Прикладная статистика. Временные ряды II.

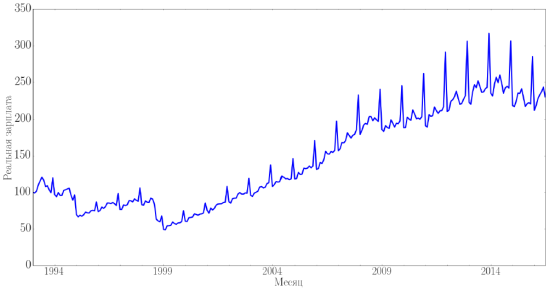
Леонид Иосипой

Программа «Математика для анализа данных»
Центр непрерывного образования, ВШЭ

21 января 2021

Временные ряды

Временным рядом называется последовательность наблюдений X_t, полученных через одинаковые временные интервалы.



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

Временные ряды

В анализе временных рядов чаще всего решаются следующие две важные задачи:

- определение природы явления, поиск тренда, сезонности и других закономерностей;
- осуществление прогноза будущих значений ряда.

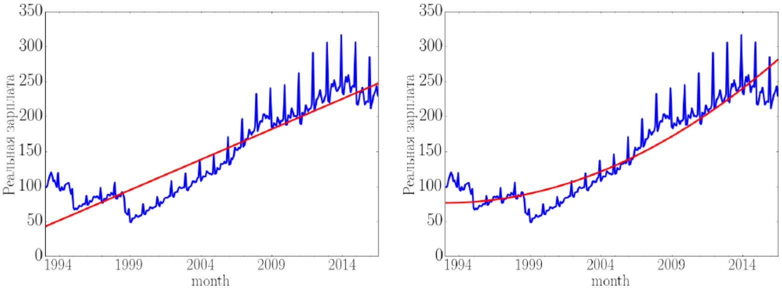
Обе задачи требуют построения модели, адекватно описывающей имеющиеся наблюдения.

Временные ряды

В отличие от задачи регрессии, временной ряд не образует независимую выборку. Наоборот, мы предполагаем, что данные в прошлом каким-то образом связаны с данными в будущем.

Выявив структуру этой зависимости, можно учесть её в модели и построить действительно точный прогноз.

Временные ряды

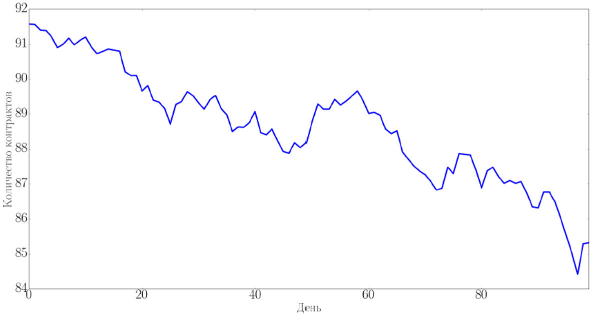


Применение модели линейной (слева) и квадратичной (справа) регрессии к задаче прогнозирования временного ряда из предыдущего примера

Временные ряды

Полезно рассмотреть несколько понятий, которыми можно описать поведение временных рядов:

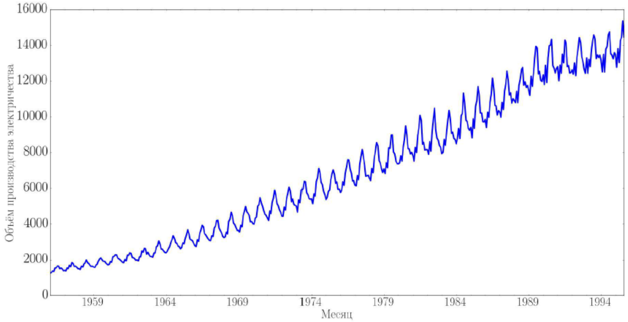
- Тренд — плавное долгосрочное изменение уровня ряда.



Количество контрактов за день в сокровищнице США

Временные ряды

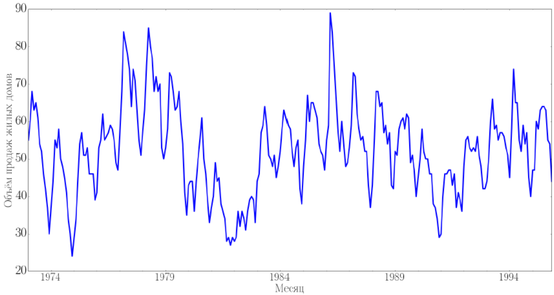
- Сезонность — циклические изменения уровня ряда с постоянным периодом.



