Прикладная статистика 11. Анализ временных рядов.

29 апреля 2013 г.

Прогнозирование временного ряда

Временной ряд: $x_1, \dots, x_T, \dots, \quad x_t \in \mathbb{R}$

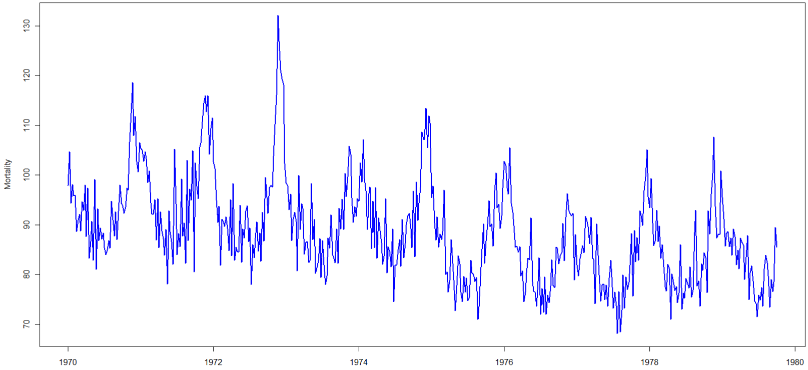
Задача прогнозирования: найти функцию f_T:

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

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

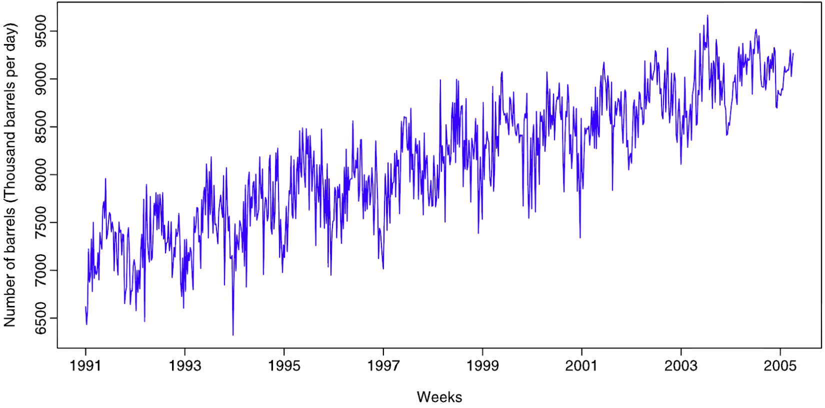
$$Q_T = \sum_{t=1}^{T} \left(x_t - \hat{x}_t\right)^2 \to \min_{f_T}.$$

Смертность от сердечно-сосудистых заболеваний



- есть годичный «профиль» — сезонность (годовая)
- есть линейное убывание — тренд

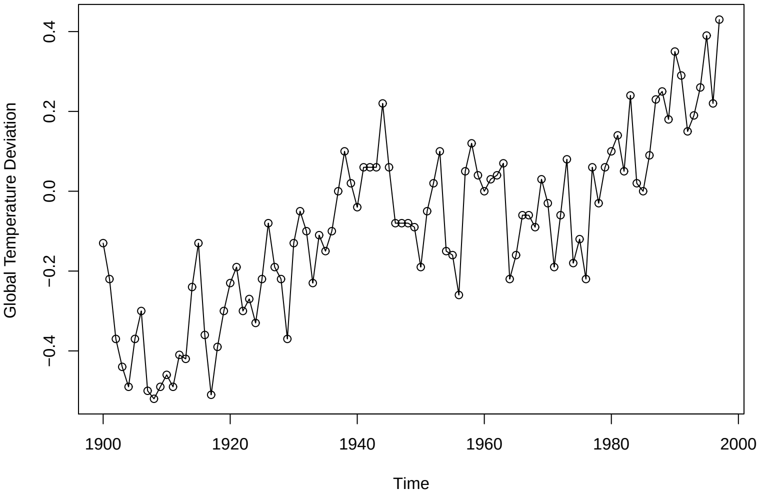
Продажи нефтепродуктов в США



- годовая сезонность
- повышающийся линейный тренд

Исходные данные

Отклонение от среднегодовой температуры в градусах Цельсия

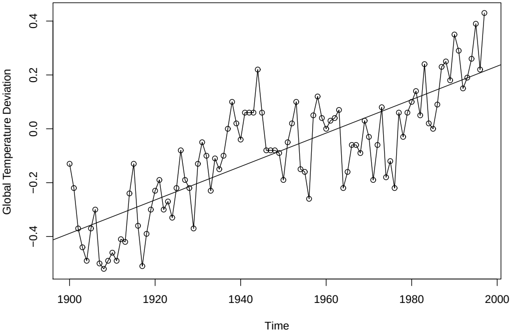


Линейный тренд: регрессия

Построим зависимость отклонения температуры от года:

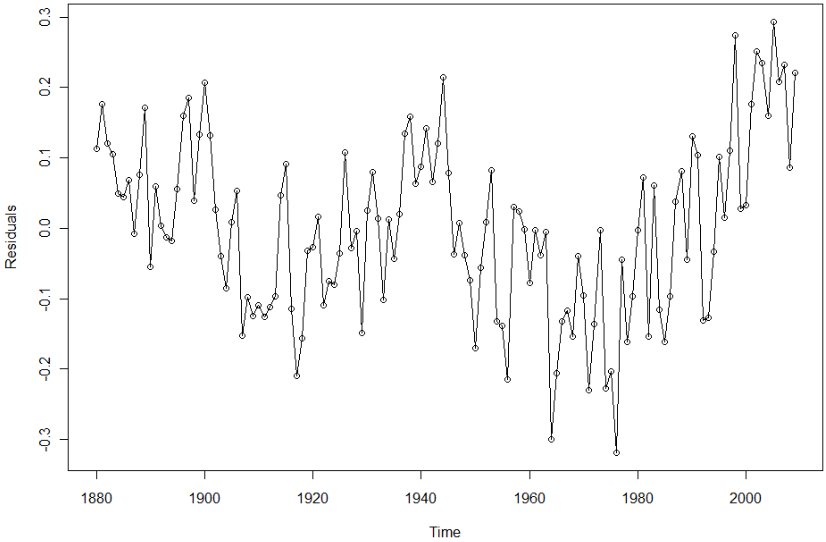
$$x_t = \beta_1 + \beta_2 t + \varepsilon_t, \quad t = 1900, \dots, 2000.$$

$\beta_1 = -12.186, \quad \beta_2 = 0.006;$
$SE(\beta_1) = 0.9, \quad SE(\beta_2) = 0.005;$
$R^2 = 0.6515, \quad R^2_A = 0.6479;$
$F = 179.5, \quad p < 2.2 \times 10^{-16}.$



Остатки

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



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, критерий Дарбина-Уотсона (автокорреляция с лагом 1), Q-критерий Льюнга-Бокса (группа лагов);
- стационарность (отсутствие зависимости от времени) — критерий KPSS.

Автокорреляция

Автокорреляционная функция:

$$r_\tau = \frac{\sum\limits_{t=\tau+1}^{T}\left(x_t - \bar{x}_{(1)}\right)\left(x_{t-\tau} - \bar{x}_{(2)}\right)}{\sqrt{\sum\limits_{t=\tau+1}^{T}\left(x_t - \bar{x}_{(1)}\right)\sum\limits_{t=\tau+1}^{T}\left(x_{t-\tau} - \bar{x}_{(2)}\right)}},$$

$$\bar{x}_{(1)} = \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} x_t, \quad \bar{x}_{(2)} = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau} x_t.$$

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

Проверка значимости отличия автокорреляции от нуля:

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

нулевая гипотеза: $H_0\colon r_\tau = 0$;

альтернатива: $H_1\colon r_\tau <\neq> 0$;

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

$T\left(X^T\right) \sim St(T-\tau-2)$ при H_0.

Критерий Дарбина-Уотсона

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

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

альтернатива: $H_1\colon r_1 \neq 0$;

статистика: $d\left(\varepsilon^T\right) = \frac{\sum\limits_{t=2}^{T}\left(\varepsilon_t - \varepsilon_{t-1}\right)^2}{\sum\limits_{t=1}^{T}\varepsilon_t^2}$;

$$d\left(\varepsilon^T\right) \in \begin{cases} (-\infty, d_l] & \text{— значимая положительная автокорреляция;} \\ [d_u, 4 - d_u] & \text{— незначимая автокорреляция;} \\ [4 - d_l, \infty) & \text{— значимая отрицательная автокорреляция.} \end{cases}$$

Значения d_l, d_u табулированы и зависят от уровня значимости и числа переменных в модели ряда.

Достигаемый уровень значимости — наименьший уровень значимости, при котором $d = d_l$ или $d = 4 - d_l$.

$$d\left(\varepsilon^T\right) \approx 2\left(1 - r_1\right).$$

Q-критерий Льюнга-Бокса

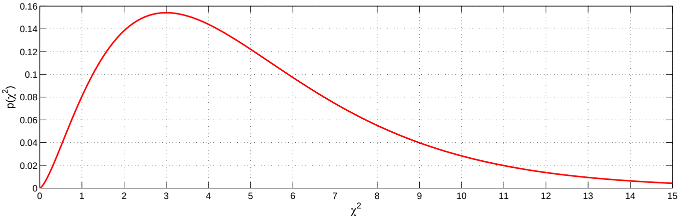
ряд ошибок прогноза: $\varepsilon^T = \varepsilon_1, \dots, \varepsilon_T$;

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

альтернатива: $H_1\colon H_0$ неверна;

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

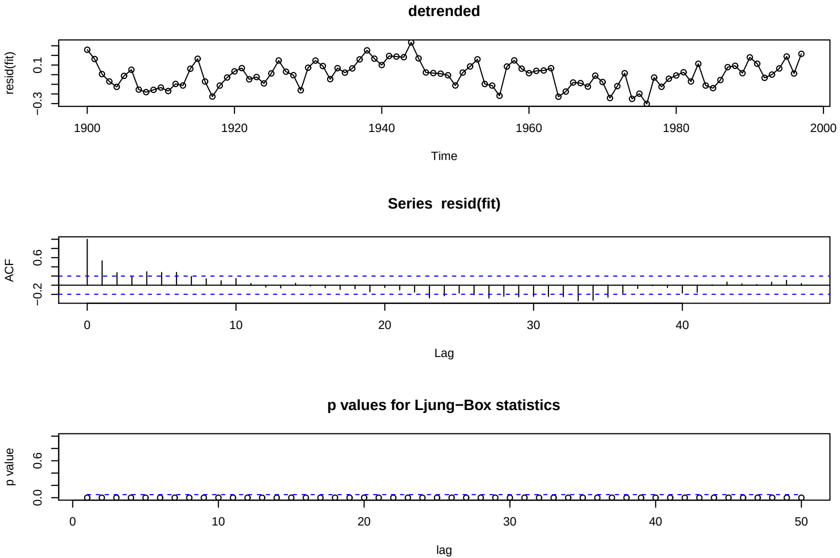
$Q\left(\varepsilon^T\right) \sim \chi_L^2$ при H_0;



