Прикладной статистический анализ данных
Модели статистического анализа временных рядов

Описательная статистика временных рядов
$+$
Модели сглаживания и выделения сезонности временных рядов

Чупраков Д. В.

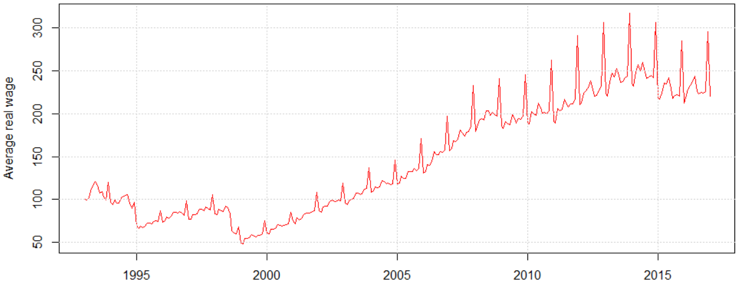
Временной ряд

Временной ряд — последовательность значений показателя, зависящего от времени:

$$y_1, y_2, \ldots, y_T, \ldots, \qquad y_t \in \mathbb{R},$$

Стохастический процесс с дискретным временем:

$$Y(\omega, t) \colon Y_1, Y_2, \ldots, Y_T, \ldots$$

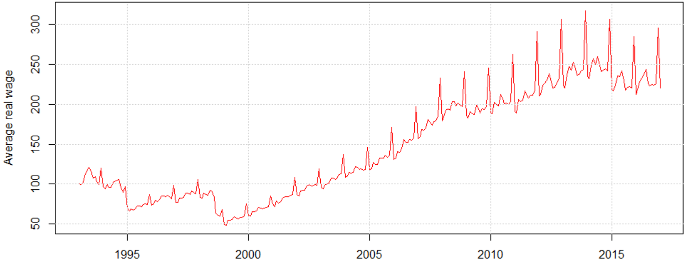


- Y_t не являются одинаково распределенными;
- y_t статистически зависимы.

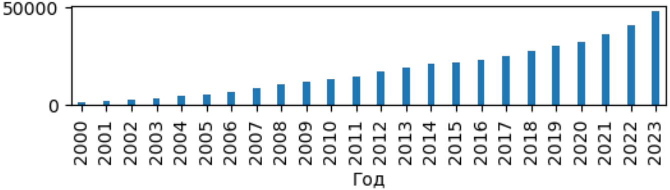
Виды временных рядов

По способу выражения

- **Ряды абсолютных величин**
- **Ряды относительных величин**



- **Ряды средних величин**



Виды временных рядов

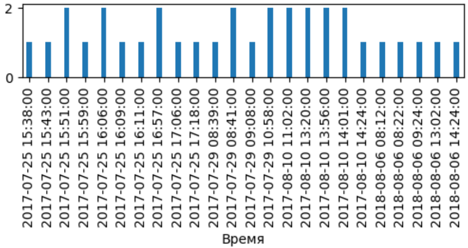
По способу регистрации

- **Моментные:** ряды значений показателя по состоянию на определенные моменты времени.
  - цена на определененный вид товаров
  - курс акций
  - численность населения
  - число посетителей
- **Интервальные:** ряды накопленных показателей за определенные время:
  - объем продаж
  - численность населения
  - число посетителей
- **Производные:** ряды, полученные из наблюдаемых данных некоторыми преобразованиями

Виды временных рядов

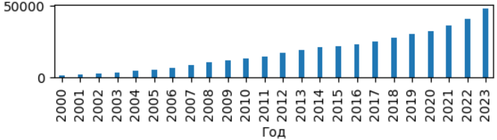
По свойству параметра времени

Нерегулярные (разнесенные во времени): сбор данных в произвольные моменты времени



- суммы средств, снятых через банкомат
- поток данных от систем мониторинга

Регулярные (равноотстоящие): сбор данных через равные промежутки времени



Дискретность — длина промежутка между соседними значениями регулярного ВР

Общая математическая модель временного ряда

Статистическая характеристика случайного дискретный процесс $Y(\omega,t)$ — совместная функция распределения случайных величин:

$$F(Y_1, Y_2, \ldots, Y_t, \ldots)$$

Бесконечное число моментов времени?

$$F_1\left(Y_{t_1}\right), \quad F_2\left(Y_{t_1}, Y_{t_2}\right), \quad F_3\left(Y_{t_1}, Y_{t_2}, Y_{t_3}\right), \ldots$$

Функции распределения согласованы:

$$F_n(y_{t_1}, \ldots, y_{t_n}) = \int\limits_{y \in Y_{t_{n+1}}} F_{n+1}(Y_{t_1}, \ldots, Y_{t_n}, Y_{t_{n+1}})$$

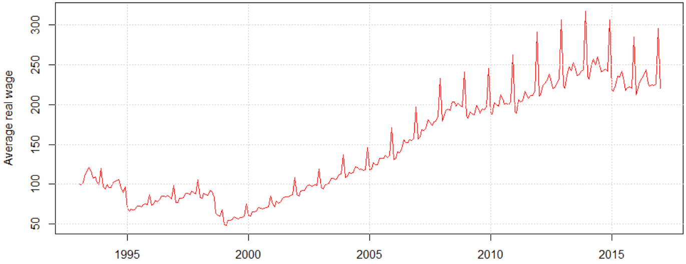
Как это получить из временного ряда?

