Mathematical Forecasting Methods

Лекция 2

МФТИ

Осень, 2024

Временной ряд

- **Временной ряд** — это совокупность значения параметра $\{x_1, x_2, ..., x_T\} = \{x_t\}_{t=1}^{T}$, изменяющегося во времени, через равные промежутки времени.
- **Задача прогнозирования**: найти функции $f_{T,d}$:

$$x_{T+d} \approx f_{T,d}(x_1, ...x_T; w) =: \hat{x}_{T+d},$$

  где $f_{T,d}$ — модель временного ряда, $d = 1, ..., D$ — горизонт прогнозирования.
- **Минимизация квадратов ошибок** (МНК):

$$Q_t(w) = \sum_{t=1}^{T} (\hat{x}_t(w) - x_t)^2 \to \min_{w}$$

Временной ряд

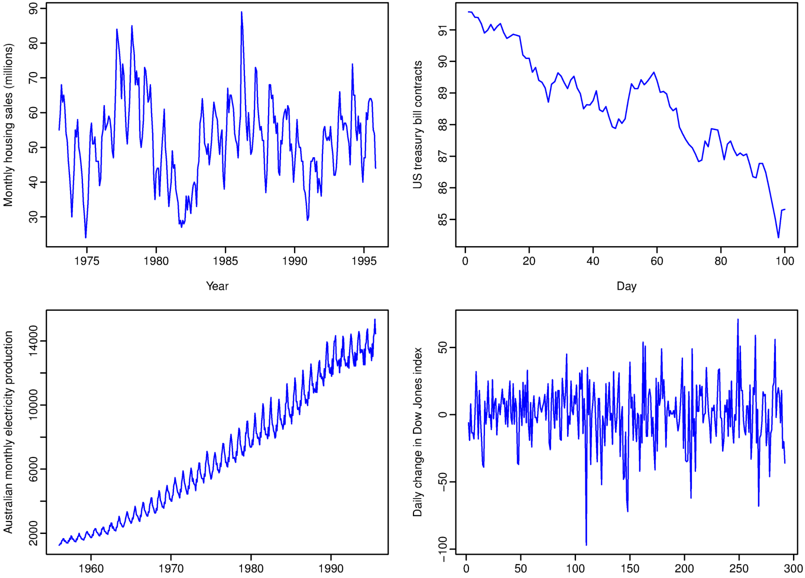
Важно:

- временной ряд — реализация последовательности случайных величин,
- совокупность случайных величин — *дискретный случайный или стохастический процесс*,
- при каждом фиксированном t значение стохастического процесса рассматривается как случайная величина.

Компоненты временного ряда

- **тренд** — плавное долгосрочное изменение временного ряда,
- **сезонность** — циклические изменения временного ряда с постоянным периодом,
- **цикл** — изменения временного ряда с **переменным** периодом (цикл жизни товара, экономические волны, периоды солнечной активности),
- **ошибка** — непрогнозируемая случайная компонента ряда.

Примеры временных рядов

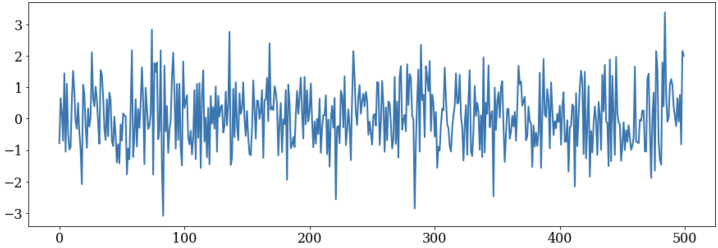


Стационарный временной ряд

Определение. Временной ряд $\{x_i\}_{i=1}^{T}$ называется слабо стационарным (или стационарным в широком смысле), если

- $\mathrm{E}[x_t] = \text{const}$ (т.е. временной ряд не имеет *тренда*),
- $\mathrm{Cov}(x_t, x_{t+k}) = \mathrm{E}[(x_t - \mathrm{E}x_t)(x_{t+k} - \mathrm{E}x_{t+k})] = \gamma(k)$ (ковариация зависит только от разницы во времени).

