Временные ряды

Временным рядом называется последовательность значений признака y, измеряемого через постоянные временные интервалы:

$$y_1, \ldots, y_T, \ldots, \quad y_t \in \mathbb{R}.$$

В этом определении нужно обратить внимание на то, что временные интервалы между измерениями признака постоянны.

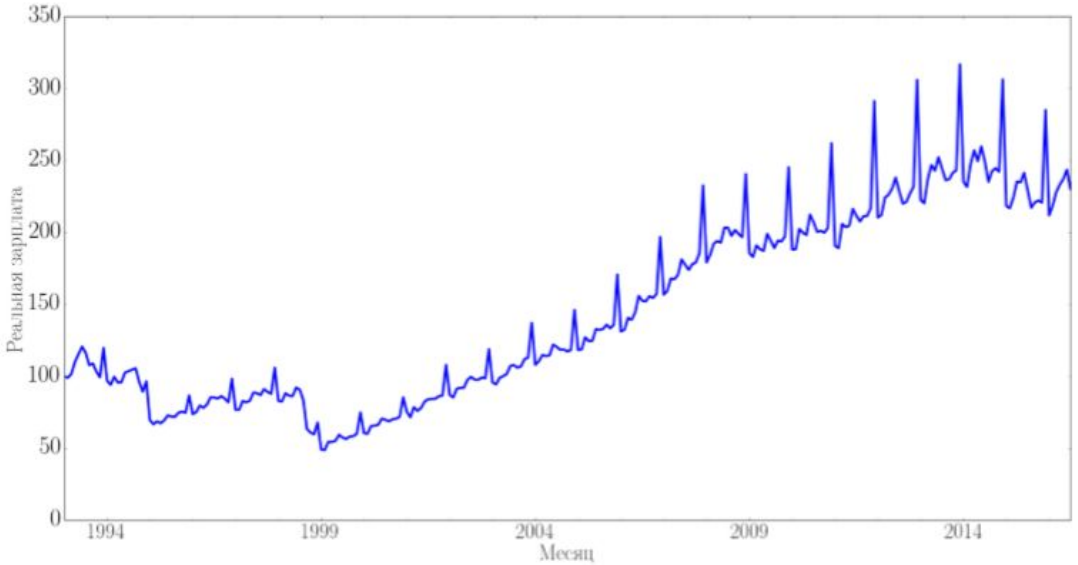


Рис. 1.1: Среднемесячная реальная заработная плата в России, выраженная в процентах от её значения в январе 1993 г.

Применение регрессионной модели в прогнозе временного ряда

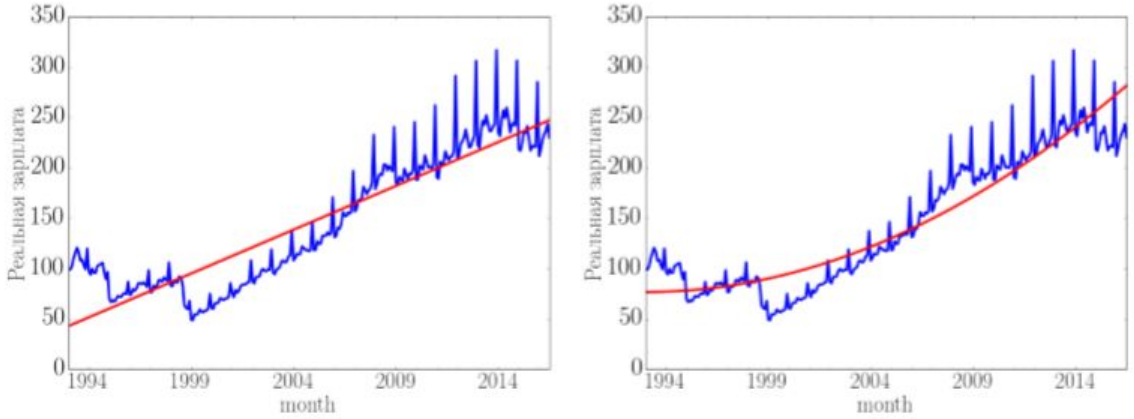


Рис. 1.2: Применение модели линейной (слева) и квадратичной (справа) регрессии к задаче прогнозирования временного ряда.

Остатки прогноза модели регрессии

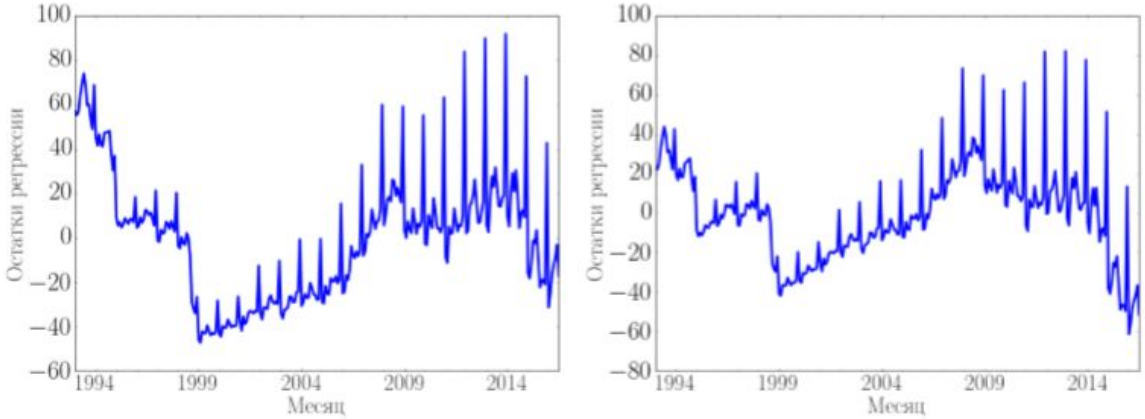


Рис. 1.3: Остатки модели линейной (слева) и квадратичной (справа) регресси в задаче прогнозирования временного ряда.

