A Comparative Study of Test Procedures uses in Assessing the Forecasting Ability of Linear Models with Applications to Crop Yield Data

Linardis, Apostolis and Panaretos, John

1998

A Comparative Study of Test Procedures used in Assessing the Forecasting Ability of Linear Models with Applications to Crop Yield data.

A.Linardis and J.Panaretos

Department of Statistics

Athens University of Economics and Bussiness

76, Patision St., 104 34 *Athens, Greece*

$e-mail\ for\ correspondence: jpan@aueb.gr$

Abstract

The choice of the appropriate linear model before this can be used for planning and decision making, has been the concern of many statistical workers. Most of the methods in the literature aim at evaluating the descriptive ability of the candidate models.

In the present paper an evaluation scheme of the predictability of a linear model based on a function of the discrepancy of the observed and the corresponding predicted values of the dependent variable is studied. Based on this statistical function, the predictability of a linear model is tested. Considering the ratio of such functions for two linear models, the predictability of these models is compared. Applications on real and simulated data are also presented.

Key words: Linear model; Model Selection; Decision Making; Predictability; χ^2 distribution; F ditribution.

1. Introduction

Most of the methods in the literature evaluate the descriptive ability of the candidate models. An alternative approach is to evaluate the predictability of the models.

Let's suppose that the data are taken as a time series and that data are given until the time-point t. Then an appropriate linear model is estimated by the least square method. The methods that evaluate the descriptive ability of the candidate models are based on the discrepancy between the observed and the estimated value of the dependent variable, for all time-points t.

With respect to methods that evaluate the predictive ability of a linear model, an alternative approach is the following. For every time-point a model is estimated by the least square method. Then the predictability of the model is evaluated, based on

the discrepancy of the predicted value for the t+1 time-point and the corresponding observed value of the dependent variable.

Under some conditions, the sum of the squared discrepancies that evaluate the predictability is χ^2-distributed. Based on this statistical function, one may test the predictability of a linear model. Considering the ratio of the squared discrepancies for two linear models, one is able to compare the predictability of two linear models. It is proved that under some conditions this ratio is distributed according to a generalized form of the F-distribution.(Panaretos et al. (1997))

In section 2 the methods that evaluate the predictive ability of one and two linear models are examined. In section 3 two applications in crop-yield data are presented as well a small simulation study. In section 4 the advantages and disadvantages of the methods proposed are discussed.

2. The methods proposed

2.1. Estimation of Predictions

Consider the linear model :

$$\mathbf{Y}_t = \mathbf{X}_t \mathbf{b} + \mathbf{e}_t \qquad (1)$$

where

$\mathbf{Y}_t$ is an $(\ell_t \times 1)$vector of observations on the dependent random variable
$\mathbf{X}_t$ is an $(\ell_t \times m)$matrix of known coefficients where $(\ell_t \geqslant m, |\mathbf{X}_t'\mathbf{X}_t| \neq 0)$
$\mathbf{b}$ is an $(m \times 1)$ vector of regression coefficients and
$\mathbf{e}_t$ is an $(\ell_t \times 1)$vector of normal error random variables with $E\,(\mathbf{e}_t) = 0$ and
$\quad V\,(\mathbf{e}_t) = \sigma^2 I_t$, where I_t is the $(\ell_t \times \ell_t)$ identity matrix.
The prediction of the $(t+1)$time-point is given by:

$$\widehat{Y}_{t+1}^{\circ} = \mathbf{X}_{t+1}^{\circ} \widehat{\mathbf{b}}_t \qquad (2)$$

where $\widehat{\mathbf{b}}_t$ is the least square estimator of b at time t given by:

$$\widehat{\mathbf{b}}_t = \left(\mathbf{X}_t'\mathbf{X}_t\right)^{-1} \mathbf{X}_t'\mathbf{Y}_t \qquad (3)$$

and $\mathbf{X}_{t+1}^{\circ}$ is an $(1 \times m)$vector of values of the regressors for the $(t+1)$time-point.
The variance of the prediction $\widehat{Y}_{t+1}^{\circ}$ is then given by:

$$V\left(\widehat{Y}_{t+1}^{\circ}\right) = S_t^2 \{\mathbf{X}_{t+1}^{\circ} \left(\mathbf{X}_t'\mathbf{X}_t\right)^{-1} \mathbf{X}_{t+1}^{\circ\,'} + 1\} \qquad (4)$$

where

$$S_t^2 = \frac{\left[\mathbf{Y}_t - \mathbf{X}_t \widehat{b}_t\right]' \left[\mathbf{Y}_t - \mathbf{X}_t \widehat{b}_t\right]}{[\ell_t - m]} \qquad (5)$$