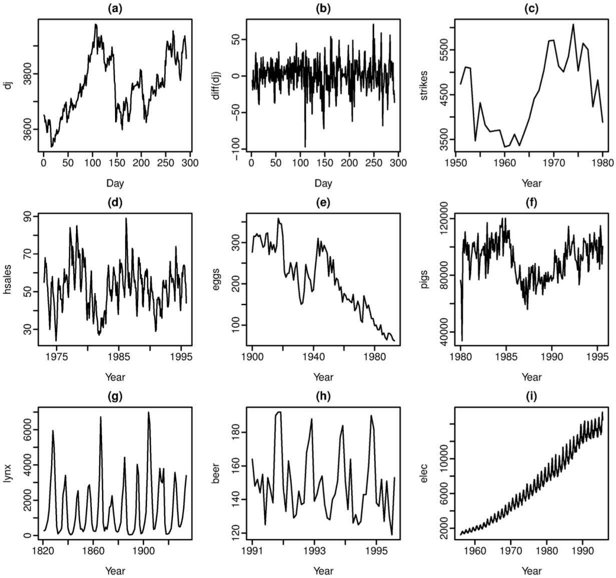
Причем $\mathrm{Cov}(x_t, x_t) = \mathrm{D}(x_t) = \gamma(0) = \gamma_0$, т.е. дисперсия стационарного временного ряда не меняется со временем.



Белый шум $u_t \sim \mathrm{WN}(0, \sigma^2)$: $\mathrm{E}u_t = 0$, $\mathrm{D}u_t = \sigma^2$, $\mathrm{Cov}(u_t, u_{t+k}) = 0$

Стационарный временной ряд

Вопрос: какие из этих рядов, вероятно, стационарные?



Автокорреляция

Определение. Функция $\rho(k)$, где k - *величина лага*, называется *автокорреляционной функцией (autocorrelation function, ACF)* стационарного временного ряда.

$$\rho(k) = \mathrm{Corr}(x_t, x_{t+k}) = \frac{\mathrm{Cov}(x_t, x_{t+k})}{\sqrt{\mathrm{D}(x_t) \cdot \mathrm{D}(x_{t+k})}} = \frac{\gamma(k)}{\sqrt{\gamma(0) \cdot \gamma(0)}} = \frac{\gamma(k)}{\gamma(0)}$$

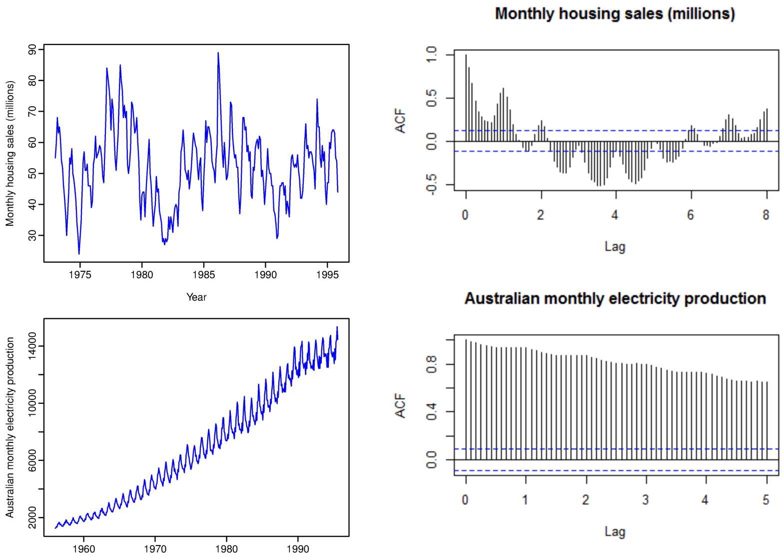
Оценка:

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{T-k}(x_t - \overline{x})(x_{t+k} - \overline{x})}{\sum_{t=1}^{T}(x_t - \overline{x})}, \quad \text{где} \quad \overline{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$$

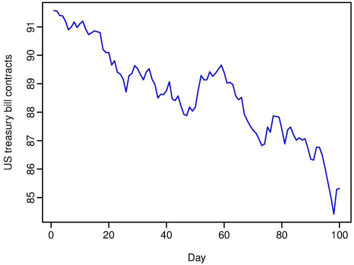
Для стационарных временных рядов верно, что

