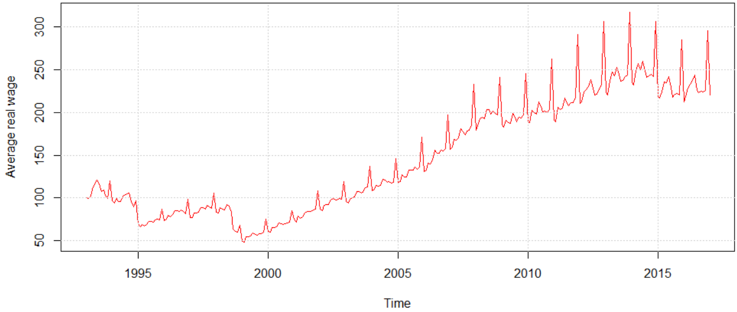
Прикладной статистический анализ данных.
9. Анализ временных рядов

Ольга Добролюбова
Юлиан Сердюк
cs.msu.psad@gmail.com

08.04.2022

Прогнозирование временного ряда

Временной ряд: $y_1, \dots, y_T, \dots, \quad y_t \in \mathbb{R}$, — значения признака, измеренные через постоянные временные интервалы.



Задача прогнозирования — найти функцию f_T:

$$y_{T+d} \approx f_T\left(y_T, \dots, y_1, d\right) \equiv \hat{y}_{T+d|T},$$

где $d \in \{1, \dots, D\}$ — отсрочка прогноза, D — горизонт прогнозирования.

Автокорреляционная функция (ACF)

Наблюдения временного ряда автокоррелированы.

Автокорреляция:

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

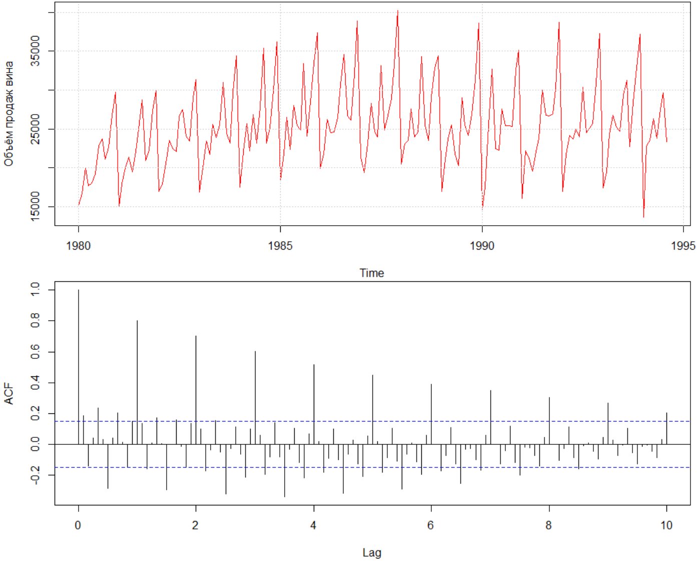
$r_\tau \in [-1, 1]$, τ — лаг автокорреляции.

Проверка значимости отличия автокорреляции от нуля:

временной ряд:	$Y^T = Y_1, \dots, Y_T$;
нулевая гипотеза:	$H_0 \colon r_\tau = 0$;
альтернатива:	$H_1 \colon r_\tau \neq 0$;
статистика:	$T\left(Y^T\right) = \frac{r_\tau \sqrt{T-\tau-2}}{\sqrt{1-r_\tau^2}}$;
нулевое распределение:	$St\left(T-\tau-2\right)$.