Задачи исследования временных рядов

- Прогнозирование развития явления
- Характеристика отдельных изменений в уровнях ряда
- Определение средних показателей
- Выявление закономерностей динамики исследуемого явления
- Выявление факторов, обусловливающих изменение изучаемого объекта во времени

Задача прогнозирования временного ряда

Регулярный временной ряд: $y_1, \ldots, y_T, \ldots, \; y_t \in \mathbb{R}$,



Задача прогнозирования — найти функцию f_T:

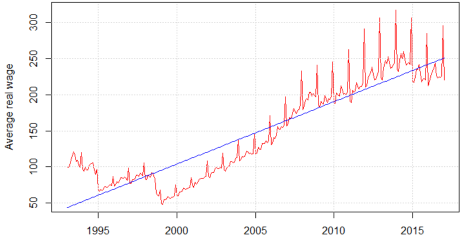
$$y_{T+d} \approx f_T\left(y_1, \ldots, y_T, d\right) \equiv \hat{y}_{T+d|T},$$

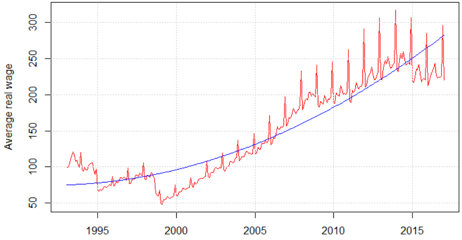
где

- $d \in \{1, \ldots, D\}$ — прогнозный лаг,
- D — горизонт прогнозирования.

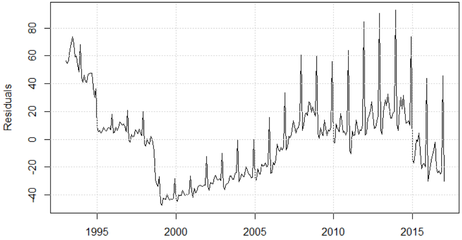
Регрессия

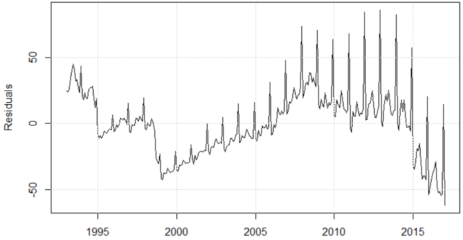
Сделаем регрессию на время?





Остатки содержат информацию!





Характеристика отдельных изменений в уровнях ряда

	Базисный	**Цепной**
Абсолютный прирост Δy_i	$\Delta y_i^{\text{б}} = y_i - y_0$	$\Delta y_i^{\text{ц}} = y_i - y_{i-1}$
Темп роста Ty_i	$Ty_i^{\text{б}} = \dfrac{y_i}{y_0}$	$Ty_i^{\text{ц}} = \dfrac{y_i}{y_{i-1}}$
Темп прироста ΔTy_i	$\Delta Ty_i^{\text{б}} = \dfrac{\Delta y_i^{\text{б}}}{y_0} = Ty_i^{\text{б}} - 1$	$\Delta Ty_i^{\text{ц}} = \dfrac{\Delta y_i^{\text{ц}}}{y_{i-1}} = Ty_i^{\text{ц}} - 1$
Темп наращивания $T_{\text{н}}y_i$	$T_{\text{н}}y_i = \dfrac{\Delta y_i^{\text{ц}}}{y_0}$	
Абсолютное значение одного процента прироста A_i	$A_i = \dfrac{\Delta y_i^{\text{ц}}}{100 \cdot \Delta Ty_i^{\text{ц}}} y_i = 0.01 y_{i-1}$	

Определение средних показателей уровней ВР

Показатели ВР

	Регулярный	**Нерегулярный**
Интервальный ВР	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$\bar{y} = \frac{\sum_{i=1}^{n} y_i \Delta t_i}{\sum_{i=1}^{n} \Delta t_i}$
Моментный ВР	$\bar{y} = \frac{1}{n-1}\sum_{i=1}^{n-1} \frac{y_i + y_{i+1}}{2}$	$\bar{y} = \frac{\sum_{i=1}^{n-1} \frac{y_i + y_{i+1}}{2} \Delta t_i}{\sum_{i=1}^{n-1} t_i}$

Средние показатели изменения уровней ВР

Средний абсолютный прирост

$$\Delta\bar{y} = \frac{1}{n}\sum_{i=1}^{n-1} \Delta y_i^{\text{ц}}$$

Средний темп роста

$$\overline{Ty} = \sqrt[n]{\prod_{i=1}^{n-1} Ty_i^{\text{ц}}} = \sqrt[n]{\frac{y_n}{y_1}}$$

