Доп. задачи анализа данных

Лекция "Временные ряды"

1. Временные ряды

Временной ряд — случайный процесс $(y_t, t \in \mathbb{N})$, являющийся значениями некоторого признака, измеренного через постоянные промежутки времени. Известны значения $y_1, ..., y_T$.

Задача прогнозирования.

Найти функцию f_T, т.ч. величина $\widehat{y}_{T+h|T} = f_T(y_1, ..., y_T, h)$

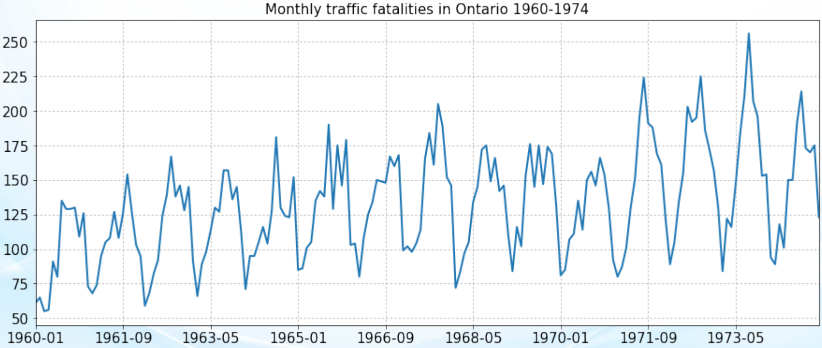
как можно лучше приближает значение y_{T+h},

где $h \in \{1, ..., H\}$,

величина H — горизонт прогнозирования.

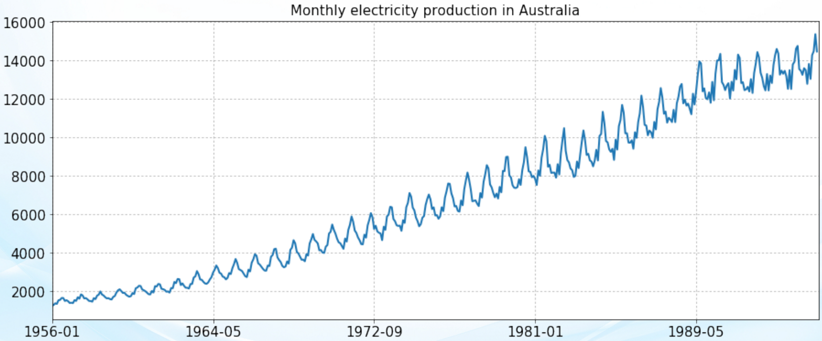
Кроме этого имеет смысл строить **предсказательный интервал**, то есть интервал $(d_{T+h|T}, u_{T+h|T})$, т.ч. $\mathsf{P}(d_{T+h|T} \leqslant y_{T+h} \leqslant u_{T+h|T}) \geqslant \alpha$.

Ежемесячное количество погибших в ДТП в провинции Онтарио (Канада) в 1960-1974 года



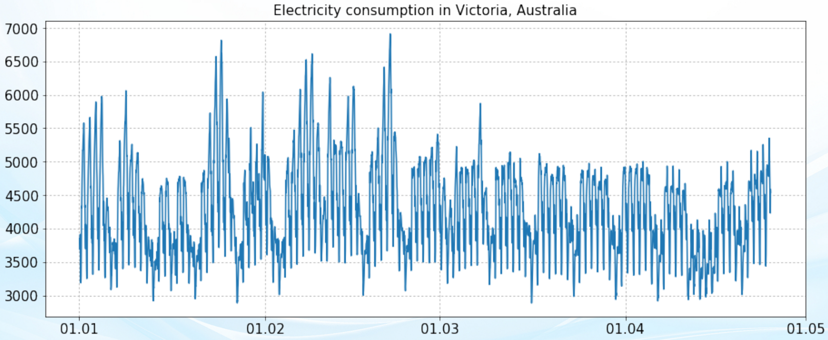
Данные

Ежемесячное производство электроэнергии в Австралии (миллион киловатт-часов). Январь 1956 — Август 1995



Данные

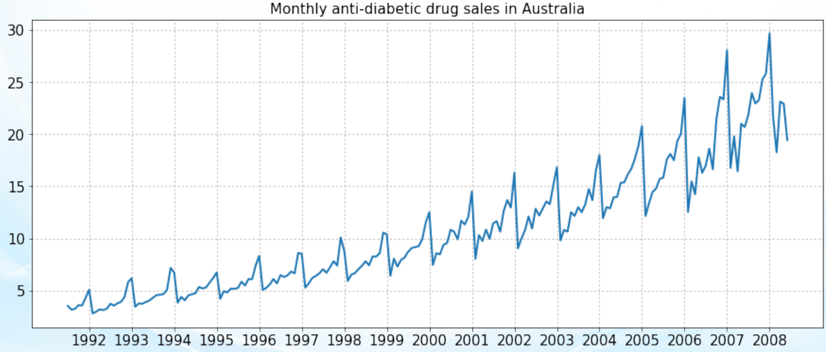
Максимальный спрос на электричество в штате Виктория (Австралия) за 30-минутные интервалы с 10 января 2000 в течении 115 дней



Данные

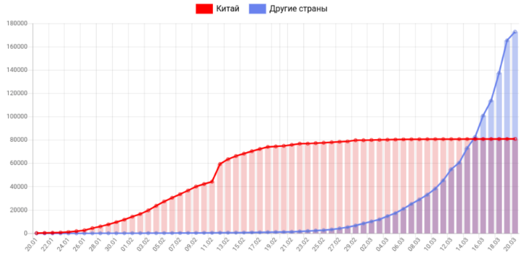
Ежемесячные продажи антидиабетических лекарств в Австралии.