Объём электричества, произведённого за месяц в Австралии

Временные ряды

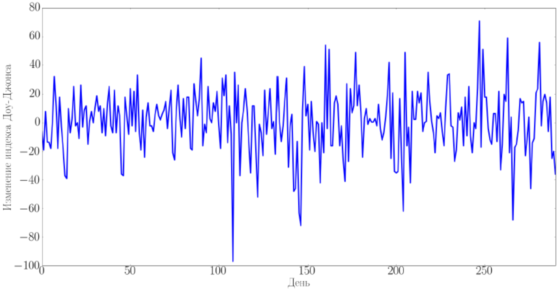
- Цикл — изменение уровня ряда с переменным периодом. Такое поведение часто встречается в рядах, связанных с продажами, и объясняется циклическими изменениями экономической активности.



Объём проданной жилой недвижимости (в млн кв. м.) в Америке за месяц

Временные ряды

- Ошибка — непрогнозируемая случайная компонента. Сюда включены все те характеристики временного ряда, которые сложно измерить.



Ежедневное изменение индекса Доу-Джонса

Перейдем теперь к численным характеристикам, которые описывают зависимость наблюдений временного ряда.

Временные ряды

Простейшей характеристикой зависимости близких значений ряда является автокорреляционная функция (или просто автокорреляция):

$$R(n) = \rho(X_n, X_0),$$

где $\rho(X_n, X_0)$ — коэффициент корреляции (Пирсона) между X_n и X_0. Аргумент n здесь называется лагом (lag).

Временные ряды

Для оценивания автокорреляций $R(n)$ по наблюдениям ряда обычно используется выборочные автокорреляции

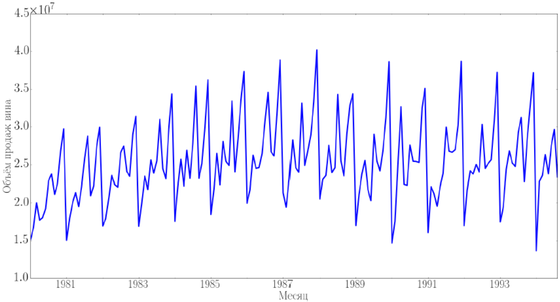
$$\widehat{R}(n) = \frac{\sum_{t=1}^{N-n}\left(X_t - \overline{X}\right)\left(X_{t+n} - \overline{X}\right)}{\sum_{t=1}^{N}\left(X_t - \overline{X}\right)^2}, \quad \overline{X} = \frac{1}{N}\sum_{t=1}^{N} X_t,$$

где N — длина ряда.

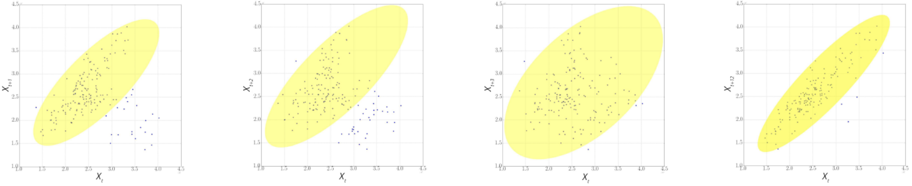
Обратите внимание, что эти оценки несколько отличаются от выборочного коэффициента корреляции между самим рядом X_t и сдвинутым на лаг n рядом X_{t+n}.

Временные ряды

Рассмотрим в качестве примера следующий ряд:



Месячный объём продаж вина в Австралии, в бутылках



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

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

Временные ряды

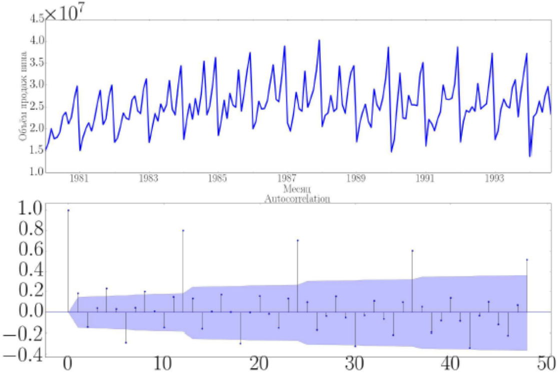


График автокорреляционной функции (ACF) для количества проданного вина в Австралии за месяц

Временные ряды

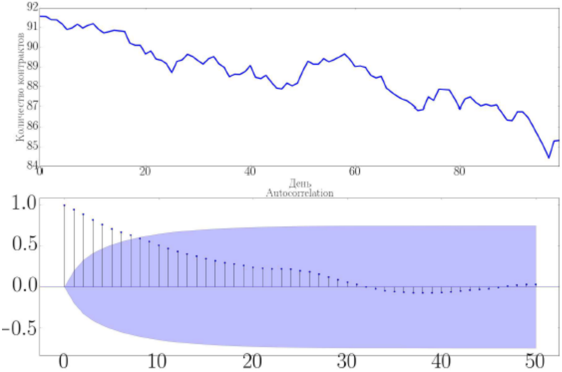


График автокорреляционной функции (ACF) для количества контрактов за день в сокровищнице США

