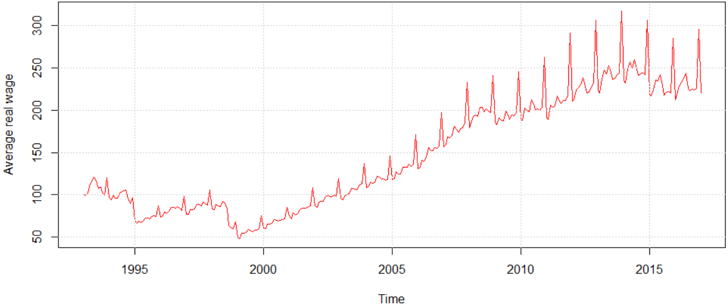
Прикладной статистический анализ данных.
9. Анализ временных рядов, часть первая

Алексей Романенко
psad.homework@gmail.com

2017

Прогнозирование временного ряда

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

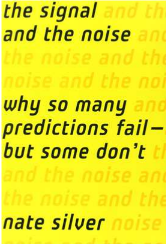


Задача прогнозирования — найти функцию f_T:

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

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

Предсказательный интервал



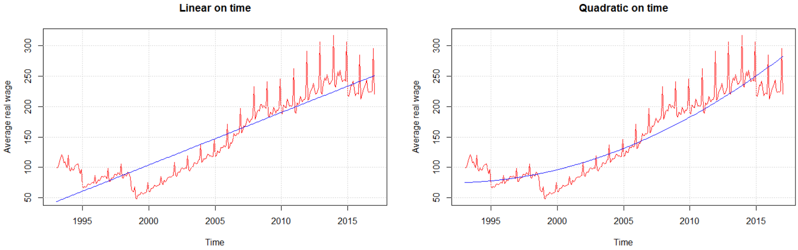
Пример: в апреле 1997 года в Гранд-Форкс, Северная Дакота, произошло наводнение. Город был защищён дамбой высотой в 51 фут; согласно прогнозу, высота весеннего паводка должна была составить 49 футов; истинная высота паводка оказалась равной 54 футам.

50000 жителей было эвакуировано, 75% зданий повреждено или уничтожено, ущерб составил несколько миллиардов долларов.

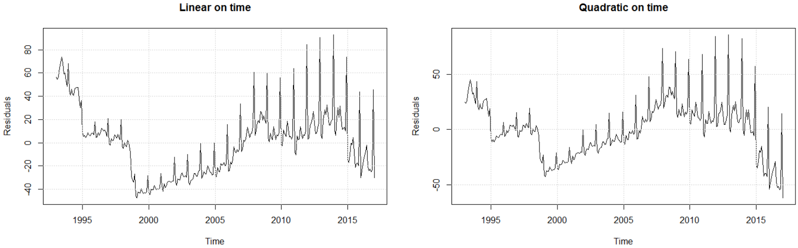
На исторических данных точность прогнозов метеорологической службы составляла ± 9 футов.

Регрессия

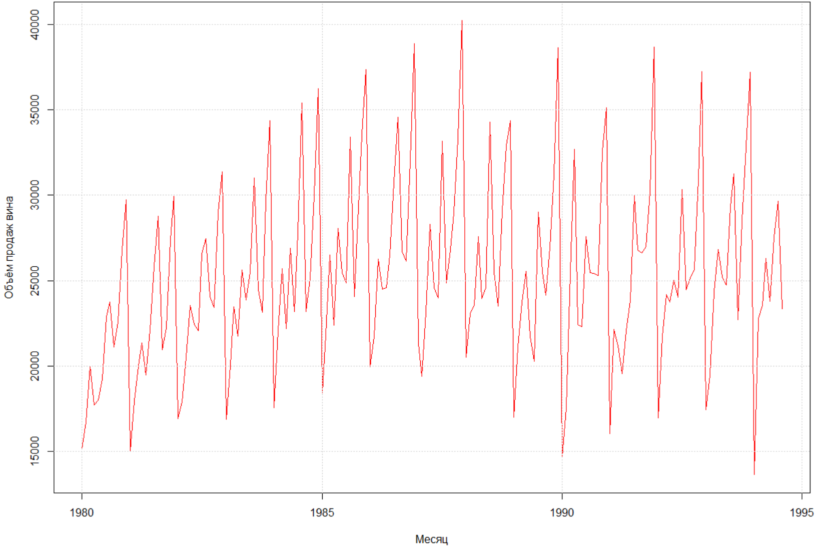
Простейшая идея: сделать регрессию на время.



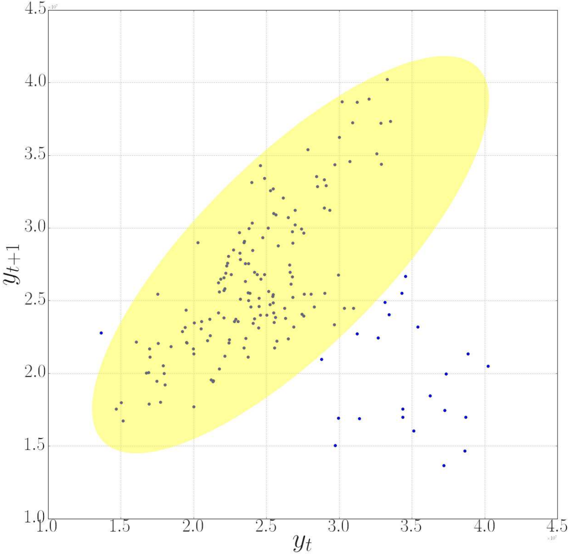
Остатки не выглядят как шум:



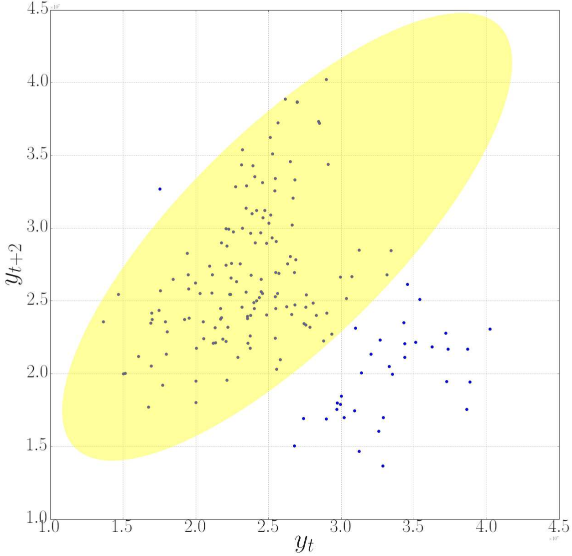
Продажи вина в Австралии



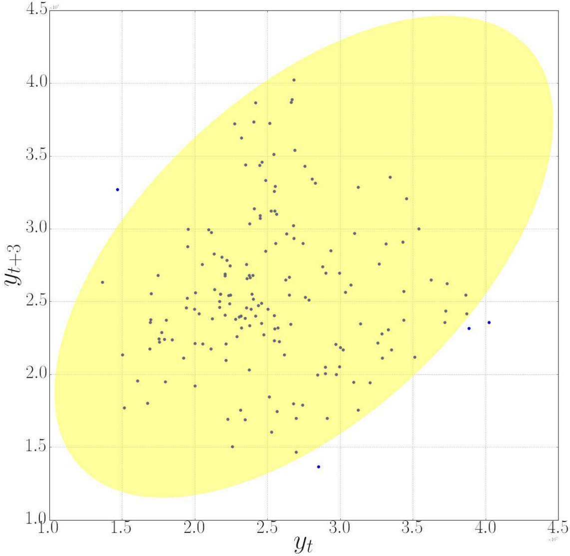
Продажи в соседние месяцы



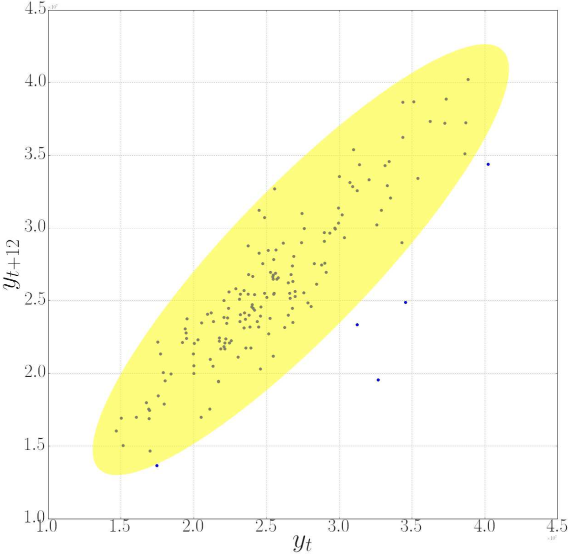
Продажи через 1 месяц



Продажи через 2 месяца



Продажи через год



Автокорреляционная функция (ACF)

Наблюдения временного ряда автокоррелированы.

Автокорреляция:

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

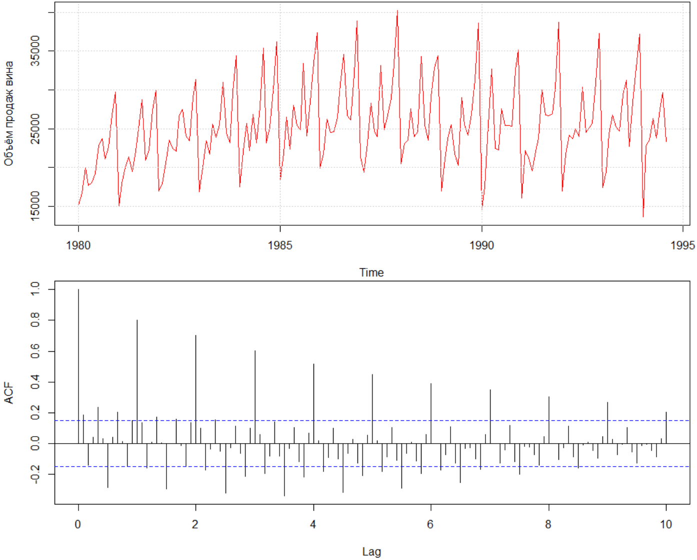
$r_\tau \in [-1, 1]$, τ — лаг автокорреляции.

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:



Компоненты временных рядов

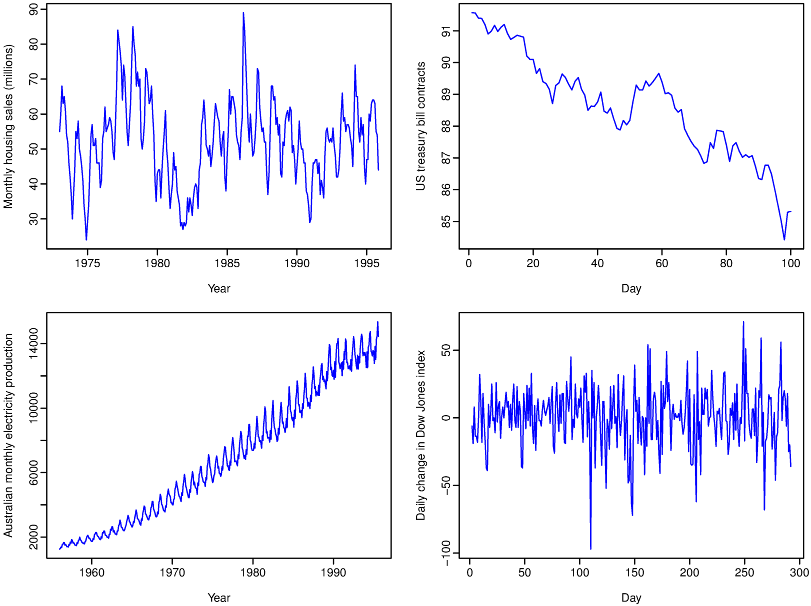
Тренд — плавное долгосрочное изменение уровня ряда.