Июль 1991 — Июнь 2008

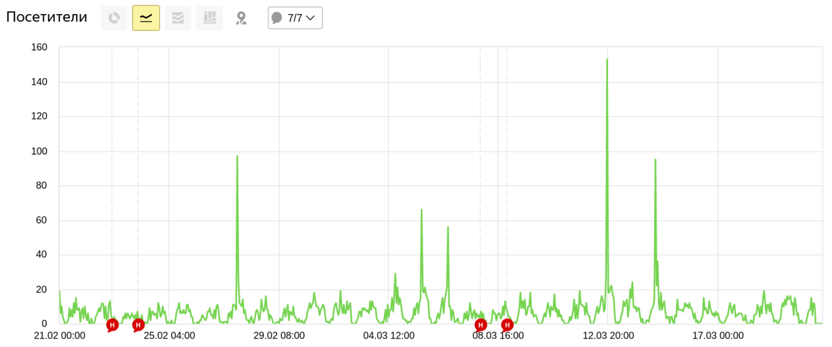


Данные

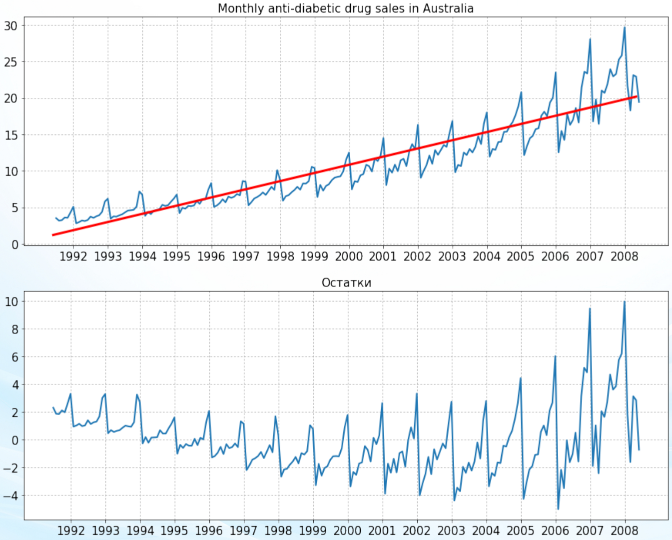
Заболевание коронавирусом



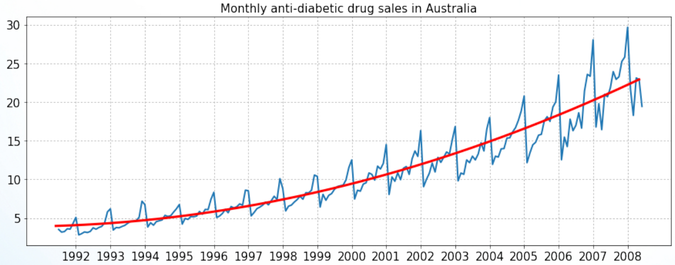
Метрика

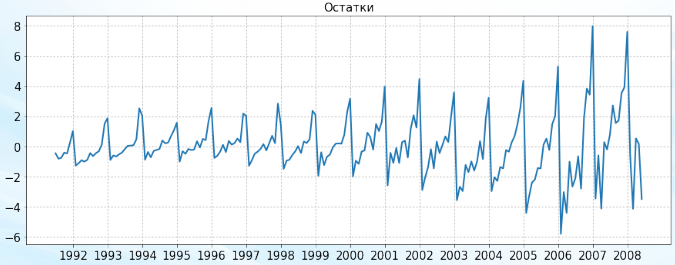

Посетители
7/7
160
140
120
100
80
60
40
20
0
21.02 00:00
25.02 04:00
29.02 08:00
04.03 12:00
08.03 16:00
12.03 20:00
17.03 00:00

Попробуем приблизить линейной регрессией



Попробуем приблизить линейной регрессией





2. Аналитика временных рядов.

2.1. Первые шаги

Автокорреляционная функция

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

где τ — лаг автокорреляции.

Формула: корреляция Пирсона ряда с самим собой со сдвигом на τ.

Смысл: степень влияния значения y_t на значение $y_{t+\tau}$.

Библиотека statsmodels

Автокорреляционная функция

statsmodels.tsa.stattools.acf(x, unbiased=False, nlags=40, qstat=False, fft=False, alpha=None, missing='none')

Возвращает: acf, confint, qstat, pvalues

Коррелограмма

statsmodels.graphics.tsaplots.plot_acf(x, ax=None, lags=None, alpha=0.05, use_vlines=True, unbiased=False, fft=False, title='Autocorrelation', zero=True, **kwargs)

