

Урок 1

Прогнозирование временных рядов

1.1. Временные ряды.

Этот урок посвящён классическим задачам, связанным с временными рядами. Задачи такого типа часто возникают в бизнес-аналитике.

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

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

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

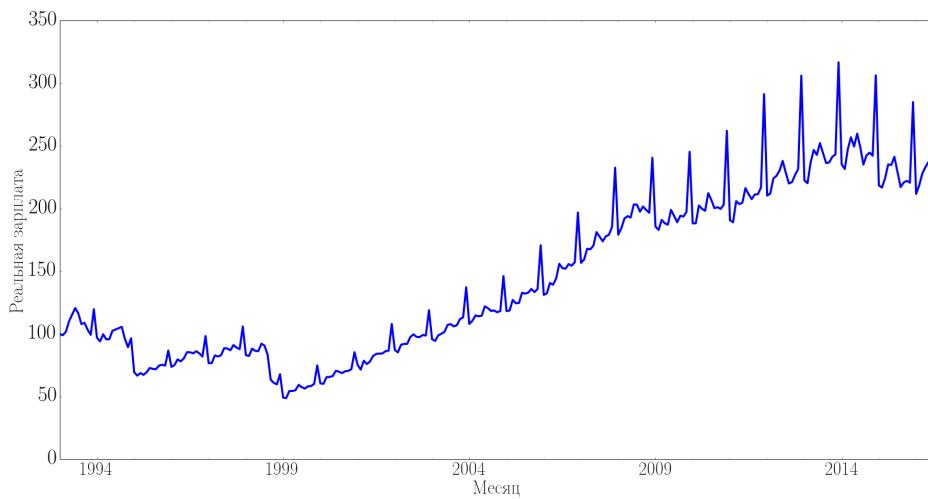


Рис. 1.1: Среднемесячная реальная заработная плата в России, выраженная в процентах от её значения в январе 1993 г.

Примеры временных рядов — это ряды среднедневных цен на акции определённой компании, среднемесячного уровня безработицы, измеренного в течение нескольких лет, среднегодового уровня производства автомобилей. Ещё один пример временного ряда (показан на рисунке 1.1) — это реальная заработная плата в России, выраженная в процентах от её значения на январь 1993 г., измеренная и усреднённая за каждый месяц, начиная с того момента.

1.1.1. Прогнозирование временного ряда.

Интерес представляет задача прогнозирования временных рядов. Подразумевается, что зная значение признака в прошлом, можно предсказать его в будущем. Формально задача ставится как поиск функции f_T :

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

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

1.1.2. Главная особенность временных рядов

До этого, на протяжении практически всей специализации, считалось, что анализируемые данные — это простые выборки, то есть независимые одинаково распределённые наблюдения. В задаче анализа временных рядов всё с точностью наоборот: предполагается, что данные в прошлом каким-то образом связаны с данными в будущем. Чем сильнее они связаны, тем больше имеется информации о поведении временного ряда в будущем и тем точнее можно сделать прогноз.

Полезно снова рассмотреть данные о реальной среднемесячной зарплате в России (рис. 1.1). Видно, что на графике изображена не простая выборка (измерения не являются независимыми и одинаково распределёнными), а сложный, структурированный процесс. Выявив структуру этого процесса, можно учесть её в прогнозирующей модели и построить действительно точный прогноз.

1.1.3. Применение модели регрессии

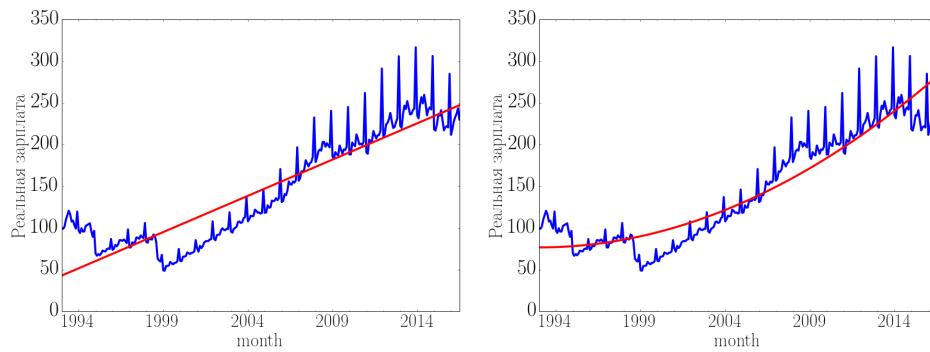


Рис. 1.2: Применение модели линейной (слева) и квадратичной (справа) регрессии к задаче прогнозирования временного ряда.