After the true value Y_{t+1}° for the $(t+1)$time-point has been observed, the model to be used for predicting the value of the $(t+2)$time-point becomes:

$$\mathbf{Y}_{t+1} = \mathbf{X}_{t+1}\mathbf{b} + \mathbf{e}_{t+1} \qquad (6)$$

where now the matrices $\mathbf{X}_{t+1}$and $\mathbf{Y}_{t+1}$ are defined as :

$$\mathbf{X}_{t+1} = \begin{pmatrix} \mathbf{X}_t \\ \mathbf{X}_{t+1}^{\circ} \end{pmatrix} \text{ and } \mathbf{Y}_{t+1} = \begin{pmatrix} \mathbf{Y}_t \\ Y_{t+1}^{\circ} \end{pmatrix}$$

with dimensions $[(\ell_t + 1) \times m]$ and $[(\ell_t + 1) \times 1]$ respectively.

The method just described is called the method of "one step-ahead prediction" (see, e.g. Atckinson (1985)).

2.2. Testing the Predictability of a Linear model

Xekalaki and Katti (1984), used the difference between the observed and the predicted value of the dependent variable on every time-point, to evaluate the predictability of a linear model. Let

Y_{t+1}° : the observed value of the dependent variable for the $(t+1)$time-point

$\widehat{Y}_{t+1}^{\circ}$: the predicted value of the dependent variable for the $(t+1)$ time-point.

Then it is known that:

$$\widehat{Y}_{t+1}^{\circ} - Y_{t+1}^{\circ} \sim N\left(0, \; S_t^2 \left(\mathbf{X}_{t+1}^{\circ} \left(\mathbf{X}_t'\mathbf{X}_t\right)^{-1} \mathbf{X}_{t+1}^{\circ'} + 1\right)\right) \qquad (7)$$

A statistical function that could be used for the evaluation of the predictive ability of the model is the function:

$$r_{t+1} = \frac{\widehat{Y}_{t+1}^{\circ} - Y_{t+1}^{\circ}}{S_t \sqrt{\left(\mathbf{X}_{t+1}^{\circ} \left(\mathbf{X}_t'\mathbf{X}_t\right)^{-1} \mathbf{X}_{t+1}^{\circ'} + 1\right)}} \qquad (8)$$

where S_t is given by (5).

Because of (7) and (8), holds : $\quad r_{t+1} \sim N(0,1) \qquad (9)$

According to Brown (1975) and Kendall (1983) the above differences are independent. So the statistical function: $\qquad \sum_{t=1}^{n} r_t^2 \sim \chi_n^2 \qquad (10)$

According to this statistical function we can test the hypothesis:

H_0 : the model is appropriate for predictions
H_A : the model presents lack of predictability

2.3. Testing the Predictability of two Linear Models.

Let A and B two linear models which are given by (1). Suppose that we have observations for n_1, n_2 time-points respectively and we want to choose the more adequate model, that is, the one that has the greater predictability. A statistical function appropriate for testing the hypothesis:

H_0 : the two models give equivalent predictions
H_A : model A predicts better than model B or the opposite
was given by Panaretos et al. (1997) as

$$Z = \frac{\sum_{t=1}^{n_1} r_t^2(A)}{\sum_{t=1}^{n_2} r_t^2(B)} \quad (11)$$

$\sum_{t=1}^{n_1} r_t^2(A)$, $\sum_{t=1}^{n_2} r_t^2(B)$ are not independent since the standardized residuals of predictions come from the same response. Panaretos et al. (1997) considered the same n number of time-points for the two models and proved that, for large n, the above ratio is distributed according to a distribution that is a generalized form of the F-distribution, the Correlated Gamma-Ratio distribution with p.d.f:

$$f(Z) = \frac{1-\rho^2}{B(\kappa,\kappa)} Z^{\kappa-1} (1+Z)^{-2\kappa} \left(1 - \left(\frac{2\rho}{Z+1}\right)^2 Z\right)^{-\frac{2\kappa+1}{2}}, 0 < Z < +\infty \quad (12)$$