Критерий Льюнга-Бокса

$y_1, \ldots, y_T$ — временной ряд

r_τ — автокорреляция с лагом τ

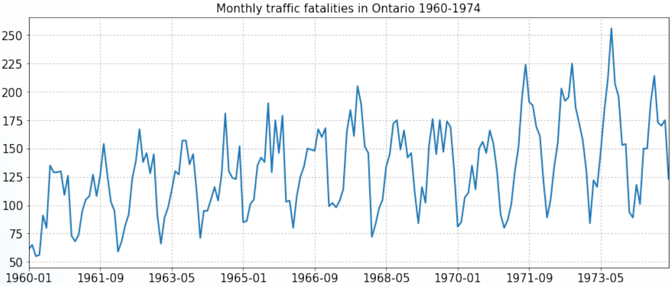
$\mathrm{H}_0 : r_1 = \ldots = r_L = 0$

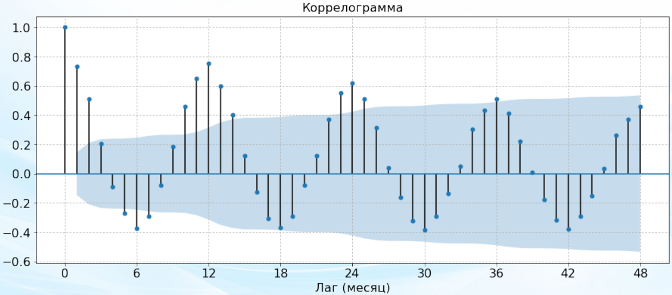
Q-статистика критерия

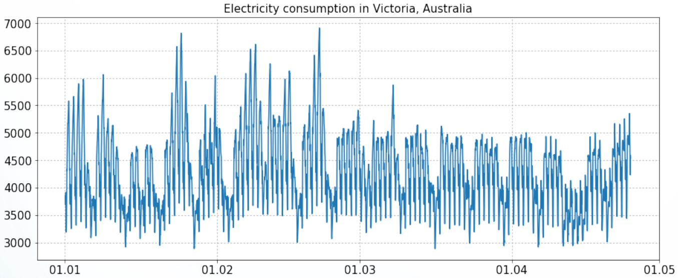
$$Q = T(T+2)\sum_{\tau=1}^{L} \frac{r_\tau^2}{T-\tau}$$

Если H_0 верна, то $Q \sim \chi^2_L$

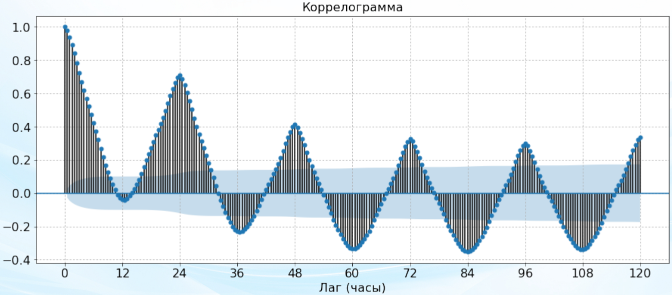
Если H_0 отвергается, то признается наличие значимой корреляции до лага L включительно.


Monthly traffic fatalities in Ontario 1960-1974
250
225
200
175
150
125
100
75
50
1960-01
1961-09
1963-05
1965-01
1966-09
1968-05
1970-01
1971-09
1973-05


Коррелограмма
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
6
12
18
24
30
36
42
48
Лаг (месяц)



Electricity consumption in Victoria, Australia
7000
6500
6000
5500
5000
4500
4000
3500
3000
01.01
01.02
01.03
01.04
01.05



Коррелограмма
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
12
24
36
48
60
72
84
96
108
120
Лаг (часы)

STL-декомпозиция

Пусть s — известный заранее период сезонности.

Разложение временного ряда на несколько компонент:

- **Тренд** — плавное долгосрочное изменение уровня ряда.

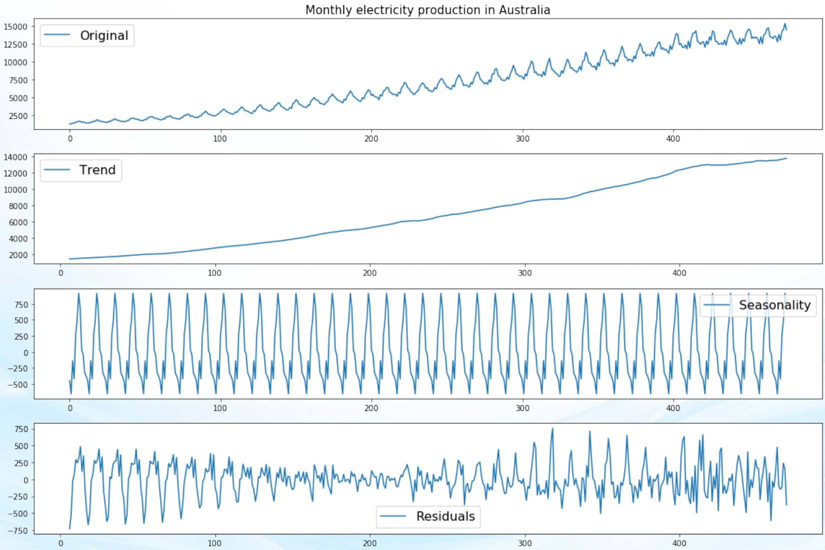
  Вычисляется с помощью скользящего окна длины s.

- **Сезонность** — циклические изменения уровня ряда

  с постоянным периодом сезонности s.

  Усреднение наблюдений для каждого сезона после удаления тренда.
  Например, усреднение значений по апрелю за все года.

- **Ошибка** — непрогнозируемая случайная компонента ряда.


Monthly electricity production in Australia
Original
15000
12500
10000
7500
5000
2500
0
100
200
300
400
Trend
14000
12000
10000
8000
6000
4000
2000
Seasonality
750
500
250
0
−250
−500
Residuals
−750

Библиотека statsmodels

statsmodels.tsa.seasonal.seasonal_decompose

(x, model='additive', filt=None, freq=None, two_sided=True)

Возвращает: results

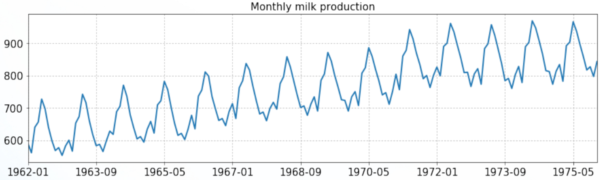
results.trend — тренд

results.seasonal — сезонность

results.resid — остатки

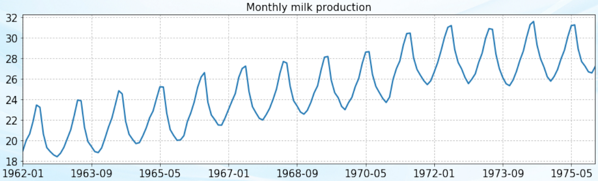
Календарные эффекты

Ежемесячное производство молока (фунты на корову) [Данные]



Интервалы времени не равны, т.к. месяцы разной длины.

