

Прикладной статистический анализ данных.  
11. Анализ временных рядов, часть вторая.

Рябенко Евгений  
riabenko.e@gmail.com

21 ноября 2014 г.

## Прогнозирование временного ряда

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

Задача прогнозирования — найти функцию  $f_T$ :

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

где  $d \in \{1, 2, \dots, D\}$  — отсрочка прогноза,  $D$  — горизонт прогнозирования.

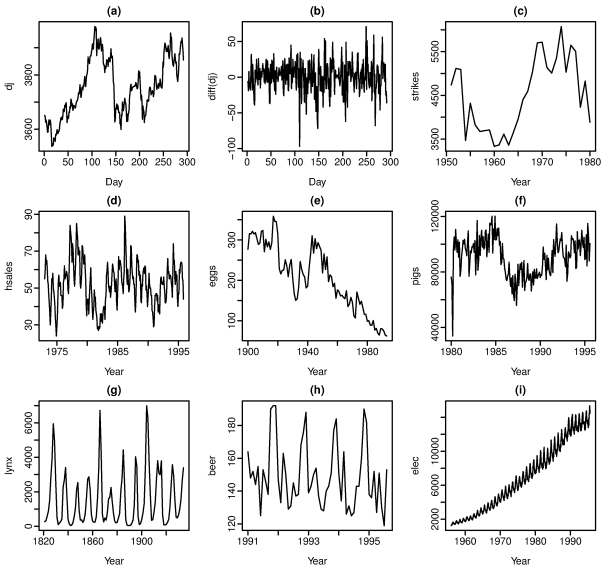
# Стационарность

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 находиться максимумы и минимумы.

# Стационарность



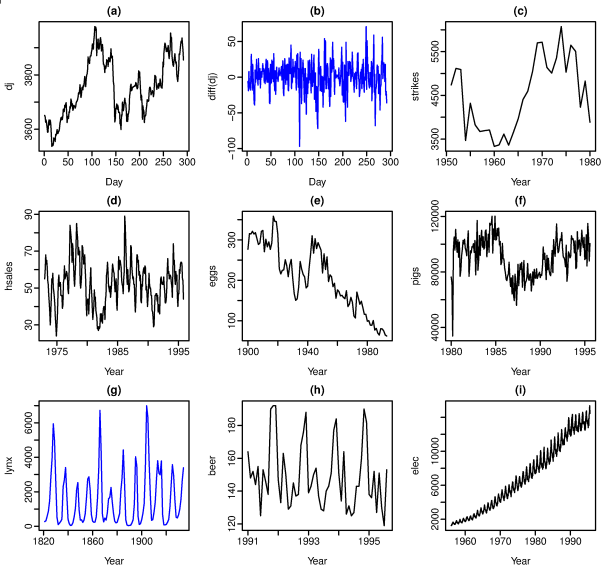






# Стационарность

Стационарны:

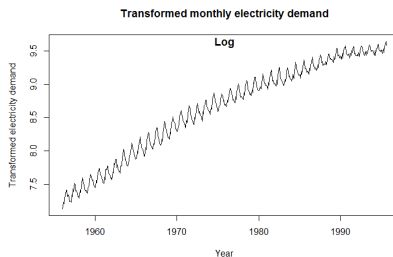
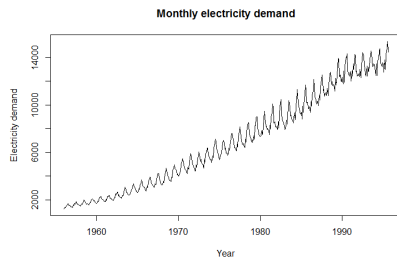




# Стабилизация дисперсии

Для рядов с монотонно меняющейся дисперсией можно использовать стабилизирующие преобразования.