where
Z is the ratio of the squared standardized residuals of predictions
ρ is the correlation coefficient of the standardized residuals of predictions and
κ is equal to n/2.
Using the Correlated Gamma-Ratio distribution we can test the hypothesis of the equivalence of two linear models in giving predictions. In the following plot the p.d.f of the Correlated Gamma Ratio distribution for $\kappa=5$ and different values of the correlation coefficient is depicted.
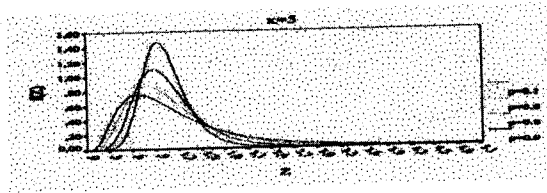
Figure 1: P.d.f of the Correlated Gamma-Ratio distribution for $\kappa = 5$ and different values of the correlation coefficient.

3. Applications.

3.1. First Application.

The data of this application concern corn crop at the state of Indiana in the USA, in the time interval 1931-1980. The crop yields for ten differerent districts (CRD) are given and two different sets of variables have been used by the USDA to predict the crop yield for the next years, for each of the district. We want to find out which model is appropriate for predictions as well as which of the two models is more appropriate for the specific district.

First we have to check if each one of the two models used, give satisfactory predictions. The p-values of the hypothesis tests for the appropriateness of each one of the linear models for predictions, based on χ^2—distribution are given in table 1. (we take into consideration only the estimations for the last 24 years).

Table 1 : *P — Values of the hypothesis test for the appropriateness of Models A and B for predictions, for the ten districts*

	Model A		Model B	
	$\sum\limits_{t=1}^{24} r_t^2(A)$	P-Value	$\sum\limits_{t=1}^{24} r_t^2(B)$	P-Value
Crd 10	58,8440	$9,3\ 10^{-5}$	92,7986	0,000
Crd 20	58,6818	$9,8\ 10^{-5}$	59,5954	$7,3\ 10^{-5}$
Crd 30	24,6387	**0,4256**	35,3542	**0,0633**
Crd 40	69,6775	$2\ 10^{-6}$	66,6917	$7\ 10^{-6}$
Crd 50	49,0058	$1,891\ 10^{-3}$	51,0283	$1,046\ 10^{-3}$
Crd 60	55,9491	$2,33\ 10^{-4}$	32,7895	**0,1086**
Crd 70	39,9333	0,0217	49,0120	$1,88\ 10^{-3}$
Crd 80	57,3969	$1,47\ 10^{-4}$	52,2323	$7,3\ 10^{-4}$
Crd 90	61,4615	$4\ 10^{-5}$	41,8104	0,01355
Crd 100	46,5158	$3,836\ 10^{-3}$	73,9439	$1\ 10^{-6}$

Rejecting the null hypothesis of satisfactory predictions if p-value<0.05, we can conclude that only model A in Crd 30 and models B in Crd 30 and Crd 60 seem to be adequate for predictions.

Further, we want to examine which of the two candidate models is more adequate for predictions. The p-values of the hypothesis tests based on the Correlated Gamma-Ratio distribution are given in table 2 (Here also we take into consideration only the estimations for the last 24 years).

Table 2 : P — Values of the hypothesis test that compares the predictability of the Models A and B for predictions, for the ten districts

	P-Value	Best Model
Crd 10	0,0355	Model A
Crd 20	0,4656	Equivalent
Crd 30	0,0337	Model A
Crd 40	0,453	Equivalent
Crd 50	0,45	Equivalent
Crd 60	0,0963	Model B
Crd 70	0,275	Equivalent
Crd 80	0,353	Equivalent
Crd 90	0,1068	Equivalent
Crd 100	0,0868	Model A

According to the results of table 1 and 2, we can conclude that we should consider only model A for Crd 30 and model B for Crd 60.

For the above models, we also estimated the coefficients R^2_{adj} to evaluate the descriptive ability of the models. According to R^2_{adj} all models A and B describe the given data in a very satisfactory way since all R^2_{adj} are greater than 0.9. Comparing these results with those of table 1 we notice that, in most cases, the models that describe the data well do not forecast well.

3.2. Second Application.

The method of testing the predictability of one and two linear models was applied to another set of real data used by Drapper and Smith (1981, p.407). The data refer to corn crop yields at the state of Iowa in the USA, in the time interval 1930-1962. The true crop-yield for every year was given and nine variables were used as possible predictors of the crop yield.