До этого в курсе большое внимание было уделено задаче обучения с учителем. Можно попробовать свести к ней задачу прогнозирования временного ряда. Процесс разворачивается во времени, поэтому кажется логичным задать признаки, связанные со временем и попробовать решить задачу, применяя модель регрессии. Регрессия может быть линейной или, например, квадратичной (рис. 1.2).

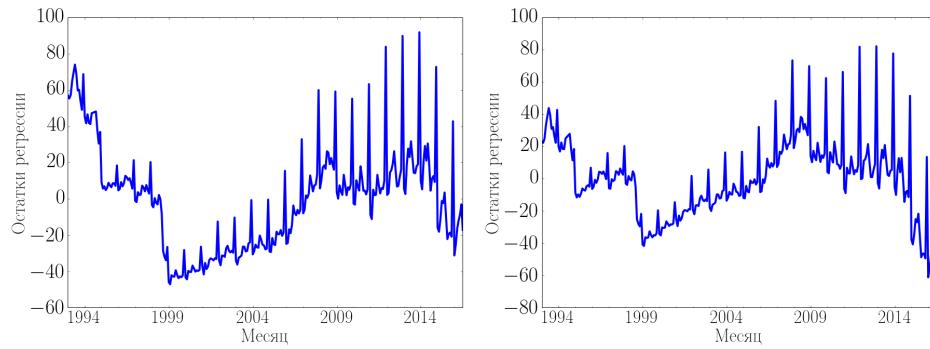


Рис. 1.3: Остатки модели линейной (слева) и квадратичной (справа) регрессии в задаче прогнозирования временного ряда.

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. Построению таких моделей будет посвящена оставшаяся часть урока.

1.1.4. Компоненты временных рядов

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

- Тренд — плавное долгосрочное изменение уровня ряда. Этую 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).

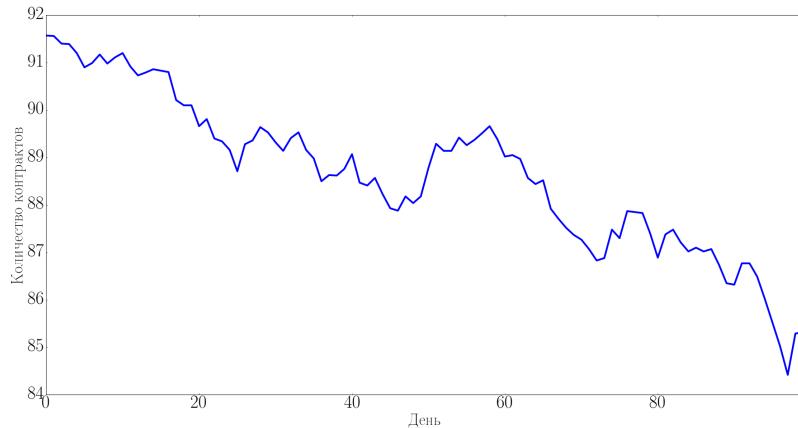


Рис. 1.4: Количество контрактов за день в сокровищнице США

В качестве примера временного ряда можно рассмотреть количество контрактов за день в сокровищнице США (рис. 1.4). На графике виден хорошо выраженный понижающийся тренд, который можно описать линейной функцией. На этом участке в данных не наблюдается ни циклов, ни сезонности. По-видимому, всё, что не удаётся описать трендом, является ошибкой.

На рисунке 1.5 показаны данные за несколько лет о суммарном объёме электричества, произведённого за месяц в Австралии. На графике, как и в предыдущем случае, виден тренд, на этот раз повышающийся. Кроме того, наблюдается годовая сезонность: значение признака совершает колебания, минимум которых всегда приходится на зиму, а максимум — на середину лета. Это легко объяснить тем, что зимой электричества необходимо меньше всего, это самый тёплый сезон в Австралии.