Часто используют логарифмирование:

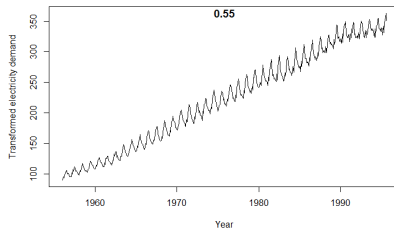
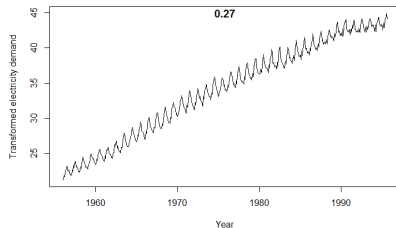


# Преобразования Бокса-Кокса

Параметрическое семейство стабилизирующих дисперсию преобразований:

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

Параметр  $\lambda$  выбирается так, чтобы минимизировать дисперсию или максимизировать правдоподобие модели.



## Преобразования Бокса-Кокса

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

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

- Если некоторые  $y_t \leq 0$ , преобразования Бокса-Кокса невозможны (нужно прибавить к ряду константу).
- Часто оказывается, что преобразование вообще не нужно.
- Стоит округлять значение  $\lambda$ , 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:

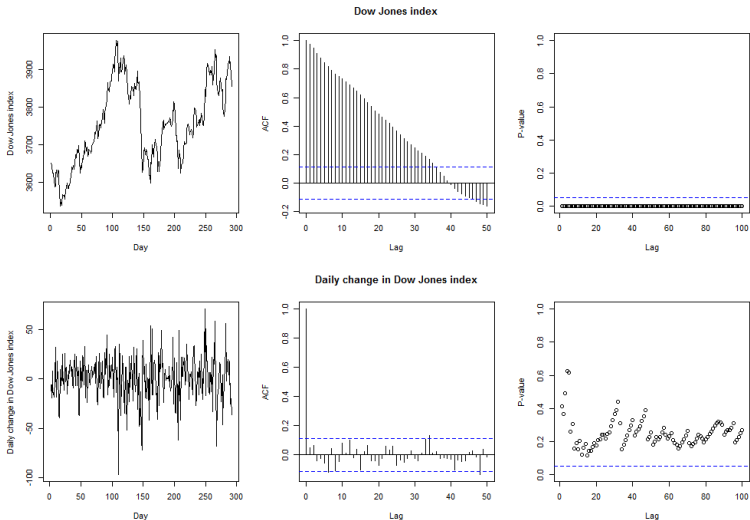
$$y_1, \dots, y_T \longrightarrow y'_2, \dots, y'_T,$$
$$y'_t = y_t - y_{t-1}.$$

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

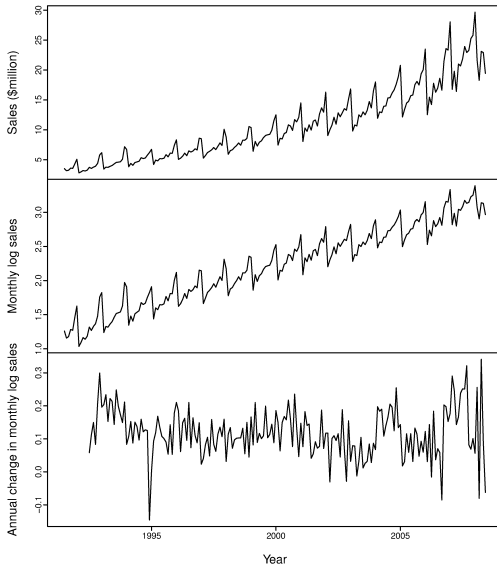
**Сезонное дифференцирование ряда** — переход к попарным разностям его значений в соседних сезонах:

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

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

# Сезонное дифференцирование

Antidiabetic drug sales



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

## Комбинированное дифференцирование

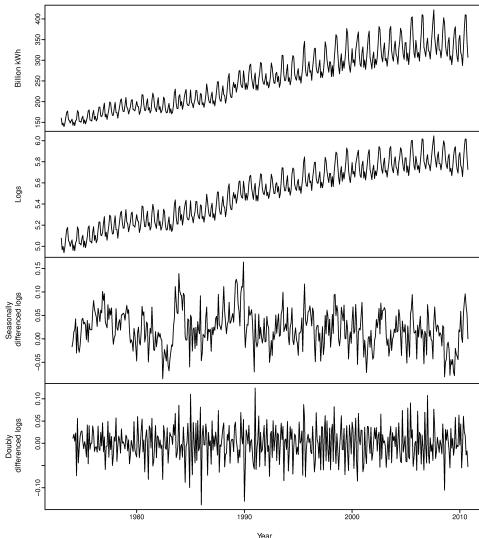
Сезонное и обычное дифференцирование может применяться к одному ряду в любом порядке.

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



# Комбинированное дифференцирование

Monthly US net electricity generation



Критерий

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

## 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 ( $\phi_p \neq 0$ ),  $\varepsilon_t$  — гауссов белый шум с нулевым средним и постоянной дисперсией  $\sigma_\varepsilon^2$ .

Если среднее равно  $\mu$ , 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  $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$ .

Другой способ записи:

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

где  $B$  — разностный оператор ( $By_t = y_{t-1}$ ).

Линейная комбинация  $p$  подряд идущих членов ряда даёт белый шум.

# Авторегрессия

Чтобы ряд  $AR(p)$  был стационарным, должны выполняться ограничения на коэффициенты. Например,

- в  $AR(1)$  необходимо  $-1 < \phi_1 < 1$ ;
- в  $AR(2)$  необходимо  $-1 < \phi_2 < 1$ ,  $\phi_1 + \phi_2 < 1$ ,  $\phi_2 - \phi_1 < 1$ .

С ростом  $p$  вид ограничений усложняется.

## Скольльзящее 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 ( $\theta_q \neq 0$ ),  $\varepsilon_t$  — гауссов белый шум с нулевым средним и постоянной дисперсией  $\sigma_\varepsilon^2$ .

Если среднее равно  $\mu$ , 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:

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

где  $B$  — разностный оператор.

Линейная комбинация  $q$  подряд идущих компонент белого шума  $\varepsilon_t$  даёт элемент ряда.

## Скользящее среднее

Чтобы ряд модель  $MA(q)$  была обратимой, должны выполняться ограничения на коэффициенты. Например,

- в  $MA(1)$  необходимо  $-1 < \theta_1 < 1$ ;
- в  $MA(2)$  необходимо  $-1 < \theta_2 < 1$ ,  $\theta_1 + \theta_2 > -1$ ,  $\theta_1 - \theta_2 < 1$ .

С ростом  $q$  вид ограничений усложняется.

## ARMA (Autogressive moving average)

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

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

Если среднее равно  $\mu$ , 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  $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$ .

Другой способ записи:

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

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

ARIMA (Autogerressive integrated moving average)<sup>1</sup>

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

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

описывается моделью  $ARMA(p, q)$ .

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

---

<sup>1</sup>Также это энергетическое имя, данное творцом первоизданным двум своим посланникам для работы планете Земля, подробности см.  
<http://light-love.ru/nasha-istoriya/ot-avtorov.html/>

# Частичная автокорреляционная функция

**Частичная автокорреляция стационарного ряда  $y_t$ :**

$$\phi_{hh} = \begin{cases} r(y_{t+1}, y_t), & h = 1, \\ r(y_{t+h} - y_{t+h}^{h-1}, y_t - y_t^{h-1}), & h \geq 2, \end{cases}$$

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

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

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



## Оценка параметров модели

- При заданных  $p, d, q$  коэффициенты модели оцениваются методом максимального правдоподобия; функционал качества — логарифм правдоподобия  $LL$ .
- $d$  выбирается так, чтобы ряд был стационарным.
- $p$  и  $q$  нельзя выбирать из принципа максимума правдоподобия:  $LL$  всегда увеличивается с ростом  $p$  и  $q$ .
- При выборе  $p$  и  $q$  помогут автокорреляционные функции ACF и PACF:
  - в модели  $ARIMA(p, d, 0)$  ACF экспоненциально затухает или имеет синусоидальный вид, а PACF значительно отличается от нуля при лаге  $p$ ;
  - в модели  $ARIMA(0, d, q)$  PACF экспоненциально затухает или имеет синусоидальный вид, а ACF значительно отличается от нуля при лаге  $q$ .