Сезонность — циклические изменения уровня ряда с постоянным периодом.

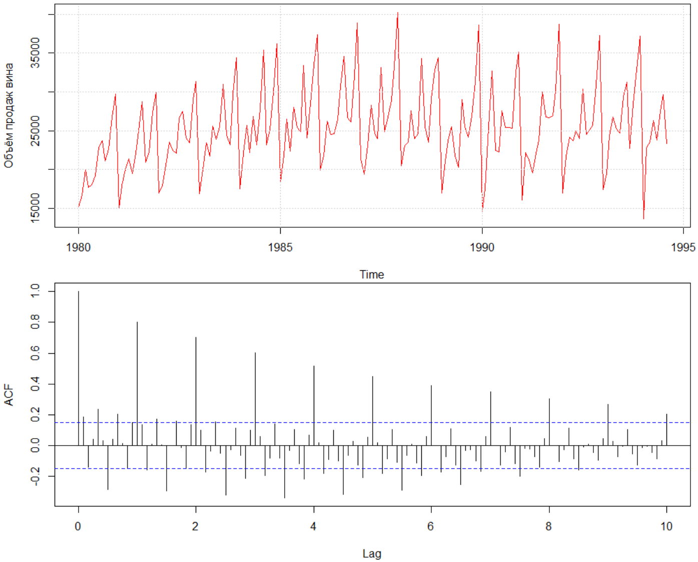
Цикл — изменения уровня ряда с переменным периодом (цикл жизни товара, экономические волны, периоды солнечной активности).

Ошибка — непрогнозируемая случайная компонента ряда.

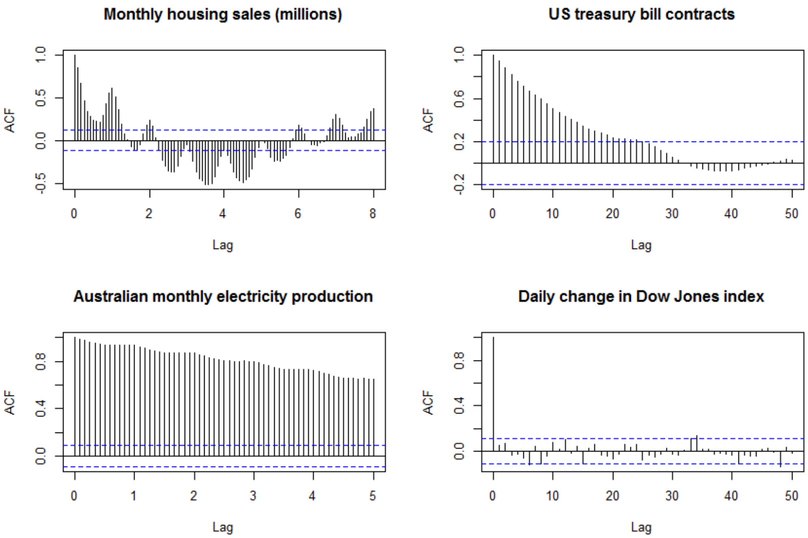
Компоненты временных рядов



Компоненты временных рядов



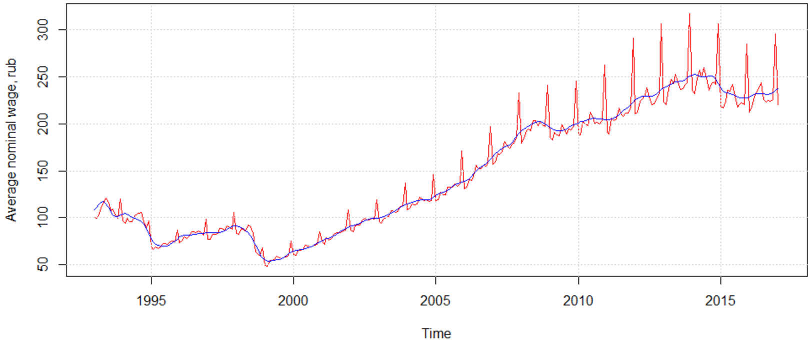
Компоненты временных рядов



STL-декомпозиция:

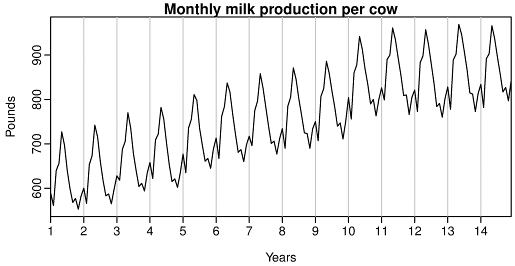
Снятие сезонности

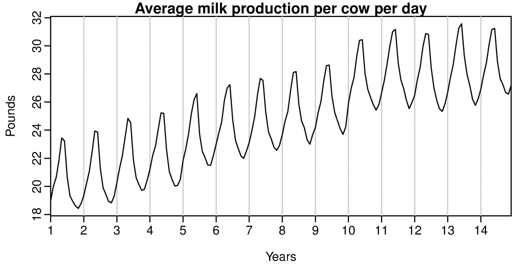
Часто для удобства интерпретации ряда сезонная компонента вычитается:



Календарные эффекты

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





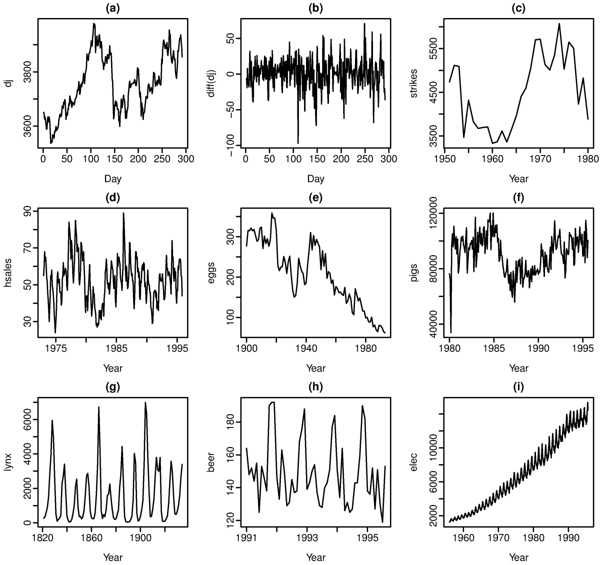
Стационарность

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.

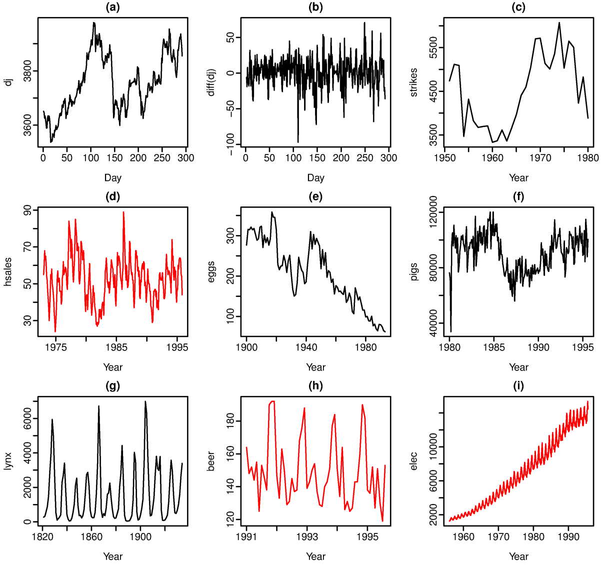
Ряды с непериодическими циклами стационарны, поскольку нельзя предсказать заранее, где будут находится максимумы и минимумы.

Стационарность



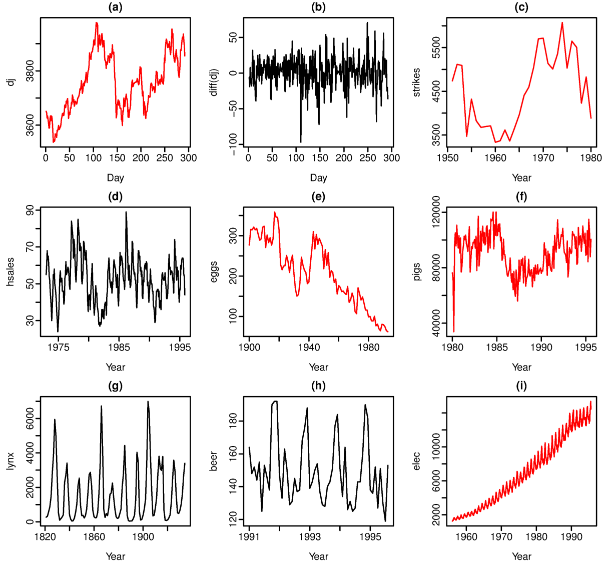
Стационарность

Нестационарны из-за сезонности:



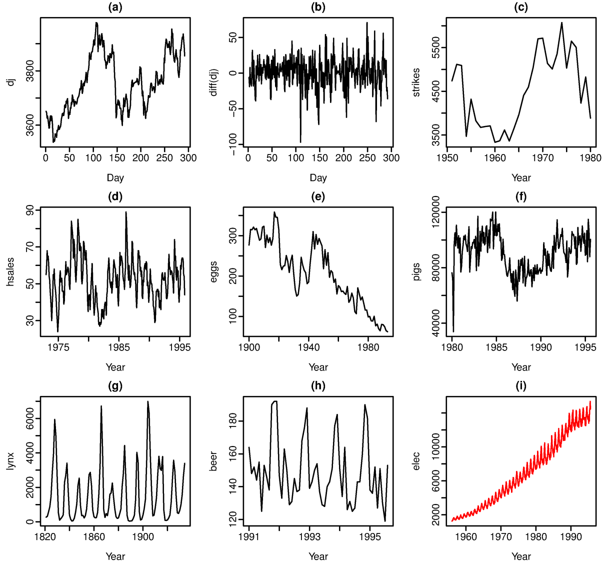
Стационарность

Нестационарны из-за тренда:



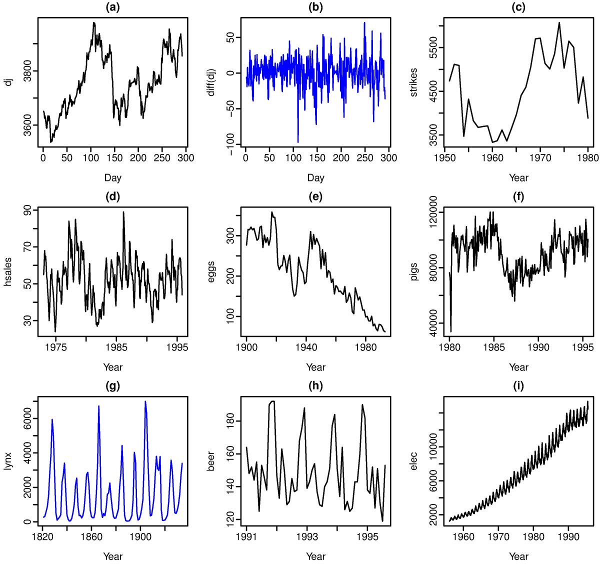
Стационарность

Нестационарны из-за меняющейся дисперсии:



Стационарность

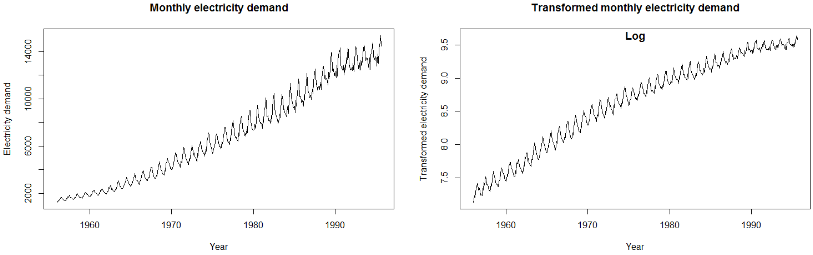
Стационарны:



Стабилизация дисперсии

Для рядов с монотонно меняющейся дисперсией можно использовать стабилизирующие преобразования.

Часто используют логарифмирование:

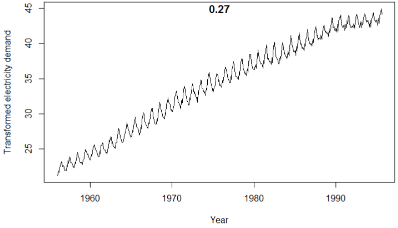


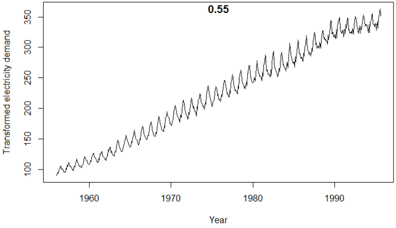
Преобразования Бокса-Кокса

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:

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

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

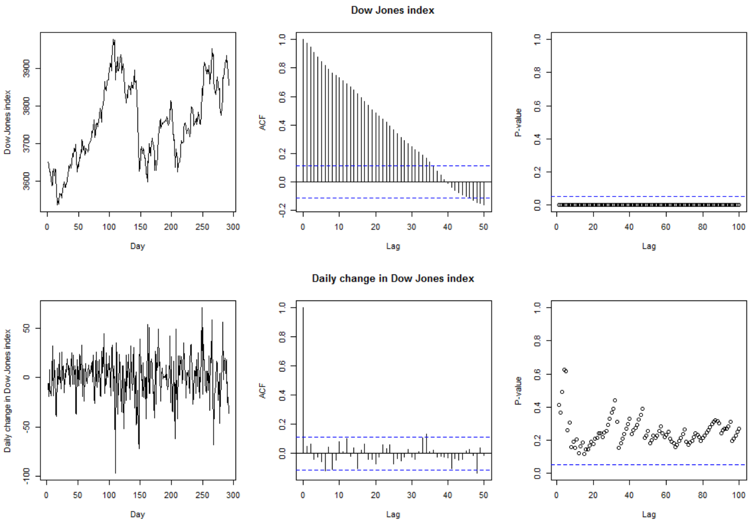
Дифференцированием можно стабилизировать среднее значение ряда и избавиться от тренда и сезонности.

Может применяться неоднократное дифференцирование; например, для второго порядка:

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

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

Дифференцирование



Критерий KPSS: для исходного ряда $p < 0.01$, для ряда первых разностей — $p > 0.1$.

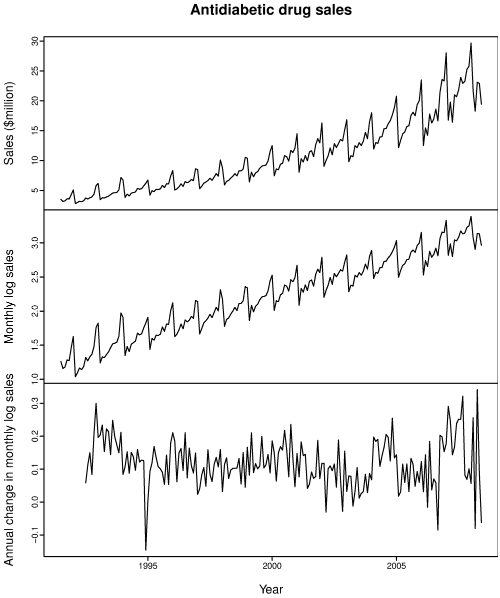
Сезонное дифференцирование

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

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

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

Сезонное дифференцирование



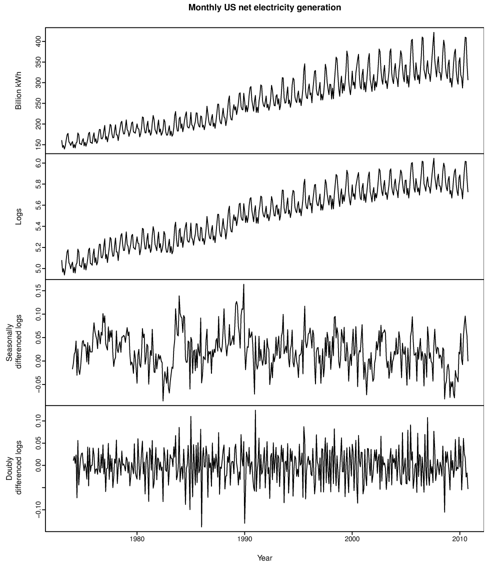
Критерий KPSS: для исходного ряда $p < 0.01$, для логарифмированного — $p < 0.01$, после сезонного дифференцирования — $p > 0.1$.