Однако это решение слишком простое, чтобы быть хорошим. Остатки такой регрессии (рис. 1.3) далеко не похожи на случайный шум, в них остаётся большая часть структуры, которая не была учтена в регрессионной модели. Чем больше структуры временного ряда учитывается в модели, тем лучшее предсказание она даёт. Вид остатков регрессии намекает на то, что можно построить более сложную модель, которая будет лучше описывать имеющиеся данные, а также давать более точные прогнозы в будущем.

Основные компоненты временного ряда

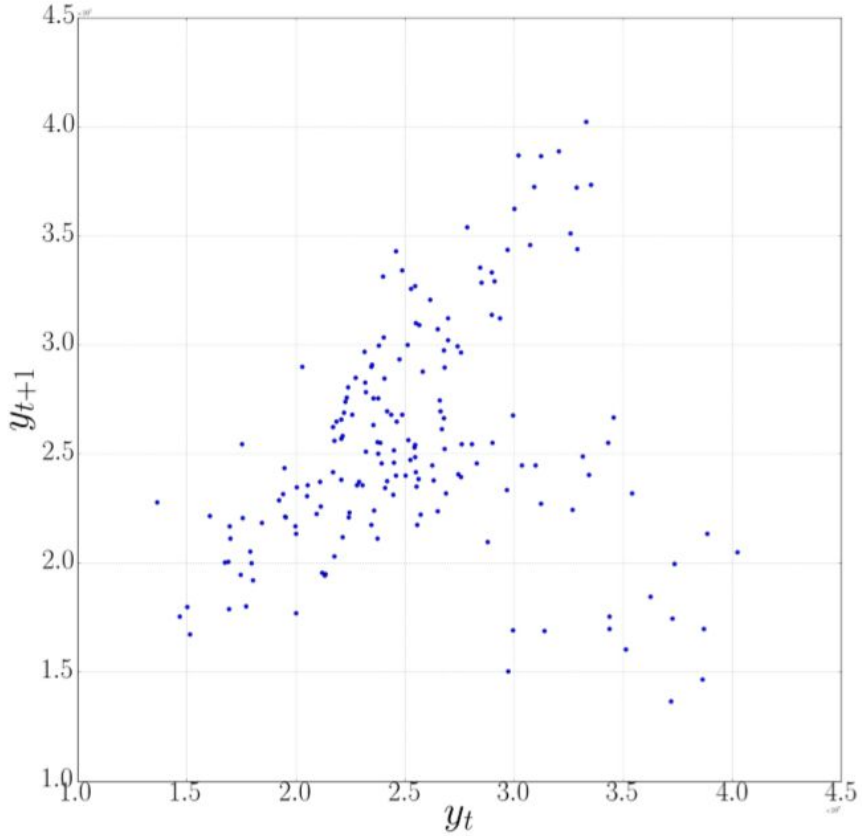
- Тренд — плавное долгосрочное изменение уровня ряда. Эту характеристику можно получить, наблюдая ряд в течение достаточно долгого времени.
- Сезонность — циклические изменения уровня ряда с постоянным периодом. В данных о средней зарплате в России (рис. 1.1) очень хорошо видны подобные сезонные колебания: признак всегда принимает максимальное значение в декабре каждого года, а минимальное — в январе следующего года. В целом профиль изменения зарплаты внутри года остаётся более-менее постоянным.
- Цикл — изменение уровня ряда с переменным периодом. Такое поведение часто встречается в рядах, связанных с продажами, и объясняется циклическими изменениями экономической активности. В экономике выделяют циклы длиной $4-5$ лет, $7-11$ лет, $45-50$ лет и т. д. Другой пример ряда с такой характеристикой — это солнечная активность, которая соответствует, например, количеству солнечных пятен за день. Она плавно меняется с периодом, который составляет несколько лет, причём сам период также меняется во времени.
- Ошибка — непрогнозируемая случайная компонента ряда. Сюда включены все те характеристики временного ряда, которые сложно измерить (например, слишком слабые).

Автокорреляционные модели

- ARMA
- ARIMA
- SARIMA

Что такое автокорреляция?