Временные ряды

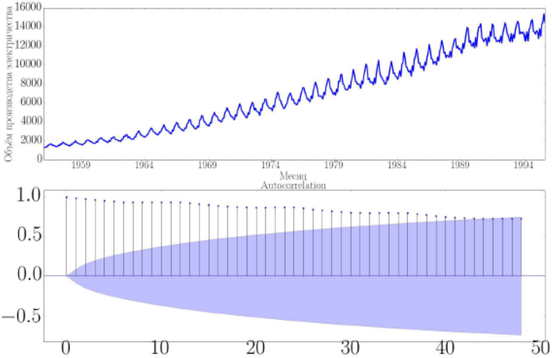


График автокорреляционной функции (ACF) для ежемесячного производства электричества в Австралии

Временные ряды

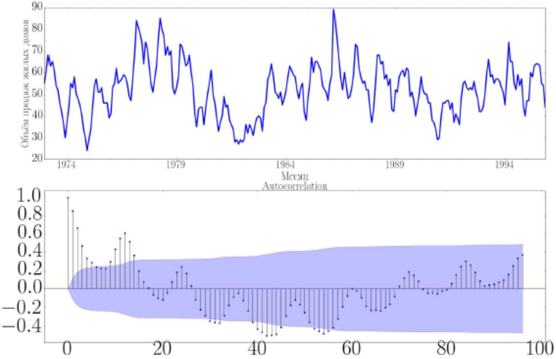


График автокорреляционной функции (ACF) для объёма проданной в Америке недвижимости за месяц

Временные ряды

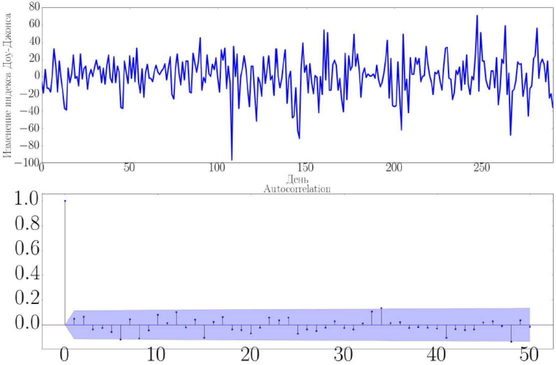


График автокорреляционной функции (ACF) для данных о ежедневном изменении индекса Доу-Джонса.

Временные ряды

На всех показанных графиках изображён фиолетовый коридор вокруг горизонтальной оси. Это коридор значимости отличия корреляции от нуля. Все автокорреляции, которые выходят за этот коридор, значимо отличаются от нуля.

Стоит учитывать, что как и для обычной корреляции Пирсона, границы доверительного интервала вычисляются в нормальных предположениях.

Временные ряды

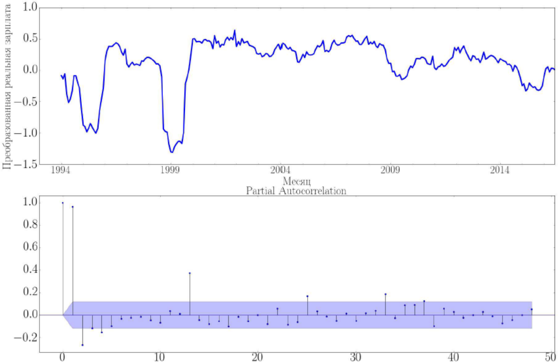
Еще одной полезной характеристикой зависимости членов временного ряда является частная автокорреляция:

$$\widetilde{R}(1) = \rho(X_1, X_0), \quad \widetilde{R}(n) = \rho(X_n, X_0 | X_1, \ldots, X_{n-1}),$$

где $\rho(X_n, X_0 | X_1, \ldots, X_{n-1})$ — коэффициент частной корреляции между X_n и X_0 при исключении влияния случайных величин $X_1, \ldots, X_{n-1}$.

Частные автокорреляции можно оценить по наблюдениям ряда. Мы не будем выписывать явные формулы.

Временные ряды



Пример графика частной автокорреляционной функции (PACF).

Временные ряды

В дальнейшем мы будем использовать графики автокорреляционной и частной автокорреляционной функции при подгонке моделей к временному ряду.

Временные ряды

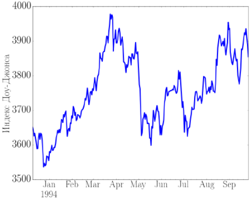
Ещё одно важное свойство временных рядов — это стационарность.