Средний темп прироста

$$\overline{\Delta Ty} = \overline{\Delta Ty} - 1$$

Размах уровней: $$\delta_i = |y_i - \bar{y}|$$

Среднее линейное отклонение ряда: $$d = \frac{1}{n}\sum_{i=1}^{n}\delta_i$$

Среднее квадратичное отклонение: $$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\delta_i^2}$$

Коэффициент вариации: $$V = \frac{\sigma}{\bar{y}}$$

Автокорреляционная функция (ACF)

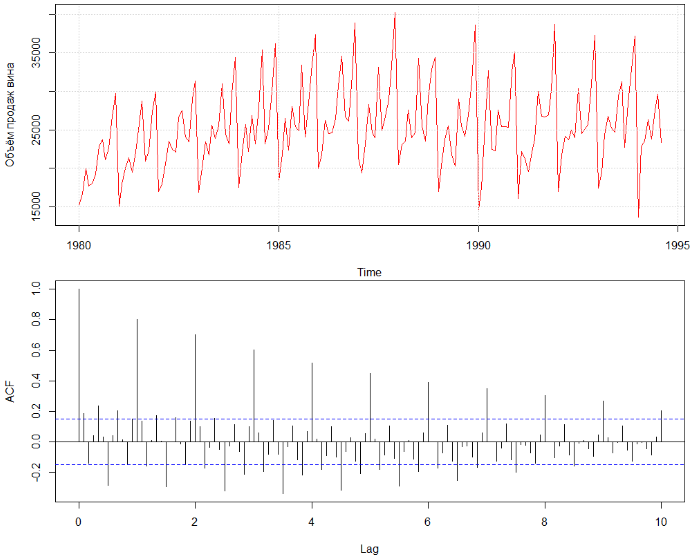
Сопоставим каждому лагу τ значение коэффициента корреляции между временным рядом Y_t и $Y_{t-\tau}$

$$\rho_\tau = \frac{\mathrm{Cov}(Y_t,Y_{t+\tau})}{D(Y_t)} = \frac{\sum\limits_{t=1}^{n-\tau} (y_t - \bar{y})(y_{t+\tau} - \bar{y})}{\sum\limits_{t=1}^{n} (y_t - \bar{y})^2} \in [-1,1],$$

Проверка значимости отличия автокорреляции от нуля:

$$\begin{aligned} \text{временной ряд:} \quad & Y^T = Y_1, \ldots, Y_n \\ \text{гипотезы:} \quad & H_0 : r_\tau = 0, \qquad H_1 : r_\tau \neq 0 \\ \text{статистика:} \quad & T\left(Y^T\right) = \frac{r_\tau \sqrt{T - \tau - 2}}{\sqrt{1 - r_\tau^2}}; \\ \text{нулевое распределение:} \quad & St\,(T - \tau - 2). \end{aligned}$$

Коррелограмма



Частная автокорреляционная функция PACF

partial autocorrelation function

$PACF(\tau)$ — МНК оценка коэффициента α_τ в регрессии:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \alpha_\tau Y_{t-\tau} + \varepsilon_t$$

Частная автокорреляция $PACF(\tau)$ показывает «чистую» зависимость между уровнями Y_t и $Y_{t-\tau}$ временного ряда при исключении влияния промежуточных значений.

Компоненты временных рядов

Тренд u_t — плавное долгосрочное изменение уровня ряда.

Сезонность s_t — циклические изменения уровня ряда с постоянным периодом.

Цикличность v_t — изменения уровня ряда с переменным периодом

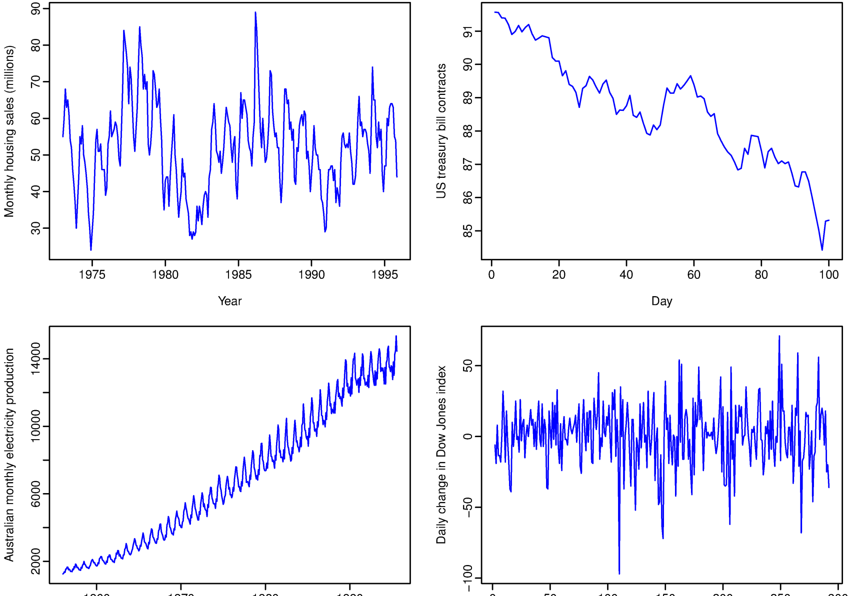
- цикл жизни товара,
- экономические волны,
- периоды солнечной активности

Остатки ε_t — непрогнозируемая случайная компонента ряда.