достигаемый уровень значимости:

$$p\left(\varepsilon^T\right) = 1 - chi2cdf\left(Q, L\right).$$

В рассматриваемой задаче



Критерий KPSS (Kwiatkowski–Phillips–Schmidt–Shin)

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

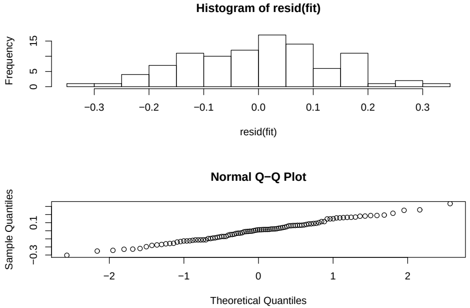
нулевая гипотеза: H_0: ряд ε^T стационарен;

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

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

$KPSS\left(\varepsilon^T\right)$ 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: $p > 0.1$.
Критерий автокоррелированности Дарбина-Уотсона: $p = 1.78 \times 10^{-10}$.

Авторегрессия

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

где x_t — стационарный ряд с нулевым средним, $\phi_1, \dots, \phi_p$ — константы $(\phi_p \neq 0)$, ω_t — гауссов белый шум с нулевым средним и постоянной дисперсией σ_ω^2.

Если среднее равно μ, модель принимает вид

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

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

Другой способ записи:

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

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

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

Скользящее среднее

$$MA\,(q):\quad x_t = \omega_t + \theta_1\omega_{t-1} + \theta_2\omega_{t-2} + \ldots \theta_q\omega_{t-q},$$

где x_t — стационарный ряд с нулевым средним, $\theta_1, \ldots, \theta_q$ — константы $(\theta_q \neq 0)$, ω_t — гауссов белый шум с нулевым средним и постоянной дисперсией σ_ω^2.

Если среднее равно μ, модель принимает вид

$$x_t = \mu + \omega_t + \theta_1\omega_{t-1} + \theta_2\omega_{t-2} + \ldots \theta_q\omega_{t-q}.$$

Другой способ записи:

$$x_t = \theta\,(B)\,\omega_t = \left(1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q\right)\omega_t,$$

где B — разностный оператор.

Линейная комбинация q компонент белого шума ω_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:

$$\phi_{hh} = \begin{cases} corr(x_1, x_0), & h = 1, \\ corr(x_h - x_h^{h-1}, x_0 - x_0^{h-1}), & h \geq 2, \end{cases}$$

где x_h^{h-1} — регрессия x_h на $\{x_{h-1}, x_{h-2}, \dots, x_1\}$:

$$x_h^{h-1} = \beta_1 x_{h-1} + \beta_2 x_{h-2} + \ldots + \beta_{h-1} x_1,$$

$$x_0^{h-1} = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{h-1} x_{h-1}.$$

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

ARMA (Autoregressive moving average)

$$ARMA\,(p,q): \quad x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots \phi_p x_{t-p} + \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \dots \theta_q \omega_{t-q},$$

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

Если среднее равно μ, модель принимает вид

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

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

Другой способ записи:

$$\phi\left(B\right) x_t = \theta\left(B\right) \omega_t.$$

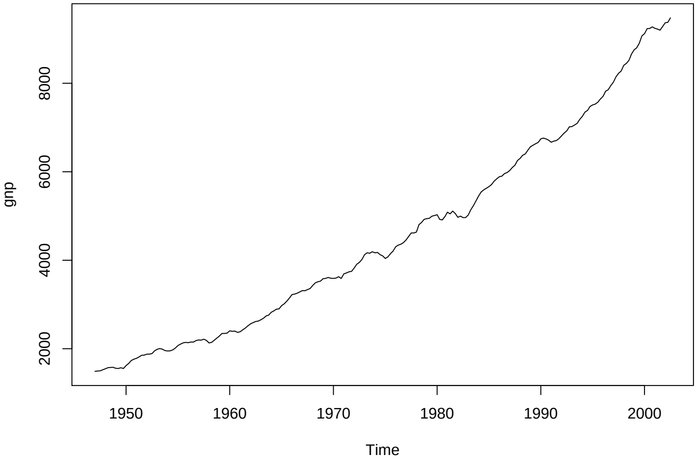
ARIMA (Autoregressive integrated moving average)

Для нестационарного ряда стационарным может оказаться ряд его разностей.