Временной ряд $X_1, \dots, X_N$ называется стационарным, если его вероятностные характеристики (среднее, дисперсия и автокорреляция) не меняются с течением времени.

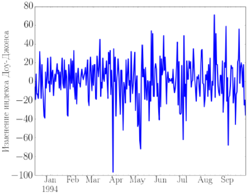
Временные ряды с трендом или сезонностью не являются стационарными. При этом ряды с циклами могут быть стационарными, поскольку положение максимумов и минимумов этого ряда заранее предсказать нельзя.

Временные ряды

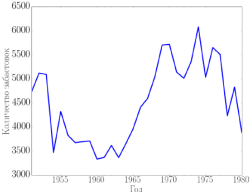
Какие из следующих рядов являются стационарными?



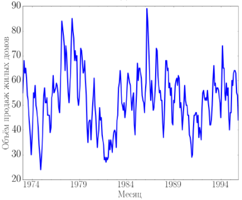
(a)



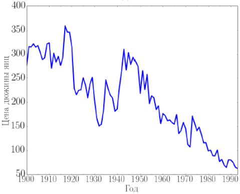
(b)



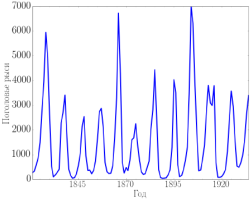
(c)



(d)



(e)



(f)

Существует множество критериев для проверки гипотезы о стационарности ряда. В Python реализован:

Критерий Дики-Фуллера

выборки:	$X_1, \ldots, X_N$ — временной ряд
нулевая гипотеза:	H_0 : ряд нестационарен
альтернатива:	H_1 : ряд стационарен
статистика:	некоторая
нулевое распределение:	табличное

Временные ряды

Рассмотрим теперь модели временных рядов.

Ранее была предпринята (неуспешная) попытка свести задачу к регрессии с признаками, которые зависели от времени линейно или квадратично.

Тривиальное обобщение: сделать регрессию не на признаки, зависящие от времени, а на собственные значения ряда в прошлом.

Временные ряды

Модель авторегрессии порядка p или AR(p):

$$X_t = \theta + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t,$$

где

- X_t — отклик;
- $X_{t-1}, \ldots, X_{t-p}$ — признаки;
- $\theta, \alpha_1, \ldots \alpha_p$ — параметры модели, которые необходимо оценить (это можно сделать с помощью МНК);
- ε_t — шум.

Временные ряды

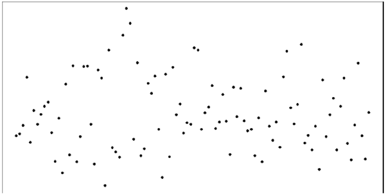
Модель скользящего среднего порядка q или MA(q):

$$X_t = \theta + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q},$$

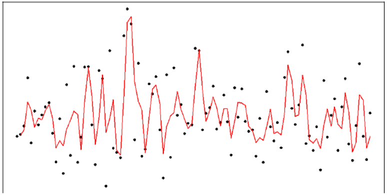
где

- X_t — отклик;
- $\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-p}$ — шум в моменты времени от $t-q$ до t;
- $\beta_1, \ldots \beta_q$ — параметры модели, которые необходимо оценить (это можно сделать с помощью МНК).

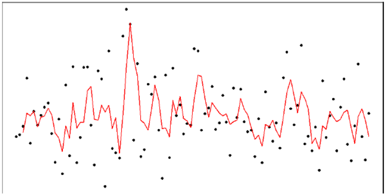
Временные ряды



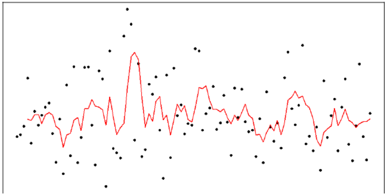
(a) Независимый, одинаково распределённый во времени случайный шум



(b) Среднее по двум соседним точкам



(c) Среднее по трём соседним точкам



(d) Среднее по четырём соседним точкам

Временные ряды

Модель ARMA порядка p, q или ARMA(p,q) получается объединением моделей $AR(p)$ и $MA(q)$:

$$\begin{aligned} X_t = \theta + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} \\ + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q}. \end{aligned}$$

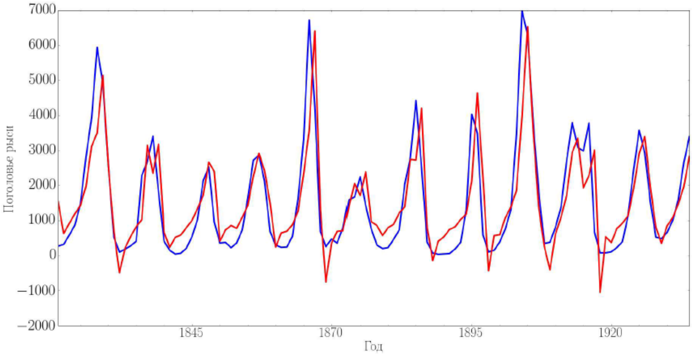
Временные ряды

Оказывается, любой стационарный временной ряд может быть хорошо описать моделью ARMA(p,q) с правильным подбором значений параметров p, q.