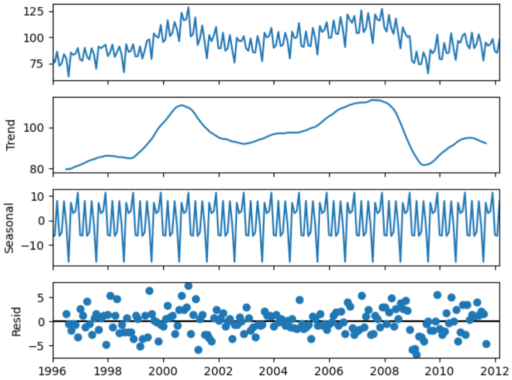
Декомпозиция ВР:

Аддитивная	**Мультипликативная**	**Смешанная**
$y_t = u_t + s_t + v_t + \varepsilon_t$	$y_t = u_t \cdot s_t \cdot v_t \cdot \varepsilon_t$	$y_t = u_t \cdot s_t \cdot v_t + \varepsilon_t$

Примеры, содержащие компоненты



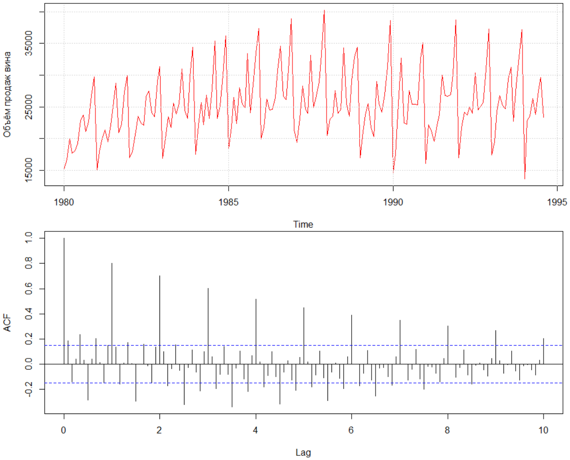
Метод STL-декомпозиции



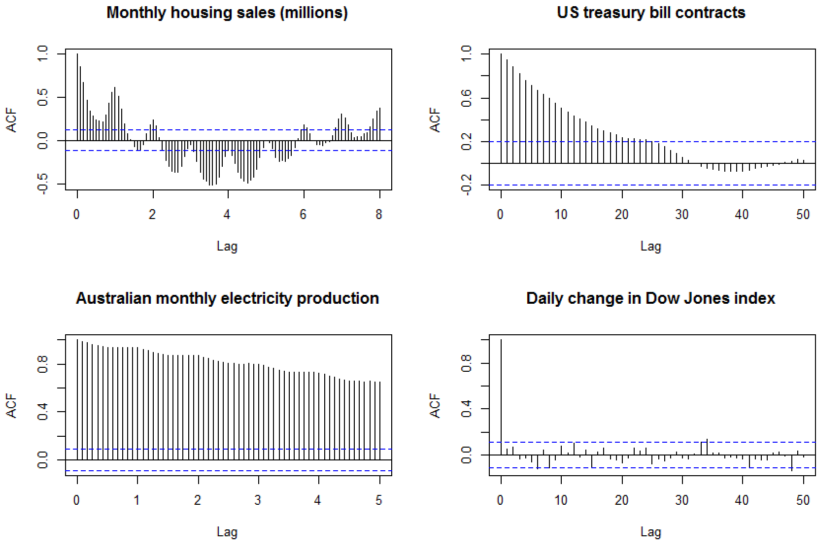

```
from statsmodels.tsa.seasonal \
 import seasonal_decompose
elecequip = pd.read_csv('elecequip.csv')
elecequip.set_index('Index', inplace=True)

decompose = seasonal_decompose(
 elecequip,
 model = 'additive',
 period=12)

trend = decompose.trend
seasonal = decompose.seasonal
residual = decompose.resid

decompose.plot();
```

https://github.com/antoinecarme/TimeSeriesData/raw/refs/heads/master/fpp2/elecequip.csv

Компоненты временных рядов



Компоненты временных рядов



Метод разности средних

Проверка наличия тренда

Предполагается, что данные временного ряда нормально распределены.

1. Временной ряд разбивают на две примерно равные по числу уровней части.
2. Вычисляются средние значения и выборочные дисперсии
3. Проверяется гипотеза о равенстве дисперсий с помощью F-критерия Фишера. Если гипотеза отвергается, то метод не применим.
4. Проверяется гипотеза о равенстве средних. Если гипотеза отвергается, то имеет место тернд.

