МСПС | Временные ряды |

Опр. Временной ряд – послед. сл. величин $\{\xi_n\}_{n\in\mathbb{N}}$, зависящих от n (грубо говоря, зависящих от времени). ✓ В дальнейшем – y_n

Задача прогнозирования: найти ф-цию f_T: $\xi_{T+d} \approx f_T(\xi_1 \dots \xi_T, d)$

Простейшие методы типа сделать регрессию на время, как правило, не помогают, потому что наблюдения сильно зависимыми от ближайших наблюдений.

Опр. Автокорреляция (оценка ковариац. ф-ции, почти) (ACF)

$r_\tau = r_{t,t+\tau} = \dfrac{\sum_{t=1}^{T-\tau}(y_t-\bar y)(y_{t+\tau}-\bar y)}{\sum_{t=1}^{T}(y_t-\bar y)^2}$, $\bar y = \frac{1}{T}\sum_{t=1}^{T} y_t$.

в предположении, что набл. образуют стац. ряд.

Тренд – плавное долгосрочное изменение ур. ряда

Сезонность – циклич. изм. уровня ряда с постоян. периодом.

Ошибка – непрогнозир. случ. компонента ряда.

Цикл – описывает повторяющиеся, но непериодич. колебания.

Есть алгоритм, который раскладывает временной ряд на 3 компоненты: тренд + циклич. составляющая, сезон. состав. и ошибки: STL-алгоритм (в R – функция stl), алгоритм авт. робастный, есть продолжение идей алгоритма LOWESS и его более ранней версии loess (сначала с помощью непараметрич. регрессии оцен. тренд, затем циклич. составляющая, а потом ошибки корректируются с помощью процедуры, похожей на ту, что применялось в LOWESS). Написать это самому тяжело, но если хочется, то гитайте оригинал. статью Кливленда STL: a seasonal-trend decomposition procedure based on Loess.

Соотв., предсказывать с помощью этой штуки тоже можно, ведь периодич. компонента не меняется, надо лишь оценить тренд.

Можно на временной ряд посмотреть с другой стороны.

Опр. Врем. ряд $(y_1 \dots y_T)$ стационарен в уз. смысле, если $\forall s\ \forall t_1 \dots t_k$ $(y_{t_1} \dots y_{t_k}) \overset{d}{=} (y_{t_1+s} \dots y_{t_k+s})$. В широком, если мат.ож. – константа и ковар. ф-ция зависит только от разности: $\gamma(t-s) = \mathrm{cov}(y_t, y_s) = \mathrm{cov}(y_{t+h}, y_{s+h})$

Ряды с трендом и сезонностью нестационарны. Для рядов с непериодич. циклами модель стац. ряда использовать можно.

Как избавиться от тренда и сезонности вручную? Использовать "дифференцирование": $y'_t = y_t - y_{t-1}$ (это от постоян. тренда), $y''_t = y'_t - y'_{t-1}$ – уже помогает от лин. тренда; пусть сезон имеет длину s, тогда преобразование $y'_t = y_t - y_{t-s}$ наз. сезонным дифференцированием. Сезонное дифференцир. лучше делать первым – после него ряд уже может оказаться стационарным (т.е. дальше понятно, как решать задачу прогноза)

Далее мы будем строить разные модели, но требования к остаткам $\hat{\varepsilon}_t = y_t - \hat{y}_t$, $\hat{y}_t$ – наше предсказание, будут одинаковыми:

1) Несмещённость (проверяем крит. Стьюдента, Уилкоксона и тр.) крит. Дики-Фуллера
2) Стационарность, т.е. независимость от времени (визуал. анализ, крит. KPSS)
3) Неавтокоррелируемость – отсут. неучтённой зависимости от пред. наблюдений. Для остатков – отсутствие кластеров полож. и отриц. значений.

Проверяется с помощью коррелограммы, Q-крит. Льюнга-Бокса.

Желательное, но не обязат. св-ва:

4) Нормальность (знаем: qq-plot и тр.), 5) Гетероскедастичность (Уайт, Бокс-Кокс, Бреуш-Пэган)

Критерий KPSS: $y_t = \delta t + \xi_t + \varepsilon_t$, ξ_t – сл. блуждание $= \sum_{i=1}^{t} \eta_t$, $\eta_t \sim iid(0, \sigma^2)$, ε_t – стац. (можно более слож. тренд)

Пусть $e = (e_1 \dots e_T)$ ряд ошибок прогноза. H_0: y_t – стац.; H_1: ~~ε опис. моделью~~ $\varepsilon_t = \alpha \varepsilon_{t-1}$, $\sigma^2 \neq 0$

Статистика $KPSS(e) = \frac{\sum_{t=1}^{T} S_t^2}{\hat{\sigma}_e^2}$, где $S_t = \sum_{i=1}^{t} e_i$; $\hat{\sigma}_\varepsilon^2 = \frac{1}{T}\sum_{i=1}^{T} e_i^2$, e_t – остатки при регрессии y_t по времени и тренду.

(если e_i не одинаково расп., то $\hat{\sigma}_e^2$ заменяется на другую статистику $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} e_t^2 + \frac{2}{T}\sum_{\tau=1}^{[\sqrt{T}]}\left(1 - \frac{\tau}{[\sqrt{T}]+1}\right)\cdot \sum_{t=\tau+1}^{T} e_t e_{t-\tau}$, подробнее см. монографию Maddala, Kim. Unit roots... (1998)). При H_0 статистика KPSS имеет табл. распределение.