Следующий пример — суммарный объём проданной жилой недвижимости в Америке за месяц (рис. 1.6), данные так же собраны за несколько лет. На графике наблюдается сочетание двух основных компонент. Первая компонента — это годовая сезонность (минимум всегда приходится на зиму, а максимум — на середину лета), а вторая — это циклы, связанные с изменением среднего уровня экономической активности (период в данном случае составляет 7-9 лет).

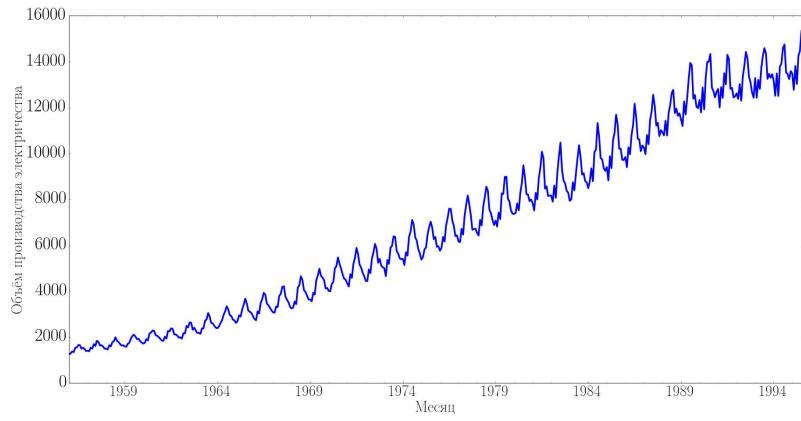


Рис. 1.5: Суммарный объём электричества, произведённого за месяц в Австралии

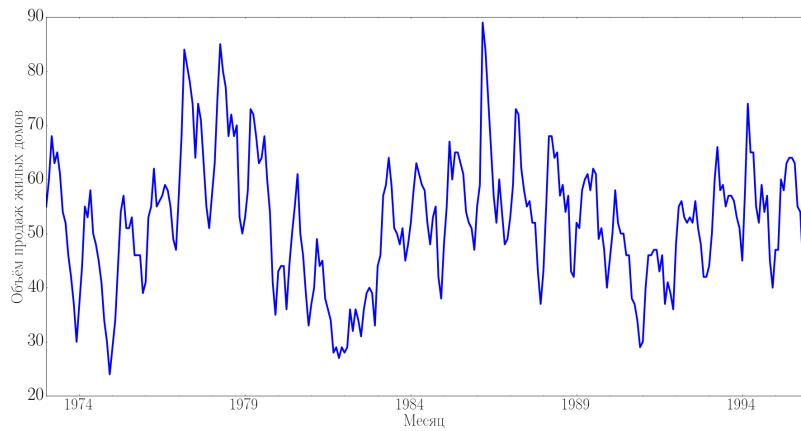


Рис. 1.6: Суммарный объем проданной жилой недвижимости (в млн кв. м.) в Америке за месяц

На рисунке 1.7 показаны ежедневные изменения индекса Доу-Джонса. Глядя на этот график, сложно сказать, присутствует ли в данных какая-то систематическая компонента: явно нет ни тренда, ни сезонности, ни цикла. По всей видимости, ряд представляет собой что-то похожее на случайную ошибку. Однако даже такие ряды можно прогнозировать.

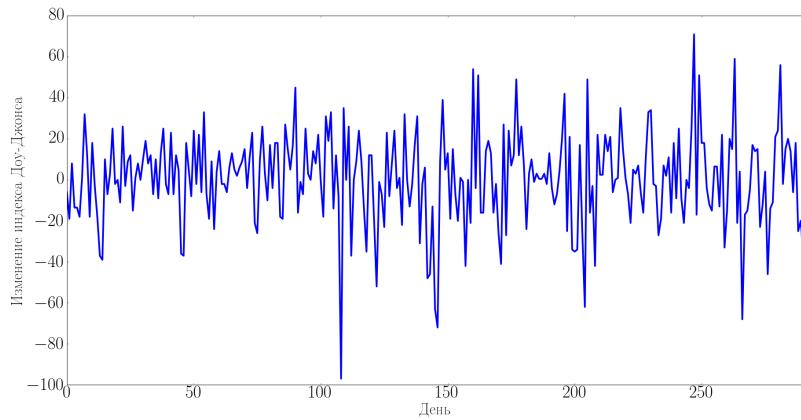


Рис. 1.7: Ежедневное изменение индекса Доу-Джонса

1.2. Автокорреляция

1.2.1. Пример: продажи вина в Австралии

Одной из важнейших характеристик временного ряда является автокорреляция. Далее суть этой характеристики будет демонстрироваться на примере данных о суммарном объёме продаж вина в Австралии за месяц на протяжении почти 15 лет (рис. 1.8).