## Информационные критерии

$AIC$  — информационный критерий Акаике:

$$AIC = -2LL + 2(p + q + k + 1),$$

где  $k = 1$  при  $c \neq 0$  и  $k = 0$  при  $c = 0$ ;

$AIC_c$  — он же с поправкой на случай небольшого размера выборки:

$$AIC_c = -2LL + \frac{2(p + q + k + 1)(p + q + k + 2)}{T - p - q - k - 2};$$

$BIC$  ( $SIC$ ) — байесовский (Шварца) информационный критерий:

$$BIC = -2LL + (\log T - 2)(p + q + k + 1).$$

## 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 AICс.
- 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 гомоскедастичны, 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 гомоскедастичность не выполняется, предсказательные интервалы генерируются с помощью симуляции.

## Seasonal multiplicative ARMA/ARIMA

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

где

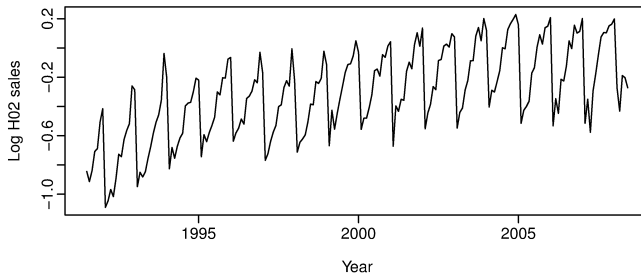
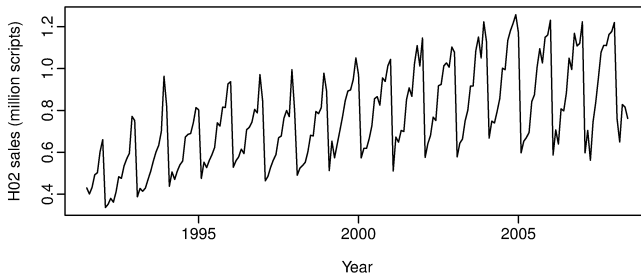
$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps},$$

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \dots + \Theta_Q B^{Ps}.$$

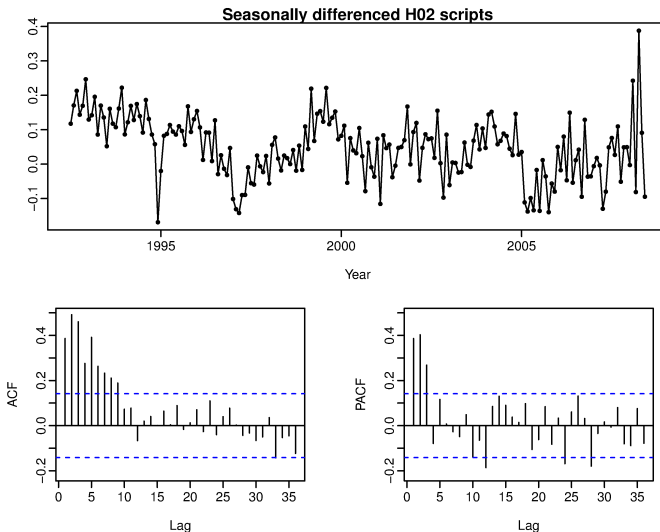
SARIMA:

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

# Объём продаж кортикостероидов в Австралии



# Сезонное дифференцирование



На PACF значимы лаги 1-3, 12, 24. Начнём с модели  $ARIMA(3,0,0)(2,1,0)_{12}$ .