Поделим значения на количество дней в месяце:



2. Аналитика временных рядов.

2.2. Стационарность

Стационарные случайные процессы

Случайный процесс $(X_t, t \in T)$ называется **стационарным**

- **в узком смысле** если $\forall\ t_1, ..., t_n, \tau$ вектор $(X_{t_1+\tau}, ..., X_{t_n+\tau})$ совпадает по распределению с $(X_{t_1}, ..., X_{t_n})$.
- **в широком смысле** если
  1. X_t является L^2-процессом, т.е. $\mathsf{E}X_t^2 < +\infty$ для любого t.
  2. $\mathsf{E}X_t$ не зависит от t.
  3. $cov(X_{t+\tau}, X_{s+\tau}) = cov(X_t, X_s)$ для любых t, s, τ.
- **для гауссовских** определения эквивалентны.

Пример. $X_t = \xi_1 \cos t + \xi_2 \sin t$, где $\xi_1, \xi_2 \sim Bern_{\pm 1}(1/2)$ и незав.

$\mathsf{E}X_1 = 0, \quad cov(X_t, X_s) = \cos t \cos s\ \mathsf{D}\xi_1 + \sin t \sin s\ \mathsf{D}\xi_2 = \cos(t-s)$
$\Longrightarrow$ есть стационарность в широком смысле.

Но при $t = 0: \quad X_0 = \xi_1 \in \{-1, 1\}$
при $t = \pi/4: \quad X_{\pi/4} = (\xi_1 + \xi_2)/\sqrt{2} \in \{-\sqrt{2}, 0, \sqrt{2}\}$
Распределения разные $\Longrightarrow$ поэтому нет стацион. в узком смысле.

Критерий KPSS (Kwiatkowski-Phillips-Schmidt-Shin)

H_0: ряд y_t стационарен

Статистика

$$KPSS(y) = \frac{1}{T^2}\sum_{t=1}^{T} S_t^2 \Big/ \widehat{\sigma}_y^2, \quad \text{где } S_t = \sum_{i=1}^{t} y_i, \quad \widehat{\sigma}_y^2 = \frac{1}{T}\sum_{t=1}^{T} y_t^2$$

При H_0 имеет табличное распределение.

```
statsmodels.tsa.stattools.kpss(y)
```

```
/usr/local/lib/python3.5/dist-packages/statsmodels/tsa/stattools.py:125
8: InterpolationWarning: p-value is smaller than the indicated p-value
  warn("p-value is smaller than the indicated p-value", InterpolationWa
rning)
```

```
(1.3136750533447588,
 0.01,
 15,
 {'1%': 0.739, '10%': 0.347, '2.5%': 0.574, '5%': 0.463})
```

Значения pvalue рассчитаны только в интервале (0.01, 0.1).

Приведение к стационарным: стабилизация дисперсии

Зачем?

Методы прогнозирования, требующие стационарность, лучше обрабатывают стационарные ряды.

Как правило, это сильно влияет на предсказательный интервал.

Преобразования:

1. Класс преобразований Бокса-Кокса:

$$z_t = \begin{cases} \ln y_t, & \lambda = 0 \\ (y_t^\lambda - 1)/\lambda, & \lambda \neq 0 \end{cases}$$

2. Если есть предположения о зависимости $\mathrm{D}y_t$ от t, то можно рассмотреть ряд $z_t = y_t \big/ \sqrt{\mathrm{D}y_t}$

После построения прогноза для преобразованного ряда нужно сделать обратное преобразование.

Приведение к стационарным: тренд и сезонность

Преобразования:

1. Дифференцирование ряда — переход к ряду $(y'_t, t \in \{2, \ldots, T\})$,

   где $y'_t = y_t - y_{t-1}$.

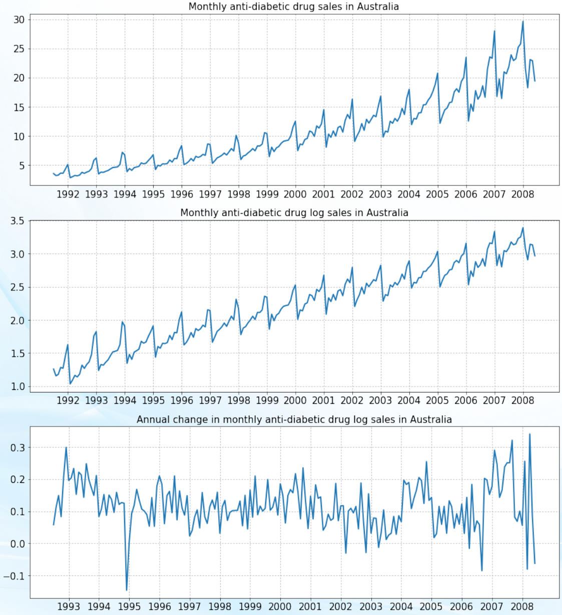
   Используется для снятие тренда.

2. Сезонное дифференцирование ряда — переход к ряду

   $(y'_t, t \in \{s+1, \ldots, T\})$, где $y'_t = y_t - y_{t-s}$, s — длина сезона.

Преобразования можно применять несколько раз.

Обычно сначала применяют сезонное дифференцирование.