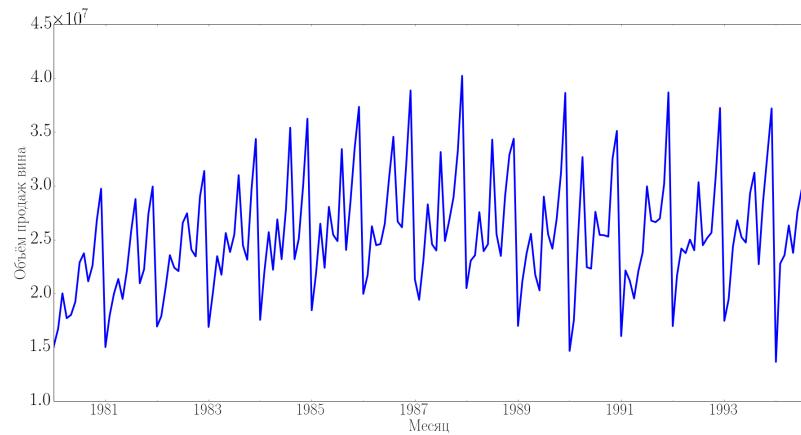


Рис. 1.8: Месячный объём продаж вина в Австралии, в бутылках

Этот ряд обладает ярко выраженной годовой сезонностью: максимум продаж за год приходится на декабрь, а затем, в январе, происходит существенное падение.

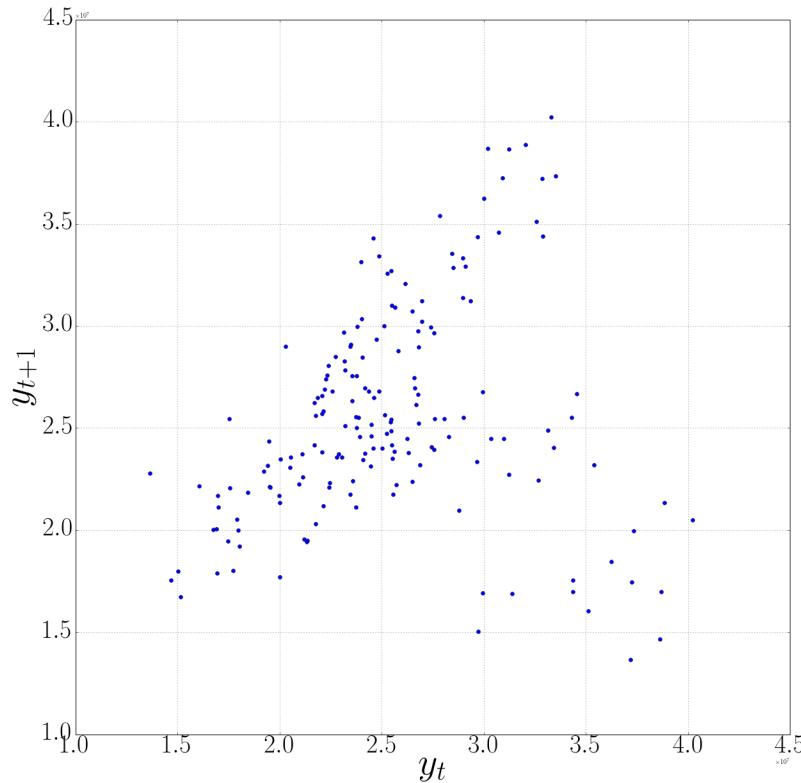


Рис. 1.9: Связь между значениями объёма продаж вина в соседние месяцы, по горизонтали отложен объём продаж в месяц t , по вертикали — в следующий месяц, $t + 1$, каждая точка задаёт продажи в 2 соседних месяца

На рисунке 1.9 показано, как связаны объёмы продаж вина в соседние месяцы. Видно, что большая часть точек на графике группируется вокруг главной диагонали. Это говорит о том, что в основном значения продаж в соседние месяцы похожи. Ещё одно подмножество точек выделяется в правом нижнем углу, оно связано с падением продаж от декабря к январю, которое было видно на предыдущем графике.

