Dynamiczne modele liniowe w badaniach okresowych

mgr Kamil Wilak

Katedra Statystyki
UE w Poznaniu

O czym będzie mowa?

- badamy zmienność pewnego parametru w czasie w pewnej populacji
- co pewien okres losujemy próbę
- na podstawie próby szacujemy badany parametr
- precyzja oszacowania zależy od wielkości próby
- co, gdy mamy za małą próbę?

Dynamiczne modele liniowe (DLM)

Specyfikacja modelu

$Y_t = \boldsymbol{F}_t' \boldsymbol{\theta}_t + \varepsilon_t$	$\varepsilon_t \sim N(0, \sigma^2_{\varepsilon,t})$	- równanie pomiaru
$\boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t$	$\boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{W}_t)$	- równanie przejścia

gdzie

- Y_t - obserwowana wartość szeregu czasowego
- $\boldsymbol{\theta}_t$ - nieobserwowany wektor parametrów stanu
- $\boldsymbol{F}_t$ i $\mathbf{G}_t$ - wektor i macierz znanych współczynników liniowych
- ε_t - zm. los. opisująca zmienność nieopisana przez parametry stanu / błąd obs.
- $\boldsymbol{\eta}_t$ - wektor losowy opisujący stratę w czasie informacji o wartości wektora $\boldsymbol{\theta}_t$
- $\mathbf{W}_t$ - macierz kowariancji macierzy wektora losowego $\boldsymbol{\eta}_t$

Dynamiczne modele liniowe (DLM)

Filtr Kalmana - filtrowanie

D_t informacje dostępne w okresie t, $\theta_t | D_t \sim N(\boldsymbol{m}_t, \mathbf{C}_t)$

- prognoza 1 krok naprzód wektora $\boldsymbol{\theta}_{t+1}$ na podstawie informacji D_t ma rozkład normalny z parametrami:

$$\boldsymbol{a}_{t+1} = E(\boldsymbol{\theta}_{t+1} | D_t) = \mathbf{G}_{t+1} \boldsymbol{m}_t$$
$$\mathbf{Z}_{t+1} = Var(\boldsymbol{\theta}_{t+1} | D_t) = \mathbf{G}_{t+1} \mathbf{C}_t \mathbf{G}'_{t+1} + \mathbf{W}_{t+1}$$

- prognoza 1 krok naprzód Y_{t+1} na podstawie informacji D_t ma rozkład normalny z parametrami:

$$f_{t+1} = E(Y_{t+1} | D_t) = \boldsymbol{F}_{t+1} \boldsymbol{a}_t$$
$$\mathbf{Q}_t = Var(Y_{t+1} | D_t) = \boldsymbol{F}_{t+1} \mathbf{Z}_{t+1} \boldsymbol{F}'_{t+1} + \sigma^2_{\varepsilon, t+1}$$

Dynamiczne modele liniowe (DLM)

Filtr Kalmana - filtrowanie cd.

$D_{t+1} = \{D_t, Y_{t+1}\}$ informacje dostępne w okresie t

- oszacowanie wektora $\boldsymbol{\theta}_{t+1}$ na podstawie informacji D_{t+1} ma rozkład normalny z parametrami:

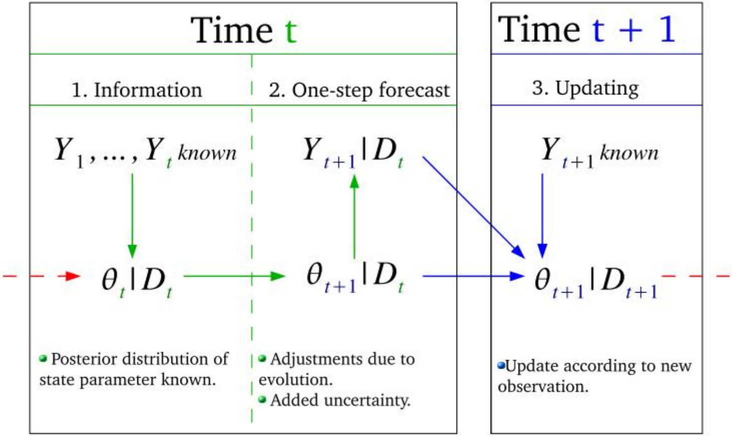
$$\boldsymbol{m}_{t+1} = E(\boldsymbol{\theta}_{t+1}|D_{t+1}) = \boldsymbol{a}_t + \mathbf{Z}_{t+1}\boldsymbol{F}'_{t+1}\mathbf{Q}^{-1}_{t+1}(Y_{t+1} - f_{t+1})$$