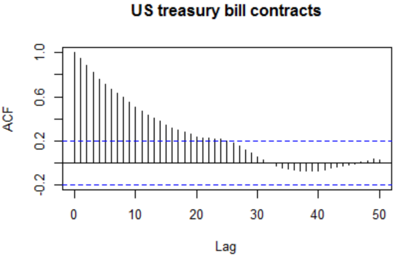
$$\lim_{k \to \infty} \rho(k) = 0$$

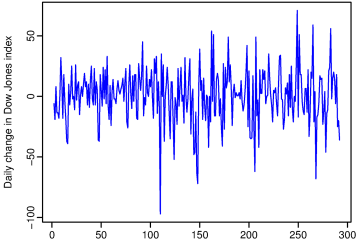
Автокорреляция

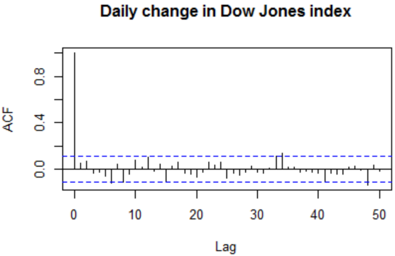


Автокорреляция









Частичная Автокорреляция (PACF)

Частичная автокорреляция лага $k > 1$ также измеряет корреляцию между x_t и x_{t+k}, но за вычетом линейных зависимостей этих величин от $x_{t+1}, ..., x_{t+k-1}$:

$$\rho_{PACF}(k) = \mathrm{Corr}(x_t - \hat{x}_t, x_{t+k} - \hat{x}_{t+k})$$

Здесь $\hat{x}_t$, $\hat{x}_{t+k}$ - это линейные комбинации $x_{t+1}, ..., x_{t+k-1}$ с коэффициентами, минимизирующими среднеквадратичную ошибку предсказания значений x_t и x_{t+k} соответственно:

$$\hat{x}_t = \beta_1^{(1)} x_{t+1} + ... + \beta_{k-1}^{(1)} x_{t+k-1}, \quad \hat{x}_{t+k} = \beta_1^{(2)} x_{t+1} + ... + \beta_{k-1}^{(2)} x_{t+k-1}$$

Для стационарных временных рядов значение коэффициента, полученного с помощью МНК, зависит только от разности временных индексов, поэтому коэффициенты для $\hat{x}_t$ и $\hat{x}_{t+k}$ одинаковые, но имеют противоположный порядок:

$$\hat{x}_t = \beta_1 x_{t+1} + ... + \beta_{k-1} x_{t+k-1}, \quad \hat{x}_{t+k} = \beta_{k-1} x_{t+1} + ... + \beta_1 x_{t+k-1}$$

Модель ARMA

Общая смешанная модель ARMA(p,q) (AutoRegression Moving Average) авторегрессии-скользящего среднего:

$$x_t = \mu + \sum_{j=1}^{p} \phi_j x_{t-j} + u_t + \sum_{s=1}^{q} \theta_s u_{t-s}, \quad u_t \sim \mathrm{WN}(0, \sigma^2), \quad \phi_p, \theta_q \neq 0$$

Составные части:

- $\mu + \sum_{j=1}^{p} \phi_j x_{t-j}$ — авторегрессионная часть AR,
- $u_t + \sum_{s=1}^{q} \theta_s u_{t-s}$ — часть скользящего среднего MA (в классическом случае гауссовский белый шум).

Согласно теорема Вольда, любой стационарный ряд может быть аппроксимирован моделью ARMA(p,q) с любой точностью.

Модель ARMA. Прогноз

$$x_t = \mu + \sum_{j=1}^{p} \phi_j x_{t-j} + u_t + \sum_{s=1}^{q} \theta_s u_{t-s}, \quad u_t \sim \mathrm{WN}(0, \sigma^2), \quad \phi_p, \theta_q \neq 0$$

Пусть известны значения ряда x_t и возмущения u_t до момента времени T включительно, а также получены веса ϕ_j, $j = \overline{1, p}$, θ_s, $s = \overline{1, q}$ модели ARMA(p, q).