Отсюда возникает следующий алгоритм подгонки модели к временному ряду:

1. С помощью преобразований сделать ряд стационарным.
2. Подогнать модель ARMA(p,q) к стационарному ряду. Если это необходимо, сделать прогноз.
3. Применить обратные преобразования к построенной модели (и прогнозу).

Временные ряды



Как сделать ряд стационарным?

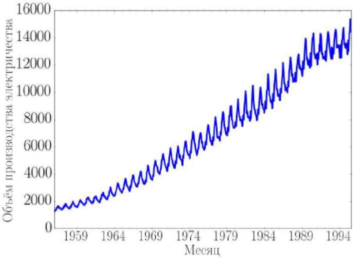
1. Стабилизация дисперсии

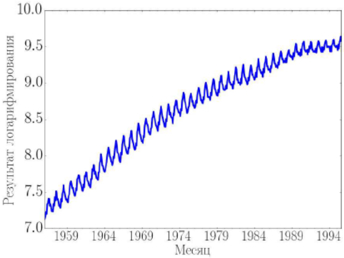
Среди множества возможных преобразований на практике хорошо работает параметрическое преобразование Бокса-Кокса:

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

где λ — это параметр преобразования.

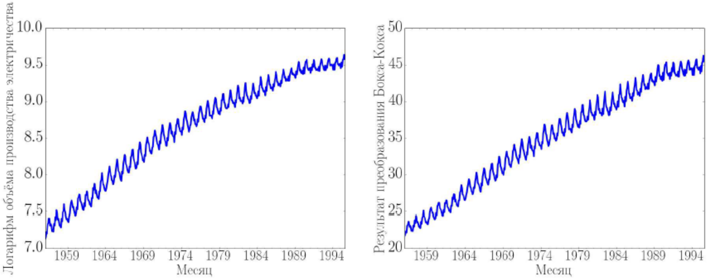
Временные ряды





Объем электричества, произведенного а Австралии, до и после преобразования Бокса-Кокса с $\lambda = 0$.

Временные ряды



Объем электричества, произведенного а Австралии,
до и после преобразования Бокса-Кокса с $\lambda = 0.27$.

Как сделать ряд стационарным?

2. Удаление тренда

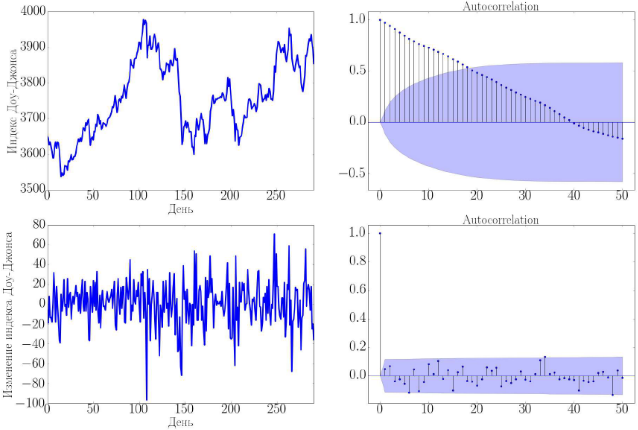
Чтобы убрать тренд, часто применяют дифференцирование — переход к попарным разностям соседних значений:

$$X_t' = X_t - X_{t-1}, \quad t = 2, \dots, N.$$

Эта операция уменьшает количество измерении на 1.

Дифференцирование можно применять неоднократно: от ряда первых разностей, продифференцировав его, можно прийти к ряду вторых разностей и так далее.

Временные ряды



Данные о значении индекса Доу-Джонса.
Сверху — исходный ряд, снизу — ряд после дифференцирования.

Как сделать ряд стационарным?

3. Удаление сезонности

Сезонность тоже можно убрать с помощью дифференцирования, но уже сезонного:

$$X'_t = X_t - X_{t-s}, \quad t = s+1, \dots, N,$$

где s — длина периода сезона.

Эта операция уменьшает количество измерении на s.

Временные ряды

Не всегда удается избавиться от сезонности в модели только с помощью дифференцирования. Поэтому часто к модели ARMA(p,q) добавляют еще сезонные компоненты.

