Прикладная статистика. Лабораторная 2.

Алексей Григорьев, 674 гр. ФУПМ МФТИ

Май 2020

№2.3 Делаем минимальный препроцессинг, ряд не стационарен, тренд очевиден. Дифференцируем ряд и сдвигаем ряд на его среднее значение $\overline{y}$, чтобы среднее видоизмененного ряда было равно нулю, тогда допуская стационарность ряда (проверка по kpss) мы можем записать для него модель ARMA(p, q):

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q} = \sum_{i=1}^{p} \phi_i x_{t-i} + u_t + \sum_{i=1}^{q} \theta_i u_{t-i},$$

здесь подразумевается нормальность и независимость шума $u_t \sim i.i.d.N(0, \sigma^2)$.

Через графики asf и pasf производим первоначальную оценку парметров модели p и q, вблизи этих значений ищем оптимум, обучая ARMA модели и рассматривая достигаемые значения AIC. Приходим к модели ARMA(6, 2).

Модель описывается $p+q+1$ параметром: $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$ и σ^2, далее будем производить оценку этих параметров. Это может быть сделано несколькими способами, здесь же предлагается использовать наиболее простой - двухшаговую регрессию.

Задача двухшаговой регрессии сводится к построении двух последовательных регрессий, которые могут быть переформулированы в виде задач оптимизации, которые будут поставлены ниже.

На первом шаге мы строим регрессию $x_{t-1}, \ldots, x_{t-p}$ на x_t. Тогда для ряда длины N получим следующую систему уравнений на $\hat{\phi}_1, \ldots, \hat{\phi}_p$.

$$\begin{bmatrix} x_p & x_{p-1} & \cdots & x_1 \\ x_{p+1} & x_p & \cdots & x_2 \\ x_{p+2} & x_{p+1} & \cdots & x_3 \\ \vdots & \vdots & \ddots & \vdots \\ x_{N-1} & x_{N-2} & \cdots & x_{N-p} \end{bmatrix} \cdot \begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_p \end{bmatrix} = \begin{bmatrix} x_{p+1} \\ x_{p+2} \\ x_{p+3} \\ \vdots \\ x_N \end{bmatrix}$$

Формально на данном шаге мы решаем задачу минимизации:

$$\|A\boldsymbol{\phi} - b\|^2 \to \min_{\boldsymbol{\phi}}$$

Систему решаем через МНК: $A\boldsymbol{\phi} = b \Rightarrow \hat{\boldsymbol{\phi}} = (A^T A)^{-1} A^T b$.Полученные оценки $\hat{\phi}_1, \ldots, \hat{\phi}_p$ используются только для оценки шума, т.е. шум в данном подходе определяется через невязки:

$$\hat{u_n} = x_n - (\hat{\phi}_1 x_{n-1} + \cdots + \hat{\phi}_p x_{n-p}), \ \forall n = p+1, \ldots, N$$

В силу того, что $u_n \sim i.i.d.N(0, \sigma^2)$ получим несмещенную оценку дисперсии:

$$\hat{\sigma}^2 = \frac{1}{N-p-1} \sum_{n=p+1}^{N} \hat{u_n}^2$$

Вторым шагом является построение регрессии $x_{t-1}, \dots, x_{t-p}, \hat{u}_{t-1}, \dots, \hat{u}_{t-q}$ на x_t, таким образом мы получим валидные оценки $\hat{\phi}_1, \dots, \hat{\phi}_p, \hat{\theta}_1, \dots, \hat{\theta}_q$, разрешив через МНК следующую систему (в допущении $p > q$).

$$\begin{bmatrix} x_p & x_{p-1} & \dots & x_1 & u_p & u_{p-1} & \dots & u_{p-q+1} \\ x_{p+1} & x_p & \dots & x_2 & u_{p+1} & u_p & \dots & u_{p-q+2} \\ x_{p+2} & x_{p+1} & \dots & x_3 & u_{p+2} & u_{p+1} & \dots & u_{p-q+3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N-1} & x_{N-2} & \dots & x_{N-p} & u_{N-1} & u_{N-2} & \dots & u_{N-q} \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_q \end{bmatrix} = \begin{bmatrix} x_{p+1} - u_{p+1} \\ x_{p+2} - u_{p+2} \\ x_{p+3} - u_{p+3} \\ \vdots \\ x_N - u_N \end{bmatrix}$$