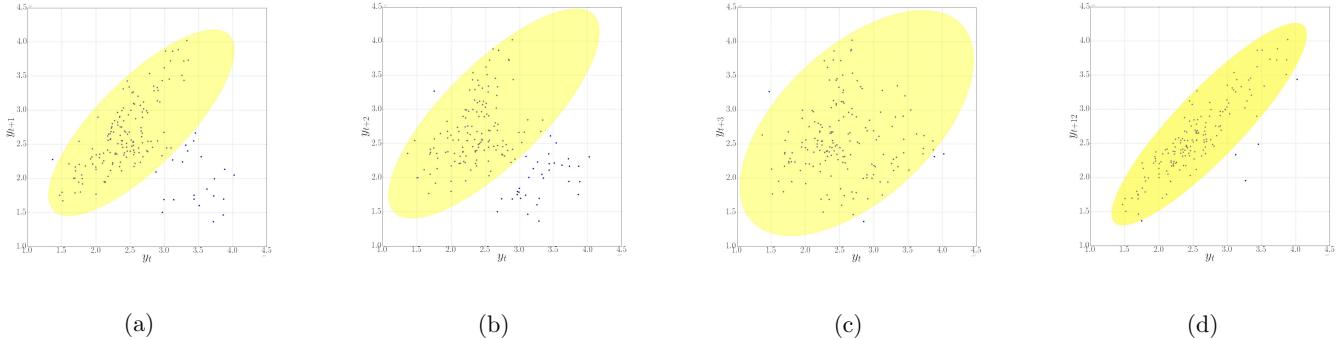


Рис. 1.10: Связь между продажами в соседние месяцы (а), через месяц (б), через два месяца (с) и через год (д).

Если построить аналогичный график, но по вертикальной оси отложить y_{t+2} (рис. 1.10б), то видно, что точки в основном облаке начинают "расплываться" вокруг главной диагонали, то есть сходство между продажами через месяц уменьшается по сравнению с соседними месяцами. Если посмотреть связь между продажами через два месяца (рис. 1.10с), то облако станет ещё шире, а сходство — ещё меньше. Однако если рассмотреть продажи в одни и те же месяцы соседних лет (рис. 1.10д), то видно, что точки на графике снова стягиваются к главной диагонали. Это значит, что значения продаж в одни и те же месяцы соседних лет очень сильно похожи.

1.2.2. Автокорреляция, её вычисление

Количественной характеристикой сходства между значениями ряда в соседних точках является автокорреляционная функция (или просто автокорреляция), которая задаётся следующим соотношением:

$$r_\tau = \frac{\mathbb{E}((y_t - \mathbb{E}y)(y_{t+\tau} - \mathbb{E}y))}{\mathbb{D}y}.$$

Автокорреляция — это уже встречавшаяся ранее корреляция Пирсона между исходным рядом и его версией, сдвинутой на несколько отсчётов. Количество отсчётов, на которое сдвинут ряд, называется лагом автокорреляции (τ). Значения, принимаемые автокорреляцией такие же, как и у коэффициента Пирсона: $r_\tau \in [-1, 1]$.

Вычислить автокорреляцию по выборке можно, заменив в формуле математическое ожидание на выборочное среднее, а дисперсию — на выборочную дисперсию:

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

1.2.3. Коррелограммы

Анализировать величину автокорреляции при разных значениях лагов удобно с помощью графика, который называется коррелограммой. По оси ординат на нём откладывается автокорреляция, а по оси абсцисс — размер лага τ . На рисунке 1.11а показан пример коррелограммы для исследуемых ранее данных о месячных продажах вина в Австралии (рисунок 1.8). На графике видно, что автокорреляция принимает большие значения в лагах, кратных сезонному периоду. Такой вид коррелограммы типичен для данных с выраженной сезонностью.