506

Firstly, we applied the classical backward elimination procedure and for the models risen we estimated the $\sum r_t^2$ statistic. The classical backward elimination procedure ives the model with variables $\{1,2,6,9\}$ as the better one ($R_{adj}^2=0.68$) while the $\sum r_t^2$ statistic gives the model with variables $\{1,2,3,4,5,6,8,9\}$ as the model with the est predictability which is also statistically equivalent with the one with variables $1,2,3,4,5,6,9\}$. So one may choose the latter because it is simpler.

We may notice that the statistics of the two procedures give different models as he most appropriate.

ín Alternative Backward Procedure

ín alternative method is to apply the Backward elimination procedure using the tatistic $\sum r_t^2$ as a criterion for eliminating variables. That is :

We estimate the statistic $\sum r_t^2$ for the full model.
We estimate the statistic $\sum r_t^2$ for all the models that arise by eliminating one pre-lictor.
We consider the model that presents the smallest statistic $\sum r_t^2$ (say, the one that rises from the eliminaton of X_1).
We test statistically if this model is appropriate for predictions according to χ^2- dis-ribution as well as whether this one provides better predictability than all the other onsidering also the full model, according to Correlated Gamma-Ratio distribution.
If the model that presents the smallest statistic is better than all the others we inally eliminate variable X_1, as "variable that reduces the predictive ability of the nodel" and we continue the procedure.
If the tests indicate an "inappropriate model for predictions", or conclude equiva-ence, we stop the procedure.

A disadvantage of the method is that it stops quite quickly, retaining many vari-ıbles in the model. Applying the above method to the Iowa data, we eliminated the ıredictors as follows: 7, 9(*), 4, 3, 8, 2, 6, 5 (from the second stage on, we continue he procedure, without testing statistically).

}.3. A Simulation Study

Γhe algorithm for generating a Correlated Gamma-Ratio variable amounts to the ollowing steps:

ʰGenerate a sample of n observations (X_i, Y_i) on the vector (X,Y) that is distrib-ıted according to the $N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$ distribution.

507

*Compute $R_X = \sum_{i=1}^{n} X_i^2$ and $R_Y = \sum_{i=1}^{n} Y_i^2$ and obtain Z=R_X / R_Y.

Simulating 150 samples for different values of κ and correlation coefficient we rejected the null hypothesis of the equivalence of the models in 17 cases.

4. Discussion.

In the present paper we studied methods that test the predictive ability of one linear model as well as methods that compare the predictive ability of two competing linear models based on the χ^2 and the Correlated Gamma-Ratio distribution.

Comparison of these methods with those that evaluate the descriptive ability of one and two linear models indicates that the first are more appropriate for data that come from a time-series. Besides, the methods described here take into consideration all the linear models that arise in every time-point while the other methods just the model containing all the observations.

The Correlated Gamma-Ratio test presents some advantages compared with the other model selection procedures. We do not have to know the functional form of the competing linear models. On the other hand it is appropriate for both nested and non-nested or overlapping linear models.

According to the results of the applications, a drawback of the use of the Correlated Gamma Ratio test is that it frequently results in equivalence of the predictive ability of the compared linear models.

Many statisticians believe that if a model describes well the data then it also predicts well. The results of the applications of this paper do not support this view. Our results indicate that a model that fits well the data may not predict well. Besides it seems that a parsimonious model is most likely to give correct predictions and that an ovefitted model does not predict well.

References

Atckinson, A.C. (1985). *Plots, Transformation and Regression.* Oxford Science Publications.

Draper, N.R. and Smith H. (1981). *Applied Regression Analysis.* John Wiley and Sons.

Linardis, A.(1998). *A Comparative Study of Test Procedures used in Assessing the Forecasting Ability of Linear Models with Applications to Crop Yield data.* M.Sc. Dissertation, Department of Statistics, Athens University of Economics and Business, Athens (ISBN : 960-7929-13-6).

Panaretos. J., Psarakis, S. and Xekalaki, E. (1997). *The Correlated Gamma-Ratio distribution in model evaluation and selection.* Technical report 33, Department of Statistics, Athens University of Economics and Business, November 1997.

Xekalaki, E. and Katti, S.K. (1981). *Rating models based on concordance with observed data.* Biometrical Journal, 26, 173-184.