

Прикладная статистика 11. Анализ временных рядов.

15 ноября 2013 г.

Прогнозирование временного ряда

Временной ряд: x_1, \dots, x_T, \dots , $x_t \in \mathbb{R}$, — значения признака, измеренные через постоянные временные интервалы.

Задача прогнозирования: найти функцию f_T :

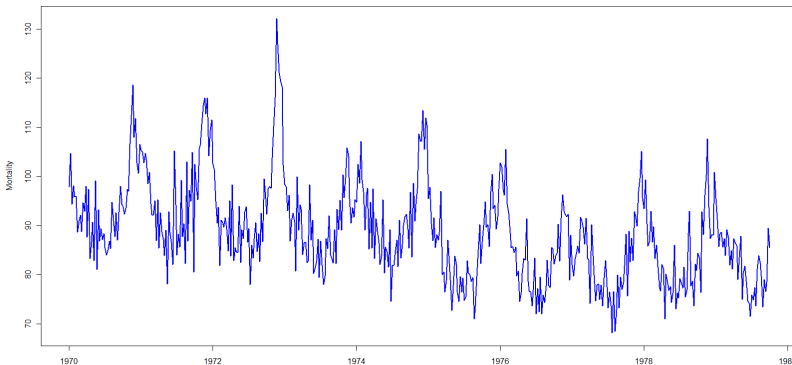
$$x_{T+d} \approx \hat{x}_{T+d} = f_T(x_T, \dots, x_1, d),$$

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

Функция выбирается следующим образом:

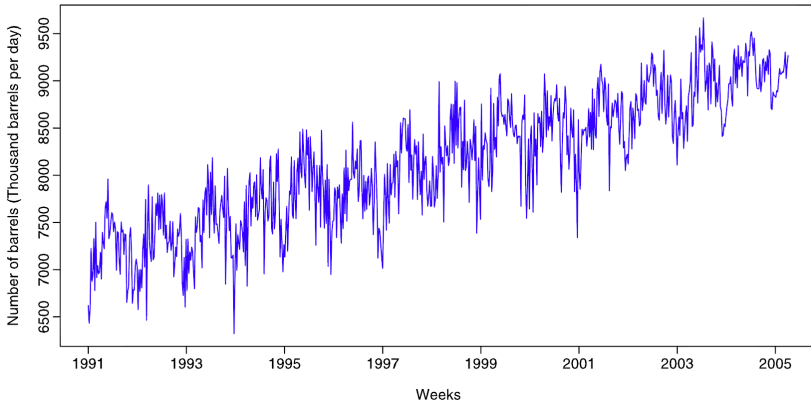
$$Q_T = \sum_{t=1}^T (x_t - \hat{x}_t)^2 \rightarrow \min_{f_T}.$$

Смертность от сердечно-сосудистых заболеваний



- есть годичный профиль — сезонность (годовая);
- есть линейное убывание — тренд.

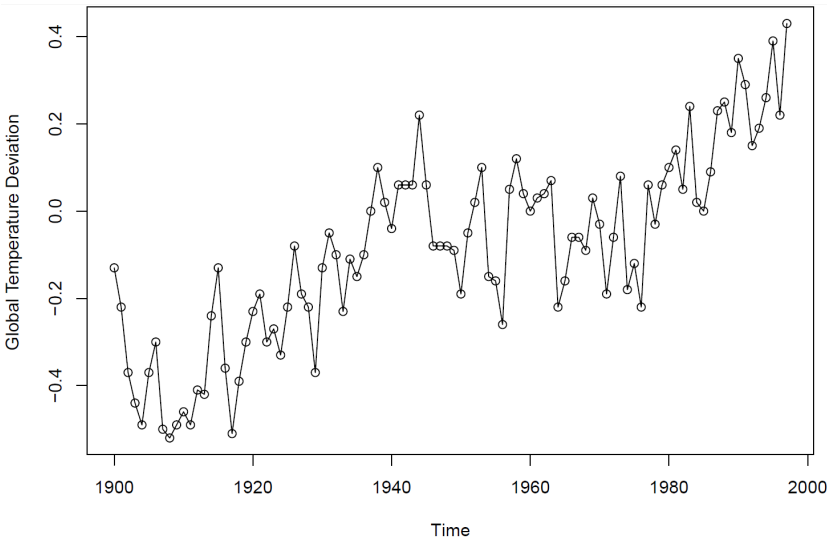
Продажи нефтепродуктов в США



- годовая сезонность;
- повышающийся линейный тренд.

Динамика среднегодовой температуры

Отклонение от среднегодовой температуры в градусах Цельсия:



Линейный тренд: регрессия

Построим зависимость отклонения температуры от года:

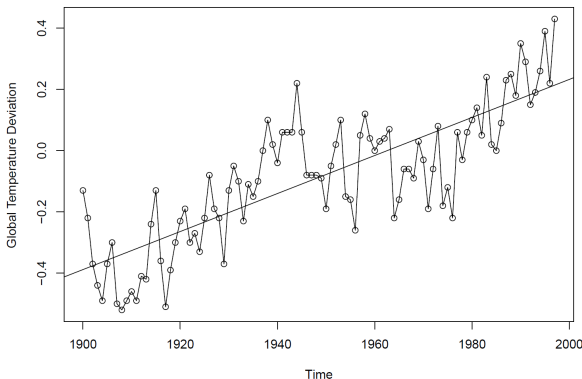
$$x_t = \theta_0 + \theta_1 t + \varepsilon, \quad t = 1900, \dots, 2000.$$

$$\theta_0 = -12.186, \quad \theta_1 = 0.006;$$

$$SE(\theta_0) = 0.9, \quad SE(\theta_1) = 0.005;$$

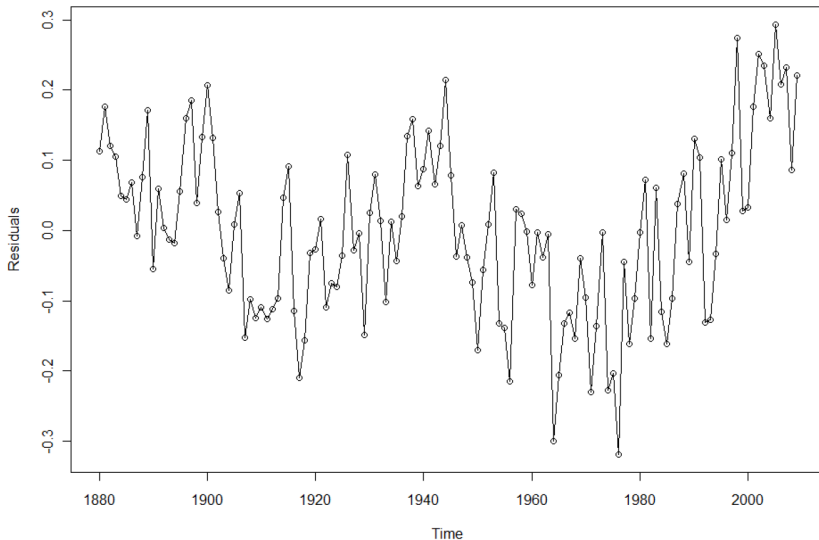
$$R^2 = 0.651, \quad R_a^2 = 0.648;$$

$$F = 179.5, \quad p = 2 \times 10^{-16}.$$



Остатки

$$\hat{\varepsilon}_t = \hat{x}_t - x_t, t = 1, \dots, T :$$



Анализ остатков

Требуемые свойства остатков и методы их проверки:

- нормальность (улучшает свойства МНК-оценки, определяет выбор критериев для проверки других гипотез) — критерий Шапиро-Уилка;
- несмещённость — критерии Стьюдента и Уилкоксона;
- гомоскедастичность — критерий Бройша-Пагана;
- неавтокоррелированность (отсутствие неучтённой линейной зависимости от предыдущих наблюдений) — коррелограмма, Q-критерий Льюнга-Бокса (группа лагов);
- стационарность (отсутствие зависимости от времени) — критерий KPSS.

Автокорреляция

Автокорреляционная функция:

$$r_\tau = \text{corr}(x_t, x_{t-\tau}) = \frac{\sum_{t=\tau+1}^T (x_t - \bar{x}^1)(x_{t-\tau} - \bar{x}^2)}{\sqrt{\sum_{t=\tau+1}^T (x_t - \bar{x}^1)^2 \sum_{t=\tau+1}^T (x_{t-\tau} - \bar{x}^2)^2}},$$

$$\bar{x}^1 = \frac{1}{T-\tau} \sum_{t=\tau+1}^T x_t,$$

$$\bar{x}^2 = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} x_t.$$

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

Проверка значимости отличия автокорреляции от нуля:

временной ряд: $X^T = X_1, \dots, X_T$;

нулевая гипотеза: $H_0: r_\tau = 0$;

альтернатива: $H_1: r_\tau \neq 0$;

статистика: $T(X^T) = \frac{r_\tau \sqrt{T-\tau-2}}{\sqrt{1-r_\tau^2}}$;

$T(X^T) \sim St(T-\tau-2)$ при H_0 .

Q-критерий Льюнга-Бокса

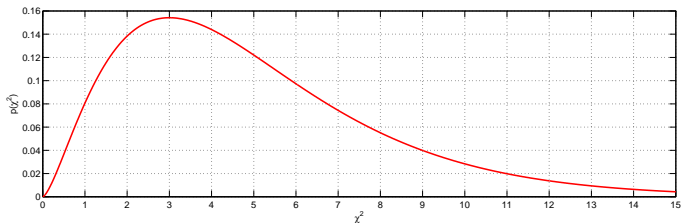
ряд ошибок прогноза: $\varepsilon^T = \varepsilon_1, \dots, \varepsilon_T$;

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

альтернатива: $H_1: H_0$ неверна;

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

$Q(\varepsilon^T) \sim \chi_L^2$ при H_0 ;

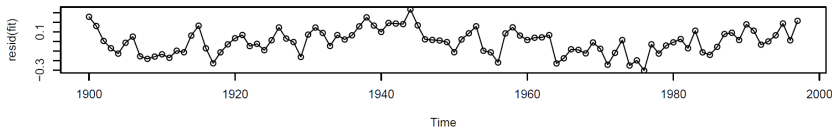


достигаемый уровень значимости:

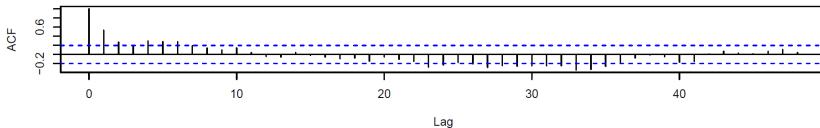
$$p(q) = 1 - \text{chi2cdf}(q, L).$$

В рассматриваемой задаче

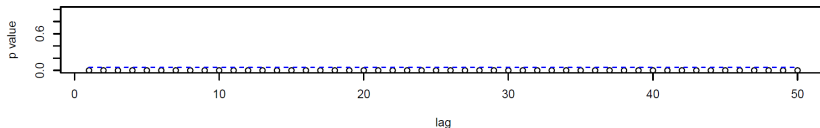
detrended



Series resid(fit)



p values for Ljung-Box statistics



Критерий KPSS (Kwiatkowski-Philips-Schmidt-Shin)

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

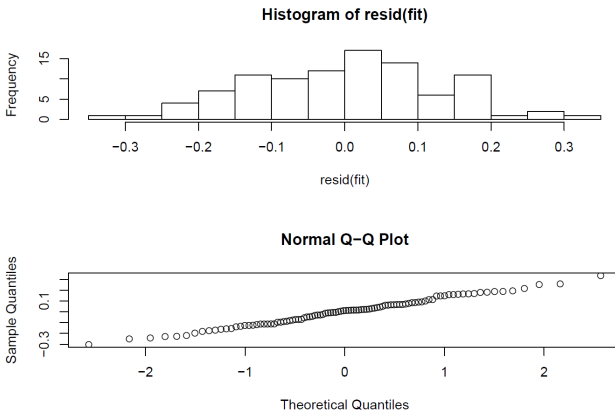
нулевая гипотеза: H_0 : ряд ε^T стационарен;

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

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

$KPSS(\varepsilon^T)$ при H_0 имеет табличное распределение.

В рассматриваемой задаче



Критерий нормальности Шапиро-Уилка: $p = 0.8618$.

Критерий Стьюдента: $p \approx 1$.

Критерий гомоскедастичности Бройша-Пагана: $p = 0.9335$.

Критерий стационарности KPSS: $p > 0.1$.

Авторегрессия

$$AR(p): \quad x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + \varepsilon_t,$$

где x_t — стационарный ряд с нулевым средним, ϕ_1, \dots, ϕ_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

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + \varepsilon_t,$$

где $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$.

Другой способ записи:

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

где B — разностный оператор ($Bx_t = x_{t-1}$).

Линейная комбинация p подряд идущих членов ряда даёт белый шум.

Скользящее среднее

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

где x_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

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

Другой способ записи:

$$x_t = \theta(B)\varepsilon_t = (1 + \theta_1B + \theta_2B^2 + \cdots + \theta_qB^q)\varepsilon_t,$$

где B — разностный оператор.

Линейная комбинация q подряд идущих компонент белого шума ε_t даёт элемент ряда.

Автокорреляции

В моделях $MA(q)$ автокорреляция ряда равна нулю при лаге, большем q , и строго больше нуля при лаге q .

Частичная автокорреляция стационарного ряда x_t :

$$\phi_{hh} = \begin{cases} \text{corr}(x_{t+1}, x_t), & h = 1, \\ \text{corr}(x_{t+h} - x_{t+h}^{h-1}, x_t - x_t^{h-1}), & h \geq 2, \end{cases}$$

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

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

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

В моделях $AR(p)$ частичная автокорреляция ряда равна нулю при лаге, большем p , и строго больше нуля при лаге p .

ARMA (Autogressive moving average)

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

где x_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 вреднее равно μ , модель принимает вид

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q},$$

где $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$.

Другой способ записи:

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

ARIMA (Autogerressive integrated moving average)¹

Для нестационарного ряда стационарным может оказаться ряд его разностей.

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

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

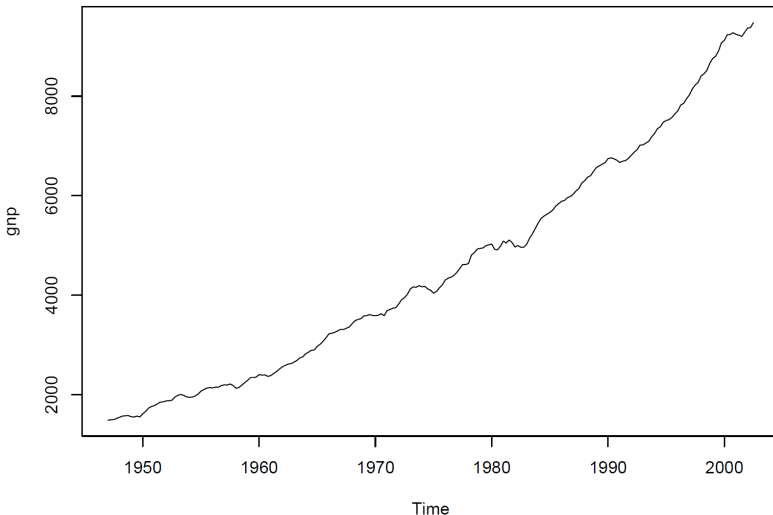
описывается моделью $ARMA(p, q)$.