Пусть ряд имеет сезонный период длины s.

Модель SARMA порядка p, q, P, Q или SARMA(p,q)x(P,Q) получается добавлением к модели ARMA(p,q):

- P авторегрессионных компонент с шагом, равным s:

$$+\alpha_s X_{t-s} + \alpha_{2s} X_{t-2s} + \ldots + \alpha_{Ps} X_{t-Ps},$$

- Q компонент скользящего среднего, также с шагом s:

$$+\beta_s \varepsilon_{t-s} + \beta_{2s} \varepsilon_{t-2s} + \ldots + \beta_{Qs} \varepsilon_{t-Qs}.$$

Временные ряды

Итак, после стабилизации дисперсии, обычного и сезонного дифференцирования, мы должны получить стационарный ряд, который должен хорошо описываться моделью SARMA(p,q)x(P,Q).

Еще немного специальных названий:

- ARIMA(p,d,q) — это модель ARMA(p; q) ряда, к которому d раз было применено обычное дифференцирование
- SARIMA(p,d,q)x(P,D,Q) — это модель SARMA(p,q)x(P,Q) для ряда, к которому d раз было применено обычное дифференцирование и D раз — сезонное.

  Иногда указывают еще длину сезонного периода: SARIMA(p,d,q)x(P,D,Q,s).

Как происходит обучение модели SARIMA(p,d,q)x(P,D,Q)?

Если параметры модели зафиксированы, то коэффициенты $\theta, \alpha_1, \dots, \alpha_p, \alpha_s, \dots, \alpha_{Ps}, \beta_1, \dots, \beta_q, \beta_s, \dots, \beta_{Qs}$ ищутся с помощью метода наименьших квадратов (МНК).

Единственный нюанс заключается в определении коэффициентов β_i, которые стоят при шумовых компонентах. Мы шум не наблюдаем, а для оценок МНК мы должны знать значения признаков. Поэтому сначала строится обычная авторегрессия и оценивается шум, а затем уже оцениваются все коэффициенты модели.

Временные ряды

Аналогично тому, как было в регрессии: если шум является одинаково распределённым и нормальным, то оценки метода наименьших квадратов являются оценками максимального правдоподобия, то есть обладают хорошими свойствами.

Как выбрать параметры SARIMA(p,d,q)x(P,D,Q)?

- Параметры d и D

  Порядки дифференцирования, необходимо подбирать так, чтобы ряд стал стационарным.

  Сезонное и обычное дифференцирование могут применяться к ряду в любом порядке. Но если у ряда есть ярко выраженный сезонный профиль, то лучше начинать с сезонного дифференцирования — уже после такого преобразования может оказаться, что ряд стационарен.

  Чем меньше порядки дифференцирования, тем лучше: с увеличением количества дифференцирований растёт дисперсия итогового прогноза.

Как выбрать параметры SARIMA(p,d,q)x(P,D,Q)?

- Параметры p, q и P, Q

  С параметрами p, q и P, Q ситуация такая же, как и в регрессии: чем больше эти параметры, тем сложнее модель. Мы обычно отдаем предпочтение простым моделям: сложные модели склонны «подстраиваться» под обучающую выборку и хуже работать на тестовой выборке (или делать хуже предсказания).

  Часто для параметров p, q и P, Q находят начальные приближения, а затем перебирают все значения меньше и выбирают те, которые минимизируют какой-то информационный критерий, например, критерий Акаике.

Информационный критерий Акаике: происходит минимизация следующей величины

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

где L — функция правдоподобия модели, а $k = P + Q + p + q + 1$ — число параметров.

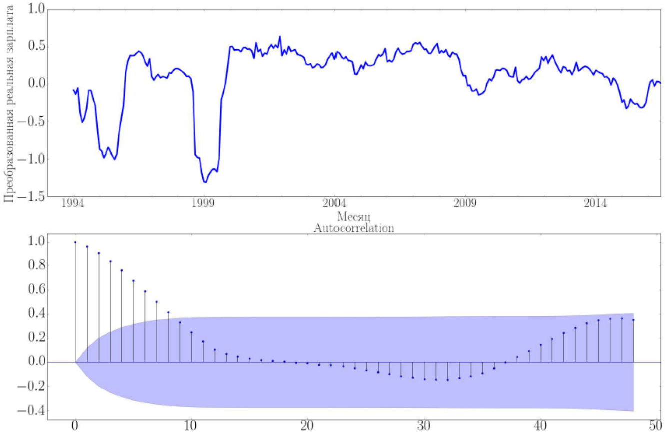
Оптимальной по критерию Акаике будет модель с наименьшим значением AIC. Такая модель, с одной стороны, будет достаточно хорошо описывать данные, а с другой — содержать не слишком большое количество параметров.