Замечание: здесь под u_n понимаются именно полученные ранее оценки $\hat{u}_n$. Так же заметим, что некоторые первые уравнения системы будут отброшены с учетом того, что у нас есть оценки $\hat{u}_n$ только для $n \geq p+1$ и другие шумы не могут фигурировать в матрице.

Здесь решаем задачу минимизации:

$$\|C \cdot (\boldsymbol{\phi}, \boldsymbol{\theta})^T - d\|^2 \to \min_{\boldsymbol{\phi}, \boldsymbol{\theta}}$$

Нахождение этих оценок заканчивает оценку параметров модели ARMA(p, q).

Окончательно выпишем все формулы в более явном виде для $p = 6$, $q = 2$, $N = 29$:

(a) $\begin{bmatrix} \hat{u}_7 \\ \vdots \\ \hat{u}_{29} \end{bmatrix} = b - A(A^TA)^{-1}A^Tb = (I - A(A^TA)^{-1}A^T)b,$

где $b = \begin{bmatrix} x_7 \\ \vdots \\ x_{29} \end{bmatrix}$, $A = \begin{bmatrix} x_6 & x_5 & \dots & x_1 \\ x_7 & x_6 & \dots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{28} & x_{27} & \dots & x_{23} \end{bmatrix}$

Численно $\begin{bmatrix} \hat{u}_7 \\ \vdots \\ \hat{u}_{29} \end{bmatrix} \approx \begin{bmatrix} -0.419 \\ -0.505 \\ 0.23 \\ 0.163 \\ 0.059 \\ -0.061 \\ -0.176 \\ -0.374 \\ -0.381 \\ 0.412 \\ 0.369 \\ 0.252 \\ 0.081 \\ -0.11 \\ -0.397 \\ 0.514 \\ -0.473 \\ -0.54 \\ 0.223 \\ 0.184 \\ 0.105 \\ 0.001 \\ -0.111 \end{bmatrix}$,

(b) $\begin{bmatrix} \hat{\phi}_1 \\ \vdots \\ \hat{\phi}_6 \\ \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = (C^T C)^{-1} C^T d,$

где с учетом исключения первых нескольких уравнений в заявленной выше системе имеем

$C = \begin{bmatrix} x_8 & x_7 & \dots & x_3 & \hat{u}_8 & \hat{u}_7 \\ x_9 & x_8 & \dots & x_4 & \hat{u}_9 & \hat{u}_8 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{28} & x_{27} & \dots & x_{23} & \hat{u}_{28} & \hat{u}_{27} \end{bmatrix}$, $d = \begin{bmatrix} x_7 - \hat{u}_7 \\ x_8 - \hat{u}_8 \\ \vdots \\ x_{29} - \hat{u}_{29} \end{bmatrix}$, $\hat{u}_7, \dots \hat{u}_{29}$ определены выше.

Численно $\begin{bmatrix} \hat{\phi}_1 \\ \vdots \\ \hat{\phi}_6 \end{bmatrix} \approx \begin{bmatrix} -0.286 \\ 0.419 \\ 0.248 \\ -0.012 \\ -0.314 \\ -0.71 \end{bmatrix}$, $\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} \approx \begin{bmatrix} 0.000 \\ 0.000 \end{bmatrix}$.

(c) $\hat{\sigma}^2 = \frac{1}{22} \sum_{n=7}^{29} \hat{u_n}^2 \approx 0.103$, где $\hat{u}_7, \dots \hat{u}_{29}$ определены выше.

Ссылки на описание данного алгоритма оценки параметров ARMA в источниках:

1. lecture by Umberto Triacca (slides 4-5)
2. guide by Michael Hauser