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

¹ 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/>

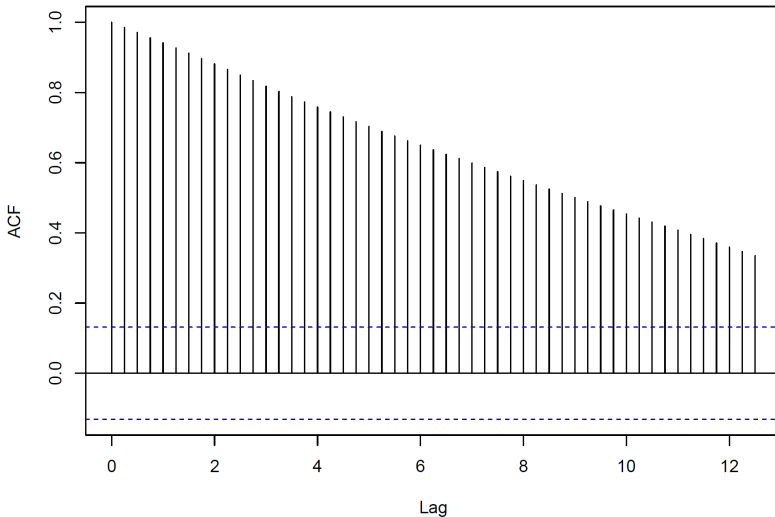
ВВП США

Поквартальные очищенные от сезонности данные о ВВП США
в миллиардах долларов 1996 года:

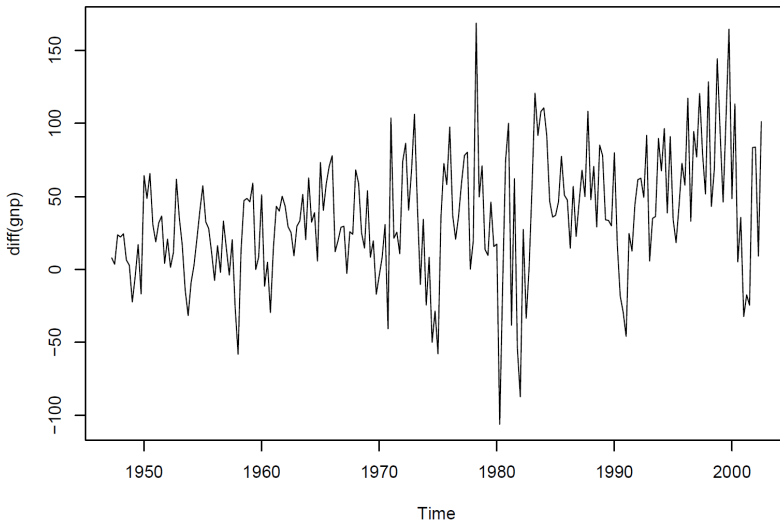


Автокорреляция

Series gnp

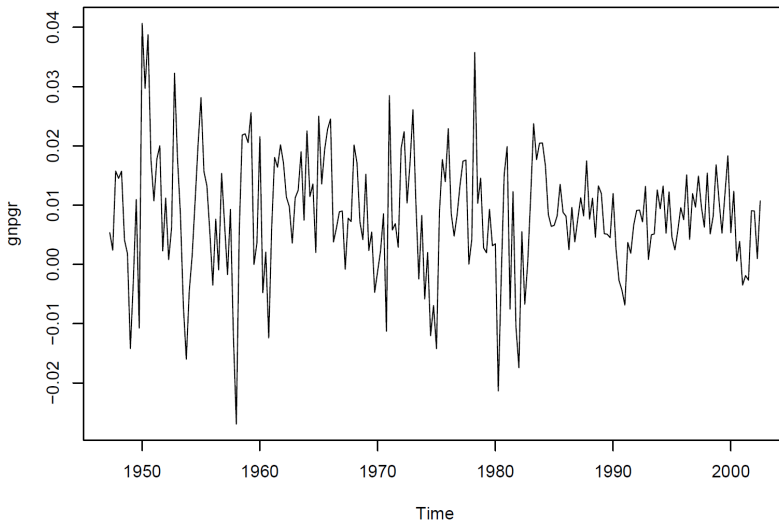


Ряд первых разностей



Нестационарен, вариация данных выше во второй половине ряда ($KPSS$ $p < 0.01$).

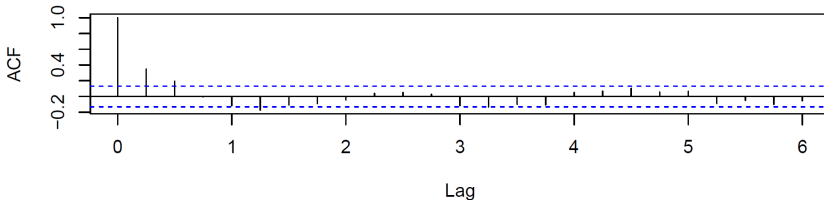
Ряд первых разностей логарифмов



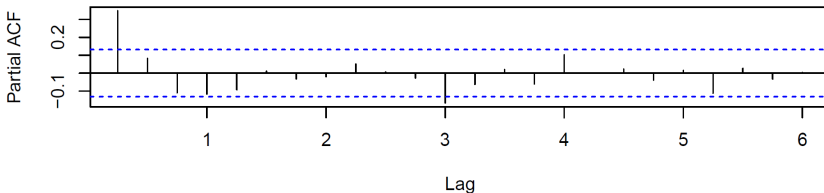
Стационарен ($KPSS\ p > 0.1$), интерпретируется как прирост ВВП в процентах.

Автокорреляция и частичная автокорреляция ряда приростов

Series gnpggr



Series gnpggr



Модели

Варианты интерпретации графиков:

- AC равна нулю после лага 2, PAC убывает — модель $MA(2)$;
- PAC равна нулю после лага 1, AC убывает — модель $AR(1)$;
- модель $ARMA(1, 2)$.