Автокорреляционная функция (ACF)

Коррелограмма:



Частичная автокорреляционная функция (PACF)

Частичная автокорреляция стационарного ряда y_t — автокорреляция остатков авторегрессии предыдущего порядка:

$$\phi_{hh} = \begin{cases} r\left(y_{t+1}, y_t\right), & h = 1, \\ r\left(y_{t+h} - \hat{y}_{t+h}, y_t - \hat{y}_t\right), & h \geq 2, \end{cases}$$

где $\hat{y}_{t+h}$ и $\hat{y}_t$ — предсказания регрессий y_{t+h} и y_t на $y_{t+1}, y_{t+2}, \ldots, y_{t+h-1}$:

$$\begin{aligned} \hat{y}_t &= \beta_1 y_{t+1} + \beta_2 y_{t+2} + \cdots + \beta_{h-1} y_{t+h-1}, \\ \hat{y}_{t+h} &= \beta_1 y_{t+h-1} + \beta_2 y_{t+h-2} + \cdots + \beta_{h-1} y_{t+1}. \end{aligned}$$

Q-критерий Льюнга-Бокса

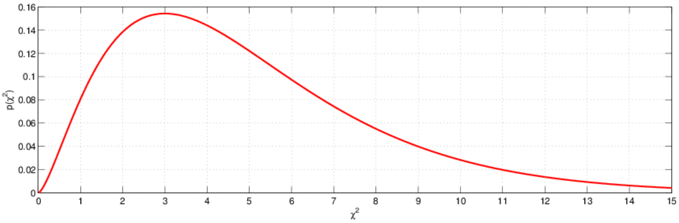
ряд ошибок прогноза: $\varepsilon^T = \varepsilon_1, \ldots, \varepsilon_T$;

нулевая гипотеза: $H_0\colon r_1 = \cdots = r_L = 0$;

альтернатива: $H_1\colon H_0$ неверна;

статистика: $Q\left(\varepsilon^T\right) = T\left(T+2\right)\sum\limits_{\tau=1}^{L}\frac{r_\tau^2}{T-\tau}$;

нулевое распределение: χ^2_{L-K}, K — число настраиваемых параметров модели ряда.



Компоненты временных рядов

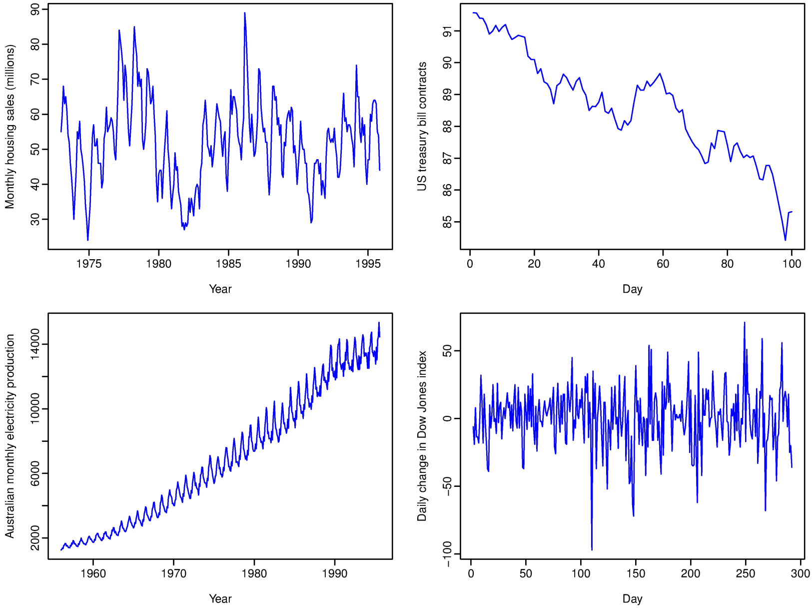
Тренд — плавное долгосрочное изменение уровня ряда.

Сезонность — циклические изменения уровня ряда с постоянным периодом.

Цикл — изменения уровня ряда с переменным периодом (цикл жизни товара, экономические волны, периоды солнечной активности).

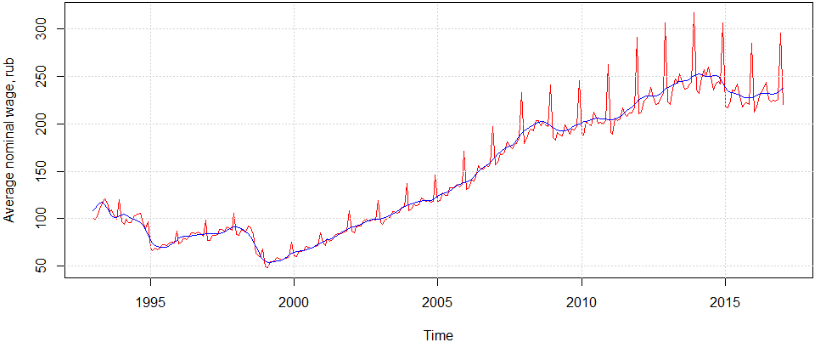
Ошибка — непрогнозируемая случайная компонента ряда.