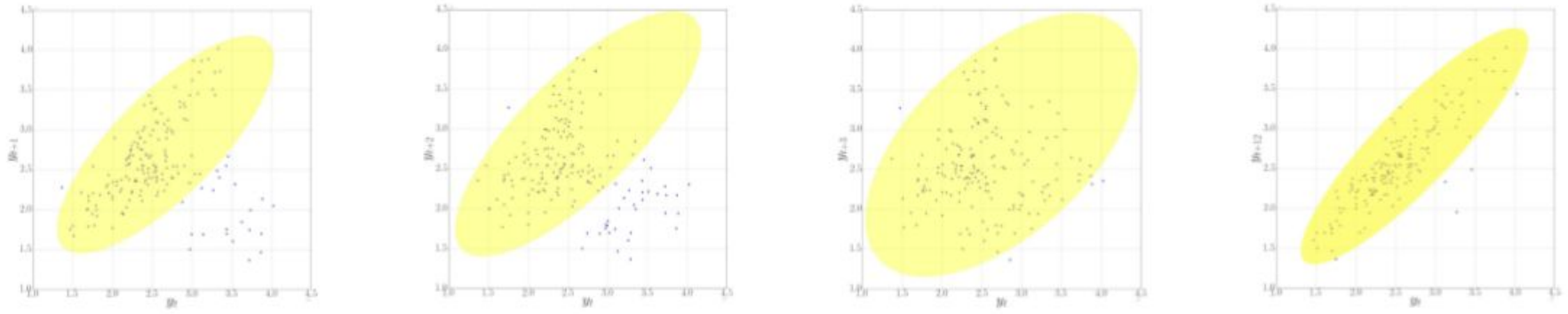
Автокорреляция



Автокорреляция объясняет связь между соседними временными интервалами с заданным шагом.

Каждая точка задает продажи в 2 месяца

Автокорреляция



(a) (b) (c) (d)

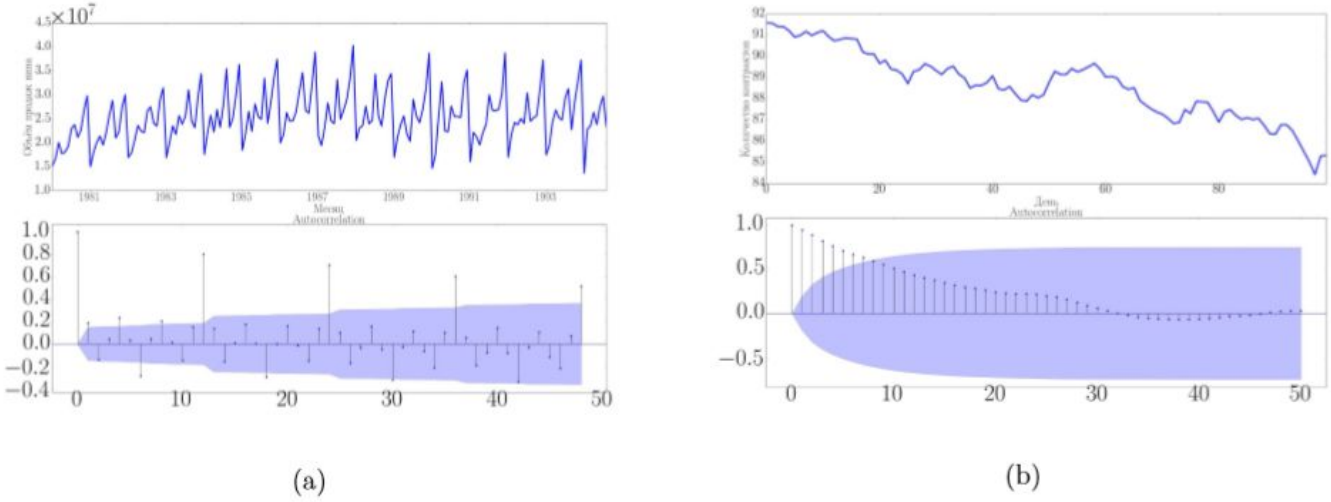
Рис. 1.10: Связь между продажами в соседние месяцы (a), через месяц (b), через два месяца(c) и через год (d).

Автокорреляция

$$r_\tau = \frac{\sum_{t=1}^{T-\tau}(y_t - \bar{y})(y_{t+\tau} - \mathbb{E}y))}{\sum_{t=1}^{T-\tau}((y_t - \bar{y}))^2}.$$

Автокоррелляция - обычная корреляция Пирсона между исходным рядом и рядом сдвинутым на несколько отсчетов. Мы хотим найти такой лаг при котором данная корреляция будет максимальна.

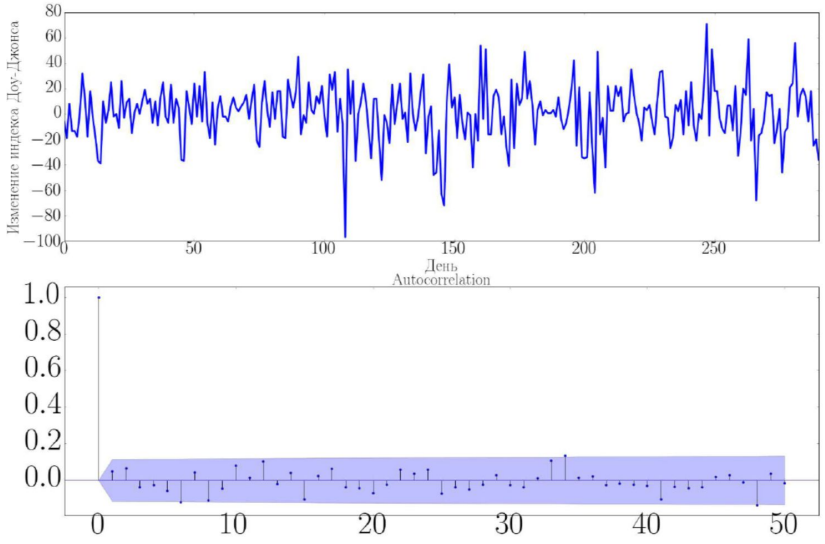
Коррелограмма



(a) (b)