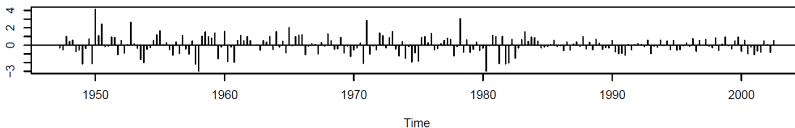
$$AR(1): x_t = 0.005 + 0.347x_{t-1} + \varepsilon_t, \hat{\sigma}_\varepsilon = 0.0095;$$

$$MA(2): x_t = 0.008 + 0.303\varepsilon_{t-1} + 0.204\varepsilon_{t-2} + \varepsilon_t, \hat{\sigma}_\varepsilon = 0.0094;$$

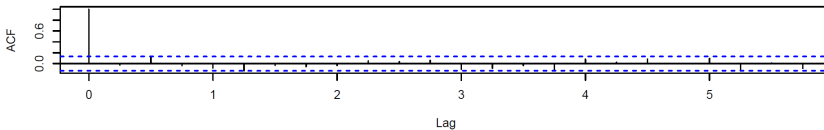
$$ARMA(1, 2): x_t = 0.008 + 0.241x_{t-1} + 0.076\varepsilon_{t-1} + 0.162\varepsilon_{t-2} + \varepsilon_t, \hat{\sigma}_\varepsilon = 0.0089.$$

Диагностика $AR(1)$

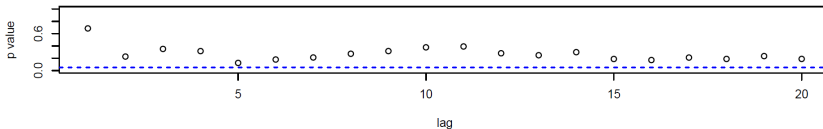
Standardized Residuals

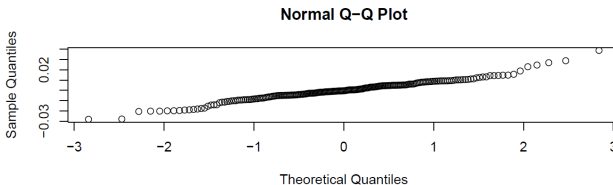
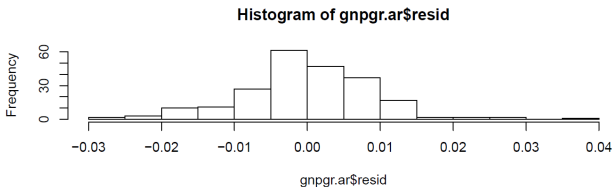


ACF of Residuals



p values for Ljung-Box statistic



Диагностика $AR(1)$ 

Критерий нормальности Шапиро-Уилка: $p = 0.0006886$.

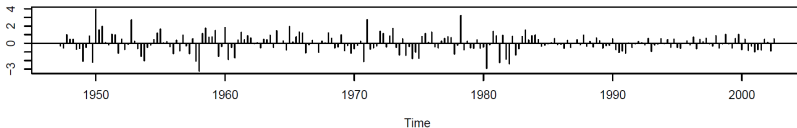
Критерий Уилкоксона: $p = 0.8665$.

Критерий гомоскедастичности Бройша-Пагана: $p = 0.5613$.

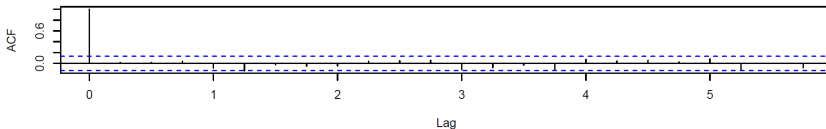
Критерий стационарности KPSS: $p > 0.1$.

Диагностика $MA(2)$

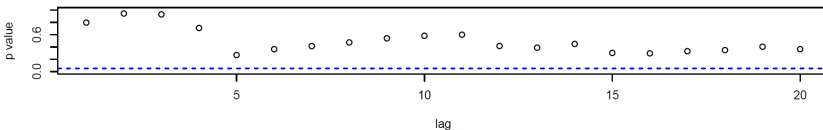
Standardized Residuals

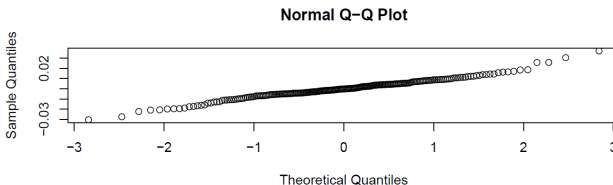
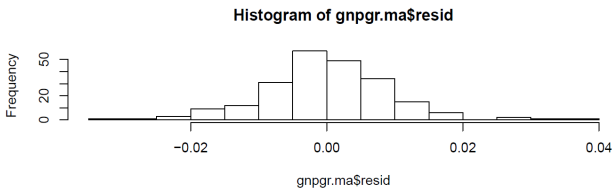


ACF of Residuals



p values for Ljung-Box statistic



Диагностика *MA(2)*

Критерий нормальности Шапиро-Уилка: $p = 0.003416$.

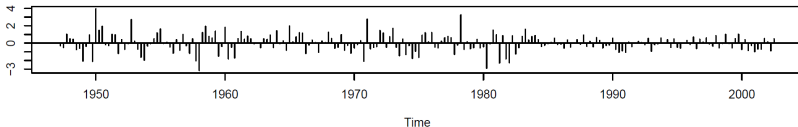
Критерий Уилкоксона: $p = 0.9917$.

