DYNAMIC ECONOMETRIC MODELS
Vol. 6 – Nicolaus Copernicus University – Toruń – 2004

Anna Szmit
Technical University of Lodz

The Analysis of the Forecast Quality Depending on the Length of Forecast Horizon

1. Introduction

The purpose of each forecaster is to obtain forecasts which will be possibly the most accurate. However, as it is well-known the prediction process is inherently connected with generating errors due to different causes. It also seems natural that the forecast accuracy becomes worse as the forecast horizon lengthens. There are several reasons of that behaviour.

In this paper an attempt to analyse the rate of worsening of forecast accuracy as their horizon lengthens has been presented on the basis of the forecasts obtained with the help of the chosen linear regression models.

The mean forecast errors, *ex ante* and *ex post*, were used as the basic measures of evaluating the forecast accuracy. Hence, the analysis of the forecast quality as dependent on the length of the forecast horizon will be considered from these two points of view.

2. Description of the problem

Among the causes of forecast errors the ones presented below deserve to be particularly underlined:

– improper form of the model,
– incorrect values of explanatory variables for the period of the forecast,
– random variability of the forecasted variable

Obviously, in choosing an appropriate forecasting method and evaluating the model quality the first of the reasons mentioned above should be reduced. However, even a model, which reflects the investigated event very well in a certain

period, can become inappropriate with a lapse of time. This is especially connected with the length of the horizon of forecast. In general, a short horizon of forecast is understand[1] as such, for which only quantitative changes occur, whereas the qualitative dependencies are preserved. On the long horizon the qualitative changes take place, e.g. the change of analytical form of existing causal relationships may occur. In consequence, the form of a description model should be changed. Therefore, a model which is suitable for a short time horizon may, in general, be inappropriate for a long-run perspective.

A lack of knowledge about the values of explanatory variables in the forecast period can also become a source of errors. Then, to generate a forecast on the basis of a descriptive econometric model at least the forecasts of those variables have to be known. However, those forecasts are calculated with an error. Hence this error will increase[2] the total forecast error.

Random variability is specific to the majority of phenomena. The range of this variability measured by the standard error of residuals can be estimated on the basis of data.

An evaluation of the forecast method, especially an econometric model, is very often decided with regard to the practical usefulness of generated forecasts, i.e. with regard to the obtained errors which are required to be possibly small. Hence, particular importance has the magnitude of root mean[3] squared error, i.e. in the form of:

$$S(y_T^P) = \sqrt{\frac{\sum_{\tau=1}^{T} (y_\tau^P - \hat{y}_\tau^P)^2}{T}}, \tag{1}$$

where

y_τ^P – real value of dependent variable at moment τ of the forecast, $\tau = 1,...T$,

$\hat{y}_\tau^P$ – theoretical value of dependent variable (generated by the model) at moment τ of the forecast.

Percentage errors are frequently applied as well, one of them being a mean absolute percentage error *MAPE*:

$$MAPE = \frac{1}{T} \sum_{\tau=1}^{T} \left| \frac{y_\tau^P - \hat{y}_\tau^P}{y_\tau^P} \right|. \tag{2}$$

[1] See: Cieślak (1999), p. 24. In the paper by Zeliaś (1997), p. 23, an agreed division was presented, according to which the forecasts up to one year are assumed as short-term ones, those for 2–5 years horizon – as medium-term, and forecasts made for more than 5 years horizon – as long-term ones.

[2] See: Zeliaś (1997), p. 68.

[3] In some cases the maximum error value is more important.

However, the *MAPE* error can be calculated only if the values of a forecasted variable y are known at the forecast period. In order to evaluate the magnitude of the forecast error in advance for different horizon periods the values of the *ex ante* prediction error are needed. For the linear regression model, for which the forecast assumptions[4] are satisfied, the expected value of prediction error equals zero and the estimate of the prediction error, i.e. the prediction standard deviation[5] is given by:

$$S(\hat{y}_\tau^P) = \sqrt{S_e^2\left(x_\tau^P (X^T X)^{-1} (x_\tau^P)^T + 1\right)}, \tag{3}$$

where

$X = \left[x_{ij}\right]_{\substack{i=1,\ldots n \\ j=0,1,\ldots k}}$ – the (n x k) matrix of k explanatory variables in a linear regression model ($x_{i0} \equiv 1$),

$x_\tau^P = \left[1 \quad x_{1\tau}^P \quad \ldots \quad x_{1\tau}^P\right]$ – the vector of the values of explanatory variables at moment τ of forecast horizon,

S_e^2 – residual variance of a model.

