Strategies for Assessing Skill and Significance of Screening Regression Models with Emphasis on Monte Carlo Techniques[1]

JOHN R. LANZANTE

Department of Meteorology and Physical Oceanography, Cook College—New Jersey Agricultural Experiment Station, Rutgers—The State University of New Jersey, New Brunswick, NJ 08903

(Manuscript received 17 March 1984, in final form 23 January 1984)

ABSTRACT

This paper reviews the considerations in evaluating the skill and significance of screening multiple linear regression (SMLR) models. Formulations and procedures are given along with relevant references to prior studies. Topics discussed include predictor selection, serial correlation, artificial skill, true skill, and Monte Carlo significance testing. New results with wide applicability in the assessment of SMLR model skill and significance are presented in graphical form. However, the results are restricted to situations involving predictors which are independent of one another and are serially uncorrelated. The methodology presented is suggested for use in both model evaluation and experimental design.

1. Introduction

Screening multiple linear regression (SMLR) is a statistical tool which has found extensive meteorological application in short and long range forecasting, both in research and operationally. Unfortunately, the simplicity in using the many "canned" statistical packages available for use in SMLR is easily matched by the additional complexity (which is a direct result of the screening process) of assessing the model skill and significance. Most statistical packages provide measures of skill and significance for SMLR models based on the assumption of *a priori* selection. Since SMLR involves *a posteriori* selection (i.e., the best M_s predictors are chosen from a pool of M candidate variables) these measures are invalid. One way in which to circumvent this problem has been to partition the data sample into a dependent part (from which the SMLR model is derived) and an independent part (used for testing the model). This approach is often unsatisfactory due to the smallness of the dependent and independent samples (particularly in climate studies) which leads to unstable estimates of model parameters and verification statistics. Often the researcher feels that the combined sample is not large enough!

In an alternative approach (dependent sample testing) skill and significance are assessed using the sample from which the SMLR equations were derived. In this case the "artificial predictability" due to both finiteness of the sample and the effect of *a posteriori* selection must be taken into account. The effects of artificial predictability on the evaluation of SMLR models have been addressed by a number of authors, most extensively by Davis (1976, 1977, 1978, and 1979, hereafter referred to as D1–D4). The purpose of this paper is to present the results of a large number of Monte Carlo simulations aimed at assessing SMLR model skill and significance. The results are summarized in such a fashion as to have wide applicability with relative simplicity. In addition, suggested strategies to be used in conjunction with the Monte Carlo results are given along with relevant formulations and their references in the literature.

2. Discussion

In essence, regression analysis is a means by which a given variable Y (the independent variable or predictand) is estimated as a linear combination of m other (dependent) variables (X_i), where both Y and the X_i are functions of time (t). This relationship can be expressed as follows:

$$Y(t) = \sum_{i=1}^{m} B_i X_i(t). \tag{1}$$

The regression coefficients (B_i) are determined by least squares in order to maximize the explained variance of Y. It is assumed here (for simplicity) and throughout the remainder of this paper that Y and the X_i have a mean value of zero (i.e., are in anomaly form). It should be noted that in the relationships presented here no distinction between population and sample parameters is made. In reality, population parameters are not known but are estimated from the available sample.

[1] Paper of the Journal Series, New Jersey Agricultural Experiment Station Cook College, Rutgers—The State University of New Jersey, New Brunswick, NJ 08903.

a. Predictor selection

The first step in building any regression model is the selection of the type and number of predictors (X_i) to be included. It is often the case in meteorology and oceanography that a large number of predictor variables (measured at a number of gridpoint locations) are needed in order to capture the spatial variability of the relevant predictor fields; this gives rise to two major complications. First, the computational inconvenience and cost increases (nonlinearly) with the number of predictors in the pool. Second, in most cases there is considerable dependence between the gridpoint data values of a given field; as a result, the usefulness of the regression coefficients (B_i) for interpretive purposes is lost.