Критерий KPSS:

Исходный ряд

$pvalue < 0.01$

логарифмированный

$pvalue < 0.01$

еще и сезонно

дифференцированный

$pvalue > 0.01$

3. Прогнозирование

3.1. Анализ остатков

Остатки

$y_1, \dots, y_T$ — временной ряд

f — прогнозирующая модель, построенная по $y_1, \dots, y_T$

$\widehat{y}_{t|t-1} = f(y_1, \dots, y_{t-1}, 1)$ — прогнозы на обучении

$\widehat{\varepsilon}_t = y_t - \widehat{y}_{t|t-1}$ — остатки модели

Необходимые свойства остатков:

1. **Несмещенность** — равенство среднего нулю:
   Критерий Стьюдента (для норм.), критерий Уилкоксона
2. **Стационарность** — отсутствие зависимости от времени:
   Визуальный анализ, критерий KPSS
3. **Неавтокоррелированность** — отсутствие завис. от значений:
   Кореллограмма, критерий Льюнга-Бокса (χ^2_{L-k})

Остатки

Желательные свойства остатков:

4. **Нормальность:**
   Q-Q plot, критерий Шапиро-Уилка
5. **Гомоскедастичность:**
   Визуальный анализ, критерий Бройша-Пагана

- Если свойства 1-3 не выполняются, прогноз можно улучшить.
- Свойства 4-5 влияют на построение доверительного интервала:
  - Выполняются $\Longrightarrow$ теоретическим способом;
  - Не выполняются $\Longrightarrow$ с помощью семплирования.

3. Прогнозирование

3.2. Модели экспоненциального сглаживания

Простое экспоненциальное сглаживание

Прогноз средним: $\hat{y}_{T+1|T} = \sum_{t=t_0}^{T} y_t$

Прогноз взвешенным средним с экспоненциально убыв. весами:

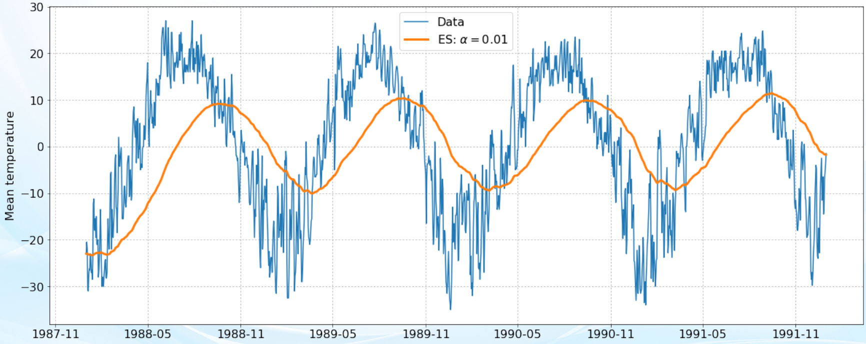
$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha) y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \ldots =$$

$$= \alpha y_T + (1-\alpha)\hat{y}_{T|T-1}$$

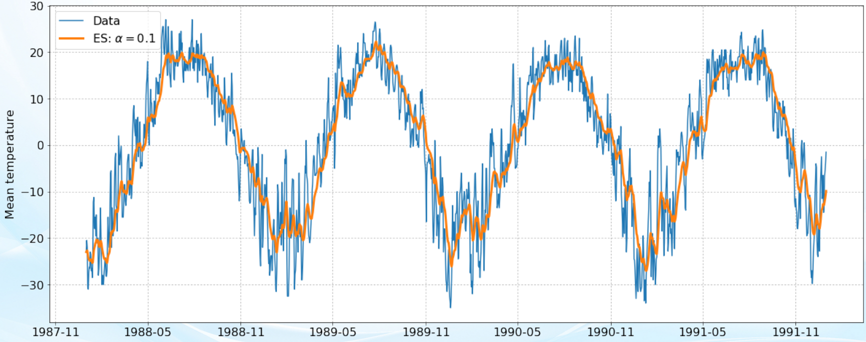
- $\alpha \approx 1 \to$ больший вес последним точкам: $\hat{y}_{T+1|T} \approx y_T$;
- $\alpha \approx 0 \to$ большее сглаживание: $\hat{y}_{T+1|T} \approx \bar{y}$.

Прогноз константный: $\hat{y}_{T+h|T} = \hat{y}_{T+1|T}$

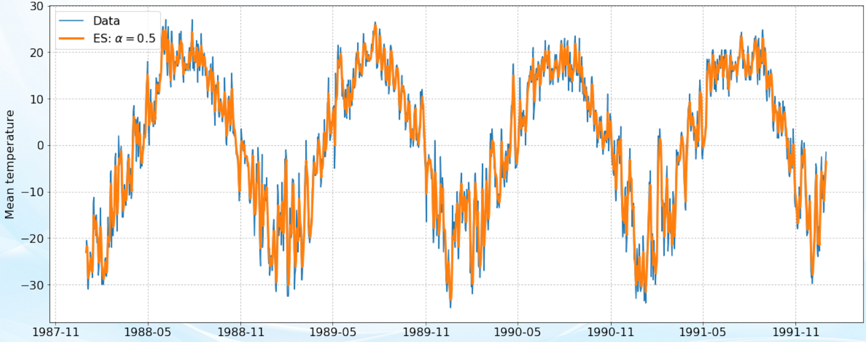
Простое экспоненциальное сглаживание



Простое экспоненциальное сглаживание



Простое экспоненциальное сглаживание



Учет тренда

Аддитивный линейный тренд (метод Хольта):

$$\widehat{y}_{t+d|t} = l_t + db_t,$$

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$

Мультипликативный линейный тренд:

$$\widehat{y}_{t+d|t} = l_t b_t^d,$$

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} b_{t-1})$$

$$b_t = \beta \frac{l_t}{l_{t-1}} + (1-\beta)b_{t-1}$$

Учет сезонности

Аддитивная сезонность с периодом длины m

(метод Хольта-Уинтерса):

$$\hat{y}_{t+d|t} = l_t + db_t + s_{t-m+(d \bmod m)},$$

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

Учет сезонности

Мультипликативная сезонность:

$$\hat{y}_{t+d|t} = (l_t + d b_t) s_{t-m+(d \bmod m)},$$

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta) b_{t-1}$$

$$s_t = \gamma \frac{y_t}{l_{t-1} - b_{t-1}} + (1-\gamma) s_{t-m}$$

Адаптивное экспоненциальное сглаживание

Пусть $\widehat{\varepsilon}_t = y_t - \widehat{y}_t$ — ошибка прогноза, сделанного на шаге $t-1$.