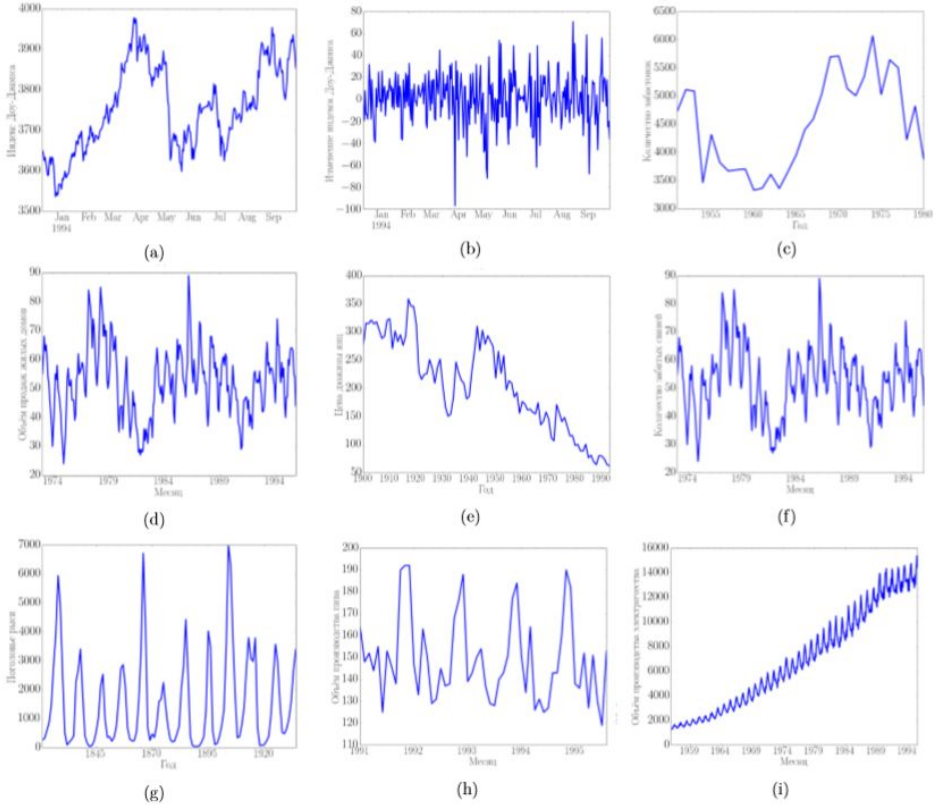
Автокорреляцию удобно анализировать на графике коррелограмм - это график на котором значение корреляций отложены по вертикали согласно кратным периодам. Согласно сезонным шагам мы видим наиболее сильные значения корреляций

Коррелограмма



Часто можно обнаружить что нет явной зависимости

Стационарность



Стационарность

Ещё одно важное свойство временных рядов — это стационарность. Временной ряд $y_1, \ldots, y_T$ называется стационарным, если $\forall s$ (ширина окна) распределение $y_t, \ldots, y_{t+s}$ не зависит от t, т.е. его свойства не зависят от времени.

Из этого определения следует, что ряды, в которых присутствует тренд, являются нестационарными: в зависимости от расположения окна изменяется средний уровень ряда. Кроме того, нестационарны ряды с сезонностью: если ширина окна меньше сезонного периода, то распределение ряда будет разным, в зависимости от положения окна. При этом интересно, что ряды, в которых есть непериодические циклы, не обязательно являются нестационарными, поскольку нельзя заранее предсказать положение максимумов и минимумов этого ряда.

Стационарность

Что можно делать:

- Проверяем критерием Дики Фуллера
- Стабилизируем дисперсию
- Логарифмируем
- Преобразование Бокса-Кокса
- Диференцируем по времени
- Повторное дифференцирование для устранения остатков