In response to these problems, atmospheric and oceanic scientists have made extensive use of Empirical Orthogonal Functions (EOFs) since their introduction by Lorenz (1956). Through EOF analysis matters can be simplified by transformation of large numbers of gridpoint variables into a small number of independent (uncorrelated) patterns. The important climatic signal is captured in those EOFs which are distinguishable from random noise. One way to determine the cutoff point which separates the climatic signal from the noise is the Monte Carlo approach discussed by Overland and Preisendorfer (1982). However, the problem of "mixing" of adjacent patterns (North *et al.*, 1982) complicates the determination of the truncation point of the EOFs. Since the Monte Carlo simulations reported in this paper are based on independent predictors, in order to use these results it is necessary to perform some transformation (EOF analysis, harmonic analysis, etc.) to insure that the predictors are independent.

From a statistical standpoint it is preferable to order and/or select the number of predictors *a priori* (D1; D3; Barnett and Hasselmann, 1979; Barnett *et al.*, 1981) in order to limit artificial predictability. The ability to do this depends largely on the complexity and on the physical insight which the researcher has regarding the particular problem. Strategies for model building based on a variety of selection criterion are given by Barnett and Hasselmann (1979) and Barnett *et al.* (1981). The central concern in these strategies is balancing the competing requirements of model skill and significance. Generally, the more predictors in a given model, the larger the hindcast skill and smaller the model significance. Unfortunately, in many studies (particularly those of an exploratory nature) the researcher does not have enough insight to specify the model *a priori*. In these cases the researcher turns to screening (*a posteriori* selection), which is the primary topic of this paper.

b. Serial correlation

Of the two components of artificial predictability discussed in the literature (i.e., the effects of serial correlation and of screening) the former has received more attention, as well as an analytic treatment, beginning with Bartlett (1946). The effect of serial correlation has been quantified through the use of the effective sample size (N^*), where $N^* \leq N$ (the actual sample size); in the absence of serial correlation $N^* = N$. Chelton (1983) presents an expression for N^* (based on Jenkins and Watts (1968), and Bendat and Piersol (1971)):

$$N_i^* = N_i / \sum_{L=-\infty}^{+\infty} [\rho_{ii}(L)\rho_{yy}(L) + \rho_{iy}(L)\rho_{yi}(L)], \quad (2)$$

where N_i^* and N_i are the effective and actual sample sizes, respectively, in the zero lag cross correlation between the predictand (Y) and predictor i (X_i); each ρ represents the cross or autocorrelation (of the two variables indicated by subscripts) at lag L. It is assumed that the sample size is large compared to the (auto- and cross-correlation) time scales of Y and the X_i. In practice, (2) is evaluated over the range $L = \pm L'$, where L' is large enough so that the $\rho(L')$'s become statistically indistinguishable from zero. The above formulation (2) is more general than those given by D1 and Sciremammano (1979) who assumed that the effect of the cross correlation terms (ρ_{iy} and ρ_{yi}) is negligible. The characteristic time scale τ (i.e., the time needed to gain another degree of freedom) used by D1 and D3 is defined as follows:

$$\tau = N\Delta t/N^*, \quad (3)$$

where Δt is the time between the (uniformly spaced) observations. Another variation on the estimation of the effective sample size was presented by Laurmann and Gates (1977) for the case of a first order Markov process.

The concept of an effective sample size (N^*) with its application to statistical significance testing has been used quite extensively in meteorological studies. Thiebaux and Zwiers (1984, hereafter referred to as TZ) have pointed out that the estimation of N^* has been based on equating the ensemble mean square of a time-averaged mean to the standard formula for the variance of the mean of independent samples. Based on an examination of other methods of computing N^*, TZ conclude that the estimates of N^* are not unique. Thiebaux and Zwiers then compared a variety of alternate methods of estimating N^*; however, even the best of those methods is not entirely satisfactory.