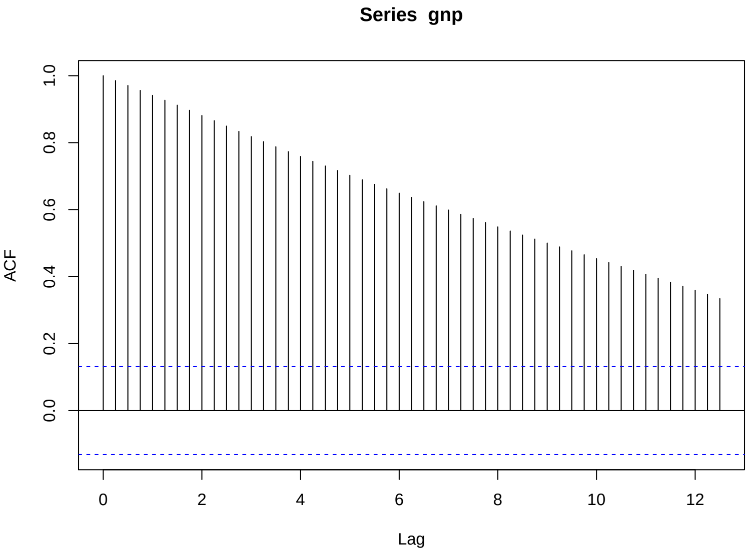
Ряд описывается моделью $ARIMA(p, d, q)$, если ряд его разностей $\nabla^d x_t = (1 - B)^d$ описывается моделью $ARMA(p, q)$.

Исходные данные

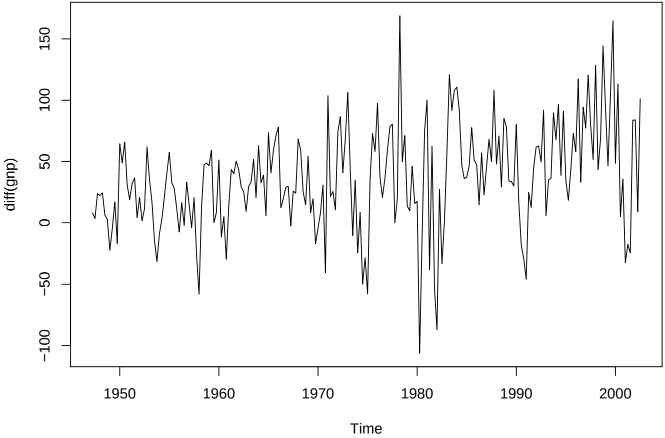
Поквартальные очищенные от сезонности данные о ВВП США в миллиардах долларов 1996 года.



Автокорреляция

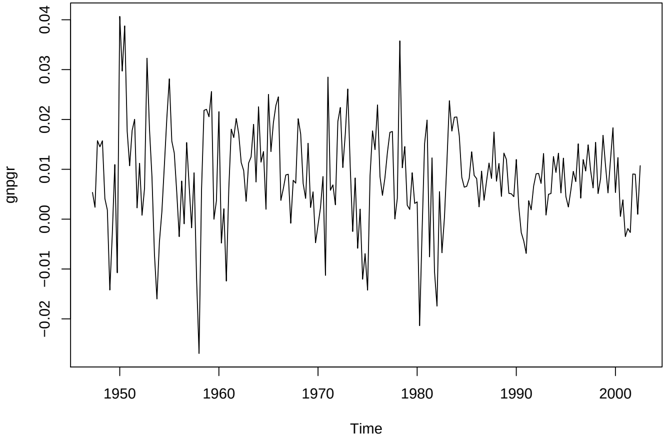


Ряд первых разностей



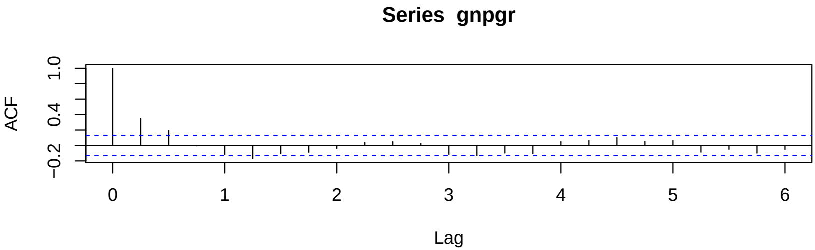
Нестационарен, вариация данных выше во второй половине ряда (KPSS $p<0.01$).