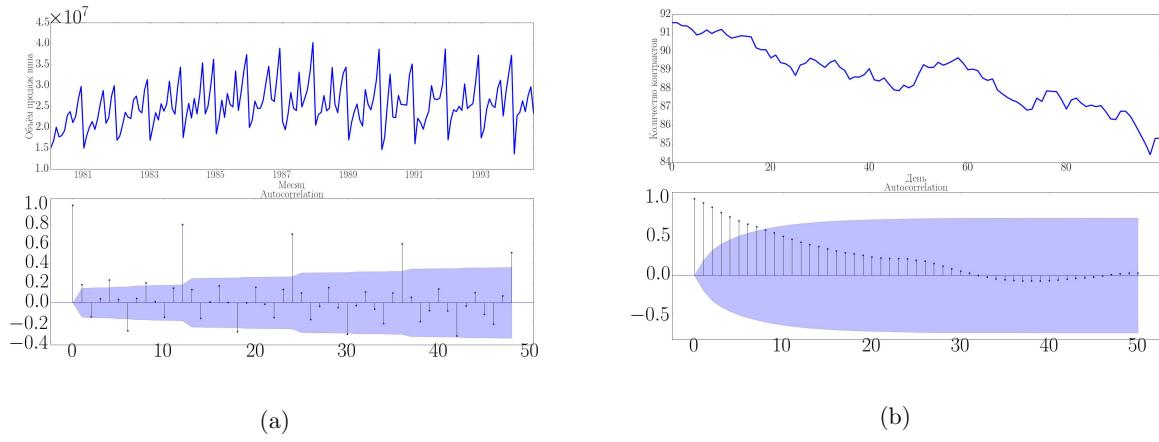


Рис. 1.11: Коррелограммы для временных рядов: (а) — количество проданного вина в Австралии за месяц, (б) — количество контрактов за день в сокровищнице США

На рисунке 1.11b показано, как выглядит коррелограмма для данных с ярко выраженным трендом. Автокорреляция тем больше, чем меньше величина лага τ , и с ростом τ она начинает постепенно убывать, при этом автокорреляция может начать колебаться вокруг горизонтальной оси, соответствующей её нулевому значению.

Коррелограмма, изображённая на рисунке 1.12a, построены для временного ряда, в котором присутствуют и тренд, и сезонность. Таким образом, на ней можно наблюдать оба описанных ранее эффекта, однако тренд настолько сильный, что практически нейтрализует влияние сезонности (следствие которой — наличие пиков в лагах, кратных периоду сезона).

На рисунке 1.12b показана типичная коррелограмма для ряда, в котором есть и сезонность, и цикл. Для самого первого лага, кратного сезонному периоду, виден пик, однако далее положение этого пика смещается: следующий пик не приходится на 2, 3 или 4 года. Это происходит, потому что в ряде есть циклы, период которых плавно меняется.

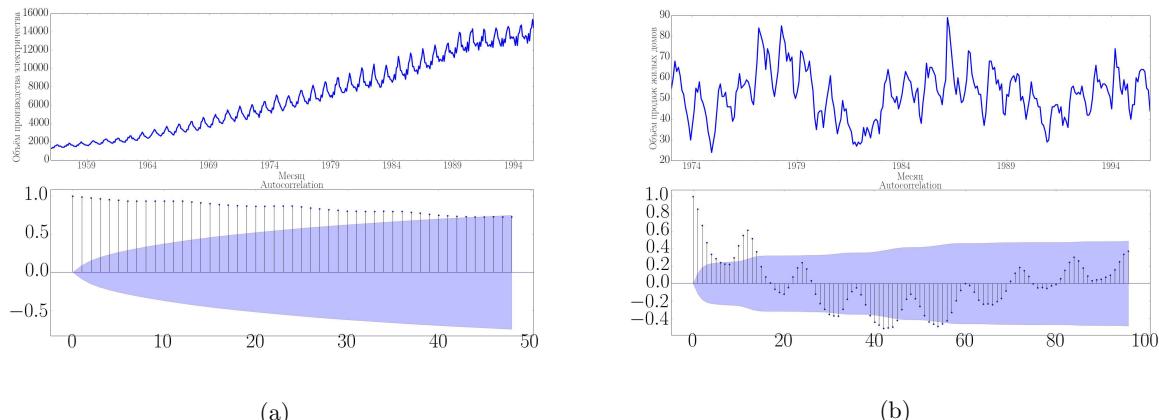


Рис. 1.12: Коррелограммы для временных рядов: (а) — ежемесячное производство электричества в Австралии, (б) — объём проданной в Америке недвижимости за месяц

На коррелограмме, соответствующей данным о ежедневном изменении индекса Доу-Джонса, все значения автокорреляции невелики, кроме первого (в данной точке лаг $\tau = 0$, и вычисляется корреляция значения ряда с самим собой, а такая корреляция всегда равна 1).