Выражение для x_{T+1} в рамках модели:

$$x_{T+1} = \mu + \sum_{j=1}^{p} \phi_j x_{T+1-j} + u_{T+1} + \sum_{s=1}^{q} \theta_s u_{T+1-s}$$

Неизвестным в правой части является только возмущение u_{T+1}. Отметим: $\mathrm{E} u_{T+1} = 0$, $\mathrm{Cov}(u_{T+1}, x_t) = 0$ для всех $t \leq T$.

Оценка на момент времени $T + 1$:

$$\hat{x}_{T+1} = \mu + \sum_{j=1}^{p} \phi_j x_{T+1-j} + \sum_{s=1}^{q} \theta_s u_{T+1-s}$$

Модель ARMA. Прогноз

Выражение для x_{T+2} в рамках модели:

$$x_{T+2} = \mu + \sum_{j=1}^{p} \phi_j x_{T+2-j} + u_{T+2} + \sum_{s=1}^{q} \theta_s u_{T+2-s}$$

Неизвестными в правой части здесь является только возмущения u_{T+1}, u_{T+2} и значение ряда x_{T+1}. Как и на предыдущем шаге, занулим неизвестные возмущения, а вместо значения x_{T+1} используем его оценку $\hat{x}_{T+1}$, получим:

$$\hat{x}_{T+2} = \mu + \phi_1 \hat{x}_{T+1} + \sum_{j=2}^{p} \phi_j x_{T+2-j} + \sum_{s=2}^{q} \theta_s u_{T+2-s}$$

Заметим, что MA часть уменьшается с каждым последующим прогнозом в будущее.

Модель ARMA. Прогноз

Последовательное построение оптимального прогноза на τ шагов для общего случая:

1. Записываем ARMA-формулу для $x_{T+\tau}$.
2. Зануляем неизвестные возмущения $u_{T+1}, \dots, u_{T+\tau}$.
3. Заменяем неизвестные значения $x_{T+1}, \dots, x_{T+\tau-1}$ на их прогнозы, полученные на предыдущих шагах.

Вопросы:

- Как получить оптимальные веса модели ARMA?
- Как в реальных данных получить возмущения $u_t,\ t \le T$, необходимые для построения прогноза?

Дифференцирование временного ряда

Дифференцирование ряда — переход к попарным разностям его соседних значений:

$$x_1, \dots, x_T \to x'_2, \dots, x'_T,$$

$$x'_t = x_t - x_{t-1}.$$

Дифференцированием можно стабилизировать среднее значение ряда и избавиться от тренда и сезонности.
Может применяться неоднократное дифференцирование; например, для второго порядка:

$$x_1, \dots, x_T \to x_2^{'}, \dots, x_T^{'} \to x_3^{''}, \dots, x_T^{''},$$

$$x_t^{''} = x_t^{'} - x_{t-1}^{'} = x_t - 2x_{t-1} + x_{t-2}.$$

Дифференцирование временного ряда

Определение: лаговыи оператор L — оператор сдвига, позволяющий получить значения элементов временного ряда на основании ряда предыдущих значений:

$$\mathrm{L}(x_t) \stackrel{\text{def}}{=} x_{t-1}.$$

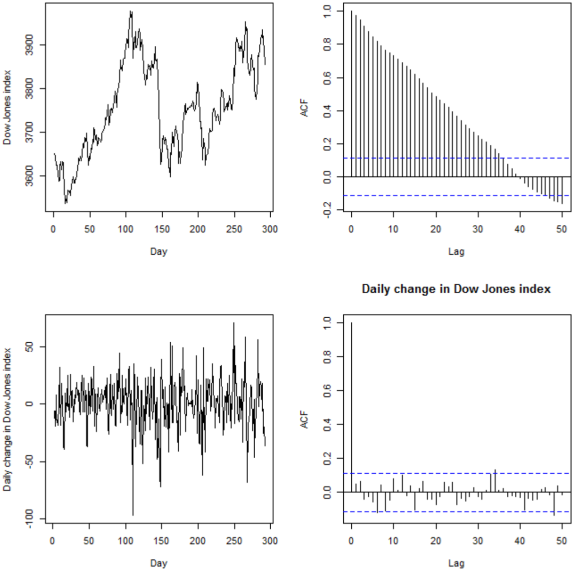
Далее $\mathrm{L}^2(x_t) = \mathrm{L}(\mathrm{L}(x_t)) = \mathrm{L}(x_{t-1}) = x_{t-2}$.