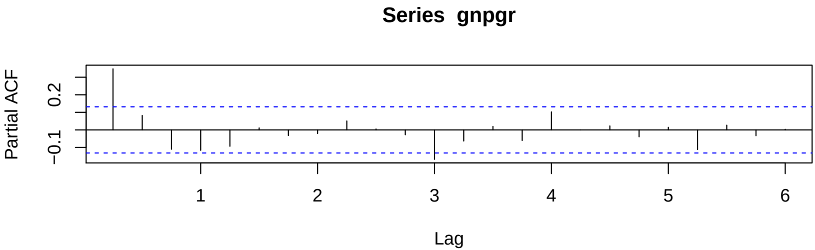
Ряд разностей логарифмов ряда



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

Автокорреляция и частная автокорреляция ряда прироста





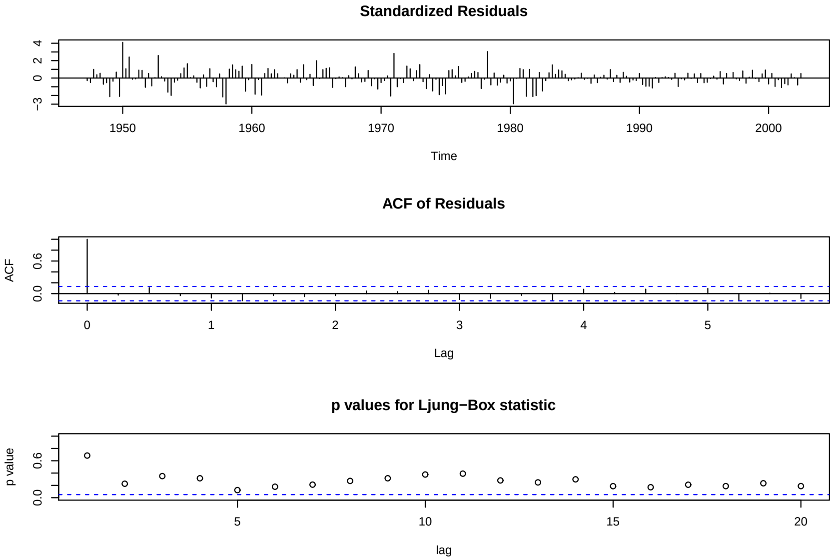
Модели

Варианты интерпретации графиков:

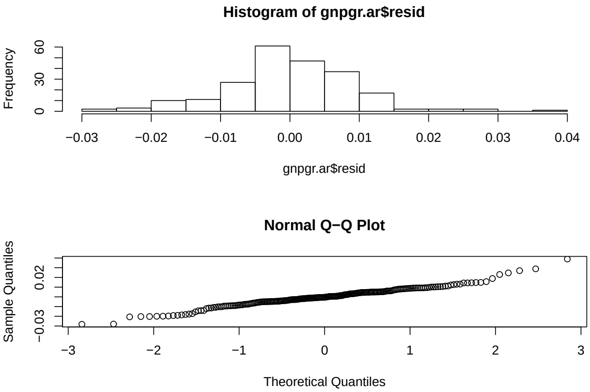
- AC равна нулю после лага 2, PAC убывает — модель $MA(2)$;
- PAC равна нулю после лага 1, AC убывает — модель $AR(1)$;
- модель $ARMA(1, 2)$.

$AR(1): \quad x_t = 0.005 + 0.347x_{t-1} + \hat{\omega}_t, \ \hat{\sigma}_\omega = 0.0095.$
$MA(2): \quad x_t = 0.008 + 0.303\hat{\omega}_{t-1} + 0.204\hat{\omega}_{t-2} + \hat{\omega}_t, \ \hat{\sigma}_\omega = 0.0094.$
$ARMA(1, 2): \quad x_t = 0.008 + 0.241x_{t-1} + 0.076\hat{\omega}_{t-1} + 0.162\hat{\omega}_{t-2} + \hat{\omega}_t, \ \hat{\sigma}_\omega = 0.0089.$

Диагностика $AR(1)$

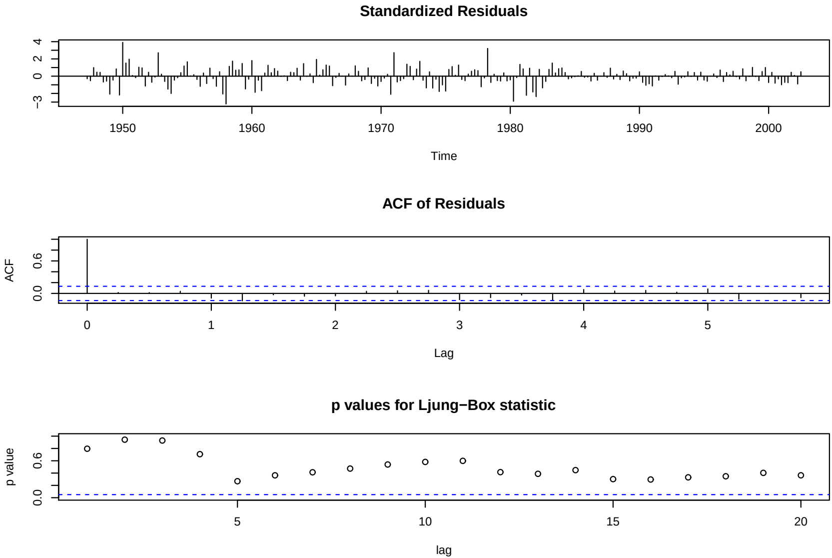


Диагностика $AR(1)$

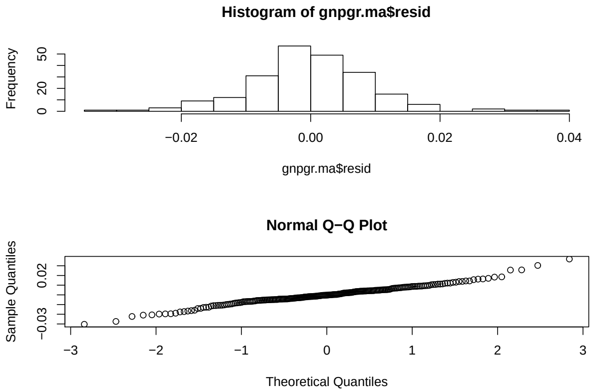


Критерий нормальности Шапиро-Уилка: $p = 0.0006886$.
Критерий Уилкоксона: $p = 0.8665$.
Критерий стационарности KPSS: $p > 0.1$.
Критерий независимости Бокса-Пирса: $p = 0.6867$.

