + (\hat{\beta}_d - \beta)'(X_d'X_d)(\hat{\beta}_d - \beta), \tag{A.8}$$

the exponential functions in (A.4), (A.5) and (A.6) cancel, leaving

$$\text{plik}(Y_f|Y_d) \propto (SSR/\sigma^2)^{-(m+n-k)/2}/(SSR_d/\sigma^2)^{-(m-k)/2}.$$

Finally, Lemma 1 yields

$$\text{plik}(Y_f|Y_d) \propto (1 + A/SSR_d)^{-(m+n-k)/2}.$$

For the case that σ^2 is known, $\hat{\beta}$ is sufficient and we can eliminate the χ^2 densities from (A.5) and (A.6). Applying (A.7) and (A.8) then yields

$$\text{plik}(Y_f|Y_d) \propto \exp\{-\tfrac{1}{2}SSR/\sigma^2\}/\exp\{-\tfrac{1}{2}SSR_d/\sigma^2\}.$$

Lemma 1 gives us the desired prediction function. □

Proof of Proposition 2. Substituting (4) and (5) into (9) yields

$$I(f^o(Z_f;\beta,\sigma^2), f^*(Z_f|Z_d)) = \int \log\left\{\frac{|J|\cdot f^*(Y_f|Y_d)}{|J|\cdot f(Y_f;\beta,\sigma^2)}\right\}|J| \cdot f(Y_f;\beta,\sigma^2)\,dZ_f.$$

Using (5) once more, this equals $I(f^o(Y_f;\beta,\sigma^2), f^*(Y_f|Y_d))$. The extension to $\bar{I}(f^o, f^*)$ is immediate. □

References

Aitchison, J., 1975, Goodness of prediction fit, Biometrika 62, 547–554.

Akaike, H., 1973, Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov and F. Csaki, eds., Second international symposium on information theory (Akademiai Kiado, Budapest).

Cooley, T.F. and W.R. Parke, 1987a, Likelihood and other approaches to prediction in dynamic models, Journal of Econometrics 35, 119–142.

Cooley, T.F. and W.R. Parke, 1987b, Asymptotic likelihood based prediction functions, Reproduced (University of Rochester, Rochester, NY).

Cooley, T.F., W.R. Parke and S. Chib, 1987, Prediction functions, Reproduced (University of Rochester, Rochester, NY).

Cox, D.R. and D.V. Hinkley, 1974, Theoretical statistics (Chapman and Hall, London).

Hinkley, D., 1979, Predictive likelihood, Annals of Statistics 7, 718–728.

Jeffreys, H., 1983, Theory of probability, 3rd ed. (Clarendon Press, Oxford).

Kullback, S., 1959, Information theory and statistics (Wiley, New York).

Larimore, W.E., 1983, Predictive inference, sufficiency, entropy and an asymptotic likelihood principle, Biometrika 70, 175–182.

Lauritzen, S.L., 1974, Sufficiency, prediction and extreme models, Scandinavian Journal of Statistics 1, 128–134.

Levy, M.S. and S.K. Perng, 1986, An optimal prediction function for the normal linear model, Journal of the American Statistical Association 81, 196–198.