Компоненты временных рядов



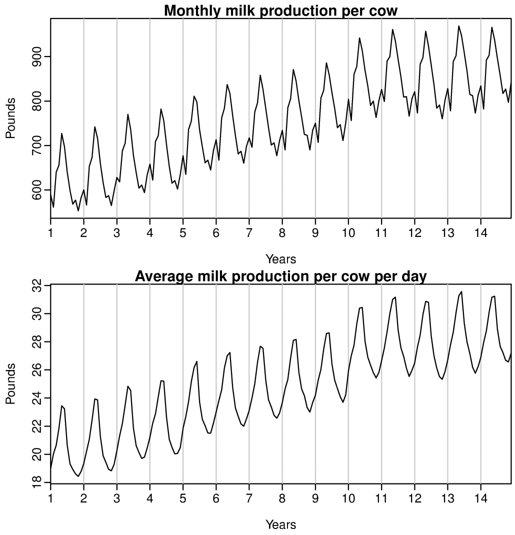
Снятие сезонности

Часто для удобства интерпретации ряда сезонная компонента вычитается:



Календарные эффекты

Иногда упростить структуру временного ряда можно за счёт учёта неравномерности отсчётов:



Стационарность

Ряд $y_1, \dots, y_T$ **стационарен**, если $\forall s$ распределение $y_t, \dots, y_{t+s}$ не зависит от t, т. е. его свойства не зависят от времени.

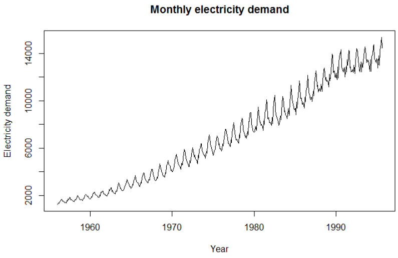
Ряды с трендом или сезонностью нестационарны.

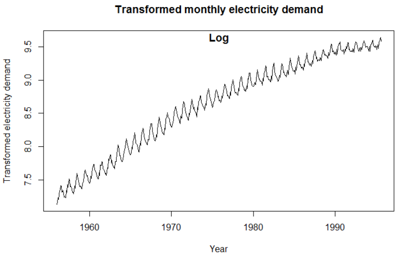
Ряды с непериодическими циклами стационарны, поскольку нельзя предсказать заранее, где будут находится максимумы и минимумы.

Стабилизация дисперсии

Для рядов с монотонно меняющейся дисперсией можно использовать стабилизирующие преобразования.

Часто используют логарифмирование:



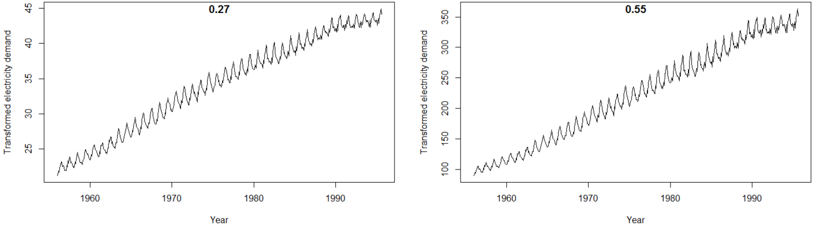


Преобразования Бокса-Кокса

Параметрическое семейство стабилизирующих дисперсию преобразований:

$$y'_t = \begin{cases} \ln y_t, & \lambda = 0, \\ \left(y_t^\lambda - 1\right)/\lambda, & \lambda \neq 0. \end{cases}$$

Параметр λ выбирается так, чтобы минимизировать дисперсию или максимизировать правдоподобие модели.



