STATISTICAL FORECASTING OF PANEL DATA BASED ON STATE SPACE MODELS

Lobach V. I., Merkulov R. I., Lobach S. V.
Mathematical Modeling and Data Analysis Department, Belarusian State University
Minsk, Republic of Belarus
E-mail: lobach@bsu.by, merkylovecom@mail.ru, lobashS@bsu.by

Panel (or longitudinal) data describes a set of objects which are observed during certain period of time, so they consist of repeated observations of the same objects in sequential time periods. The following examples of panel data can be mentioned: annual household studies, monthly performance indicators for economic institutions and many others. In this study we provide another approach to forecasting cross-sectional data based on state space models together with Kalman filtering procedure.

Introduction

In economic researches regression models are widely used within large number of applications [1]. Regression models for panel data allow usage of two indices to describe the data:

$$y_{i,t} = \alpha + X_{i,t}\beta + x_{i,t},$$

where i defines object index (household, company, country, etc), t denotes timestamp of an observation, α is an unknown intercept, β is a $(n \times 1)$-vector of unknown parameters, $X_{i,t}$ is a known matrix denoting factors which influence observations. Uncontrollable factors $x_{i,t}$ are modeled with the following equation:

$$x_{i,t} = \mu_i + \epsilon_{i,t},$$

where μ_i is an unobservable individual effect of $i-$th object, $\epsilon_{i,t}$ is a random variable which defines random uncontrollable effect.

Statistical analysis of panel data can be carried out using state space models. In order to express panel data in a state space form it is necessary to introduce one more index i for state parameters vector x_t in classic state space model formulation. This resutls in $x_{i,t}$, where $t = 1, \cdots, T_i$, $i = 1, \cdots, K$, t denotes timestamp, i denotes object index. It means that the mathematical model for panel data is a random field $\{x_{i,t}\}$, $t = \overline{1, T_i}$, $i = \overline{1, K}$.

Based on linear state space models [2] we express panel data in a state space form:

$$x_{i,t} = Fx_{i,t-1} + \omega_{1,t},$$

$$y_{i,t} = Hx_{i,t} + \omega_{2,t},$$

where $x_{i,t}$ is an unobserved state of i-th object at moment t, $y_{i,t}$ is an observation for the object at the same moment. In common case $x_{i,t} \in R_1^n$, $y_{i,t} \in R_2^n$, $\{\omega_{1,t}\}$ and $\{\omega_{2,t}\}$ are sequences of i.i.d. random variables $\omega_{1,t}$, $\omega_{2,t} \sim N(0, Q)$, $x_{i,0} \sim N(\mu, P)$. The parameters of the model are F, H, μ, P. And the problem is to estimate future observations $x_{i,t+h}$, $y_{i,t+h}$ based on previous observations $y_{i,s}$, $s = 1, \cdots, t$, $h > 0$.

I. Kalman Filter

Kalman Filter [3] allows to build optimal in mean-squared sense forecasts if they are introduced in linear state space form. Let us consider the following

$$x_{i,t}^t = E\{x_{i,t}|y_{i,0}^t\},$$

$$P_{i,t_1,t_2}^t = E\{(x_{i,t_1} - x_{i,t_1}^t)(x_{i,t_2} - x_{i,t_2}^t)|y_{i,0}^t\},$$

where $y_{i,0}^t = \{y_{i,j}, j = 1, \cdots, t\}$.

Kalman Filter can be expressed using the following equations [4]

$$x_{i,t}^t = x_{i,t}^{t-1} + K_{i,t}(y_{i,t} - H_{i,t}x_{i,t}^{t-1}, \quad (1)$$

$$P_{i,t}^t = (1 - K_{i,t}H_{i,t})P_{i,t}^{t-1}, \quad (2)$$

$$K_{i,t} = P_{i,t}^{t-1}H_{i,t}^T(H_{i,t}P_{i,t}^{t-1}H_{i,t}^T + R)^{-1}, \quad (3)$$

where $i = \overline{1, K}$, $t = \overline{1, T_i}$, $x_{i,0} = \mu$, $P_{i,0}^0 = P$.

In order to compute forecasts for $x_{i,t}$ for h lags forward equations $(1) - (3)$ are used with initial values $x_{i,t}^T$, $P_{i,t}^T$ instead of $x_{i,0}^0$, $P_{i,0}^0$.