An even more serious problem pointed out by TZ is that serial correlation in data violates some assumptions underlying the use of the Student's t and F distributions. As a result, it is not valid to compute N^* and use this value in significance testing. Thus, the results of this study are applicable only in the case of serially uncorrelated data.

c. Artificial skill and significance

By expanding on the work of Lorenz (1956), Davis (D1–D4) has been foremost in addressing the problem

of artificial predictability through the quantification of the artificial skill (S_A). He defines the true skill (S) as the amount by which the hindcast skill (S_H, the explained variance of the model) exceeds the artificial skill:

$$S = S_H - S_A. \tag{4}$$

Furthermore, the forecast skill (S_F) or skill in applying the regression equation to an independent sample can be estimated by

$$S_F \approx S_H - 2S_A. \tag{5}$$

According to Chelton (1983), the artificial skill (S_A) depends on the effective sample size (N^*), the number of predictors used in the regression model (m), and the true skill (S). Assuming that N^* is large,

$$S_A = m(1 - S)/N^*. \tag{6}$$

Davis's formulation is the special case of (6) for small true skill ($S \sim 0$):

$$S_A = \frac{m}{N^*}. \tag{7}$$

By combining (4) and (6) the true skill (S) can be expressed as follows:

$$S = \frac{S_H - m/N^*}{1 - m/N^*}. \tag{8}$$

Similarly, by combining (5) and (6) the forecast skill (S_F) can be estimated:

$$S_F \approx \frac{S_H - 2m/N^*}{1 - 2m/N^*}. \tag{9}$$

Again the reader is reminded that these formulations may be erroneous (following TZ) when $N^* \neq N$.

It should be noted that (4)–(9) apply to *a priori* selection (nonscreening regression). For SMLR m lies between M_s and M, where the M_s best predictors are selected from a pool of M variables. Because of the difficulty in deriving an analytic expression for S_A as applied to screening regression, much reliance has been placed on Monte Carlo simulations. A general relationship for the artificial skill as a function of M_s and M is given in Fig. 1 of D2. Lanzante and Harnack (1982) corrected their S_A values to account for screening by using the appropriate value from this figure as a multiplier. More discussion of this topic is found in the next subsection.

The assessment of model significance has received less analytic treatment than that of model skill. One approach which has been used is to apply the classical F-test or χ^2 test using N^* as the sample size; however, this does not account for screening. The Miller equivalent F-test (Miller, 1958), which has sometimes been mis-applied, is not in general an adequate solution since it is only applicable for the case in which *one* predictor is selected from the pool. As a result, model significance has been assessed predominantly through the use of Monte Carlo experiments, either by replacing some of the observed data with random numbers, or by randomly shuffling the predictand (e.g., Lund, 1970; Neumann *et al.*, 1977). However, when either spatial relationships or serial correlation are important care must be taken to impose these conditions in the randomization process.

d. Monte Carlo results

Given the widespread use of Monte Carlo techniques in the evaluation of regression models, general empirical relationships for assessing skill and significance would be of considerable value in terms of reducing human effort and computational costs. As an added bonus they could aid in experimental design (i.e., *a priori* decisions on predictor and pool sizes). With these aims in mind, a large number of SMLR Monte Carlo simulations were carried out using normally distributed random numbers as predictors and predictands. A variety of pool-predictor size (1–6, 8, 10, 12, 14, 16, 18 and 20) and case size (10–100 by 5) combinations were used, with each combination yielding an R^2 value (explained variance) for each of 1000 trials. From the R^2 distribution the mean value ($\overline{R^2}$) and the 5% tail value (R^2_{crit} or R^2 value above which lies 5% of all values) were computed. The R^2_{crit} and R^2 values are intended for use in assessing model significance (at the 5% significance level) and artificial skill (i.e., the R^2 value expected as a result of the screening process).

Two sets of charts (one set each for the $\overline{R^2}$ and R^2_{crit} values) were constructed for each pool size, depicting the variation of R^2 or R^2_{crit} as a function of number of predictors selected and number of cases. The values for the nonscreening cases (using all predictors in the pool) were estimated based on the relationship between the F-statistic and the R^2 value which can be derived from the definition of the F-test for regression found in any elementary statistics book:

$$R^2 = \left(\frac{\nu_2}{\nu_1 F} + 1\right)^{-1}, \tag{10}$$

where $\nu_2 = n - m - 1$, $\nu_1 = m$, n is the number of cases, m is the number of variables in the model, and F is the critical F value from an F table. The critical F value for a one tailed test of 5% significance was used to compute R^2_{crit}, while $F = 1$ (the expected value of F) was used to compute $\overline{R^2}$.