The relative prediction standard deviation in the form of

$$V_\tau = \frac{S(\hat{y}_\tau^P)}{\hat{y}_\tau^P} \tag{4}$$

is also applied.

From formula (3) results that the greater is the mean prediction error, the more distant is the vector x_τ^P from the vector $\bar{x}$. In particular, partial derivatives of $S(\hat{y}_\tau^P)$ with regard to variable x_i are given by the formula

$$\frac{dS(\hat{y}_\tau^P)}{dx_i} = S_e \frac{b_{ii}x_i + \sum\limits_{\substack{j=0..k \\ j\neq i}} b_{ij}x_j}{\sqrt{b_{ii}x_i^2 + 2\sum\limits_{\substack{j=0..k \\ j\neq i}} b_{ij}x_j + \sum\limits_{\substack{j=0..k \\ j\neq i}}\sum\limits_{\substack{l=0..k \\ l\neq i}} b_{jl}x_j x_l + 1}},$$

where b_y are the elements of the matrix $(X^T X)^{-1}$.

Then, e.g. in case of $k = 1$, we have

[4] See: Zeliaś (1997), p. 50–52, Cieślak (1999), p. 115–116, Witkowska (2002), p. 260–261.

[5] In the case when the estimates of parameters are used (instead of unknown parameter values), not taking into account a stochastic nature of a vector x_τ^P.

$$\frac{dS(\hat{y}_\tau^P)}{dx} = \frac{S_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \cdot \frac{x - \bar{x}}{\sqrt{(x-\bar{x})^2 + \left(1+\frac{1}{n}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and

$$\frac{d^2S(\hat{y}_\tau^P)}{dx^2} = \frac{S_e\left(1+\frac{1}{n}\right)\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\left((x-\bar{x})^2 + \left(1+\frac{1}{n}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}},$$

then $S(\hat{y}_\tau^P)$ is a convex function with respect to x, which means that the mean prediction error increases even more and more faster as the distance form $\bar{x}$ increases.

The above mentioned remarks do not mean, however, that the prediction error grows automatically for each linear regression model as the forecast horizon lengthens. There are a lot of models for which the values of explanatory variables do not have a monotonic character in the forecast period. Particularly the models not dependent on a time variable t belong to them. One of such examples may be the seasonality model without a trend, for which the matrix of explanatory variable values (cyclical also within the forecast period) can be written in the form similar to presented below:

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

In such a case the mean prediction error is constant throughout the whole forecast period.

3. An example: models with and without lagged variables

A time series on the consumption of electric energy with the one hour frequency of observations was chosen as the dependent (and forecasted) variable

in the presented study. The results of the analysed values of forecast errors for two linear regression models[6] will be presented below. In both models as explanatory variables a lot of dummy variables[7] were used which represent the variants of qualitative variables describing a current moment in time (these variants are known with any leads).

In the first of analysed models (referred to as M3) as explanatory variables only the dummies were used. In the latter model, the M7 model, apart from the dummies the block of lagged dependent variables was also used, which represent the past of dependent variables with lags of 1 (one hour), 2, 3, 24, 25, 26 and 168, 169 and 170 hours.

These models do not include a trend component. Hence the mean prediction error[8] (see formula (3)) has not been changing significantly at the whole forecast period. In case of the M3 model the formula (3) represents the estimate of mean prediction error at the whole forecast horizon, while for the M7 model – only at one ahead forecast horizon, because at longer horizon instead of the unknown values of lagged dependent variables their forecasts obtained on the basis of model M7 were used. Hence the mean prediction error of forecast obtained form the M7 model increases as the forecast horizon lengthens. This behaviour is also observed for the *ex post* errors (see Fig 1).