In order to predict observed values $y_{i,t}$ for h future lags we provide the following procedure:

$$y_{i,t+h} = E\{y_{i,t+h}|y_{i,0}^T\},$$

$$B_{i,T+h}^T = E\{(y_{i,T+h} - y_{i,T+h}^T)|y_{i,0}^T\},$$

Using Kalman Filter $(1) - (3)$ the following equations for forecasting statistics are provided:

$$x_{i,T+h}^T = Fx_{i,T+h-1}^T, \quad (4)$$

$$y_{i,T+h}^T = H_i x_{i,T+h}^T, \quad (5)$$

$$P_{i,T+h}^T = FP_{i,T+h-1}^T F^T + Q, \quad (6)$$

$$B_{i,T+h}^T = H_i P_{i,T+h}^T H_i + R. \quad (7)$$

II. Panel data in linear state space form

Classic linear mixed regression model in a compact form can be expressed in the following way:

$$y = X\beta + Z\gamma + \epsilon, E\{\gamma, \epsilon)\} = (0, 0),$$
$$cov(\gamma, \epsilon) = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}$$

where y is observed variable with the following expectation and covariance $E\{y\} = XB$, $cov(y, y) = ZQZ^T + R$. Matrices X and Z describe determined and stochastic effects in observations respectively. For panel data modification of a linear mixed regression model observations for i-th object $y_i = (y_{i,1}|, \cdots, y_{i,T_i})^T$, $i = \overline{1, K}$ are aggragated for $t = \overline{1, T_i}$ which results in the following model:

$$y_i = X_i\beta + Z_i\gamma_i + \epsilon_i,$$
$$\gamma_i \sim N(0, G),$$
$$\epsilon_i = (\epsilon_{i,1}, \epsilon_{i,T_i})^T \sim N(0, \Sigma),$$

which leads to $y_i \sim N(X_i\beta, Z_iGZ_i^T + \Sigma_i)$.

One of the possible ways of expressing longitudinal modification of mixed regression model in state space form can be expressing observations $y_{i,t}$ as a single vector of higher dimensionality, then the state and observation equations can be formulated as following

$$y_{i,t} = x_{i,t}^T\beta_{i,t} + Z_{i,t}^T\gamma + \epsilon_{i,t}, \quad (8)$$

$$\beta_{i,t} = \beta_{i,t-1}, \quad (9)$$

where $\epsilon_{i,t} \sim N(0, \sigma^2)$.

Then we apply Kalman filtering procedure $(1) - (3)$ to the panel data model $(8) - (9)$ and finally construct forecasting statistics $(4) - (7)$.

III. Computational experiments

Let us consider the case described with the model $(8) - (9)$. Let the observation vector be a constant vector with additive errors defined by $AR(1)$ process:

$$y_{i,t} = \beta_i + \epsilon_i,$$
$$\epsilon_i \sim N(0, \Sigma_t),$$
$$\Sigma_t(i, j) = \frac{\sigma^2\phi^{|i-j|}}{1 - \phi^2}, |\phi| < 1.$$

One of possible state space models for this case can be the following:

$$y_{i,t} = \beta_i + \epsilon_t,$$
$$\epsilon_t = \phi\epsilon_t + \omega_t, \omega_t \sim N(0, \sigma^2)$$

with the initial conditions $\epsilon_t = N(0, \frac{\sigma^2}{1-\phi^2})$.

The task is to estimate model parameters which can be non-trivial due to nonlinear relationships between parameters. After parameters' estimates are built they can be used to construct forecasts $x_{i,t+h}$, $y_{i,t+h}$. In order to avoid this problem we construct another state space form

$$y_{i,t} = \mu + \beta_i + \epsilon_i,$$
$$x_{i,t} = \begin{pmatrix} \epsilon_t \\ \beta_i \end{pmatrix} = \begin{bmatrix} \phi & 1 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \epsilon_{t-1} \\ \beta_i \end{pmatrix} + \begin{pmatrix} \omega_t \\ 0 \end{pmatrix},$$
$$y_{i,t} = (1, 1)x_{i,t},$$
$$\omega_t \sim N(0, \Omega),$$

with the initial condition $(\epsilon_0, \beta_i)^T \sim N(0, G)$, where

$$G = \begin{bmatrix} \frac{\sigma^2}{1-\phi^2} & 0 \\ 0 & 0 \end{bmatrix}$$

$$\Omega = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Conclusion

Finally this results in linear state space model and we can apply Kalman filtering procedure $(1) - (7)$. For computational experiments we generated two-dimensional time series according to model described above. The experients were carried out with the following parameters: $\mu = 0$, $\sigma^2 = 1$, $\phi = 0.5$, $\beta_i = 1$. To construct forecasting statistics equations $(4) - (7)$ were used. Forecasting horizon with $h = 10$ was used. We observed mean absolute percentage error below 2.1% which indicates possibility of modeling panel data using the described approach.

1. Baltagi, Badi H. Econometric analysis of panel data / B.H. Baltagi // John & Sons. New York, 2004. – 284 p.
2. Ivchenko G.I. Mathematical Statistics / G. I. Ivchenko // High, Sch., p. 248.
3. Harvey A. C. Forecasting, Structural Time Series Models and the Kalman Filter / A. C. Harvey // Cambridge: Cambridge University Press. p. 227.
4. Liptser R. Sp. Stochastic Processes Statistics / R. Sp. Liptser // Nauka, p. 696.