$$\mathbf{C}_{t+1} = Var(\boldsymbol{\theta}_{t+1}|D_{t+1}) = \mathbf{Z}_{t+1} - \mathbf{Z}_{t+1}\boldsymbol{F}'_{t+1}\mathbf{Q}^{-1}_{t+1}\boldsymbol{F}_{t+1}\mathbf{Z}_{t+1}$$

Potrzebna jest macierz $\mathbf{W}_t$
Można ją oszacować za pomocą Metody Największej Wiarygodności

Dynamiczne modele liniowe

Filtr Kalmana - filtrowanie cd.



Źródło: Solhjell Ida Kjersem [2009] *Bayesian Forecasting and Dynamic Models Applied to Strain Data from Gota River Bridge*

Filtr Kalmana - wygładzanie

Jeżeli $\boldsymbol{\theta}_{t+1}|D_t \sim N(s_{t+1}, S_{t+1}$, wtedy $\boldsymbol{\theta}_t|D_t \sim N(s_t, S_t)$, gdzie:

$$s_t = m_t + \mathbf{C}_t\mathbf{G}'_{t+1}\mathbf{R}^{-1}_{t+1}(s_{t+1} - a_{t+1})$$

$$S_t = \mathbf{C}_t + \mathbf{C}_t\mathbf{G}'_{t+1}\mathbf{R}^{-1}_{t+1}(\mathbf{S}_{t+1} - \mathbf{R}_{t+1})\mathbf{R}^{-1}_{t+1}\mathbf{G}_{t+1}\mathbf{C}_t$$

Idea wykorzystania DLM w badaniach okresowych

Niech

- Y_t - badany parametr populacji
- $\hat{Y}_t$ - oszacowanie bezpośrednie parametru populacji
- e_t - błąd oszacowania, $e_t \sim N(0, V_t)$, $V_t = Var(\hat{Y}_t)$

Wtedy

$$\hat{Y}_t = Y_t + e_t \tag{1}$$

Załóżmy, że:

$$Y_t = \boldsymbol{F}_t'\boldsymbol{\theta}_t + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2_{\varepsilon,t}) \tag{2}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{W}_t) \tag{3}$$

(2)-(3)$\rightarrow$(1), stąd

$$\hat{Y}_t = \boldsymbol{F}_t'\boldsymbol{\theta}_t + \varepsilon_t + e_t = \boldsymbol{F}_t'\boldsymbol{\theta}_t + v_t \qquad v_t \sim N(0, \sigma^2_v V_t), \quad \sigma^2_v \approx 1 \tag{4}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{W}_t) \tag{5}$$

Przykład

Dane

- dane jednostkowe z Badania Aktywności Ekonomicznej Ludności 2000-2005

Populacja

- osoby biorące udział w BAEL-u

Badany parametr

- miesięczna stopa bezrobocia $Y_t = \frac{N_t^B}{N_t^A}$

N_t^B - liczba osób bezrobotnych w miesiącu t

N_t^A - liczba osób aktywnych zawodowo w miesiącu t

Losowanie próby

- proste bez zwracania, wielkość próby - 500

Przykład

Oszacowanie bezpośrednie

$$\hat{Y}_t = \frac{n_t^B}{n_t^A}$$

n_t^B - liczba osób bezrobotnych w próbie wylosowanej w miesiącu t

n_t^A - liczba osób aktywnych zawodowo w próbie wylosowanej w miesiącu t

Precyzja oszacowania bezpośredniego

$$V_t = Var(\hat{Y}_t) \approx \frac{(N_t - n_t)n_t}{N_t(n_t - 1)} \frac{n_t^B(n_t^A - n_t^B)}{(n_t^A)^3}$$