Следовательно, $\mathrm{L}^k(x_t) = x_{t-k}$, причем $\mathrm{L}^0(x_t) = x_t$.

Тогда дифференцирование временного ряда представимо в виде

$$x'_t = x_t - x_{t-1} = (1 - \mathrm{L})(x_t).$$

Дифференцирование временного ряда



Модель ARIMA

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

$$\nabla^d x_t = (1 - \mathrm{L})^d (x_t)$$

описывается моделью ARMA(p,q):

$$\nabla^d x_t = \mu + \sum_{j=1}^{p} \phi_j (\nabla^d x_{t-j}) + u_t + \sum_{s=1}^{q} \theta_s u_{t-s}.$$

Модель Seasonal additive ARMA

Для учета сезонной компоненты в моделью ARMA(p,q), добавляют авторегрессионные части и скользящее среднее по сезонным компонентам периода S.

Модель ARMA(p,q):

$$x_t = \mu + \sum_{j=1}^{p} \phi_j x_{t-j} + u_t + \sum_{s=1}^{q} \theta_s u_{t-s}$$

с авторегрессией с P сезонными компонентами:

$$+\phi_S x_{t-S} + \phi_{2S} x_{t-2S} + ... + \phi_{PS} x_{t-PS}$$

и с скользящим средним с Q сезоными компонентами:

$$+\theta_S u_{t-S} + \theta_{2S} u_{t-2S} + ... + \theta_{QS} u_{t-QS}.$$

Модель SARIMAX

К модели SARIMA(p,d,q)(P,D,Q) добавляюся *экзогенные*переменные, значение которых формируется вне модели. Экзогенные переменные являются в модели независимыми величинами, а их изменение называется автономным изменением.

$$x_t = \mu + \sum_{j=1}^{p} \phi_j x_{t-j} + u_t + \sum_{s=1}^{q} \theta_s u_{t-s} + \ldots + \sum_{i=1}^{r} \beta_i x_i^{\text{exog}}$$

Модель SARIMAX

Из документации statsmodels.tsa.statespace.sarimax.SARIMAX Python:

The SARIMA model is specified $(p, d, q) \times (P, D, Q)_s$.

$$\phi_p(L)\tilde{\phi}_P(L^s)\Delta^d\Delta_s^D y_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\zeta_t$$

In terms of a univariate structural model, this can be represented as

$$y_t = u_t + \eta_t$$
$$\phi_p(L)\tilde{\phi}_P(L^s)\Delta^d\Delta_s^D u_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\zeta_t$$

where η_t is only applicable in the case of measurement error (although it is also used in the case of a pure regression model, i.e. if p=q=0).

In terms of this model, regression with SARIMA errors can be represented easily as

$$y_t = \beta_t x_t + u_t$$
$$\phi_p(L)\tilde{\phi}_P(L^s)\Delta^d\Delta_s^D u_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\zeta_t$$

this model is the one used when exogenous regressors are provided.

---

credit: https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html

Информационные критерии

Информационные критерии — это разные виды регуляризованного правдоподобия.

Критерий Акаике (Akaike's information criterion, AIC) — критерий выбора из класса параметризованных регрессионных моделей, оценивающий модели с разным числом параметров. Содержит функцию штрафа, линейно зависящую от числа параметров:

$$AIC = 2\frac{p+q}{T} + \ln\left(\frac{\sum_{t=1}^{T}(x_t - \hat{x}_t)^2}{n}\right)$$

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.
6. Остатки полученной модели исследуются на несмещённость, стационарность и неавтокоррелированность; если предположения не выполняются, исследуются модификации модели.
7. В финальной модели t заменяется на $T+\tau$, будущие наблюдения — на их прогнозы, будущие ошибки — на нули, прошлые ошибки — на остатки.

Резюме

- временные ряды представляются как случайные процессы,
- стационарный временной ряд по Теореме Вольда представим с помощью ARMA модели,
- модель ARMA — линейная комбинация предыстории и шумов,
- модель SARIMAX — это объединение авторегрессии с некоторыми эвристиками (сезонность, тренд) для обеспечения стационарности временного ряда.