Временные ряды

Начальные значения для p, q и P, Q обычно выбирают так:

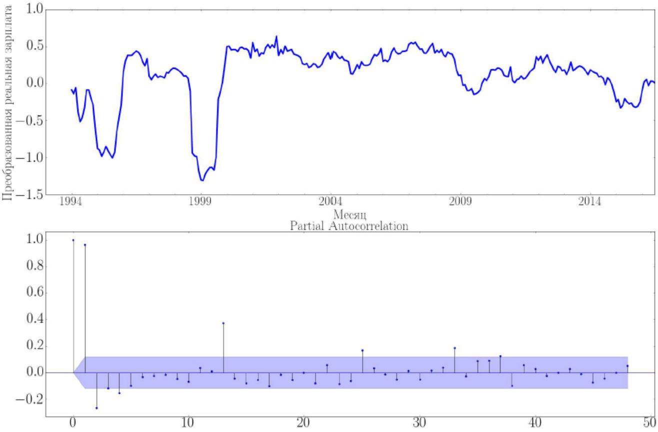
- начальное значение для $Q \cdot s$ — номер последнего сезонного лага, при котором автокорреляция значима;
- начальное значение для q — номер последнего несезонного лага, при котором автокорреляция значима;
- начальное значение для $P \cdot s$ — номер последнего сезонного лага, при котором частная автокорреляция значима;
- начальное значение для p — номер последнего несезонного лага, при котором частная автокорреляция значима.

Временные ряды



Сверху — ряд реальной з/п в России после преобразования Бокса-Кокса и сезонного дифференцирования, снизу — автокорреляционная функция этого ряда.

Временные ряды

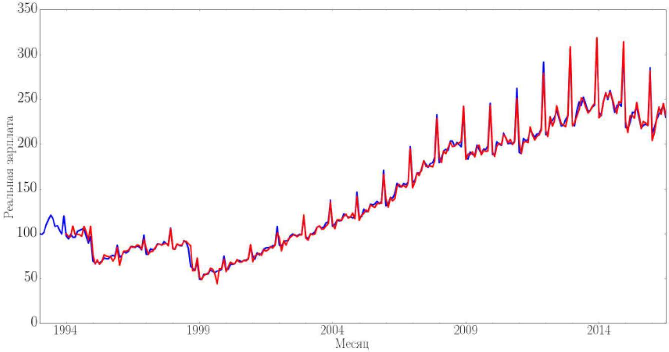


Сверху — тот же ряд реальной з/п в России после преобразований, снизу — частная автокорреляционная функция этого ряда.

График автокорреляции: сезонных лагов (s= 12) со значимой корреляцией нет, значит, начальное приближение для $Q = 0$; последняя значимая автокорреляция наблюдается у лага 8, поэтому начальное приближение для $q = 8$.

График частной автокорреляции: последний значимый сезонный лаг — 24, значит, начальное приближение для $P = 24/s = 2$; последний значимый несезонный лаг — 2, поэтому начальное приближение для $p = 2$.

Временные ряды



Модель SARIMA(2,0,1)x(2,1,2), которая, согласно критерию Акаике, является наилучшей для ряда реальной з/п в России.

Алгоритм подбора модели SARIMA(p,d,q)x(P,D,Q):

1. Визуальный анализ и предобработка.

Уже из визуального анализа можно сделать некоторые выводы: есть ли в данных сезонность, какой сезонный период, есть ли в ряде пропуски и выбросы, необходимо ли стабилизировать дисперсию, стоит ли исключить из рассмотрения начало ряда, потому что значения в начале совсем не похожи на значения в конце.

Алгоритм подбора модели SARIMA(p,d,q)x(P,D,Q):

2. Стабилизация дисперсии.

Если необходимо, стабилизируем дисперсию с помощью метода Бокса-Кокса или логарифмированием, что является частным случаем Бокса-Кокса.

Алгоритм подбора модели SARIMA(p,d,q)x(P,D,Q):

3. Дифференцирование.

Если исследуемый ряд нестационарен, то необходимо подобрать порядок дифференцирования, при котором он становится стационарным. Таким образом фиксируются параметры d и D SARIMA.

Начинаем дифференцирование с сезонного.

Алгоритм подбора модели SARIMA(p,d,q)x(P,D,Q):

4. Выбор начальных значений для p, q и P, Q.

Необходимо построить графики автокорреляционной функции (ACF) и частной автокорреляционной функции (PACF) и из этих графиков определить начальные приближения, с которых начинается перебор моделей SARIMA.

Алгоритм подбора модели SARIMA(p,d,q)x(P,D,Q):

5. Перебор и сравнение моделей для всех p, q и P, Q, которые меньше или равны начальным значениям.