$E_t = \gamma \widehat{\varepsilon}_t + (1-\gamma) E_{t-1}$ — среднее значение ошибки

$A_t = \gamma \left| \widehat{\varepsilon}_t \right| + (1-\gamma) A_{t-1}$ — средний разброс ошибки

Берем

$$\alpha_t = \min \left(\frac{|E_t|}{|A_t|}, 1 \right)$$

Обычно берут $\gamma \in (0.05, 0.1)$.

3. Прогнозирование

3.3. Модели типа ARIMA

Модель скользящего среднего MA(q)

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

где y_t — стационарный ряд со средним μ;

ε_t — гауссовский белый шум ($\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ и независимы).

Модель авторегрессии AR(p)

$$y_t = \alpha + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \varepsilon_t$$

где y_t — стационарный ряд;

ε_t — гауссовский белый шум ($\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ и независимы).

Для стационарности нужны условия на коэффициенты $\varphi_1, \ldots, \varphi_p$.

Связь модели авторегрессии с линейной регрессией

Модель авторегрессии AR(p)

$$y_t \;=\; \alpha + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \varepsilon_t$$

где y_t — стационарный ряд;

ε_t — гауссовский белый шум ($\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ и независимы).

Это модель линейной регрессией для которой

- **Отклик:** y_t — знач. ряда в мом. времени t
- **Признаки:** $y_{t-1}, \ldots, y_{t-p}$ — знач. ряда в пред. мом. времени

Стационарность моделей

L — **оператор сдвига**: $\quad Ly_t = y_{t-1}, \quad L\varepsilon_t = \varepsilon_{t-1}.$

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q} = \mu + b(L)\varepsilon_t$$

$$y_t = \alpha + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \varepsilon_t \implies a(L)y_t = \alpha + \varepsilon_t$$

где $b(z) = 1 + \theta_1 z + \ldots + \theta_q z^q,$

$a(z) = 1 - \varphi_1 z - \ldots - \varphi_p z^p.$

Теорема:

1. Любой стационарный процесс $AR(p)$ представим в виде $MA(\infty)$.
2. Процесс $AR(p)$ стационарен $\iff$

   все комплексные корни $a(z) = 0$ лежат вне единичного круга.

Модель ARMA(p, q)

Соединяем $AR(p)$ и $MA(q)$

$$\begin{aligned} y_t &= \alpha + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} \\ &\quad + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \end{aligned}$$

Эквивалентная запись:

$$a(L) y_t = \alpha + b(L)\varepsilon_t \qquad \text{или} \qquad y_t = \mu + \frac{b(L)}{a(L)}\varepsilon_t,$$

где $a(z) = 1 - \varphi_1 z - \ldots - \varphi_p z^p$,

$b(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$,

L — оператор сдвига: $L y_t = y_{t-1}$, $L\varepsilon_t = \varepsilon_{t-1}$.

Модель ARIMA(p, d, q)

Модель ARIMA(p, d, q) для y_t

$=$ модель ARMA(p, q) для ряда разностей порядка d ряда y_t.

Что такое разность порядка 1? $\quad y_t - y_{t-1} = (1 - L)y_t$

Получаем формулу модели ARIMA:

$$a(L)(1 - L)^d y_t = \alpha + b(L)\varepsilon_t \qquad \text{или} \qquad (1 - L)^d y_t = \mu + \frac{b(L)}{a(L)}\varepsilon_t.$$

То есть многочлен $\widetilde{a}(z) = a(z)(1 - z)^d$ имеет d единичных корней.

Свойство: Позволяет учесть нестационарности, в частности, тренд.

Оценка коэффициентов в ARIMA

Пусть p, d, q фиксированы.

Предполагаем, что ε_t — гауссовский белый шум.

$\Longrightarrow$ y_t — гауссовский случайный процесс.

Даны значения временного ряда $y_1, ..., y_T$, и их совместная плотность согласно модели ARIMA(p, d, q) равна $p_{\theta,\varphi,\alpha}(a_1, ..., a_T)$.

Тогда $L_y(\theta, \varphi, \alpha) = p_{\theta,\varphi,\alpha}(y_1, ..., y_T)$ — функция правдоподобия.

В качестве оценок коэффициентов берется оценка макс. правд..

Почему p, d, q нельзя оценивать с помощью ОМП?

Максимизация $L_y(\theta, \varphi, \alpha)$ по p, d, q приводит к переобучению.

Это гиперпараметры.