Forecasts were made for horizon from 1 to 8000 hours (it means up to eleven months), however for the presentation purposes the mean forecast errors only at horizon of 240 hours were shown at Fig. 1. It can be seen that the forecast errors from the M3 model without lagged dependent variables are almost constant. However, for the forecast errors from the M7 model (with lagged dependent variables) the moments of including the forecasts instead of original values can be clearly noticed. The forecast errors start at first to increase quickly, then as the forecast horizon grows the forecast errors is more and more stabilized. Therefore, as the main cause of the increase of the forecast error as the forecast horizon lengthens should be treated the inaccuracy of the values of lagged dependent variable.

It is worth noticing that the M7 model should not be used at forecast horizon longer than 7 days, because for the longer horizon the quality of the model considerably decreases and lower values of the forecast errors are obtained from the M3 model.

[6] A full description of the models can be found in the paper Szmit (2002).

[7] Both models contained over 100 of explanatory variables.

[8] The values of prediction errors were not estimated because of calculation problems: X matrix had 27048 rows and over 100 columns in case of both models.

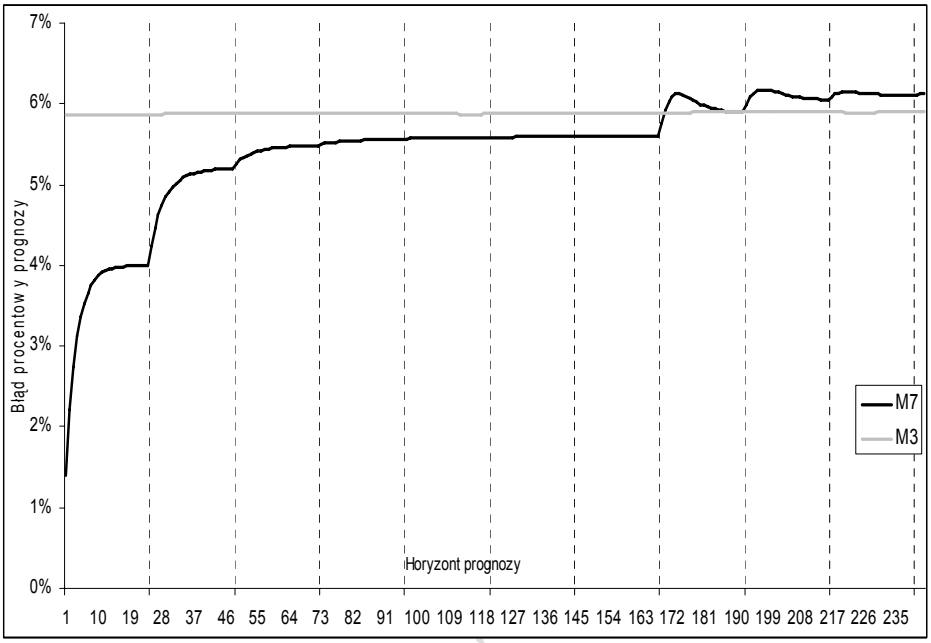

Fig. 1. Values of mean forecast error (MAPE) at the moment τ for forecasts from the M3 and M7 models
Source: Author's calculations.

The forecast horizon used in the study (up to 11 month) is the short-term horizon which means that no qualitative changes of a described phenomenon are revealed. Electric power consumption belongs to the phenomenon with a great regularity and inertia of time series.

References

Cieślak, M. (red.) (1999), *Prognozowanie gospodarcze. Metody i zastosowania*, PWN Publishing House, Warsaw.

Szmit, A. (2002), *Prognozowanie zapotrzebowania na energię elektryczną. Studium empiryczne dla regionu łódzkiego*, PhD Thesis, Technical University of Łódź 2002 (manuscript).

Witkowska, D. (2002), *Sztuczne sieci neuronowe i metody statystyczne*, C. H. Beck Publishing House, Warsaw.

Zeliaś, A. (1997), *Teoria prognozy*, Polish Economic Publishing House, Warsaw.