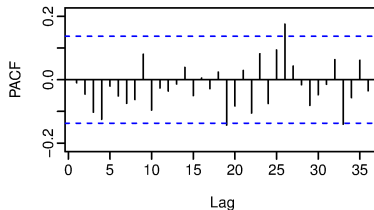
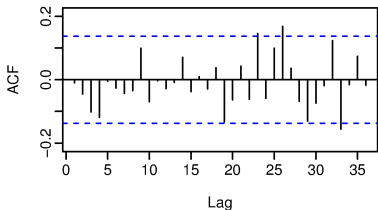
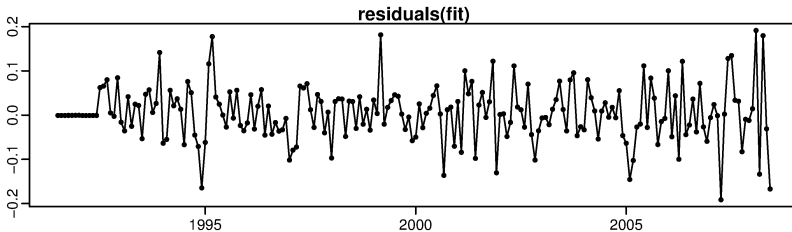
## Сравнение моделей

	<i>AICc</i>
ARIMA(3,0,0)(2,1,0) <sub>12</sub>	-475.12
ARIMA(3,0,1)(2,1,0) <sub>12</sub>	-476.31
ARIMA(3,0,2)(2,1,0) <sub>12</sub>	-474.88
ARIMA(3,0,1)(1,1,0) <sub>12</sub>	-463.40
ARIMA(3,0,1)(0,1,1) <sub>12</sub>	-483.67
ARIMA(3,0,1)(0,1,2) <sub>12</sub>	-485.48
ARIMA(3,0,1)(1,1,1) <sub>12</sub>	-484.25

Лучшая из рассматриваемых моделей — ARIMA(3,0,1)(0,1,2)<sub>12</sub>.



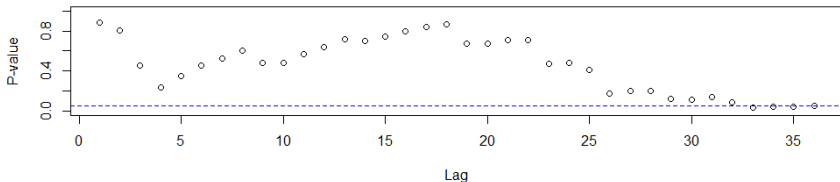
# Остатки ARIMA(3,0,1)(0,1,2)<sub>12</sub>



На ACF и PACF есть значимые лаги.

Остатки ARIMA(3,0,1)(0,1,2)<sub>12</sub>

Критерий Льюнга-Бокса:



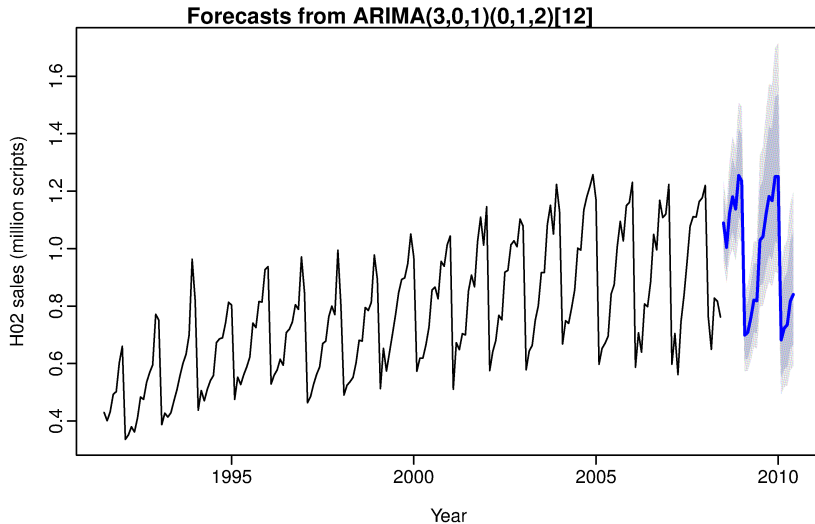
Критерий стационарности KPSS:  $p > 0.1$ .