Оценка p и q

Частичная автокорреляция (PACF) — корреляция ряда с собой после снятия линейной зависимости от предыдущих элементов ряда.

$$\gamma_h = \begin{cases} corr(y_{t+1}, y_t), & h = 1; \\ corr\left(y_{t+h} - y_{t+h}^{h-1}, y_t - y_t^{h-1}\right), & h \geqslant 2, \end{cases}$$

где y_t^{h-1} — линейная регрессия на $y_{t-1}, y_{t-2}, \dots, y_{t-(h-1)}$:

$$y_t^{h-1} = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_{h-1} y_{t-(h-1)}$$

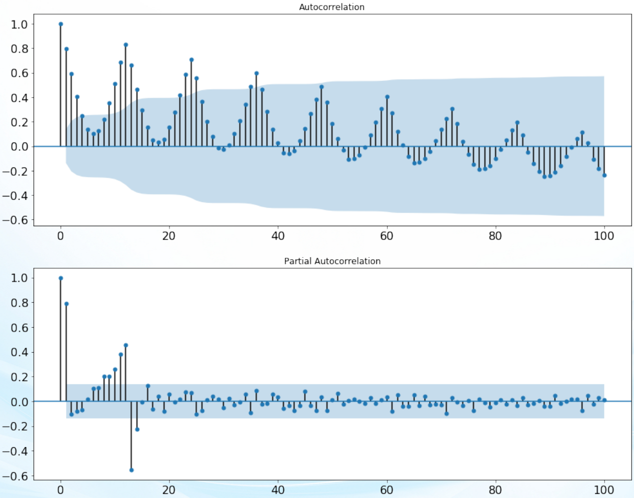
$$y_{t+h}^{h-1} = \beta_1 y_{t+h-1} + \beta_2 y_{t+h-2} + \dots + \beta_{h-1} y_{t+1}$$

Пример для $h = 2$:

$\gamma_2 = corr\left(y_{t+2} - \beta_1 y_{t+1}, y_t - \beta_1 y_{t-1}\right)$

где β_1 — МНК-оценка в модели $y_t = \beta y_{t-1}$

Оценка p и q



Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
20
40
60
80
100
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
20
40
60
80
100

Оценка p и q

Начальное приближение p: последний значимый пик у ACF.

Начальное приближение q: последний значимый пик у PACF.

Далее используем поиск по сетке вокруг подобранных значений, минимизируя информационный критерий:

$AIC = -2l^* + 2(p + q + 1)$ — критерий Акаике;

$AIC_c = -2l^* + \frac{2(p+q+1)(p+q+2)}{T-p-q-2}$ — критерий Акаике (короткие ряды);

$BIC = -2l^* + (\log T - 2)(p + q + 1)$ — критерий Шварца,

где $l^* = \ln L_y\left(\widehat{\theta}, \widehat{\phi}, \widehat{\alpha}\right)$ — логарифм функции правд. (для ОМП).

Итог: прогнозирование с помощью ARIMA

1. Анализ выбросов: замена нерелевантных выбросов на `NA` или "усреднение" по соседним элементам.
2. Стабилизация дисперсии (преобразования).
3. Дифференцирование, если ряд не стационарен.
4. Выбор пилотных p и q по ACF и PACF.
5. Вокруг этих параметров подбираем оптим. модель по AIC/AIC_c.
6. Если для полученной модели не выполняются необходимые свойства остатков, модель можно улучшить.

Итог: прогнозирование с помощью ARIMA

7. Пошаговое построение прогноза:
   - для $t \leqslant T$: $\varepsilon_t \Longrightarrow \widehat{\varepsilon}_t = y_t - \widehat{y}_t$;
   - для $t > T$: $\varepsilon_t \Longrightarrow 0$;
   - для $t > T$: $y_t \Longrightarrow \widehat{y}_t$.
8. Построение предсказательного интервала:
   - если остатки модели нормальны и гомоскедастичны, то строится теоретический предсказательный интервал;
   - иначе интервалы строятся с помощью симуляции.

Слишком просто...

Давайте усложнять!

Модель SARIMA

ARIMA(p, d, q):

$$(1-L)^d y_t = \mu + \frac{b(L)}{a(L)} \varepsilon_t$$

Пусть s — период сезонности ряда. Добавим в модель ARIMA(p, d, q) компоненты на значения в предыдущие сезоны...

SARIMA$(p, d, q) \times (P, D, Q)_s$:

$$(1-L)^d (1-L^s)^D y_t = \mu + \frac{b(L)B(L^s)}{a(L)A(L^s)} \varepsilon_t,$$

где $a(z) = 1 - \varphi_1 z - \ldots - \varphi_p z^p, \qquad b(z) = 1 + \theta_1 z + \ldots + \theta_q z^q,$

$$A(z) = 1 - \varphi_1^s z - \ldots - \varphi_P z^P, \qquad B(z) = 1 + \theta_1^s z + \ldots + \theta_Q^s z^Q.$$

Еще усложним?

Модель ARIMAX

ARIMA(p, d, q):

$$(1 - L)^d y_t = \mu + \frac{b(L)}{a(L)} \varepsilon_t$$

Пусть $x_t \in \mathbb{R}^n$ — процесс регрессоров, *известный до начала прогноза*.

Простой вариант:

$$(1 - L)^d y_t = \mu + \sum_{i=1}^{n} \frac{1}{a(L)} x_t^i + \frac{b(L)}{a(L)} \varepsilon_t$$

Общий случай:

$$(1 - L)^d y_t = \mu + \sum_{i=1}^{n} \frac{u_i(L)}{v_i(L)} x_t^i + \frac{b(L)}{a(L)} \varepsilon_t$$

Пример: $x_t = I\{\text{в момент времени } t \text{ праздник}\}$

Модель SARIMAX

ARIMA(p, d, q):

$$(1-L)^d y_t = \mu + \frac{b(L)}{a(L)}\varepsilon_t$$

Соединим...