По этому методу (в первой части n1 первых уровней исходного ряда, во второй остальные n2=n-n1 уровней). Каждая из частей рассматривается как самостоятельная выборочная совокупность, имеющая нормальное распределение. Для каждой из этих частей вычисляются средние значения и дисперсии:

Метод Фостера—Стьюарта

Проверка наличия тренда

1. Для $t = \overline{2,n}$ определяются две числовые последовательности:
$$g_t = [\forall i < t\ (y_t > y_i)], \qquad l_t = [\forall i < t\ (y_t < y_i)],$$
2. Вычисляются две величины:
$$d = \sum_{t=2}^{n}(g_t - l_t), \qquad s = \sum_{t=2}^{n}(g_t + l_t)$$
Обе величины асимптотически нормальны и имеют независимые распределения.
3. Для обнаружения тренда в среднем проверяется $H_0\colon d = 0$:
$$t_d = \frac{|d|}{\sigma_2} \sim St(\alpha, n-1), \qquad \sigma_2 = \sqrt{2\sum_{t=2}^{n}\frac{1}{t}}$$
4. Для обнаружения тренда в дисперсии проверяется гипотеза $H_0\colon s = \bar{y}$:
$$t_s = \frac{|s - \bar{y}|}{\sigma_1} \sim St(\alpha, n-1), \qquad \sigma_1 = \sqrt{2\sum_{t=2}^{n}\frac{1}{t} - 4\sum_{t=2}^{n}\frac{1}{t^2}}$$

Метод аналитического выравнивания

Выделение тренда

Идея: Задать трендовую компоненту функцией.

Вид зависимости	Формула
Линейная	$u(t) = a + bt$
Квадратическая	$u(t) = a + bt + ct^2$
Обратная	$u(t) = a + \frac{b}{t}$
Степенная	$u(t) = a + t^b$
Показательная	$u(t) = a + b^t$
Экспоненциальная	$u(t) = a + e^{bt}$

Как выбрать?

- метод Тинтнера[1]
- метод характеристик прироста

Алгоритм:

1. Выбирается вид функциональной зависимости
2. формируется модель $y_t = u(t) + \varepsilon_t$
3. Путем решения задачи регрессии $y_t = u(t)$ определяются коэффициенты u_t.
4. Выделяется тренд $u_t = u(t)$

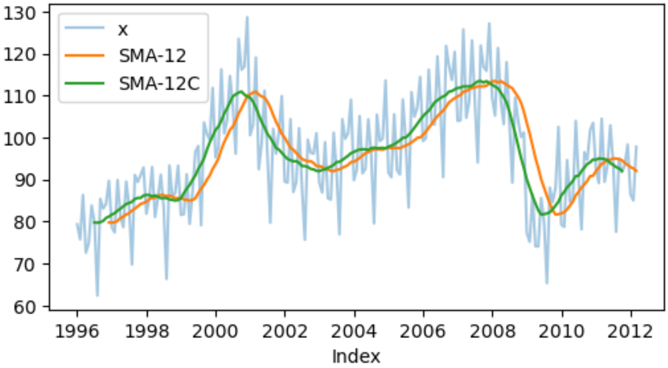
[1] Саженкова Т.В. Пономарёв И.В., Пронь С.П. Методы анализа временных рядов: учебно-методическое пособие. Барнаул: Изд-во Алт. ун-та. 2020. С. 34

Метод скользящего среднего МСС (SMA - simple moving average)

Сглаживание ряда; Выделение тренда

Идея: сопоставить значению уровня y_i — среднее арифметическое предыдущих значений.[2]

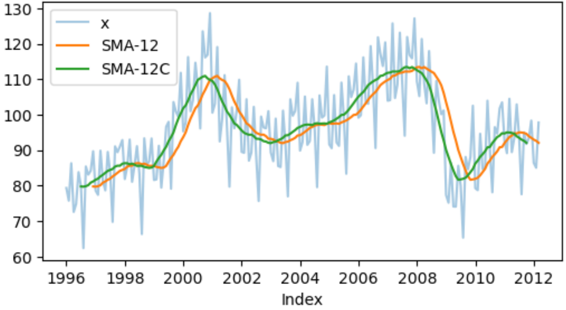
$$M_t = \frac{1}{m} \sum_{i=t-m+1}^{t} y_i \quad \text{или} \quad M_t = \frac{1}{m} \sum_{i=t-\lceil m/2 \rceil}^{t+\lfloor m/2 \rfloor} y_i$$

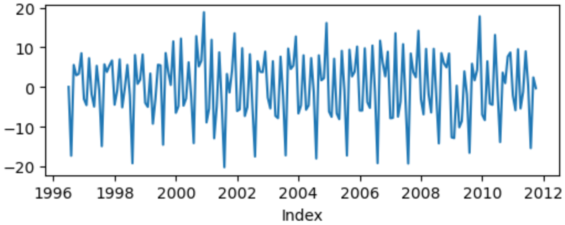


[2] pandas.DataFrame.rolling().mean()

Выравнивание ВР

Выравнивание временного ряда — удаление трендовой компоненты:





Простейшие методы прогнозирования

- средним:

$$\hat{y}_{T+d} = \frac{1}{T}\sum_{t=1}^{T} y_t;$$

- средним за последние k отсчётов:

$$\hat{y}_{T+d} = \frac{1}{k}\sum_{t=T-k}^{T} y_t;$$

- наивный:

$$\hat{y}_{T+d} = y_T;$$

- наивный сезонный (s — период сезонности):

$$\hat{y}_{T+d} = y_{T+d-ks},\ k = \lfloor (d-1)/s \rfloor + 1;$$

- экстраполяции тренда:

$$\hat{y}_{T+d} = y_T + d\frac{y_T - y_1}{T-1}.$$

Идея метода Брауна

Метод взвешенного среднего:

$$M_t = \sum_{i=0}^{m-1} w_i y_{t-i}, \qquad \sum_{i=0}^{m-1} w_i = 1.$$

Возьмем в качестве весов геометрическую убывающую последовательность:

$$\hat{y}_t = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \dots$$

Рекуррентная форма **экспоненциального сглаживания**:

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

α **— параметр сглаживания**

- $\alpha \to 1$ — больший вес последним точкам,
- $\alpha \to 0$ — большее сглаживание.