Все модели необходимо обучить и сравнить по информационному критерию Акаике. Выбираем ту модель, которая минимизирует AIC.

Алгоритм подбора модели SARIMA(p,d,q)x(P,D,Q):

6. Анализ остатков.

Необходимо посмотреть на остатки получившейся модели, чтобы понять, насколько хорошей она получилась и можно ли её улучшить или нет. Мы сейчас обсудим, какими должны быть остатки.

Временные ряды

Аналогично тому, как было в регрессии, если остатки модели не удовлетворяют некоторым предположениям, то оценки коэффициентов могут не обладать такими теоретическими свойствами, как несмещенность, состоятельность и так далее.

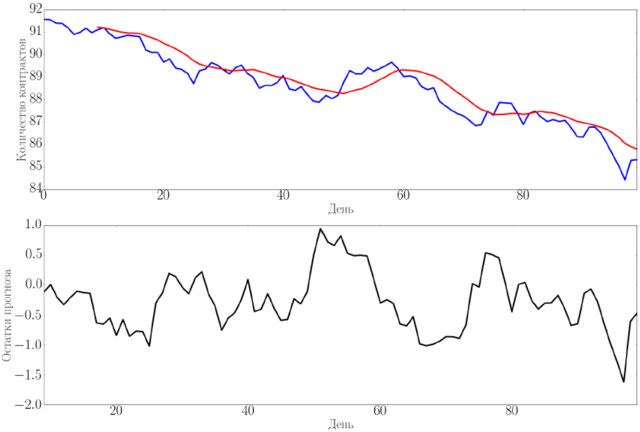
В анализе временных рядов важно обращать на следующие характеристики остатков: несмещенность, стационарность и неавтокоррелированность.

1. Несмещённость.

Остатки модели должны быть несмещенными, то есть их среднее должно быть равно нулю.

Лучше всего проверять несмещенность визуально, но можно и с помощью изученных критериев (знаков или Уилкоксона).

Временные ряды



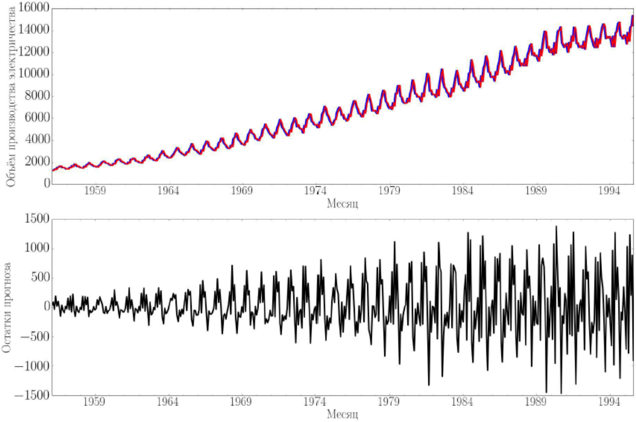
Количество контрактов в сокровищнице США и прогноз, выполненный методом скользящего среднего по 10 точкам. Снизу — остатки модели.

2. Стационарность.

Остатки модели должны быть стационарными, то есть не обладать какой-либо зависимостью от времени.

Лучше всего проверять стационарность визуально, но можно и с помощью критерия Дики-Фуллера.

Временные ряды



Количество произведённого в Австралии электричества и наивный прогноз (построенный по значению в предыдущий месяц). Снизу — остатки модели.

3. Неавтокоррелированность.

Остатки модели должны образовывать независимую выборку, то есть не обладать значимой автокорреляцией.

Лучше всего проверять отсутствие зависимости по графику автокорреляционной функции (ACF), но можно и с помощью Q-критерия Льюнга-Бокса.

Временные ряды

Q-критерий Льюнга-Бокса проверяет гипотезу о равенстве нулю одновременно нескольких автокорреляций. Если этот критерий применяется для остатков, то обычно проверяется гипотеза о том, что все автокорреляций незначимы.

Q-критерий Льюнга-Бокса

выборки:	$\varepsilon_1, \dots, \varepsilon_N$ — остатки модели
нулевая гипотеза:	H_0 : все автокорреляции равны нулю
альтернатива:	H_1 : есть значимые автокорреляции
статистика:	некоторая
нулевое распределение:	хи-квадрат

Временные ряды

Автокоррелированность остатков — признак того, что в данных присутствует информация, которая не вошла в модель. Если в остатках есть структура, то можно попытаться её внести в модель явным образом.

Но, конечно, это сделать можно далеко не всегда — возможности модели SARIMA не безграничны.

Таким образом, автокоррелированность остатков только указывает на потенциальную возможность улучшить модель.

Спасибо за внимание!