SARIMAX$(p, d, q)\times(P, D, Q)_s$:

$$(1-L)^d (1-L^s)^D y_t = \mu + \sum_{i=1}^{n} \frac{u_i(L)}{v_i(L)} x_t^i + \frac{b(L)B(L^s)}{a(L)A(L^s)}\varepsilon_t$$

Аналогия с линейной регрессией

SARIMAX$(p, d, q)\times(P, D, Q)_s$:

$$(1-L)^d(1-L^s)^D y_t = \mu + \sum_{i=1}^{n} \frac{u_i(L)}{v_i(L)} x_t^i + \frac{b(L)B(L^s)}{a(L)A(L^s)}\varepsilon_t$$

Это линейная по признакам модель, в которой

- **Отклик:** y_t — знач. ряда в мом. времени t
- **Признаки:**
  - $y_{t-1}, \dots, y_{t-p}$ — знач. ряда в пред. мом. времени
  - Знач. ряда за предыдущие сезоны
  - Значения признаков в пред. мом. времени
  - Значения признаков в пред. сезоны
- **Ошибка:** сумма шума за пред. мом. времени и пред. сезоны.

Пакет forecast в R

Построение модели:

```
auto.arima(y, d = NA, D = NA, max.p = 5, max.q = 5, max.P
= 2, max.Q = 2, max.order = 5, max.d = 2, max.D = 1,
start.p = 2, start.q = 2, start.P = 1, start.Q = 1,
stationary = FALSE, seasonal = TRUE, ic = c(”aicc”, ”aic”,
”bic”), stepwise = TRUE, trace = FALSE, approximation =
(length(x) > 150 | frequency(x) > 12), truncate = NULL,
xreg = NULL, test = c(”kpss”, ”adf”, ”pp”), seasonal.test
= c(”ocsb”, ”ch”), allowdrift = TRUE, allowmean = TRUE,
lambda = NULL, biasadj = FALSE, parallel = FALSE,
num.cores = 2, x = y, ...)
```

Пакет forecast в R

Прогнозирование:

```
forecast(object, h = ifelse(frequency(object) > 1, 2 *
frequency(object), 10), level = c(80, 95), fan = FALSE,
robust = FALSE, lambda = NULL, find.frequency = FALSE,
allow.multiplicative.trend = FALSE, model = NULL, ...)
```

3. Прогнозирование

3.4. Нелинейные модели

Идея

Модель AR(p):

$$y_t \;=\; \alpha + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \varepsilon_t = f(y_{t-1}, \ldots, y_{t-p}, \varepsilon_t),$$

где f — линейная функция.

Идея: функция f строится произв. методом машинного обучения.

Примеры признаков:

1. несколько предыдущих значений ряда;
2. несколько последних сезонных компонент;
3. максимальное и минимальное значение ряда за период;
4. степень изменчивости ряда за период;
5. экзогенные факторы.

Построение модели

Способ 1.

Для каждого $t_0 \leqslant t \leqslant T$ создается объект обучающей выборки:

- *признаковое описание* — история ряда до мом. времени $t-1$;
- *целевая метка* — значение y_t.

Предсказание строится на шаг вперед, далее рекурсивно.

Способ 2.

Для каждого $t_0 \leqslant t \leqslant T-h+1$ создается h объектов обуч. выборки:

- *признаковое описание* — история ряда до мом. времени $t-1$;
- *целевая метка* — значение y_t;
- *число* $j \in \{1, ..., h\}$ — насколько шагов строить прогноз.

3. Прогнозирование

3.5. Нейронные сети

Позже :))

3. Прогнозирование

3.5. Качество моделей

Метрики качества

Средняя квадратичная ошибка

$$MSE = \frac{1}{T-R+1} \sum_{t=R}^{T} (\widehat{y}_t - y_t)^2 .$$

Средняя абсолютная ошибка

$$MAE = \frac{1}{T-R+1} \sum_{t=R}^{T} |\widehat{y}_t - y_t| .$$

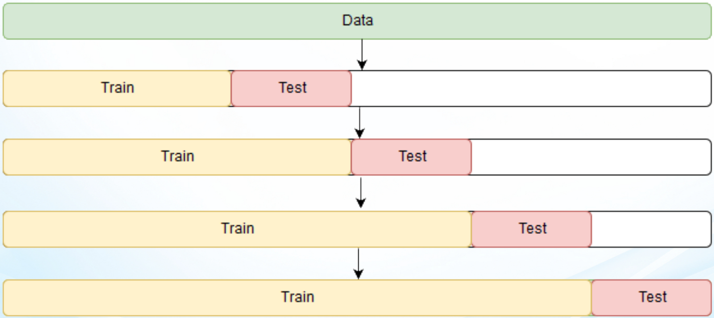
Средняя абсолютная ошибка в процентах

$$MAPE = \frac{100}{T-R+1} \sum_{t=R}^{T} \left| \frac{\widehat{y}_t - y_t}{y_t} \right| .$$

Симметричная средняя абсолютная ошибка в процентах

$$SMAPE = \frac{200}{T-R+1} \sum_{t=R}^{T} \left| \frac{\widehat{y}_t - y_t}{\widehat{y}_t + y_t} \right| .$$

Кросс-валидация для временных рядов



Кросс-валидация для временных рядов

Как выбрать лучшую модель среди тех, которые обладают хорошими свойствами?

1.1 Считаем качество прогнозов $\widehat{y}_{t_0+1|t_0}, \dots, \widehat{y}_{t_0+\Delta t|t_0}$

1.2 Считаем качество прогнозов $\widehat{y}_{t_0+\Delta t+1|t_0+\Delta t}, \dots, \widehat{y}_{t_0+2\Delta t|t_0+\Delta t}$

...

1.k Считаем качество прогнозов $\widehat{y}_{t_0+k\Delta t+1|t_0+k\Delta t}, \dots, \widehat{y}_{t_0+(k+1)\Delta t|t_0+k\Delta t}$

2. Усредняем полученные значения качества.
3. Выбираем модель, которая показывает наилучшее усредненное качество.


ВСЁ!