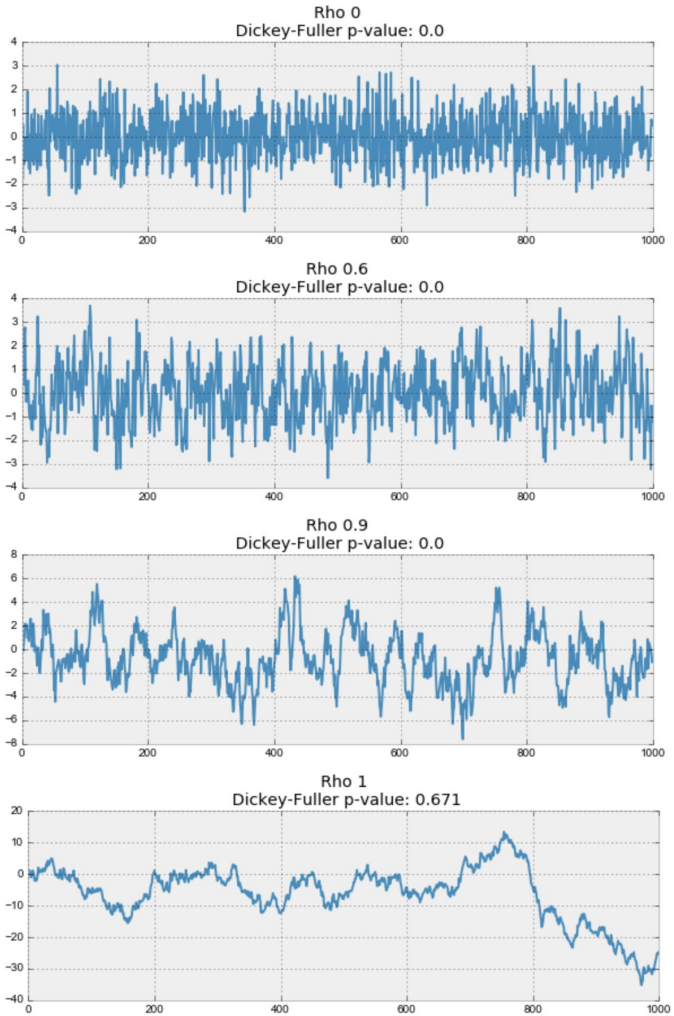
Критерий Дики-Фуллера

$$x_t = \rho x_{t-1} + e_t$$

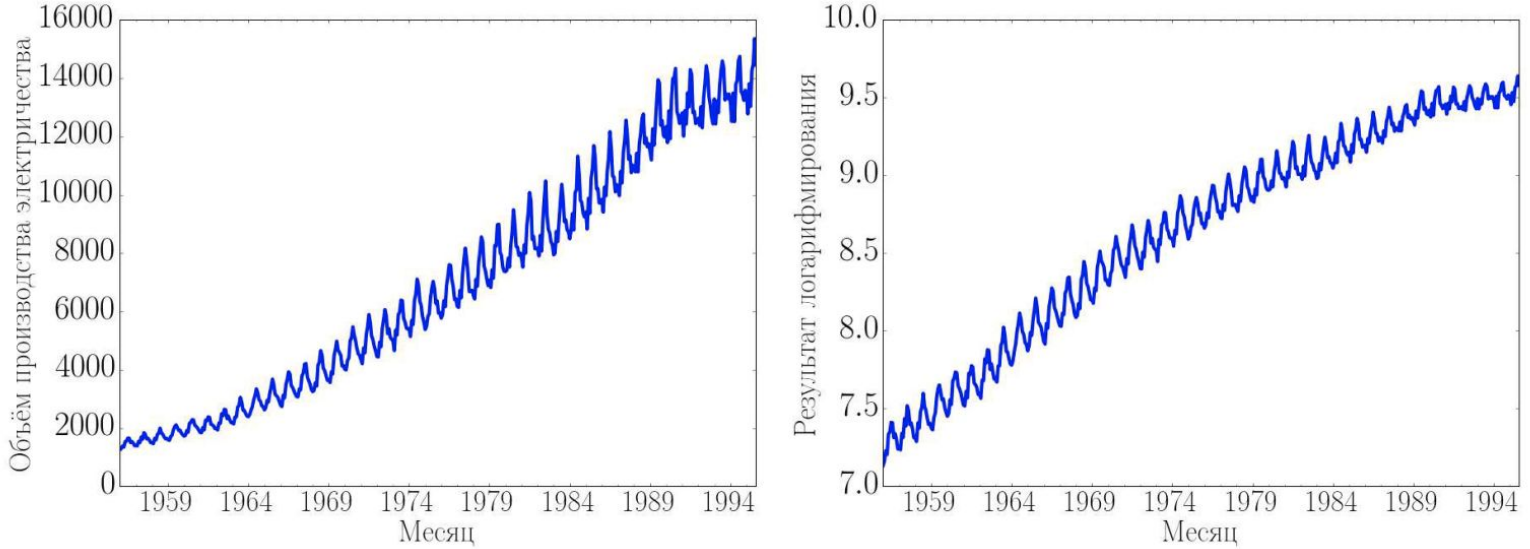
значение ρ равное единице дало процесс случайного блуждания — ряд не стационарен. Происходит это из-за того, что при достижении критической единицы, ряд перестаёт возвращаться к своему среднему значению.

$$x_t - x_{t-1} = (\rho - 1)x_{t-1} + e_t$$

Тест Дики-Фуллера проверяет какой уровень ρ получаем для получения нестационарного ряда, значение при котором распределение значений ряда в окне фиксированной ширины не постоянно.
Если критерий Дики-Фуллера не отвергает нулевую гипотезу о наличие единичного корня, то ряд не стационарен.



Логарифм ряда



Эффективно работает для рядов с монотонно изменяющейся дисперсией. Частный случай преобразований Бокса -Кокса

Преобразование Бокса-Кокса

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

Лямбда определяет как сильно преобразуем наш ряд. Логарифмирование по сути и есть логарифмирование.
При 1 тождественное преобразование ряда, получаем тот же ряд.
При других значениях это преобразование по сути степенное.

Мы можем подбирать лямбда чтобы дисперсия была стабильна во времени. Важно помнить - Лямбда-дисперсия значительно более стабильна во времени нежели обычная дисперсия временного ряда.