Преобразования Бокса-Кокса

После построения прогноза для трансформированного ряда его нужно преобразовать в прогноз исходного:

$$\hat{y}_t = \begin{cases} \exp\left(\hat{y}'_t\right), & \lambda = 0, \\ \left(\lambda \hat{y}'_t + 1\right)^{1/\lambda}, & \lambda \neq 0. \end{cases}$$

- если некоторые $y_t \leq 0$, преобразования Бокса-Кокса невозможны (нужно прибавить к ряду константу)
- часто оказывается, что преобразование вообще не нужно
- можно округлять значение λ, чтобы упростить интерпретацию
- как правило, стабилизирующее преобразование слабо влияет на прогноз и сильно — на предсказательный интервал

Дифференцирование

Дифференцирование ряда — переход к попарным разностям его соседних значений:

$$y_1, \dots, y_T \longrightarrow y'_2, \dots, y'_T,$$

$$y'_t = y_t - y_{t-1}.$$

Дифференцированием можно стабилизировать среднее значение ряда и избавиться от тренда и сезонности.

Может применяться неоднократное дифференцирование; например, для второго порядка:

$$y_1, \dots, y_T \longrightarrow y'_2, \dots, y'_T \longrightarrow y''_3, \dots, y''_T,$$

$$y''_t = y'_t - y'_{t-1} = y_t - 2y_{t-1} + y_{t-2}.$$

Сезонное дифференцирование

Сезонное дифференцирование ряда — переход к попарным разностям его значений в соседних сезонах:

$$y_1, \ldots, y_T \longrightarrow y'_{s+1}, \ldots, y'_T,$$

$$y'_t = y_t - y_{t-s}.$$

Комбинированное дифференцирование

Сезонное и обычное дифференцирование может применяться к одному ряду в любом порядке.

Если ряд имеет выраженный сезонный профиль, рекомендуется начинать с сезонного дифференцирования — после него ряд уже может оказаться стационарным.

Остатки

Остатки — разность между фактом и прогнозом:

$$\hat{\varepsilon}_t = y_t - \hat{y}_t.$$

Прогнозы $\hat{y}_t$ могут быть построены с фиксированной отсрочкой:

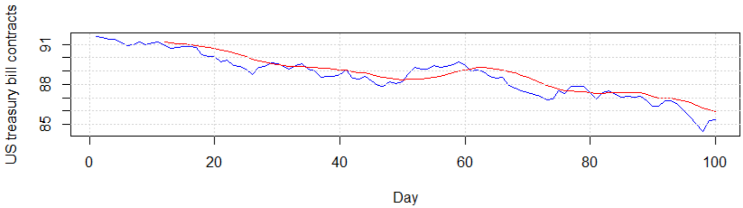
$$\hat{y}_{R+d|R}, \dots, \hat{y}_{T|T-d},$$

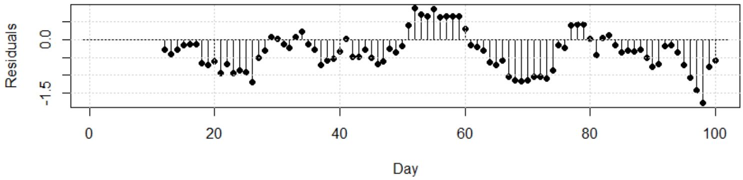
или с фиксированным концом истории при разных отсрочках:

$$\hat{y}_{T-D+1|T-D}, \dots, \hat{y}_{T|T-D}.$$

Необходимые свойства остатков прогноза

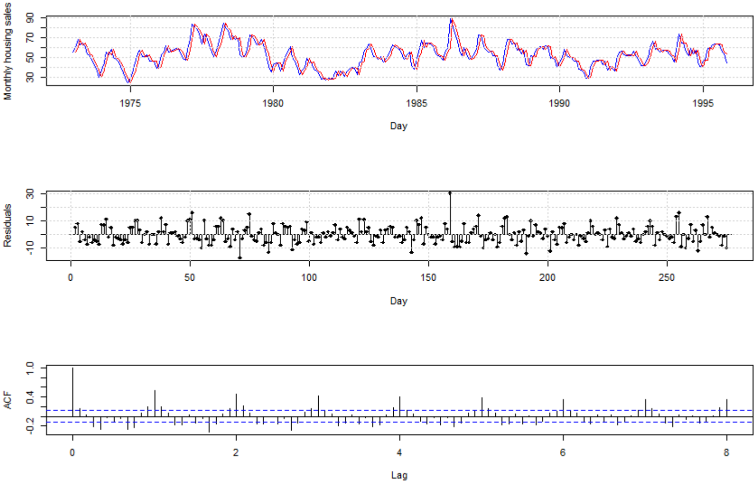
- Несмещённость — равенство среднего значения нулю:





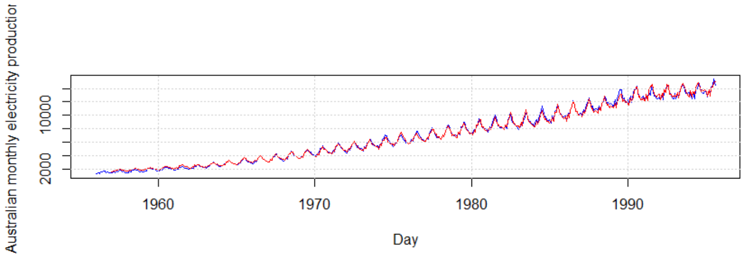
Необходимые свойства остатков прогноза

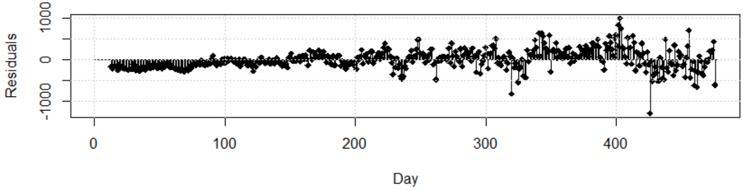
- Неавтокоррелированность — отсутствие неучтённой зависимости от предыдущих наблюдений:



Необходимые свойства остатков прогноза

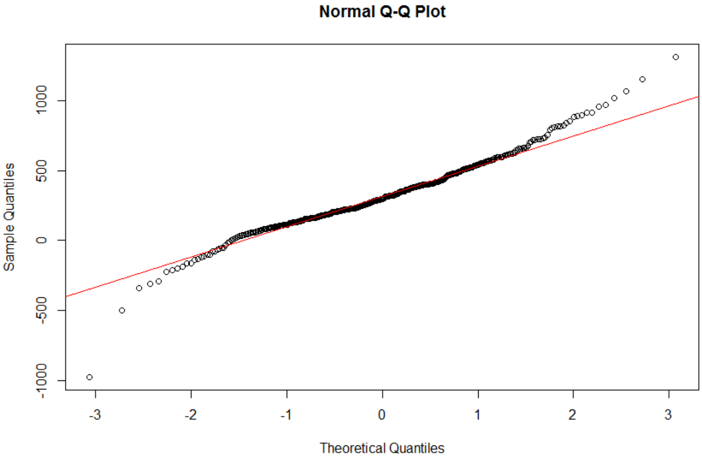
- Стационарность — отсутствие зависимости от времени:





Желательные свойства остатков прогноза

- Нормальность:



Проверка свойств остатков

- Несмещённость — критерий Стьюдента или Уилкоксона.
- Стационарность — визуальный анализ, критерий KPSS.
- Неавтокоррелированность — коррелограмма, Q-критерий Льюнга-Бокса.
- Нормальность — q-q plot, критерий Шапиро-Уилка.

Критерий KPSS (Kwiatkowski-Philips-Schmidt-Shin)

ряд ошибок прогноза:	$\varepsilon^T = \varepsilon_1, \ldots, \varepsilon_T;$
нулевая гипотеза:	$H_0:$ ряд ε^T стационарен;
альтернатива:	$H_1:$ ряд ε^T описывается моделью вида $\varepsilon_t = \alpha\varepsilon_{t-1};$
статистика:	$KPSS\left(\varepsilon^T\right) = \frac{1}{T^2}\sum\limits_{i=1}^{T}\left(\sum\limits_{t=1}^{i}\varepsilon_t\right)^2 \Big/ \lambda^2;$
нулевое распределение:	табличное.