Наблюдение	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
y_T	0.2	0.4	0.6	0.8
y_{T-1}	0.16	0.24	0.24	0.16
y_{T-2}	0.128	0.144	0.096	0.032
y_{T-3}	0.1024	0.0864	0.0384	0.0064
y_{T-4}	0.08192	0.05184	0.01536	0.00128
y_{T-5}	0.065536	0.031104	0.006144	0.000256

Прогнозированипе методом Брауна

Для прогнозирования рекуррентная формула может быть переписана в следующем виде:

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

- Метод подходит для прогнозирования рядов без тренда и сезонности:

$$\begin{aligned} \hat{y}_{t+1|t} &= l_t, \\ l_t &= \alpha y_t + (1-\alpha)\, l_{t-1} = \hat{y}_{t|t-1} + \alpha \cdot e_t. \end{aligned}$$

  $e_t = y_t - \hat{y}_{t|t-1}$ — ошибка прогноза отсчёта времени t

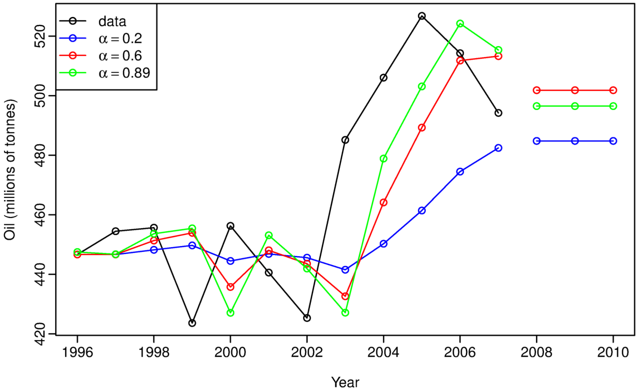
- Прогноз зависит от l_0:

$$\hat{y}_{T+1|T} = \sum_{j=1}^{T-1} \alpha\,(1-\alpha)^j\, y_{T-j} + (1-\alpha)^T\, l_0.$$

  Можно взять $l_0 = y_1$ или оптимизировать его.

- Прогноз получается *плоский*, т. е. $\hat{y}_{t+d|t} = \hat{y}_{t+1|t}$.

Метод Брауна. Пример



Простое экспоненциальное сглаживание в применении к данным о добыче нефти в Саудовской Аравии (1996–2007)[3].

[3] https://www.statsmodels.org/stable/examples/notebooks/generated/exponential_smoothing.html

Метод Хольта (двухпараметрический метод сглаживания)

1957, statsmodels.tsa.holtwinters.Holt(data, damped_trend=False)

Идея: Учтем тренд

Аддитивный линейный тренд (метод Хольта):

$$
\begin{aligned}
\hat{y}_{t+d|t} &= l_t + d b_t, \\
l_t &= \alpha y_t + (1-\alpha)\left(l_{t-1} + b_{t-1}\right), \\
b_t &= \beta\left(l_t - l_{t-1}\right) + (1-\beta)\, b_{t-1}.
\end{aligned}
$$

Мультипликативный линейный (экспоненциальный) тренд:

$$
\begin{aligned}
\hat{y}_{t+d|t} &= l_t b_t^d, \\
l_t &= \alpha y_t + (1-\alpha)\left(l_{t-1} b_{t-1}\right), \\
b_t &= \beta \frac{l_t}{l_{t-1}} + (1-\beta)\, b_{t-1}.
\end{aligned}
$$

- $\alpha \in [0,1]$ — параметр сглаживания уровня
- $\beta \in [0,1]$ — параметр сглаживания тренда

Если $\alpha = \beta$, то это *двойное экспоненциальное сглаживание Брауна*.