Дифференцирование

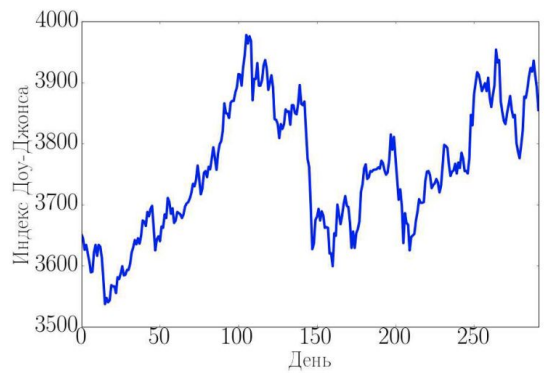
$$y' = y_t - y_{t-1}.$$

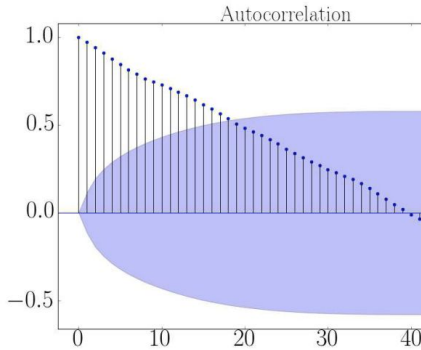
Ещё один важный трюк, который позволяет сделать ряд стационарным, — это дифференцирование, переход к попарным разностям соседних значений.

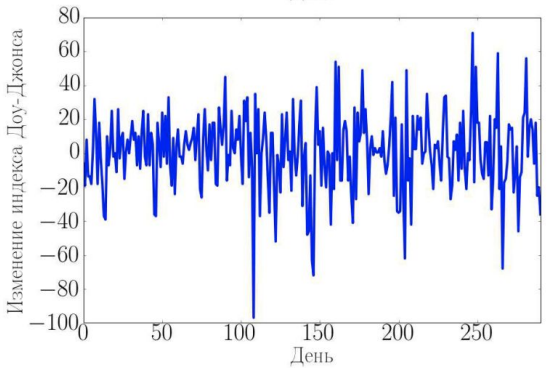
Для нестационарных рядов очень эффективно применять дифференцирование, при этом получаем стационарный ряд, избавляемся от влияния тренда.

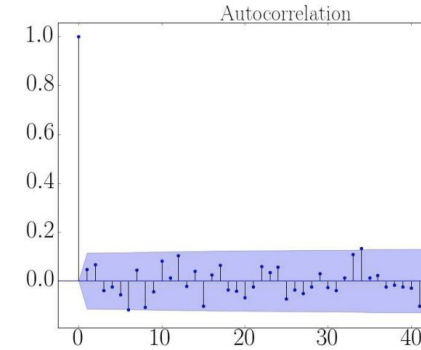
$$y'_t = y_t - y_{t-s}.$$

Также можно делать не единичное дифференцирование а сезонное









Анализ остатков

$$\hat{\varepsilon}_t = y_t - \hat{y}_t.$$

Анализ остатков — это техника, которая помогает понять, есть ли у прогнозирующей модели небольшие недостатки, которые можно устранить доработкой, или же фундаментальные проблемы.

Их можно вычислять двумя способами. Во-первых, прогнозы, которые участвуют в остатках, можно строить с фиксированной отсрочкой. Например, начиная с момента R прогноз всегда делается на одну точку вперёд, затем происходит переход в момент R + 1, получается новое истинное значение ряда, которое сравнивается с прогнозом, затем следующий прогноз делается ещё на одну точку вперёд, и так далее до самого конца ряда:

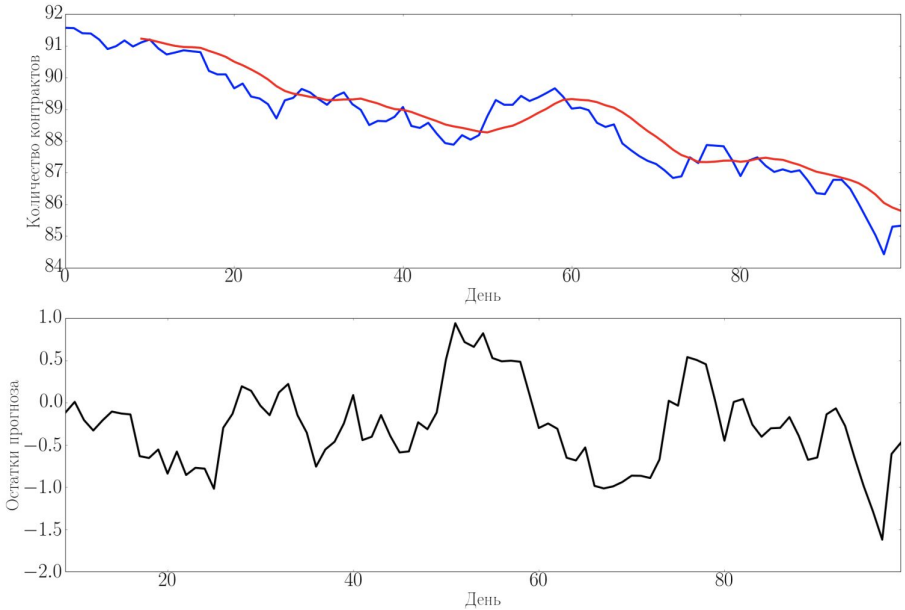
$$\hat{y}_{R+d|R}, \cdots, \hat{y}_{T|T-d}.$$

Во-вторых, остатки можно строить с фиксированным концом истории при разных отсрочках. Например, берётся начальная часть ряда от 0 до T − D, и далее делаются прогнозы

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

Несмещенность

Во-первых, остатки должны быть несмещёнными, то есть в среднем они должны быть равны нулю



Гипотезу о несмещённости остатков H0 : ε = 0 можно формально проверить с помощью какого-либо стандартного одновыборочного критерия (например, критерия Стьюдента или Уилкоксона).