Другие критерии для проверки стационарности: Дики-Фуллера, Филлипса-Перрона и их многочисленные модификации (см. Patterson K. *Unit root tests in time series, volume 1: key concepts and problems*. — Palgrave Macmillan, 2011).

Авторегрессия

$$AR(p): \quad y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t,$$

где y_t — стационарный ряд с нулевым средним, $\phi_1, \dots, \phi_p$ — константы $(\phi_p \neq 0)$, ε_t — гауссов белый шум с нулевым средним и постоянной дисперсией σ_ε^2.

Если среднее равно μ, модель принимает вид

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t,$$

где $\alpha = \mu\left(1 - \phi_1 - \dots - \phi_p\right)$.

Другой способ записи:

$$\phi(B)\, y_t = \left(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p\right) y_t = \varepsilon_t,$$

где B — разностный оператор $\left(B y_t = y_{t-1}\right)$.

Линейная комбинация p подряд идущих членов ряда даёт белый шум.

Скользящее среднее

$$MA(q): \quad y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q},$$

где y_t — стационарный ряд с нулевым средним, $\theta_1, \ldots, \theta_q$ — константы ($\theta_q \neq 0$), ε_t — гауссов белый шум с нулевым средним и постоянной дисперсией σ_ε^2.

Если среднее равно μ, модель принимает вид

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}.$$

Другой способ записи:

$$y_t = \theta\left(B\right)\varepsilon_t = \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right)\varepsilon_t,$$

где B — разностный оператор.

Линейная комбинация q подряд идущих компонент белого шума ε_t даёт элемент ряда.

ARMA (Autogerressive moving average)

$$ARMA(p,q): \quad y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q},$$

где y_t — стационарный ряд с нулевым средним, $\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q$ — константы ($\phi_p \neq 0$, $\theta_q \neq 0$), ε_t — гауссов белый шум с нулевым средним и постоянной дисперсией σ_ε^2.

Если среднее равно μ, модель принимает вид

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q},$$

где $\alpha = \mu\left(1 - \phi_1 - \dots - \phi_p\right)$.

Другой способ записи:

$$\phi\left(B\right) y_t = \theta\left(B\right) \varepsilon_t.$$

Теорема Вольда: любой стационарный ряд может быть аппроксимирован моделью $ARMA(p,q)$ с любой точностью.

ARIMA (Autogerressive integrated moving average)

Ряд описывается моделью $ARIMA(p, d, q)$, если ряд его разностей

$$\nabla^d y_t = (1 - B)^d \, y_t$$

описывается моделью $ARMA(p, q)$.

$$\phi\,(B)\,\nabla^d y_t = \theta\,(B)\,\varepsilon_t.$$

Seasonal multiplicative ARMA/ARIMA

$$ARMA(p,q)\times(P,Q)_s:\quad \Phi_P\left(B^s\right)\phi\left(B\right)y_t=\alpha+\Theta_Q\left(B^s\right)\theta\left(B\right)\varepsilon_t,$$

где

$$\Phi_P\left(B^s\right)=1-\Phi_1B^s-\Phi_2B^{2s}-\dots-\Phi_PB^{Ps},$$
$$\Theta_Q\left(B^s\right)=1+\Theta_1B^s+\Theta_2B^{2s}+\dots+\Theta_QB^{Qs}.$$

SARIMA:

$$\Phi_P\left(B^s\right)\phi\left(B\right)\nabla_s^D\nabla^d y_t=\alpha+\Theta_Q\left(B^s\right)\theta\left(B\right)\varepsilon_t.$$

q, Q, p, P

- В модели $ARIMA(p, d, 0)$ ACF экспоненциально затухает или имеет синусоидальный вид, а PACF значимо отличается от нуля при лаге p
- В модели $ARIMA(0, d, q)$ PACF экспоненциально затухает или имеет синусоидальный вид, а ACF значимо отличается от нуля при лаге q

$\Rightarrow$ начальные приближения для p, q, P, Q:

- q: номер последнего лага $\tau < S$, при котором автокорреляция значима
- $Q * S$: номер последнего сезонного лага, при котором автокорреляция значима
- p: номер последнего лага $\tau < S$, при котором частичная автокорреляция значима
- $P * S$: номер последнего сезонного лага, при котором частичная автокорреляция значима

Прогнозирование с помощью ARIMA

1. Строится график ряда, идентифицируются необычные значения.
2. При необходимости делается стабилизирующее дисперсию преобразование.
3. Если ряд нестационарен, подбирается порядок дифференцирования.
4. Анализируются ACF/PACF, чтобы понять, можно ли использовать модели AR(p)/MA(q).
5. Обучаются модели-кандидаты, сравнивается их AIC/AICc.
6. Остатки полученной модели исследуются на несмещённость, стационарность и неавтокоррелированность; если предположения не выполняются, исследуются модификации модели.
7. В финальной модели t заменяется на $T+h$, будущие наблюдения — на их прогнозы, будущие ошибки — на нули, прошлые ошибки — на остатки.

Построение предсказательного интервала

Если остатки модели нормальны и стационарны, предсказательные интервалы определяются теоретически.
Например, для прогноза на следующую точку предсказательный интервал — $\hat{y}_{T+1|T} \pm 1.96\hat{\sigma}_{\varepsilon}$.

Если нормальность или стационарность не выполняется, предсказательные интервалы генерируются с помощью симуляции.

auto_arima (SARIMA(p, d, q)(P, D, Q))

```
pmdarima.arima.auto_arima(y, X=None, start_p=2, d=None, start_q=2,
  max_p=5, max_d=2, max_q=5, start_P=1,
  D=None, start_Q=1, max_P=2, max_D=1,
  max_Q=2, max_order=5, m=1, seasonal=True,
  stationary=False, information_criterion='aic',
  alpha=0.05, test='kpss', seasonal_test='ocsb',
  stepwise=True, n_jobs=1, start_params=None,
  trend=None, method='lbfgs', maxiter=50,
  offset_test_args=None, seasonal_test_args=None,
  suppress_warnings=True, error_action='trace',
  trace=False, random=False, random_state=None,
  n_fits=10, return_valid_fits=False, out_of_sample_size=0,
  scoring='mse', scoring_args=None, with_intercept='auto',
  sarimax_kwargs=None, **fit_args)
```

Построить прогноз можно с помощью функции predict:

```
.predict(n_periods=test.shape[0], return_conf_int=True)
```

regARIMA

Эффекты плавающих праздников, краткосрочных маркетинговых акций и других нерегулярно повторяющихся событий с известной датой удобно моделировать с помощью regARIMA:

$$\Phi_P\left(B^s\right)\phi\left(B\right)\nabla_s^D\nabla^d z_t = \Theta_Q\left(B^s\right)\theta\left(B\right)\varepsilon_t$$

$$+$$

$$y_t = \sum_{j=1}^{k}\beta_j x_{jt} + z_t$$

$$=$$

$$\Phi_P\left(B^s\right)\phi\left(B\right)\nabla_s^D\nabla^d\left(y_t - \sum_{j=1}^{k}\beta_j x_{jt}\right) = \Theta_Q\left(B^s\right)\theta\left(B\right)\varepsilon_t.$$

Требования к решению задачи прогнозирования временных рядов

- визуализация данных, анализ распределения признака (оценка необходимости трансформации), оценка наличия выбросов;
- анализ автокорреляционной и частичной автокорреляционной функций;
- настройка модели ARIMA: автоматический подбор модели, проверка её соответствия особенностям ряда, при необходимости — корректировка модели, анализ остатков (нормальность, несмещённость, гомоскедастичность, неавтокоррелированность, стационарность);
- визуальный анализ, при необходимости — формальная проверка наличия структурных изменений в моделях;
- сравнение и выбор лучшей модели по критерию Диболда-Мариано;
- выводы.

Литература

Hyndman R.J., Athanasopoulos G. *Forecasting: principles and practice*. — OTexts, https://www.otexts.org/book/fpp