After checking that these charts were reasonable [internally consistent and consistent with the values computed from (10) for the nonscreening cases] the charts were summarized in two figures (based on the representation in Fig. 1 of D2) by averaging over all case sizes. In these two graphs (Figs. 1–2) the abscissa (M_s/M) represents the fraction of the pool size (M) that is selected (M_s); the ordinate is the *square* of the ratio of the R^2 values in screening M_s predictors from

a pool of M to that of using all M predictors (non-screening). This representation is the same as used in Fig. 1 of D2 except that here the ordinate is $\mathcal{F}^2$ (instead of $\mathcal{F}$) in order to provide a larger graphical separation of the family of curves. The $M = 10$ curve of D2 and this study are nearly identical; also, if the $M = 80$ curve of D2 was plotted in Fig. 1 of this study, the $M = 20$ curve would be roughly equidistant between it and the $M = 10$ curve. Figure 2 of this study (5% significance values) has no counterpart in D2 but is equivalent to the relationships expressed in Fig. 7 of Neumann *et al.* (1977); the major difference is the fact that they computed values for much larger pool and sample sizes than were used in this study, and presented less general results (they used only four values of M_s).

The methodology described below can be applied after the fact to assess skill and significance, or in the preliminary stages of research. In the latter case, typical R^2 and R^2_{crit} values can be estimated *before* performing any regression analyses; if, in the judgement of the researcher the values seem higher than could be expected, an adjustment of the M_s and M values might be in order. This should be done within the framework presented by Barnett *et al.* (1981).

The best way in which to illustrate the method is through an example. Suppose that N is 30 and that

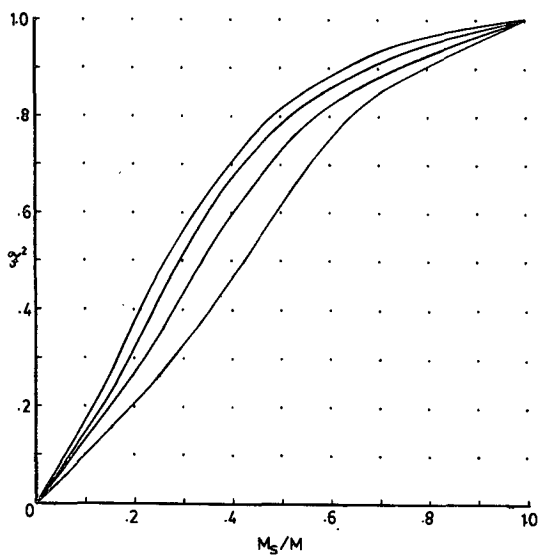

FIG. 1. The relationship between $\mathcal{F}^2$ and M_s/M for the artificial skill in a SMLR model in which M_s predictors are selected from a pool of M independent predictors. $\mathcal{F}$ is the ratio of artifical skills of a model in which the best M_s predictors are selected from a pool of M, to that of a model containing all M predictors. The family of curves represent (from top to bottom) pool sizes of 20, 10, 5, and 3. Find the $\mathcal{F}^2$ by cross indexing M_s/M with the appropriate M curve. After taking the square root, multiply $\mathcal{F}$ times the R^2 value computed from (10), where $\nu_2 = n - m - 1$, $\nu_1 = m$, n = the sample size (N), $m = M$ (pool size), and $F = 1$ (the expected F value). Finally, multiply this result by the appropriate sample size correction factor from Table 1. The value computed is an estimate of the artificial skill (S_A or R^2).