Критерий нормальности Шапиро-Уилка:  $p = 0.01079$ .

Критерий Уилкоксона:  $p = 0.3143$ .

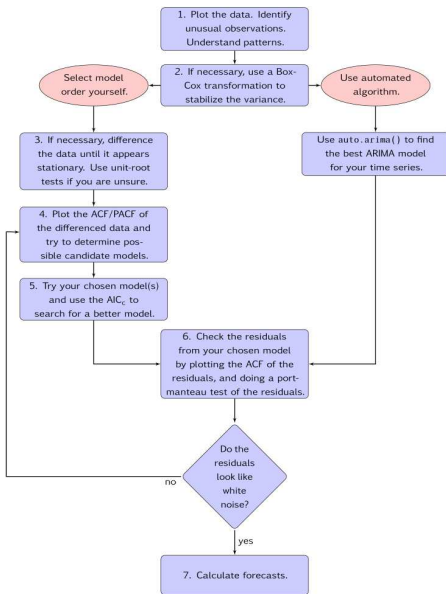
Критерий гомоскедастичности Бройша-Пагана:  $p = 0.001723$ .

# Прогноз ARIMA(3,0,1)(0,1,2)<sub>12</sub>

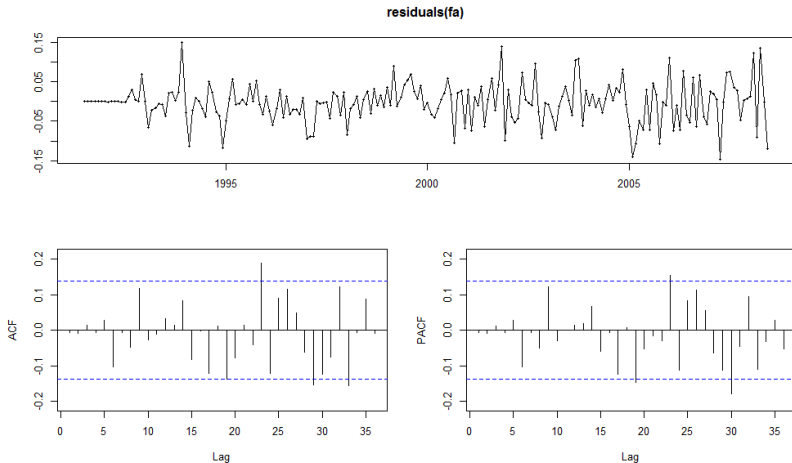


```
auto.arima(x, d=NA, D=NA, max.p=5, max.q=5,  
  max.P=2, max.Q=2, max.order=5, max.d=2, max.D=1,  
  start.p=2, start.q=2, start.P=1, start.Q=1,  
  stationary=FALSE, seasonal=TRUE,  
  ic=c("aicc","aic", "bic"), stepwise=TRUE, trace=FALSE,  
  approximation=(length(x)>100 | frequency(x)>12), xreg=NULL,  
  test=c("kpss","adf","pp"), seasonal.test=c("ocsb","ch"),  
  allowdrift=TRUE, lambda=NULL, parallel=FALSE, num.cores=2)
```

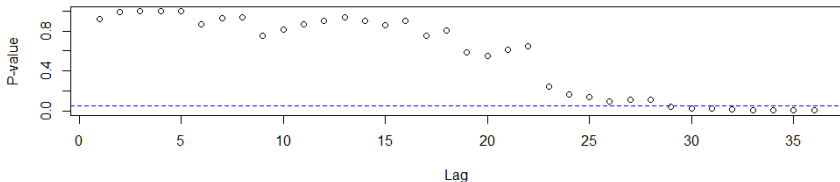
# auto.arima



В автоматическом режиме подбирается модель  $ARIMA(3,1,3)(0,1,1)_{12}$ ; её остатки:



Критерий Льюнга-Бокса:



Критерий стационарности KPSS:  $p > 0.1$ .