Если выясняется, что остатки смещены, значит с моделью что-то не так.
Если никаких улучшений явно найти не получается и модель не видит скрытый шум, можно вычитать матожидание значения остатков ε из прогноза для получения несмещенного прогноза.

В зависимости от задачи мы можем хотеть предсказывать бОльшие или меньшие значения

Немного о моделях - ARMA

Можно проделать следующий трюк: взять авторегрессионную модель порядка p (AR(p)) и модель скользящего среднего порядка q (MA(q)) и сложить то, что находится у них в правых частях. Результат — это модель ARMA(p, q), она выглядит следующим образом:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}.$$

Главное, что нужно знать об этой модели: теорема Вольда утверждает, что любой стационарный временной ряд может быть описать моделью ARMA(p, q) с правильным подбором значений параметров p, q

Немного о моделях - ARMA

Можно проделать следующий трюк: взять авторегрессионную модель порядка p (AR(p)) и модель скользящего среднего порядка q (MA(q)) и сложить то, что находится у них в правых частях. Результат — это модель ARMA(p, q), она выглядит следующим образом:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}.$$

Главное, что нужно знать об этой модели: теорема Вольда утверждает, что любой стационарный временной ряд может быть описать моделью ARMA(p, q) с правильным подбором значений параметров p, q

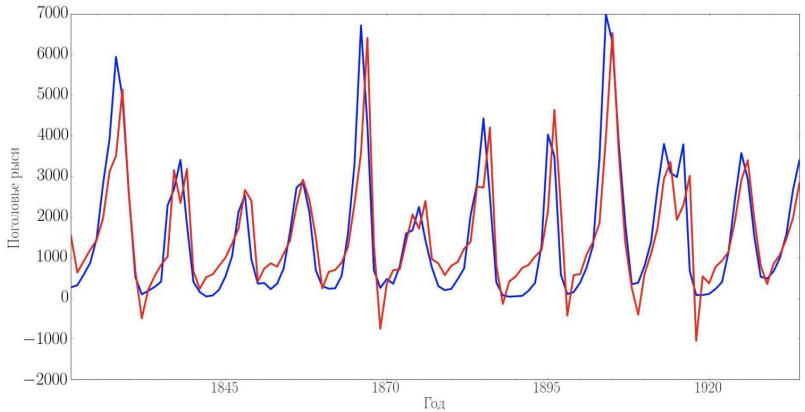


Рис. 1.20: Результат применения модели $ARMA(2, 2)$ к данным о поголовье рыси.

Немного о моделях - ARIMA

При помощи дифференцирования нестационарный ряд можно сделать стационарным.
Модель ARIMA(p, d, q) — это модель ARMA(p, q) для d раз продифференцированного ряда.

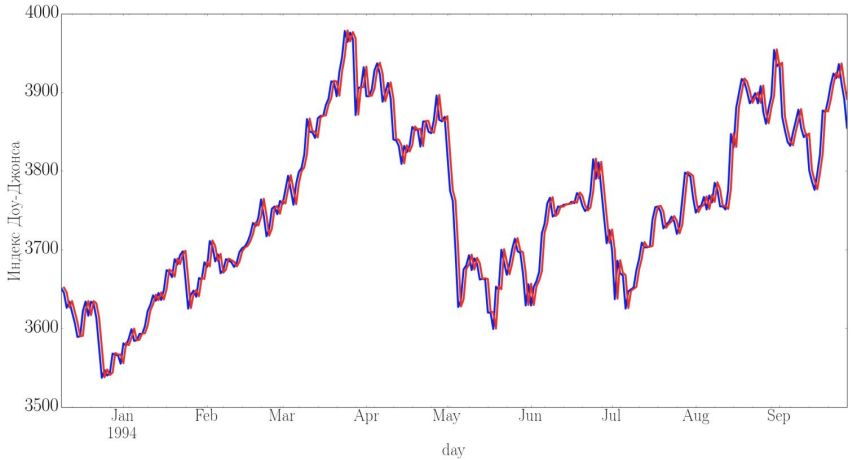


Рис. 1.23: Модель $ARIMA(0, 1, 0)$, применённая к ряду значений индекса Доу-Джонса.

Немного о моделях - SARMA

Настало время разобраться с сезонностью. Пусть ряд имеет сезонный период длины S. Тогда можно взять модель ARMA(p, q):

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

добавить к этой модели P авторегрессионных компонент, но не предыдущих, а с шагом, равным периодом сезонности:

$$+\phi_S y_{t-S} + \phi_{2S} y_{t-2S} + \cdots + \phi_{PS} y_{t-PS}$$

и Q компонент скользящего среднего, также с шагом, равным периодом сезонности:

$$+\theta_S \varepsilon_{t-S} + \theta_{2S} \varepsilon_{t-2S} + \cdots + \theta_{PS} \varepsilon_{t-QS}.$$

Общее об авторегрессионных моделях

У моделей класса ARIMA есть несколько групп параметров. Параметры d, D, q, Q, p, P можно считать гиперпараметрами, поскольку они определяют структуру и количество коэффициентов в самой модели ARIMA. Остальные параметры, α, φ, θ, являются коэффициентами в регрессионном уравнении.