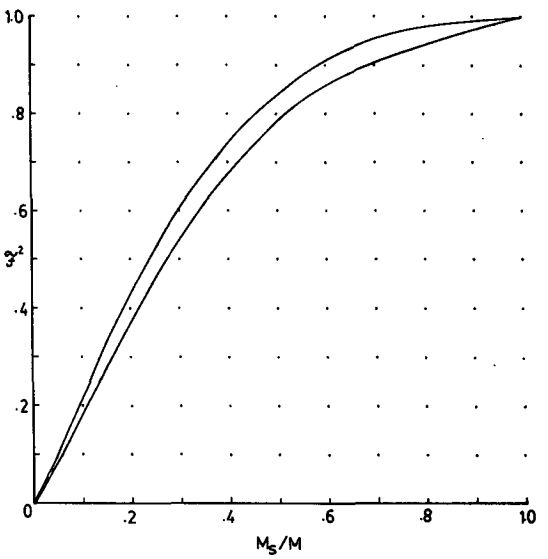

FIG. 2. As in Fig. 1 except that the relationship displayed is for R^2_{crit} (critical R^2 value used in assessing significance at the 5% level). In addition, the sample size correction factors from Table 2 and the F value for 5% significance (with ν_1 and ν_2 degrees of freedom) should be used. The two curves are for $M = 20$ (top) and $M = 3$ (bottom).

5 independent predictors were to be screened from a pool of 10. To assess the artificial skill apply the value of M_s/M (5/10 = 0.5) to the second curve from the top in Fig. 1 ($M = 10$) to get $\mathcal{F}^2 = 0.78$ (or $\mathcal{F} = 0.88$). Next, applying (10) with $n = 30$, $m = 10$, and $F = 1$ (for significance, use Fig. 2 and the 5% critical value of F with 10 and 19 degrees of freedom) yields $R^2 = 0.34$. Finally, multiplying R^2 and $\mathcal{F}$ we get 0.30; correct for the sample size by multiplying this result by the appropriate value from Table 1 (0.99) yielding $R^2 = S_A = 0.30$. Apply the same procedure using (10), Fig. 2, Table 2, and the one tailed value of F (10, 19, 0.05) to estimate R^2_{crit}, which is 0.49 in this case.

While this completes the procedure for the assessment of SMLR skill and significance, other factors may have to be considered. For example, if a model hierarchy scheme suggested by Barnett *et al.* (1981) was used, it would be necessary to make the distinction

TABLE 1. Sample size correction factors (multipliers) for use with Fig. 1. After computing an R^2 value from (10) and multiplying by $\mathcal{F}$ from Fig. 1, multiply by the correction factor above (which is a function of the sample size). These factors were computed by averaging the Monte Carlo results over all pool and predictor sizes, for a given sample size, and normalizing to a mean of 1. Note that while these correction factors have only a small effect, the sample size has a much larger effect on the F value used in (10).

Sample size									
10	20	30	40	50	60	70	80	90	100
0.91	0.97	0.99	1.00	1.01	1.02	1.02	1.02	1.02	1.02

TABLE 2. As in Table 1 except for use with Fig. 2.

Sample size									
10	20	30	40	50	60	70	80	90	100
0.98	0.98	0.98	0.99	1.00	1.00	1.01	1.02	1.02	1.03

between single candidate and multiple candidate selection criterion. Accounting for this normally involves the use of an individual model significance level which is different from 5%, in order to have 95% confidence in the overall scheme. By noting that there is only a small difference between the corresponding curves in Figs. 1–2 (which represent the 50% and 5% significance levels, respectively) and the values in Tables 1 and 2, it is possible to make a good estimate of $\mathcal{F}$ for other significance levels; the effect of the variation of F with M will account for most of the variation of R^2_{crit} with M.

Finally, when a given experiment is composed of multiple models (perhaps representing different locations) it is important to make an assessment of the *overall* significance of the experiment. In the case of a grid with a separate model for each location, a simple count of the number of significant models could be misleading due to the dependence between gridpoints. This count (adjusted to the grid spacing) should be compared to the effective number of independent gridpoints in order to assess *field significance*. An application of the binomial distribution and Monte Carlo simulation for this purpose is discussed by Livezey and Chen (1983).