Критерий гомоскедастичности Бройша-Пагана: $p = 0.4947$.

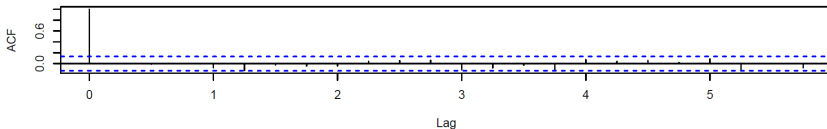
Критерий стационарности KPSS: $p > 0.1$.

Диагностика $ARMA(1,2)$

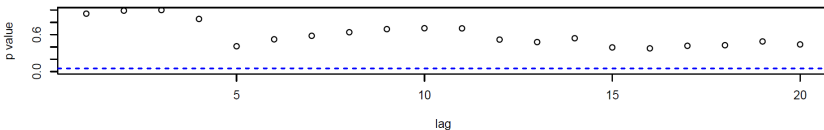
Standardized Residuals

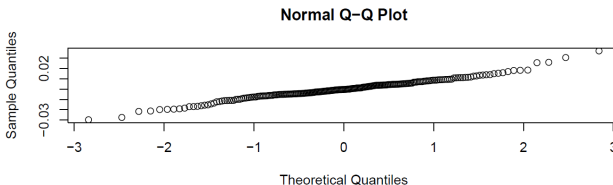
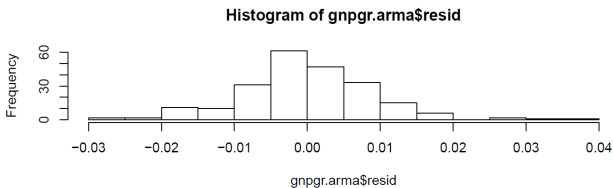


ACF of Residuals



p values for Ljung-Box statistic



Диагностика $ARMA(1,2)$ 

Критерий нормальности Шапиро-Уилка: $p = 0.003497$.

Критерий Уилкоксона: $p = 0.9817$.

Критерий гомоскедастичности Бройша-Пагана: $p = 0.5165$.

Критерий стационарности KPSS: $p > 0.1$.

Сравнение моделей

AIC — информационный критерий Акаике:

$$AIC = \log \left(\frac{1}{T} \sum_{t=1}^T (x_t - \hat{x}_t)^2 \right) + \frac{T + 2k}{T},$$

где k — число параметров модели;

AIC — он же с поправкой на случай небольшого размера выборки:

$$AIC = \log \left(\frac{1}{T} \sum_{t=1}^T (x_t - \hat{x}_t)^2 \right) + \frac{T + k}{T - k - 2};$$

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

$$BIC = \log \left(\frac{1}{T} \sum_{t=1}^T (x_t - \hat{x}_t)^2 \right) + \frac{k \log T}{T}.$$

	AIC	AIC_c	BIC
$AR(1)$	-1431.22	-8.284	-9.264
$MA(2)$	-1431.93	-8.297	-9.276
$ARMA(1, 2)$	-1430.95	-8.301	-9.281

Виды сезонных эффектов

ESS Guidelines on Seasonal Adjustment

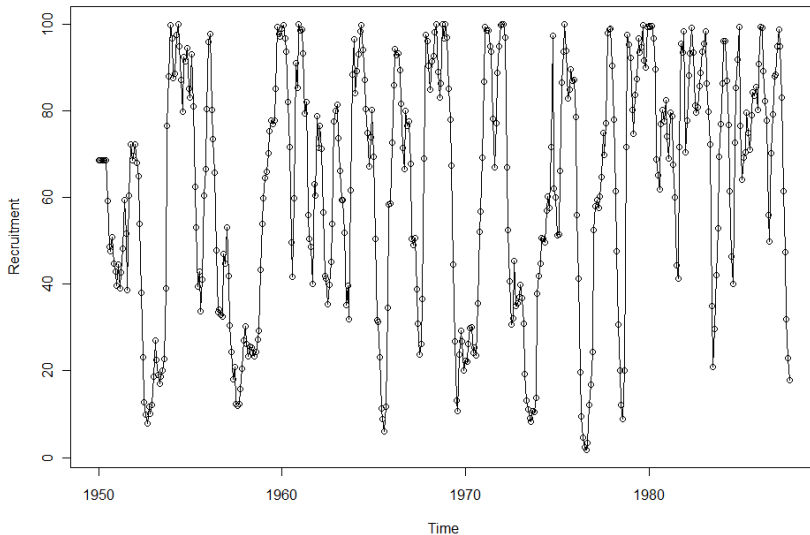
Сезонность — циклические изменения уровня ряда внутри повторяющегося периода, достаточно устойчивые между периодами.