Метод Хольта с затуханием тренда

statsmodels.tsa.holtwinters.Holt(data, damped_trend=True)

Аддитивный затухающий тренд:

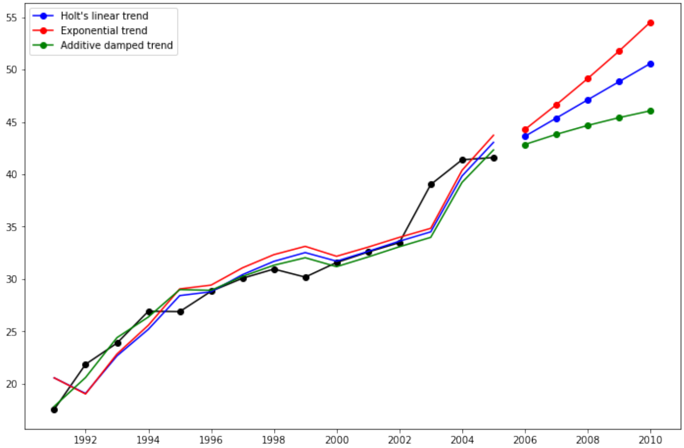
$$\begin{aligned} \hat{y}_{t+d|t} &= l_t + \left(\phi + \phi^2 + \dots + \phi^d\right) b_t, \\ l_t &= \alpha y_t + (1 - \alpha)\left(l_{t-1} + \phi b_{t-1}\right), \\ b_t &= \beta\left(l_t - l_{t-1}\right) + (1 - \beta)\,\phi b_{t-1}. \end{aligned}$$

Мультипликативный затухающий тренд:

$$\begin{aligned} \hat{y}_{t+d|t} &= l_t b_t^{\left(\phi + \phi^2 + \dots + \phi^d\right)}, \\ l_t &= \alpha y_t + (1 - \alpha)\, l_{t-1} b_{t-1}^{\phi}, \\ b_t &= \beta \frac{l_t}{l_{t-1}} + (1 - \beta)\, b_{t-1}^{\phi}. \end{aligned}$$

- $\alpha \in [0,1]$ — параметр сглаживания уровня
- $\beta \in [0,1]$ — параметр сглаживания тренда
- $\phi \in (0,1)$ — коэффициент затухания.

Метод Хольта. Пример



Прогнозы поголовья овец в Азии с учётом тренда.[4]

[4] https://www.statsmodels.org/stable/examples/notebooks/generated/exponential_smoothing.html

Метод Хольта—Винтерса (Трехпараметрическое сглаживание)

1960, statsmodels.tsa.holtwinters.ExponentialSmoothing

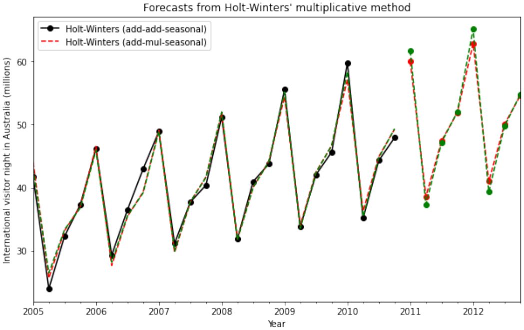
Аддитивная сезонность с периодом длины m (метод Тейла—Веджа):

$$\begin{aligned}
\hat{y}_{t+d|t} &= l_t + db_t + s_{t-m+(d \mod m)}, \\
l_t &= \alpha\left(y_t - s_{t-m}\right) + (1-\alpha)\left(l_{t-1} + b_{t-1}\right), \\
b_t &= \beta\left(l_t - l_{t-1}\right) + (1-\beta)\, b_{t-1}, \\
s_t &= \gamma\left(y_t - l_{t-1} - b_{t-1}\right) + (1-\gamma)\, s_{t-m}.
\end{aligned}$$

Мультипликативная сезонность (Хольта—Винтерса):

$$\begin{aligned}
\hat{y}_{t+d|t} &= \left(l_t + db_t\right) s_{t-m+(d \mod m)}, \\
l_t &= \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)\left(l_{t-1} + b_{t-1}\right), \\
b_t &= \beta\left(l_t - l_{t-1}\right) + (1-\beta)\, b_{t-1}, \\
s_t &= \gamma \frac{y_t}{l_{t-1} + b_{t-1}} + (1-\gamma)\, s_{t-m}.
\end{aligned}$$

Метод Хольта—Винтерса. Пример



Прогнозы с учётом тренда и сезонности количества ночей, проведённых туристами в Австралии.[5]

[5] https://www.statsmodels.org/stable/examples/notebooks/generated/exponential_smoothing.html

Сезонная компонента периода T:

$$S_{t+T} = S_t \quad \text{очень грубо}$$

Как выявить?

- График
- Коррелограмма
- Спектральные методы (Фурье)
- Статистические критерии. Проверка гипотезы о случайности ряда: $l_t = (y_t - u_t) - s_t$.

Как выделить?

- Регрессия
- Спектр
- Итерационные методы[6]

[6] Саженкова Т.В. Пономарёв И.В., Пронь С.П. Методы анализа временных рядов: учебно-методическое пособие. Барнаул: Изд-во Алт. ун-та. 2020. С. 55

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:

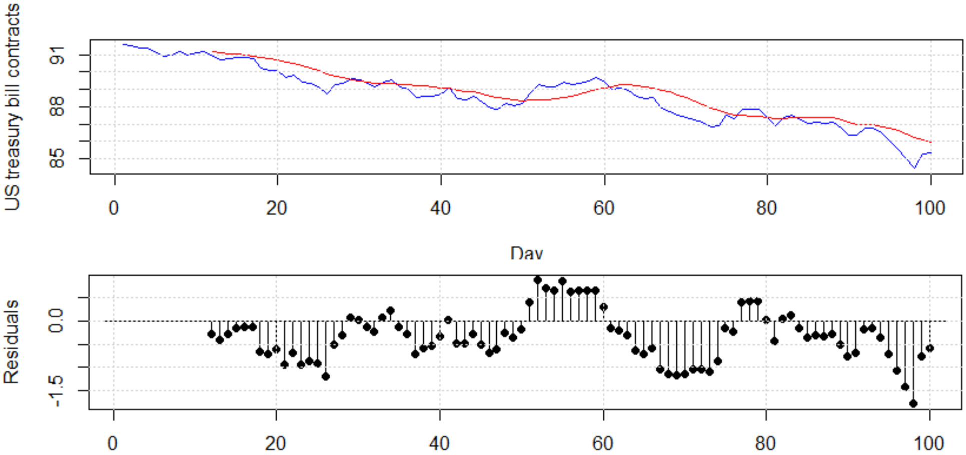
$$\hat{y}_{R+d|R}, \ldots, \hat{y}_{T|T-d},$$

или с фиксированным концом истории при разных отсрочках:

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

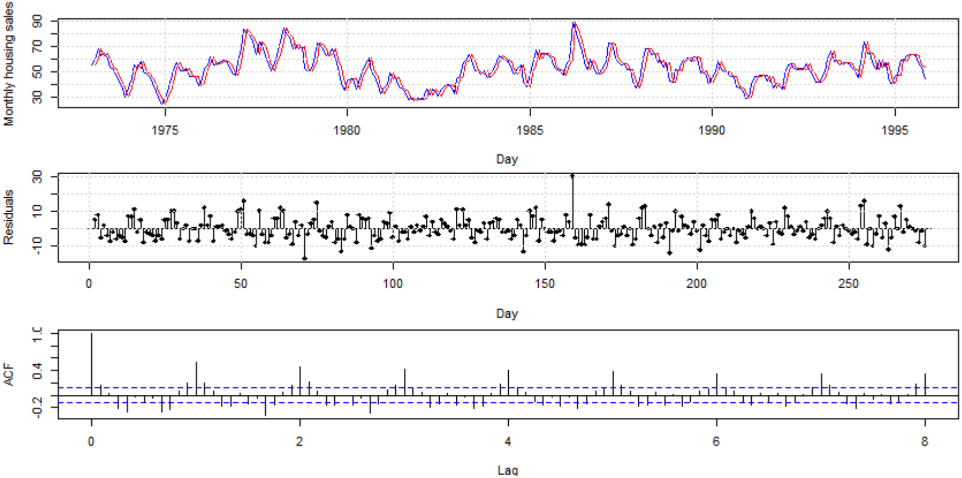
Необходимые свойства остатков прогноза

- Несмещённость — равенство среднего значения нулю:



Необходимые свойства остатков прогноза

- Неавтокоррелированность — отсутствие неучтённой зависимости от предыдущих наблюдений:



Q-критерий Льюнга—Бокса

Проверим гипотезу о том, что автокорреляция равна нулю сразу при всех лагах от 1 до L

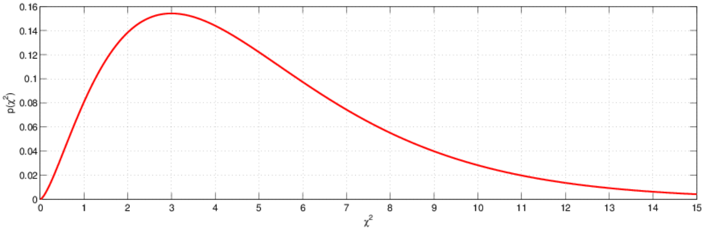
ряд ошибок прогноза: $\varepsilon^T = \varepsilon_1, \dots, \varepsilon_T$;

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

альтернатива: $H_1\colon H_0$ неверна;

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

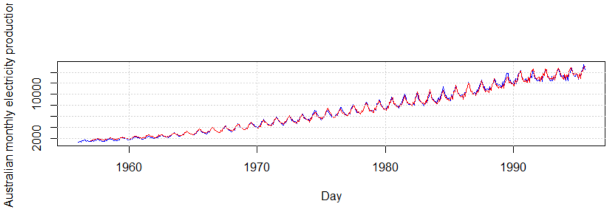
нулевое распределение: χ^2_{L-K}, K — число настраиваемых параметров модели ряда.

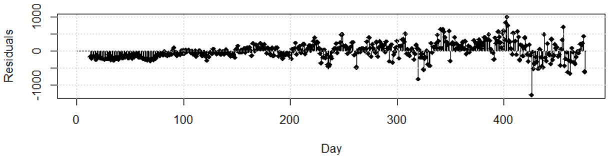


Отсутствие зависимости от времени

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

- Стационарность — отсутствие зависимости от времени:





Критерий KPSS (Kwiatkowski-Philips-Schmidt-Shin)

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

Есть и другие критерии для проверки стационарности: Дики-Фуллера, Филлипса-Перрона и их многочисленные модификации[7]

[7] Patterson K. *Unit root tests in time series, volume 1: key concepts and problems.* — Palgrave Macmillan, 2011

Как проверить свойства остатков прогноза?

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