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.

Комбинированное дифференцирование



Критерий KPSS: для исходного ряда $p < 0.01$, для логарифмированного — $p < 0.01$, после сезонного дифференцирования — $p = 0.0355$, после ещё одного дифференцирования — $p > 0.1$.

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:

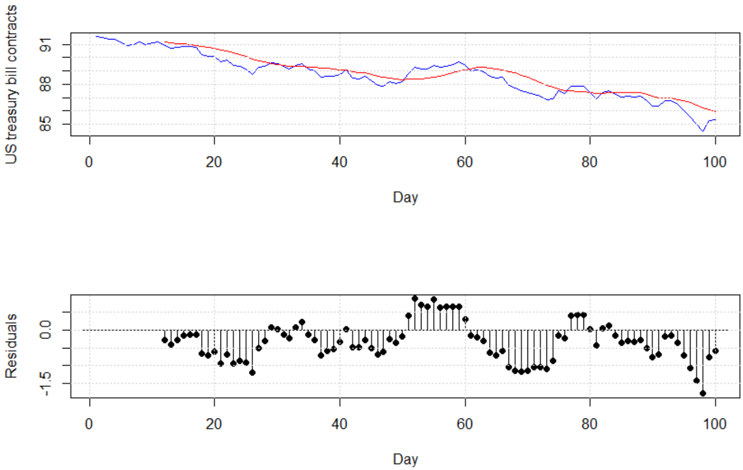
$$\hat{y}_{R+d|R}, \dots, \hat{y}_{T|T-d},$$

или с фиксированным концом истории при разных отсрочках:

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

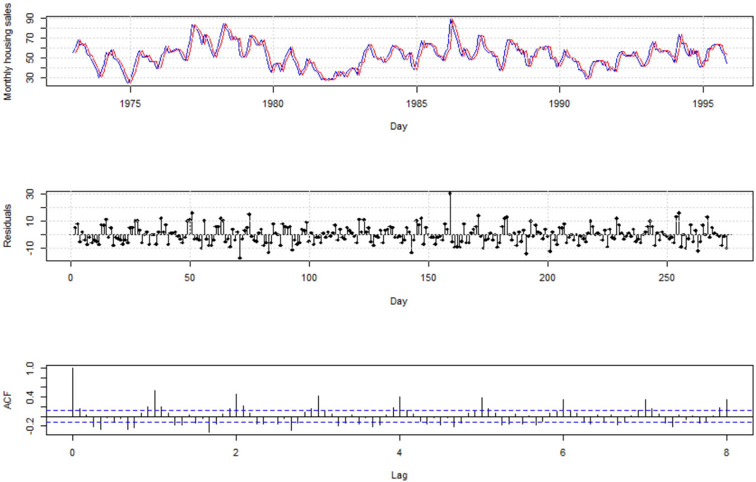
Необходимые свойства остатков прогноза

- Несмещённость — равенство среднего значения нулю:



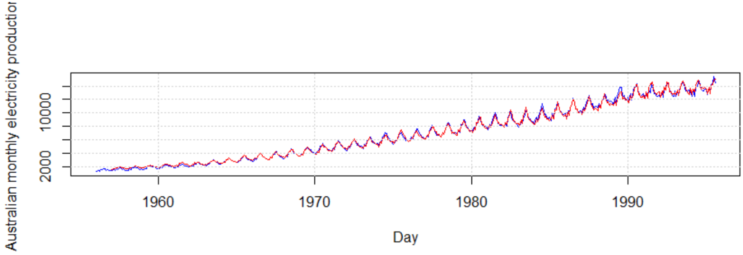
Необходимые свойства остатков прогноза

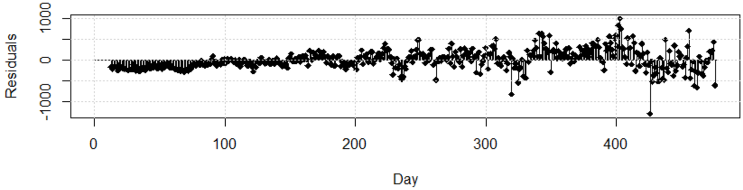
- Неавтокоррелированность — отсутствие неучтённой зависимости от предыдущих наблюдений:



Необходимые свойства остатков прогноза

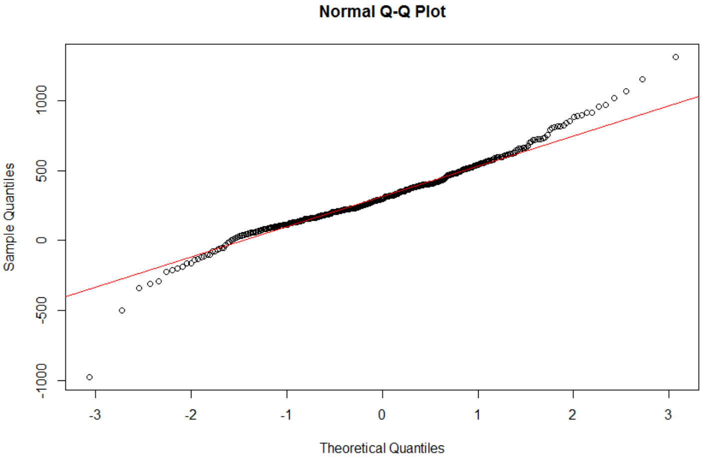
- Стационарность — отсутствие зависимости от времени:





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

нулевая гипотеза: H_0: ряд ε^T стационарен;

альтернатива: H_1: ряд ε^T описывается моделью вида $\varepsilon_t = \alpha\varepsilon_{t-1}$;

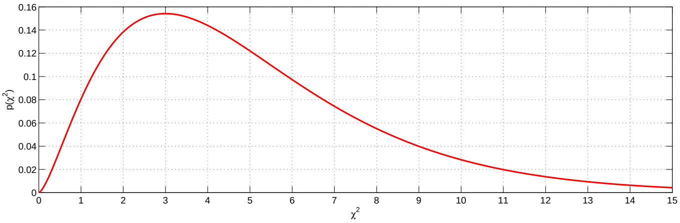
статистика: $KPSS\left(\varepsilon^T\right) = \frac{1}{T^2}\sum\limits_{i=1}^{T}\left(\sum\limits_{t=1}^{i}\varepsilon_t\right)^2 \Big/ \lambda^2$;

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).