Причины возникновения сезонности:

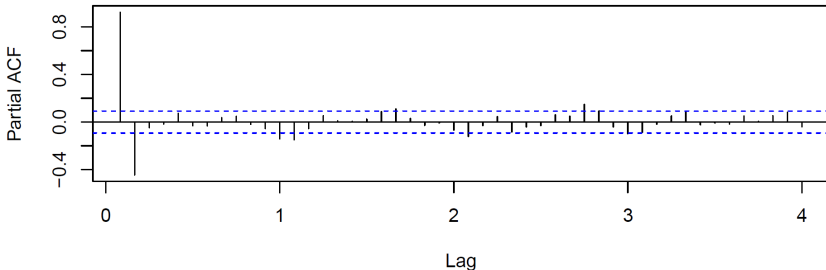
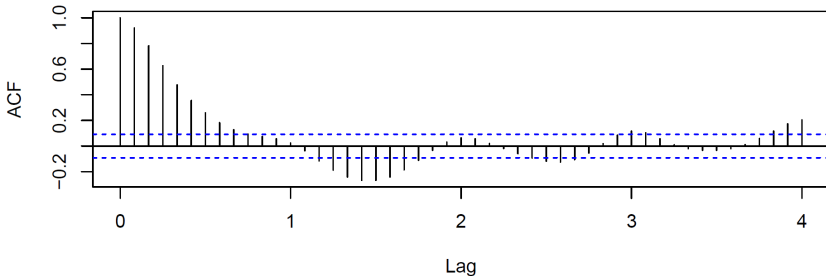
- природные факторы;
- административные и юридические факторы;
- календарные эффекты: число рабочих дней, эффекты фиксированных праздников.

Размер популяции рыб

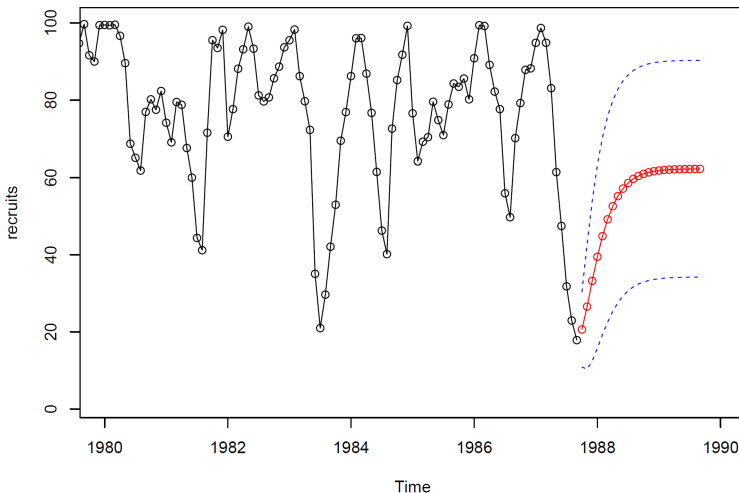
Recruitment — число новых особей рыбы — за каждый месяц 1950-1987:



Корреляции



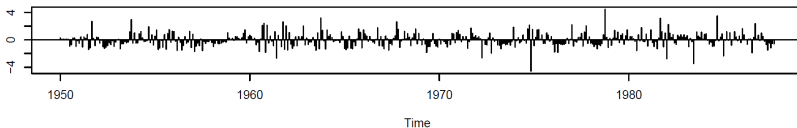
Прогноз

Выбор — модель $AR(2)$.

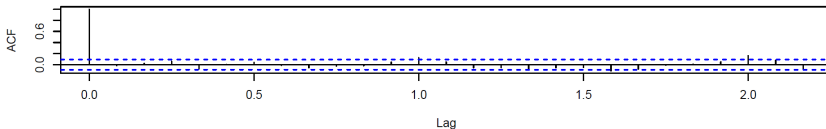
В моделях $ARMA(p, q)$ с увеличением горизонта прогноз всё больше похож на константу.

Диагностика $AR(2)$

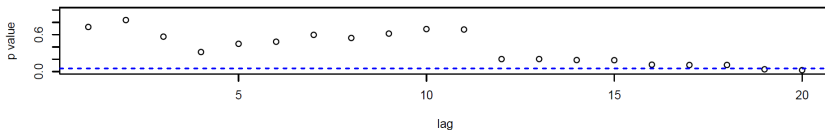
Standardized Residuals

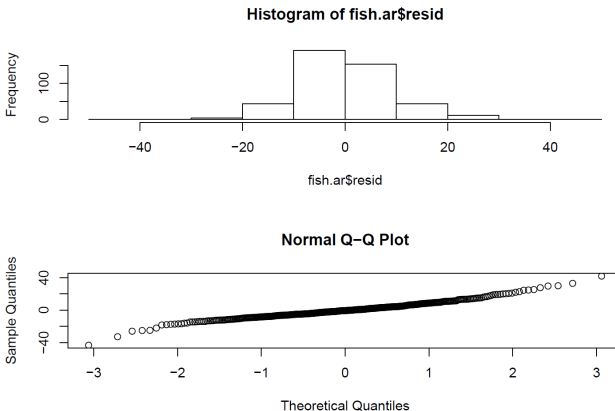


ACF of Residuals



p values for Ljung-Box statistic



Диагностика $AR(2)$ 

Критерий нормальности Шапиро-Уилка: $p = 2.7 \times 10^{-7}$.

Критерий Уилкоксона: $p = 0.4167$.

Критерий гомоскедастичности Бройша-Пагана: $p = 0.3783$.

Критерий стационарности KPSS: $p > 0.1$.