3. Summary

Some of the various concerns which must be addressed in the evaluation of screening regression models were given in this paper along with a review of formulations and procedures for estimating model skill and significance, including the effective sample size, artificial skill, and true skill. Additionally, the results of a large number of Monte Carlo simulations were presented as an aid in assessing model skill and significance for screening regression involving predictors that are independent of one another and are serially uncorrelated. These results were presented in such a form as to have wide applicability, and can be used in experimental design as well as model evaluation.

Acknowledgments. New Jersey Agricultural Experiment Station Publication No. D-13507-2-84 was supported by state funds and by the Climate Dynamics Program, Division of Atmospheric Sciences, National Science Foundation, under Grant ATM-8217215. The efforts of those who have developed and introduced statistical concepts (without which this paper would not have been possible) are greatly appreciated. Special thanks to Dr. Robert Livezey for his help in improving this manuscript.

REFERENCES

Barnett, T., and K. Hasselmann, 1979: Techniques of linear prediction, with application to oceanic and atmospheric fields in the tropical Pacific. *Rev. Geophys. Space Phys.*, **17**, 949–968.

——, R. Preisendorfer, L. Goldstein and K. Hasselmann, 1981: Significance tests for regression model hierarchies. *J. Phys. Oceanogr.*, **11**, 1150–1154.

Bartlett, M., 1946: On the theoretical specification and sampling properties of autocorrelated time series. *J. Roy. Stat. Soc.*, **B8**, 27–41.

Bendat, J., and A. Piersol, 1971: *Random Data: Analysis and Measurement Procedures*, Wiley Interscience, 407 pp.

Chelton, D., 1983: Effects of sampling errors in statistical estimation. *Deep-Sea Res.*, **30**, 1083–1103.

Davis, R., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **6**, 249–266.

——, 1977: Techniques for statistical analysis and prediction of geophysical fluid systems. *Geophys. Astrophys. Fluid Dyn.*, **8**, 245–277.

——, 1978: Predictability of sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **8**, 233–246.

——, 1979: A search for short range climate predictability. *Dyn. Atmos. Oceans*, **3**, 485–497.

Jenkins, G., and D. Watts, 1968: *Spectral Analysis and its Applications*, Holden-Day, 525 pp.

Lanzante, J., and R. Harnack, 1982: Specification of United States summer season precipitation. *Mon. Wea. Rev.*, **110**, 1843–1850.

Livezey, R., and W. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111**, 46–59.

Laurmann, J., and L. Gates, 1977: Statistical considerations in the evaluation of climate experiments with atmospheric general circulation models. *J. Atmos. Sci.*, **34**, 1187–1199.

Lorenz, E., 1956: Empirical orthogonal functions and statistical weather prediction. Sci. Rep. No. 1, Statistical Forecasting Project, Dept. Meteor., MIT, Cambridge, MA.

Lund, I., 1970: A Monte Carlo method for testing the statistical significance of a regression equation. *J. Appl. Meteor.*, **9**, 330–332.

Miller, R., 1958: The screening procedure, Part II of studies in statistical weather prediction. Final Rep., Contract AF 19(604)-1590, Travelers Research Center, Hartford, CT, 86–136.

Neumann, C., M. Lawrence and E. Caso, 1977: Monte Carlo significance testing as applied to statistical tropical cyclone models. *J. Appl. Meteor.*, **16**, 1165–1174.

North, G., T. Bell, R. Cahalan and F. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **110**, 699–706.

Overland, J., and R. Preisendorfer, 1982: A significance test for principal components applied to cyclone climatology. *Mon. Wea. Rev.*, **110**, 1–4.

Sciremammano, F., 1979: A suggestion for the presentation of correlations and their significance levels. *J. Phys. Oceanogr.*, **9**, 1273–1276.

Thiébaux, H., and F. Zwiers, 1984: The interpretation and estimation of effective sample size. *J. Clim. Appl. Meteor.*, **23**, 800–811.