Диагностика $MA(2)$

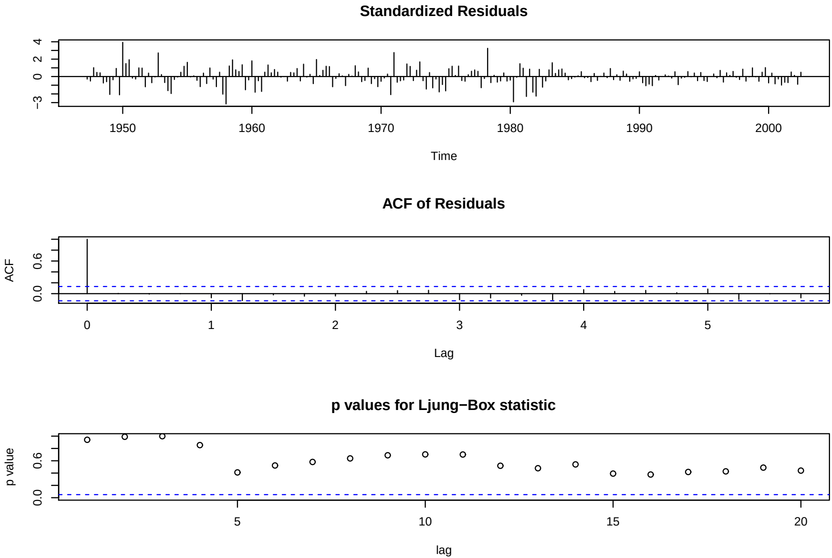


Диагностика $MA(2)$

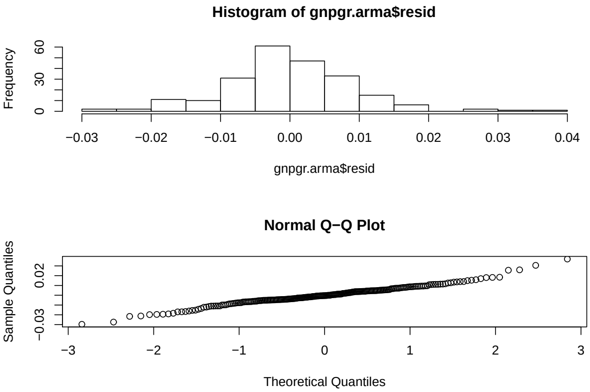


Критерий нормальности Шапиро-Уилка: $p = 0.003416$.
Критерий Уилкоксона: $p = 0.9917$.
Критерий стационарности KPSS: $p > 0.1$.
Критерий независимости Бокса-Пирса: $p = 0.797$.

Диагностика $ARMA(1, 2)$



Диагностика $ARMA(1,2)$



Критерий нормальности Шапиро-Уилка: $p = 0.003497$.
Критерий Уилкоксона: $p = 0.9817$.
Критерий стационарности KPSS: $p > 0.1$.
Критерий независимости Бокса-Пирса: $p = 0.9411$.

Сравнение моделей

AIC — информационный критерий Акаике:

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

где k — число параметров модели;
AICc — он же с поправкой на случай небольшого размера выборки:

$$AICc = \log\left(\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2\right) + \frac{T+k}{T-k-2};$$

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

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

	AIC	AICc	BIC (SIC)
$AR(1)$	-1431.221	-8.284898	-9.263748
$MA(2)$	-1431.929	-8.297199	-9.276049
$ARMA(1,2)$	**-1430.948**	**-8.301886**	**-9.280737**

Виды сезонных эффектов

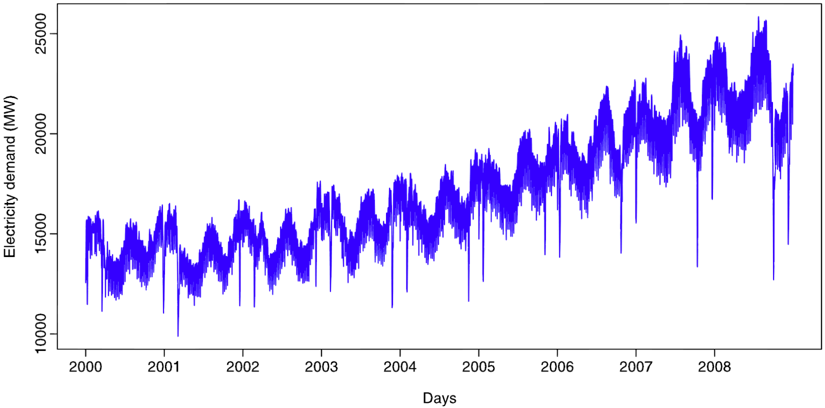
ESS Guidelines on Seasonal Adjustment, 2009

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

Причины возникновения сезонности:

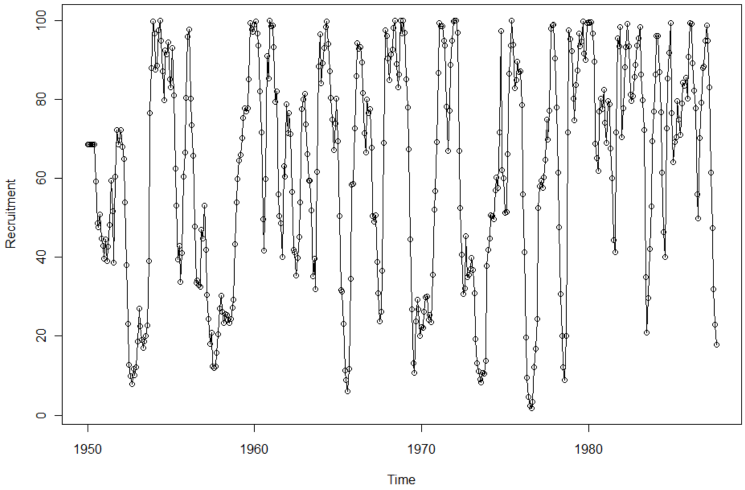
- природные факторы;
- административные и юридические факторы;
- календарные эффекты: число рабочих дней, эффекты фиксированных и плавающих праздников (национальные праздники, Пасха, Рамадан и т. д.).

Потребление электричества в Турции



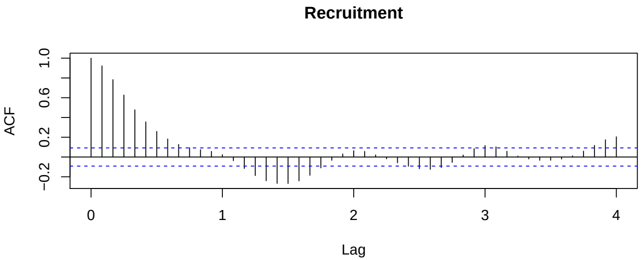
- недельная сезонность
- годовая сезонность
- праздники по исламскому календарю (год примерно на 11 дней короче, чем в грегорианском)

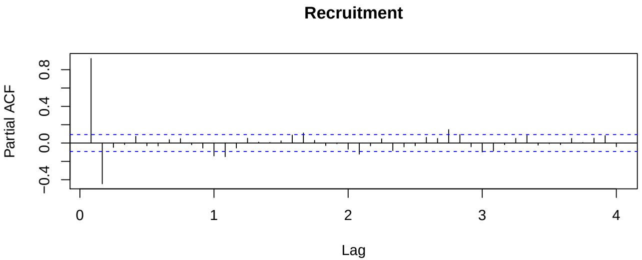
Исходные данные



Recruitment — число новых особей рыбы.

Корреляции

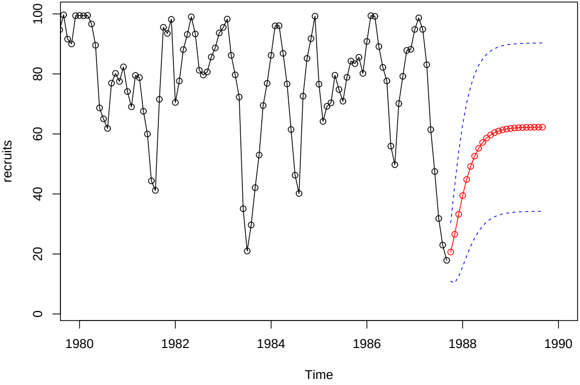

Recruitment
ACF
1.0
0.6
0.2
−0.2
0
1
2
3
4
Lag


Recruitment
Partial ACF
0.8
0.4
0.0
−0.4
0
1
2
3
4
Lag

Прогнозирование ряда

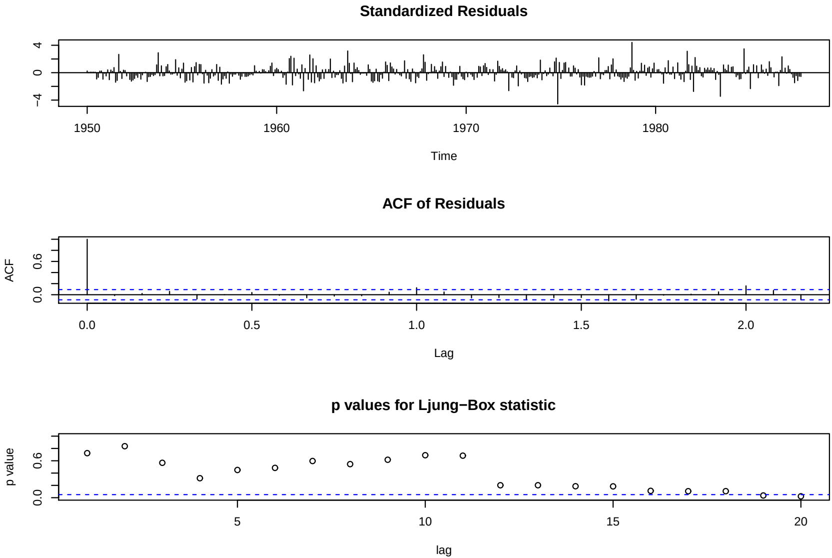
Выбор — модель $AR(2)$:

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \omega_t.$$

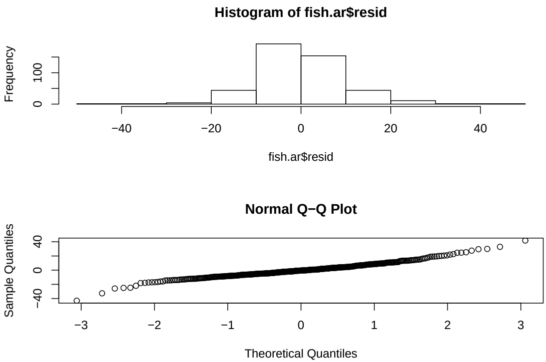


В моделях $ARMA(p,q)$ с увеличением горизонта прогноз всё больше похож на константу.