Seasonal multiplicative ARMA/ARIMA

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

где

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

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

SARIMA:

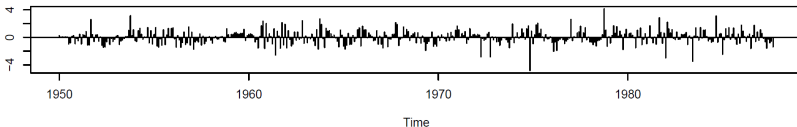
$$\Phi_P(B^s) \phi(B) \nabla_s^D \nabla^d x_t = \alpha + \Theta_Q(B^s) \theta(B) \varepsilon_t.$$

Сравнение моделей

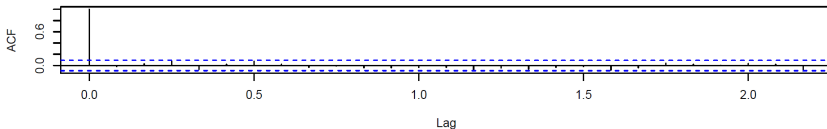
	<i>AIC</i>
$ARMA(2, 0) \times (3, 0)_{12}$	3308.515
$ARMA(2, 0) \times (2, 0)_{12}$	3316.283
$ARMA(2, 0) \times (1, 0)_{12}$	3325.706
$ARMA(2, 0) \times (0, 1)_{12}$	3327.352
$ARMA(2, 0) \times (0, 2)_{12}$	3321.880
$ARMA(2, 0) \times (0, 3)_{12}$	3314.787
$ARMA(2, 0) \times (1, 1)_{12}$	3283.717

Диагностика $ARMA(2,0) \times (1,1)_{12}$

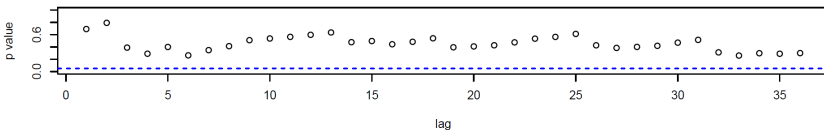
Standardized Residuals

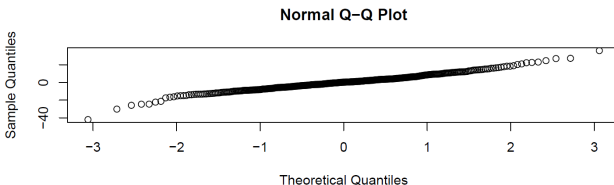
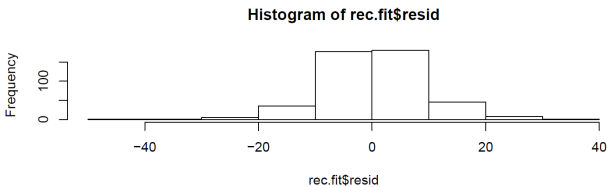


ACF of Residuals



p values for Ljung-Box statistic



Диагностика $ARMA(2,0) \times (1,1)_{12}$ 

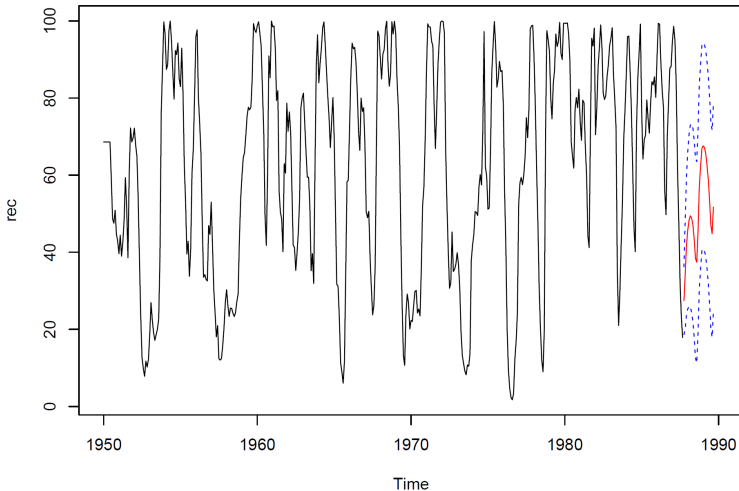
Критерий нормальности Шапиро-Уилка: $p = 6.8 \times 10^{-6}$.

Критерий Уилкоксона: $p = 0.7387$.

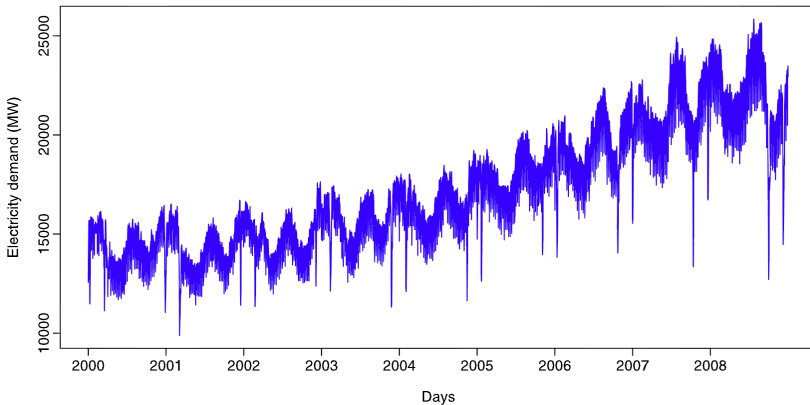
Критерий гомоскедастичности Бройша-Пагана: $p = 0.2546$.

Критерий стационарности KPSS: $p > 0.1$.

Прогноз



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 $\hat{\varepsilon}_t$.

Источник: Hyndman, Athanasopoulos. Forecasting: principles and practice.
<https://www.otexts.org/book/fpp>

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: пакет X12.

Прикладная статистика
11. Анализ временных рядов.

Рябенко Евгений
riabenko.e@gmail.com