Критерий нормальности Шапиро-Уилка:  $p = 0.00176$ .

Критерий Уилкоксона:  $p = 0.3888$ .

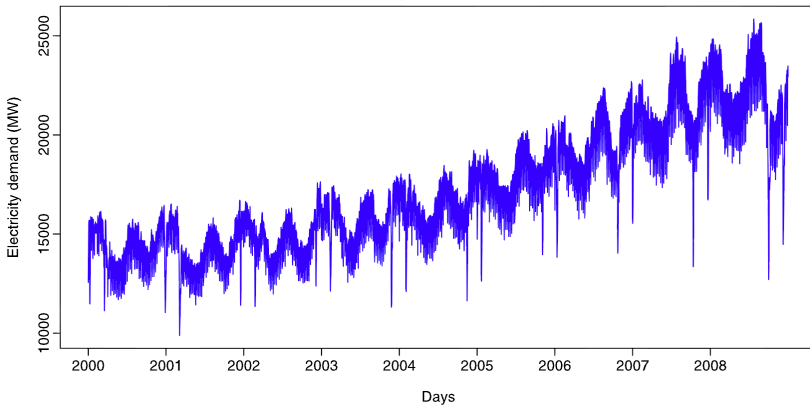
Критерий гомоскедастичности Бройша-Пагана:  $p = 0.0001466$ .

Качество на контрольной выборке:

	<i>RMSE</i>
ARIMA(3,1,3)(0,1,1) <sub>12</sub>	0.0641
ARIMA(3,0,0)(2,1,0) <sub>12</sub>	0.0661
ARIMA(3,0,1)(2,1,0) <sub>12</sub>	0.0646
ARIMA(3,0,2)(2,1,0) <sub>12</sub>	0.0645
ARIMA(3,0,1)(1,1,0) <sub>12</sub>	0.0679
ARIMA(3,0,1)(0,1,1) <sub>12</sub>	0.0644
ARIMA(3,0,1)(0,1,2) <sub>12</sub>	0.0622
ARIMA(3,0,1)(1,1,1) <sub>12</sub>	0.0630
ARIMA(4,0,3)(0,1,1) <sub>12</sub>	0.0648
ARIMA(3,0,3)(0,1,1) <sub>12</sub>	0.0640
ARIMA(4,0,2)(0,1,1) <sub>12</sub>	0.0648
ARIMA(3,0,2)(0,1,1) <sub>12</sub>	0.0644
ARIMA(2,1,3)(0,1,1) <sub>12</sub>	0.0634
ARIMA(2,1,4)(0,1,1) <sub>12</sub>	0.0632
ARIMA(2,1,5)(0,1,1) <sub>12</sub>	0.0640



## 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 удобно моделировать с помощью regARIMA:

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

+

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

=

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

## 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  $SARMA(2, 0, 0) \times (1, 0)_s$ ).
- 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>

## Реализации

- US Census Bureau: X-12-ARIMA, X-13-ARIMA-SEATS (<http://www.census.gov/srd/www/x13as/>, доступен через иностранные прокси-серверы);
- Matlab: regARIMA (2013b);
- R: параметр xreg в функциях auto.arima и Arima.

## 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);
- настройка модели экспоненциального сглаживания: автоматический подбор модели, проверка её соответствия особенностям ряда, корректировка, анализ остатков;
- визуальный анализ и 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, 2013. <https://www.otexts.org/book/fpp>

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D. *Forecasting with Exponential Smoothing: The State Space Approach*. — Berlin: Springer, 2008.