N_t - wielkość populacji w miesiącu t

n_t - wielkość próby wylosowanej w miesiącu t

$$\sqrt{Var(\hat{Y}_t)} \approx 0.023$$

$$\frac{\sqrt{Var(\hat{Y}_t)}}{Y_t} \approx 0.125$$

Przykład

$$\begin{aligned} \hat{Y}_t &= Y_t + e_t = & e_t &\sim N(0, \sigma^2_{e,t}) \\ &= L_t + S_t + \varepsilon_t + e_t = & \varepsilon_t &\sim N(0, \sigma^2_{\varepsilon,t}) \\ &= L_t + S_t + v_t & v_t &= e_t + \varepsilon_t \sim N(0, \sigma^2_v V_t) \end{aligned}$$

e_t - błąd oszacowania, ε_t - reszta modelu

$$\left.\begin{array}{l} L_t = L_{t-1} + R_{t-1} \\ R_t = R_{t-1} + \omega_{R,t} \\ E(\omega_{R,t}) = 0 \\ Var(\omega_{R,t}) = \sigma^2_R \end{array}\right\} \text{trend lokalnie liniowy}$$

$$\left.\begin{array}{l} S_t = \sum_{j=1}^{6} S_{t,j} \\ S_{t,j} = \cos\left(\frac{j\pi}{6}\right) S_{t-1,j} + \sin\left(\frac{j\pi}{6}\right) S^*_{t-1,j} + \omega_{S,t,j} \\ S^*_{t,j} = -\sin\left(\frac{j\pi}{6}\right) S_{t-1,j} + \cos\left(\frac{j\pi}{6}\right) S^*_{t-1,j} + \omega^*_{S,t,j} \\ E(\omega_{S,t,j}) = E(\omega^*_{S,t,j}) = 0 \\ Cov(\omega_{S,t,j}, \omega^*_{S,t\prime,j\prime}) = \begin{cases} \sigma^2_{S,j}, & t = t\prime, \quad j = j\prime \\ 0, & \text{wpp} \end{cases} \end{array}\right\} \begin{array}{l}\text{sezonowość} \\ \text{trygonometryczna}\end{array}$$

Dynamiczne modele liniowe

$$\left.\begin{array}{l}
\boldsymbol{F}_t = (1,0,1,0,1,0,1,0,1,0,1,0,1) \\
\boldsymbol{\theta}_t = (L_t, R_t, S_{t,1}, S^*_{t,1}, \ldots, S_{t,6})^T \\
\mathbf{G}_t = blockdiag(\mathbf{G}_t^R, \mathbf{G}_t^{S_1}, \ldots, \mathbf{G}_t^{S_6}) \\
\mathbf{G}_t^R = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\
\mathbf{G}_t^{S_j} = \begin{pmatrix} \cos(j\pi/6) & \sin(j\pi/6) \\ -\sin(j\pi/6) & \cos(j\pi/6) \end{pmatrix} \\
\boldsymbol{\eta}_t = (0, \omega_{R,t}, \omega_{S,t,1}, \omega^*_{S,t,1}, \ldots, \omega_{S,t,6})
\end{array}\right\} \Longrightarrow \begin{array}{l}
\hat{Y}_t = \mathbf{Z}_t \boldsymbol{\theta}_t + v_t \\
\boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\eta_t}
\end{array}$$

Dynamiczne Modele Liniowe w R

Pakiet *dlm*

- Giovanni Petris, Sonia Petrone, Partizia Campagnoli [2007/2011] *Dynamic Linear Models with R*
- Giovanni Petris [2010] *dlm - an R package for Bayesian analysis of Dynamic Linear Models*

Dziękuję za uwagę