Диагностика модели $AR(2)$



Диагностика модели $AR(2)$



Критерий нормальности Шапиро-Уилка: $p = 2.72 \times 10^{-7}$.
Критерий Уилкоксона: $p = 0.4167$.
Критерий стационарности KPSS: $p > 0.1$.
Критерий независимости Бокса-Пирса: $p = 0.7248$.

Seasonal multiplicative ARMA/ARIMA

$$ARMA(p,q)\times(P,Q)_s:\quad \Phi_P(B^s)\phi(B)x_t=\Theta_Q(B^s)\theta(b)\omega_t,$$

где

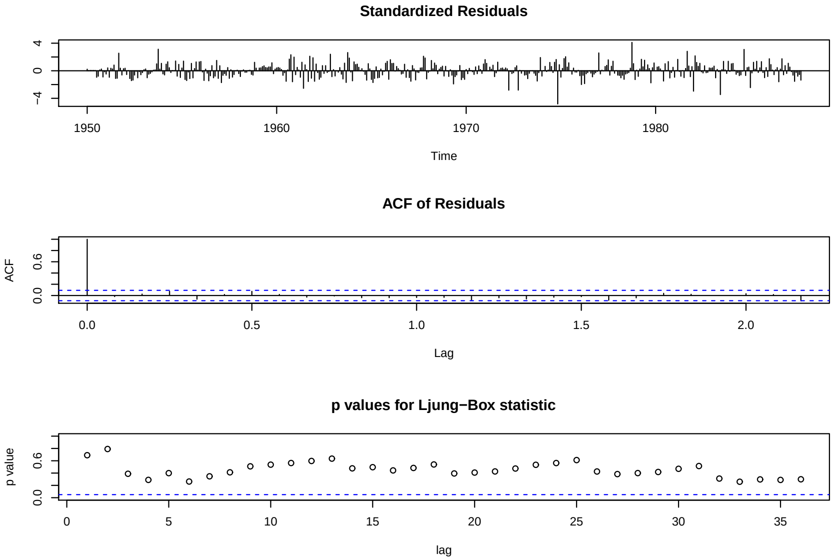
$$\Phi_P(B^s)=1-\Phi_1B^s-\Phi_2B^{2s}-\ldots-\Phi_PB^{Ps},$$
$$\Theta_Q(B^s)=1+\Theta_1B^s+\Theta_2B^{2s}+\ldots+\Theta_QB^{Qs}.$$

$$SARIMA:\quad \Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t=\alpha+\Theta_Q(B^s)\theta(b)\omega_t,$$

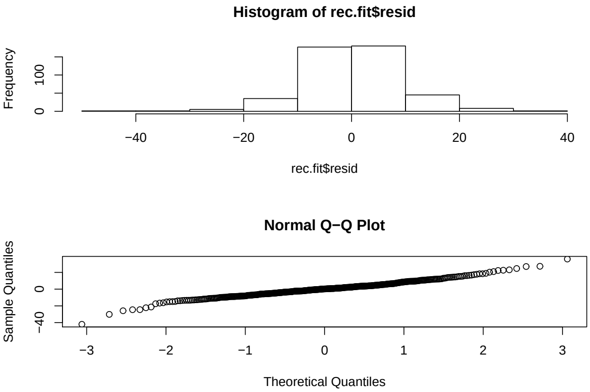
Сравнение моделей

	AIC
$ARMA(2,0) \times (3,0)_{12}$	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
$ARMA(2,0) \times (0,2)_{12}$	3321.88
$ARMA(2,0) \times (0,3)_{12}$	3314.787
$ARMA(2,0) \times (1,1)_{12}$	3283.717

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

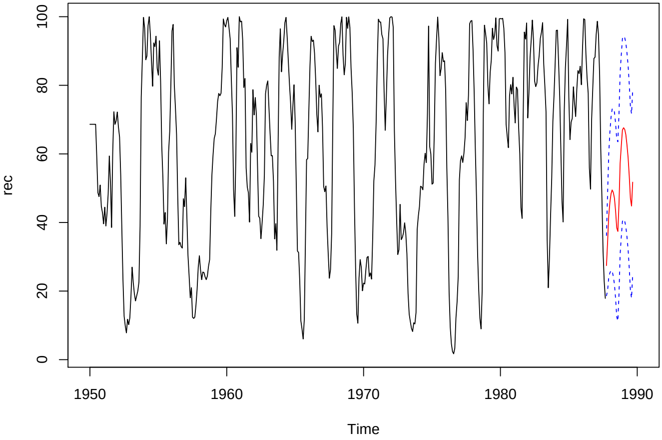


Диагностика $ARMA(2,0)\times(1,1)_{12}$



Критерий нормальности Шапиро-Уилка: $p = 6.781 \times 10^{-6}$.
Критерий Уилкоксона: $p = 0.7387$.
Критерий стационарности KPSS: $p > 0.1$.
Критерий независимости Бокса-Пирса: $p = 0.6915$.

Прогнозирование ряда



Прикладная статистика
11. Анализ временных рядов.

Рябенко Евгений
riabenko.e@gmail.com