Q-критерий Льюнга-Бокса

Пусть $\varepsilon = (\varepsilon_1 \dots \varepsilon_n)$ – ряд ошибок прогноза. H_0: $r_1 = \dots = r_L = 0$, H_1: H_0 неверна

Статистика $Q = T(T+2)\sum_{\tau=1}^{L} \frac{r_\tau^2}{T-\tau}$, при верной H_0 $Q \sim \chi^2_{L-k}$, k – число настраиваемых пар-ров модели ряда.

Коррелограмма – это график автокорреляцион. ф-ции.

Теперь к авторегрессиям:

Опр. Процесс авторегрессии порядка p (обозн. $AR(p)$) – это $D\varepsilon_t = \sigma_\varepsilon^2$

$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \dots + a_p X_{t-p} + \varepsilon_t$, $a_p \neq 0$, где ε_t – проц. белого шума, не корр. с X_t.

Введём оператор запаздывания L: $LX_t = X_{t-1}$, тогда соотношение $AR(p)$ можно записать так: $a(L)X_t = \varepsilon_t$, где $a(L) = 1 - (a_1 L + a_2 L^2 + \dots + a_p L^p)$

Теорема (Условие стац. $AR(p)$) $a(L)X_t = \varepsilon_t$ стационарен $\Leftrightarrow$ корни $a(z) = 0$ лежат вне круга $|z| \leq 1$ на комплексной плоскости.

Опр. Процесс скользящего среднего порядка q ($MA(q)$) – это

$X_t = \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q}$, $b_q \neq 0$, где ε_t – проц. белого шума с $D\varepsilon_t = \sigma_\varepsilon^2$.

Такая модель всегда ава. стац. процессом

Утв. $\forall$ стац. процесс $AR(p)$, задаваемый соотн. $a(L)(X_t - \mu) = \varepsilon_t$, можно представить в виде проц. $MA(\infty)$. Если X_t имеет $MA(q)$-представ. $X_t - \mu = b(L)\varepsilon_t = \sum_{j=0}^{q} b_j \varepsilon_{t-j}$, $b_0 = 1$, и все корни ур-ния $b(z) = 0$ по модулю > 1 (в комп. плоскости) (это нар. условие обратимости), то $\exists$ представ. X_t в виде $AR(\infty)$ модели.

Опр. X_t – авторегрессии, если $EX_t = 0$ и

$a(L)X_t = b(L)\varepsilon_t$, т.е. $X_t = \sum_{j=1}^{p} a_j X_{t-j} + \sum_{j=0}^{q} b_j \varepsilon_{t-j}$, $a_p \neq 0$, $b_q \neq 0$.

Теорема (разложение Вольда) $\forall$ стац. в широком смысле центрир. процесс X_t может быть представлен в виде $X_t = \sum_{j=0}^{+\infty} c_j \varepsilon_{t-j} + \sum_{i=1}^{+\infty} a_i X_{t-i}$, где ε_t – проц. белого шума, z_t – стац., центрир. сл. проц., ~~не зависит~~ с ε_t.

Вывод: $\forall$ стац. ряд можно приблизить моделью ARMA(p,q) сколь угодно точно.

Опр. Ряд опис. моделью ARIMA(p,d,q), если ряд его разностей (т.е. дифференцирований) $\nabla^d X_t = (1-L)^d X_t$ описывается моделью ARMA(p,q):

$a(L)\nabla^d X_t = b(L)\varepsilon_t$

Можно еще сезон. ~~комп.~~ добавить – будет SARIMA. дифференц.

Теперь к выбору коэфф. p, d, q и оценке пар-ров модели:

Определим PACF (частич. автокоррел. стац. ряда X_t)

$\varphi_h = \begin{cases} r(X_{t+1}, X_t), & h=1 \\ r(X_{t+h} - X_{t+h}^{h-1}, X_t - X_t^{h-1}) \end{cases}$ ← оценка корреляции X_{t+1} и X_t, см. стр. 1

где X_t^{h-1} – регрессия X_t на $X_{t+1}, \dots X_{t+h-1}$, т.е. $X_t^{h-1} = \beta_1 X_{t+1} + \dots + \beta_{h-1} X_{t+h-1}$ и, наоборот, $X_{t+h}^{h-1} = \alpha_1 X_{t+h-1} + \dots + \alpha_{h-1} X_{t+1}$.

Оценки пар-ров модели:

1) При выборе p и q можно использ. ACF и PACF:
   - в модели ARIMA(p,d,0) ACF экспоненц. затухает или имеет синус. вид, а PACF значимо отлич. от 0 при $h = p$.
   - в модели ARIMA(0,d,q), наоборот, PACF экспон. затухает или имеет ~~экс~~синус. вид, а ACF значимо отлич. от 0 при $\tau = q$.

2) d выбирается так, чтобы ряд был похож на стационарный

3) Пусть знаем распред. ε (напр., нормальное) – при заданных p, d, q коэфф. модели оцен. методом макс. пр-я.

Информационные критерии (чем меньше, тем лучше модель)

- $AIC_c = -2L + \frac{2(p+q+1)(p+q+2)}{T-p-q-2}$, где L – ф-ция логарифм. пр-я
- $BIC = -2L + \frac{(\log T - 2)(p+q+1)}{T}$.

Если модель – AR(p), то ~~в~~ лучше заменить $-2L$ на $\ln \sigma_\varepsilon^2$.

В R существует 2 полез. ф-ции: auto.arima и forecast.

И, наконец, как же получить прогноз $\hat{X}_{T+h}$?

Надо сначала оценить последовательно $\hat{X}_{T+1}$, $\hat{X}_{T+2}$ и тд, каждый раз сдвигая время на 1, коэфф. при X_i и ε_i оставляя теми же, новые ошибки ε_{T+1}, ... приравнивая к 0, а старые заменяя на остатки.

Стоит еще обратить внимание на метод SSA (и Гусеница-).