Параметры α, φ, θ

Если зафиксированы параметры d, D, q, Q, p, P, то есть зафиксирована структура модели ARIMA, то параметры α, φ, θ можно подобрать с помощью метода наименьших квадратов. Фактически происходит на16 страивание привычной регрессии методом минимизации квадратичной ошибки.

Параметры d, D

Параметры d, D, которые задают порядки дифференцирования, необходимо подбирать так, чтобы ряд стал стационарным. Ранее уже упоминалось, что всегда рекомендуется начинать с сезонного дифференцирования, потому что уже после него ряд может оказаться стационарным

Параметры q, Q, p, P

К сожалению, гиперпараметры q, Q, p, P нельзя выбирать из принципа максимума правдоподобия. Например, чем больше значение параметра p, тем больше авторегрессионных компонент в итоговом уравнении, тем больше параметров φ и тем лучше это уравнение описывает данные. Можно подбирать по минимизации критерия Акаике, об этом поговорим чуть дальше

Метрики

$$\text{Mean absolute error: MAE} = \text{mean}(|e_t|),$$

$$\text{Root mean squared error: RMSE} = \sqrt{\text{mean}(e_t^2)}.$$

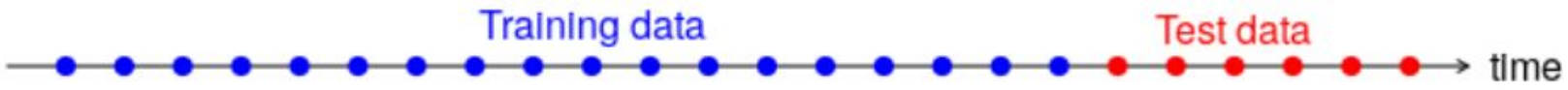
$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y_i}$$

Акаике

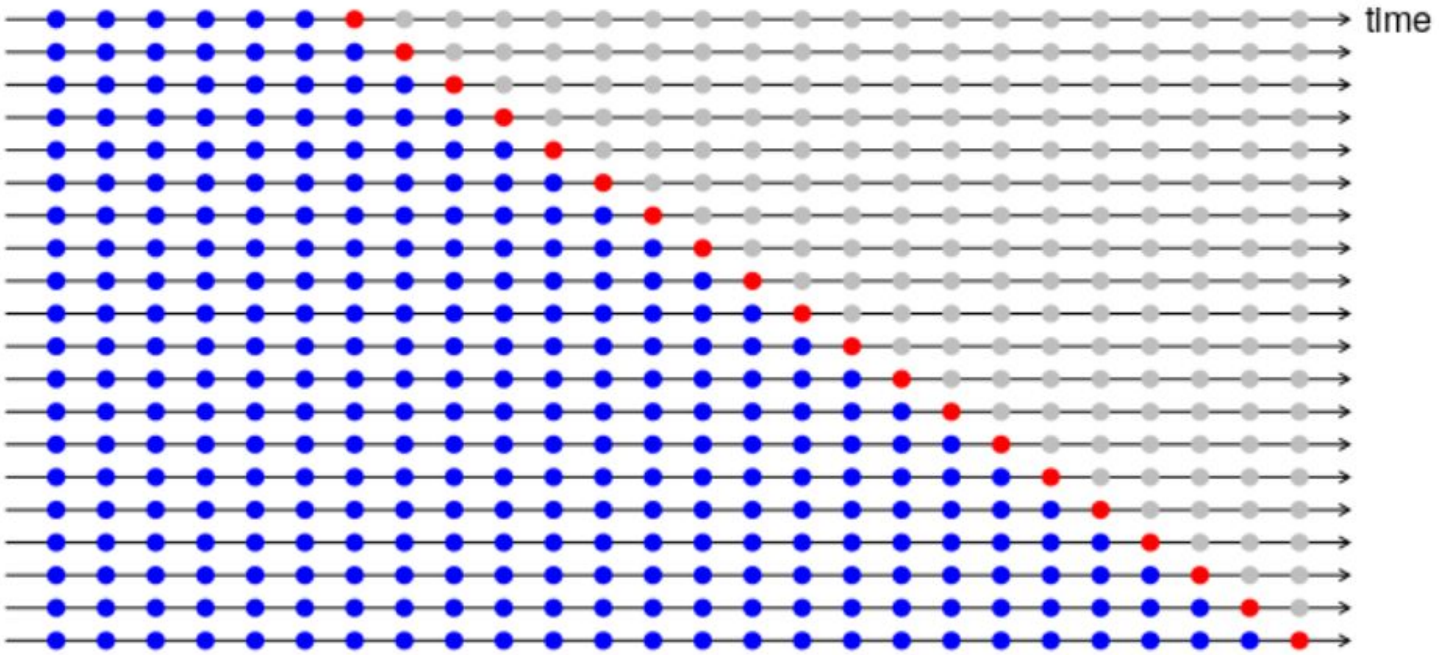
$$AIC = -2 \ln L + 2k,$$

где L — правдоподобие, k = P + Q + p + q + 1 — число параметров в модели.
Оптимальной по критерию Акаике будет модель с наименьшим значением этого критерия. Такая модель, с одной стороны, будет достаточно хорошо описывать данные, а с другой — содержать не слишком большое количество параметров

Как тестировать



Как тестировать



Как тестировать