Q-критерий Льюнга-Бокса

ряд ошибок прогноза:	$\varepsilon^T = \varepsilon_1, \dots, \varepsilon_T;$
нулевая гипотеза:	$H_0: r_1 = \dots = r_L = 0;$
альтернатива:	$H_1: H_0$ 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.



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:

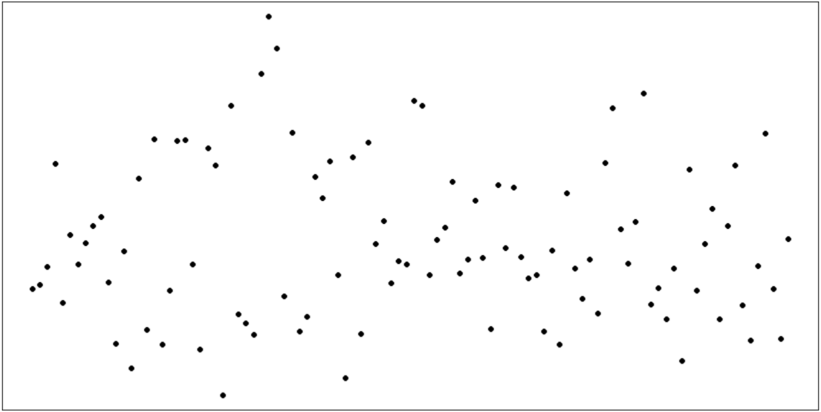
$$\phi\left(B\right) y_t = \left(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p\right) y_t = \varepsilon_t,$$

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

Линейная комбинация p подряд идущих членов ряда даёт белый шум.

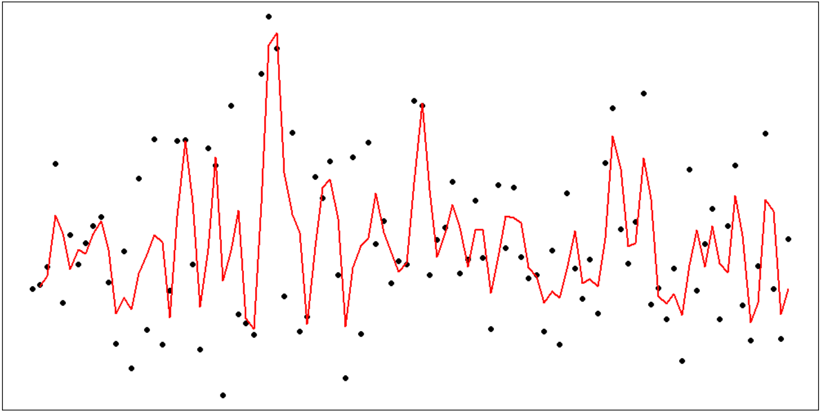
Скользящее среднее

Пусть у нас есть независимый одинаково распределённый во времени шум ε_t:



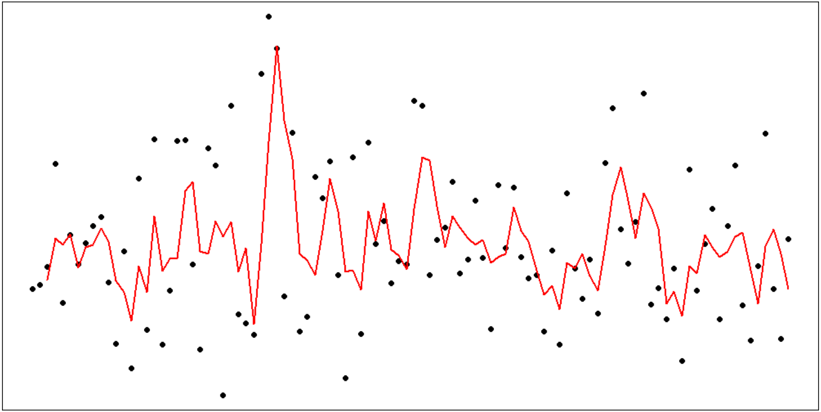
Скользящее среднее

Среднее по двум соседним точкам:



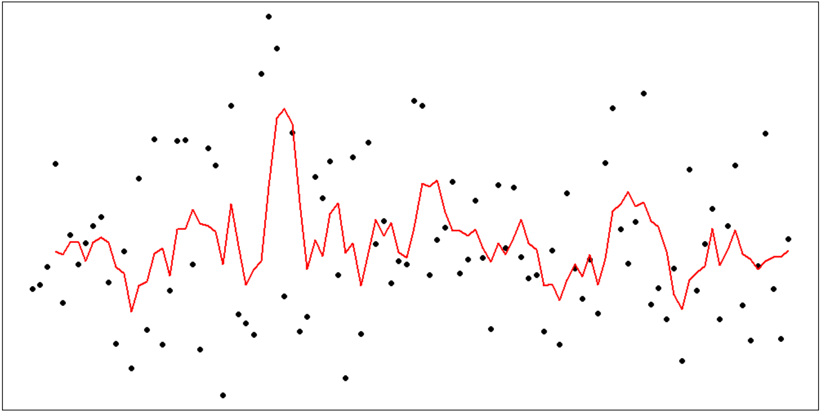
Скользящее среднее

Среднее по трём соседним точкам:



Скользящее среднее

Среднее по четырём соседним точкам:



Скользящее среднее

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

где y_t — стационарный ряд с нулевым средним, $\theta_1, \dots, \theta_q$ — 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

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

Другой способ записи:

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

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

$$\Phi_P\left(B^s\right) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps},$$
$$\Theta_Q\left(B^s\right) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \dots + \Theta_Q B^{Qs}.$$

SARIMA:

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

α, ϕ, θ

- Если все остальные параметры фиксированы, коэффициенты регрессии подбираются методом наименьших квадратов
- Чтобы найти коэффициенты θ, шумовая компонента предварительно оценивается с помощью остатков авторегрессии
- Если шум белый (независимый одинаково распределённый гауссовский), то МНК даёт оценки максимального правдоподобия

d, D

- Порядки дифференцирования подбираются так, чтобы ряд стал стационарным
- Ещё раз: если ряд сезонный, рекомендуется начинать с сезонного дифференцирования
- Чем меньше раз мы продифференцируем, тем меньше будет дисперсия итогового прогноза

q, Q, p, P

- Гиперпараметры нельзя выбирать из принципа максимума правдоподобия: L всегда увеличивается с их ростом
- Для сравнения моделей с разными q, Q, p, P можно использовать информационные критерии
- Начальные приближения можно выбрать с помощью автокорреляций

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 $y_{t+1}, y_{t+2}, \ldots, y_{t+h-1}$:

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

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.

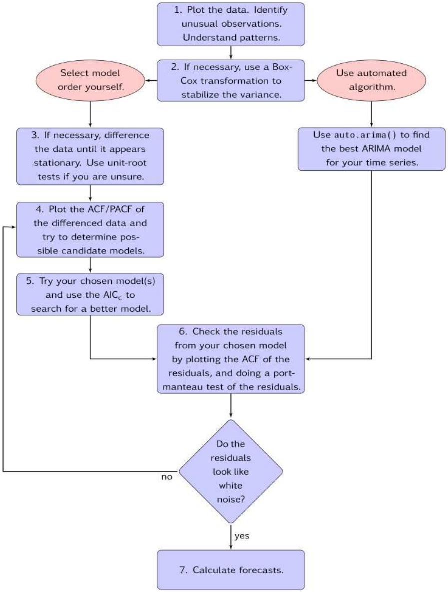
auto.arima

```
auto.arima(x, 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)>100 | frequency(x)>12),
           truncate=NULL, xreg=NULL, test=c("kpss","adf","pp"),
           seasonal.test=c("ocsb","ch"), allowdrift=TRUE,
           allowmean=TRUE, lambda=NULL, parallel=FALSE,
           num.cores=2, ...)
```

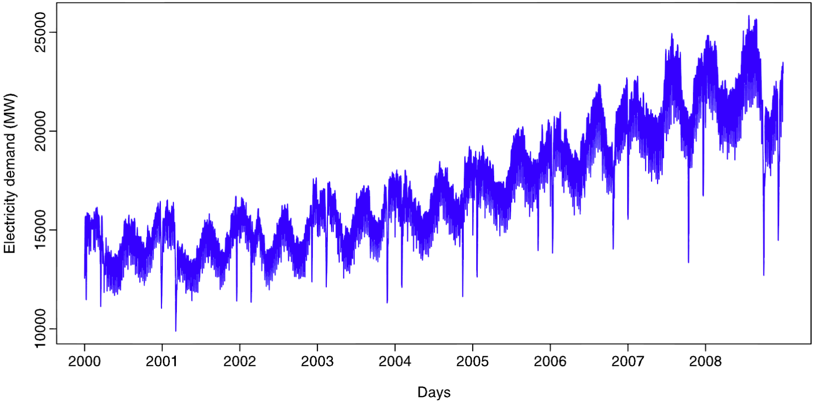
Построить прогноз можно с помощью функции forecast:

```
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, ...)
```

auto.arima


1. Plot the data. Identify unusual observations. Understand patterns.
2. If necessary, use a Box-Cox transformation to stabilize the variance.
Select model order yourself.
Use automated algorithm.
3. If necessary, difference the data until it appears stationary. Use unit-root tests if you are unsure.
Use auto.arima() to find the best ARIMA model for your time series.
4. Plot the ACF/PACF of the differenced data and try to determine possible candidate models.
5. Try your chosen model(s) and use the AICc to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals.
Do the residuals look like white noise?
no
yes
7. Calculate forecasts.

Потребление электричества в Турции



- недельная сезонность
- годовая сезонность
- праздники по исламскому календарю (год примерно на 11 дней короче, чем в грегорианском)

regARIMA

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

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

$$+$$

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

$$=$$

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

Оценка параметров модели

1. Проверить стационарность признаков, если её нет, перейти к разностям. Для лучшей интерпретируемости разностный оператор следует применять и к признакам тоже.
2. Для ряда разностей строится регрессия в предположении, что ошибки описываются моделью начального приближения (как правило, $AR(2)$ или $SARMA(2,0) \times (1,0)_s$).
3. Для остатков регрессии $\hat{z}_t$ подбирается подходящая модель $ARMA\,(p_1, q_1)$.
4. Регрессия перестраивается в предположении, что ошибки описываются моделью $ARMA\,(p_1, q_1)$.
5. Анализируются остатки $\hat{\varepsilon}_t$.

Для подзадачи регрессии формальная проверка значимости признаков неприменима, для отбора признаков необходимо сравнивать значения AIC моделей со всеми подмножествами x_j.

Пример: `https://www.otexts.org/fpp/9/1`

Реализация: параметр xreg в функциях auto.arima и Arima.

Литература

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