1.2.4. Значимость автокорреляции

На всех показанных коррелограммах помимо значений автокорреляции также изображен синий коридор вокруг горизонтальной оси. Это коридор значимости отличия корреляции от нуля. Как правило, его выводят

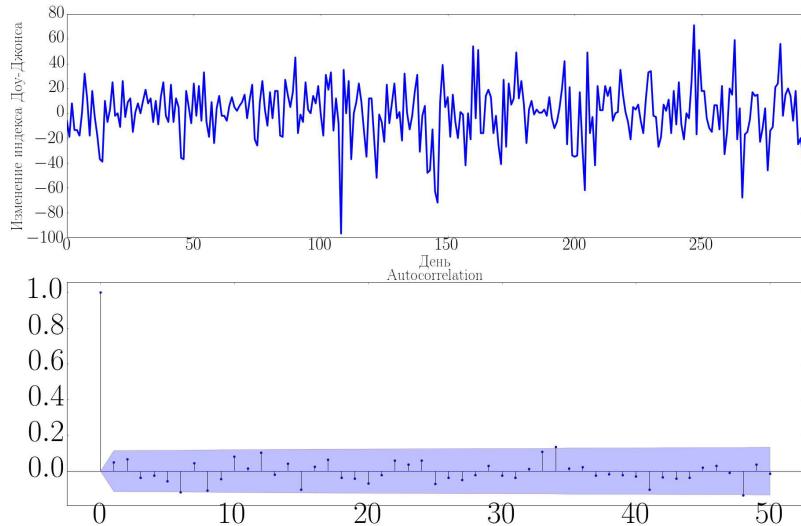


Рис. 1.13: Коррелограмма для данных о ежедневном изменении индекса Доу-Джонса.

на график все стандартные средства для работы с временными рядами. Фактически, все автокорреляции, которые изображены вне этого коридора, значимо отличаются от нуля. Как и для обычной корреляции Пирсона, значимость вычисляется с помощью критерия Стьюдента (таблица 1.1). Альтернатива чаще всего двусторонняя, потому что при анализе временных рядов крайне редко имеется гипотеза о том, какой должна быть корреляция, положительной или отрицательной.

временной ряд:	$y^T = y_1, \dots, y_T$
нулевая гипотеза:	$H_0: r_\tau = 0$
альтернатива:	$H_1: r_\tau < \neq > 0$
статистика:	$T(y^T) = \frac{r_\tau \sqrt{T-\tau-2}}{\sqrt{1-r_\tau^2}}$
нулевое распределение:	$T(y^T) \sim St(T-\tau-2)$.

Таблица 1.1: Описание статистического критерия Стьюдента

Вернувшись к коррелограмме по данным о ежедневном изменении индекса Доу-Джонса, теперь можно заметить, что ни одна из корреляций не выходит за пределы коридора значимости, а значит ни одна из них не является значимо отличающейся от нуля.

1.3. Стационарность

1.3.1. Понятие стационарности временного ряда

Ещё одно важное свойство временных рядов — это стационарность. Временной ряд y_1, \dots, y_T называется стационарным, если $\forall s$ (ширина окна) распределение y_t, \dots, 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.

В качестве примера стационарных и нестационарных можно рассмотреть временные ряды, показанные на рисунке 1.14. Ряды 1.14a, 1.14c, 1.14e, 1.14f, 1.14i не являются стационарными из-за довольно выраженного тренда. В рядах 1.14d, 1.14h, 1.14i сильно выражена сезонность, поэтому они также не стационарны. В ряду 1.14i, помимо всего прочего, меняется и дисперсия (размах колебаний в начале ряда намного меньше, чем в конце), то есть присутствует ещё одно свойство, не постоянное по времени. Остаются два ряда, 1.14b и 1.14g.

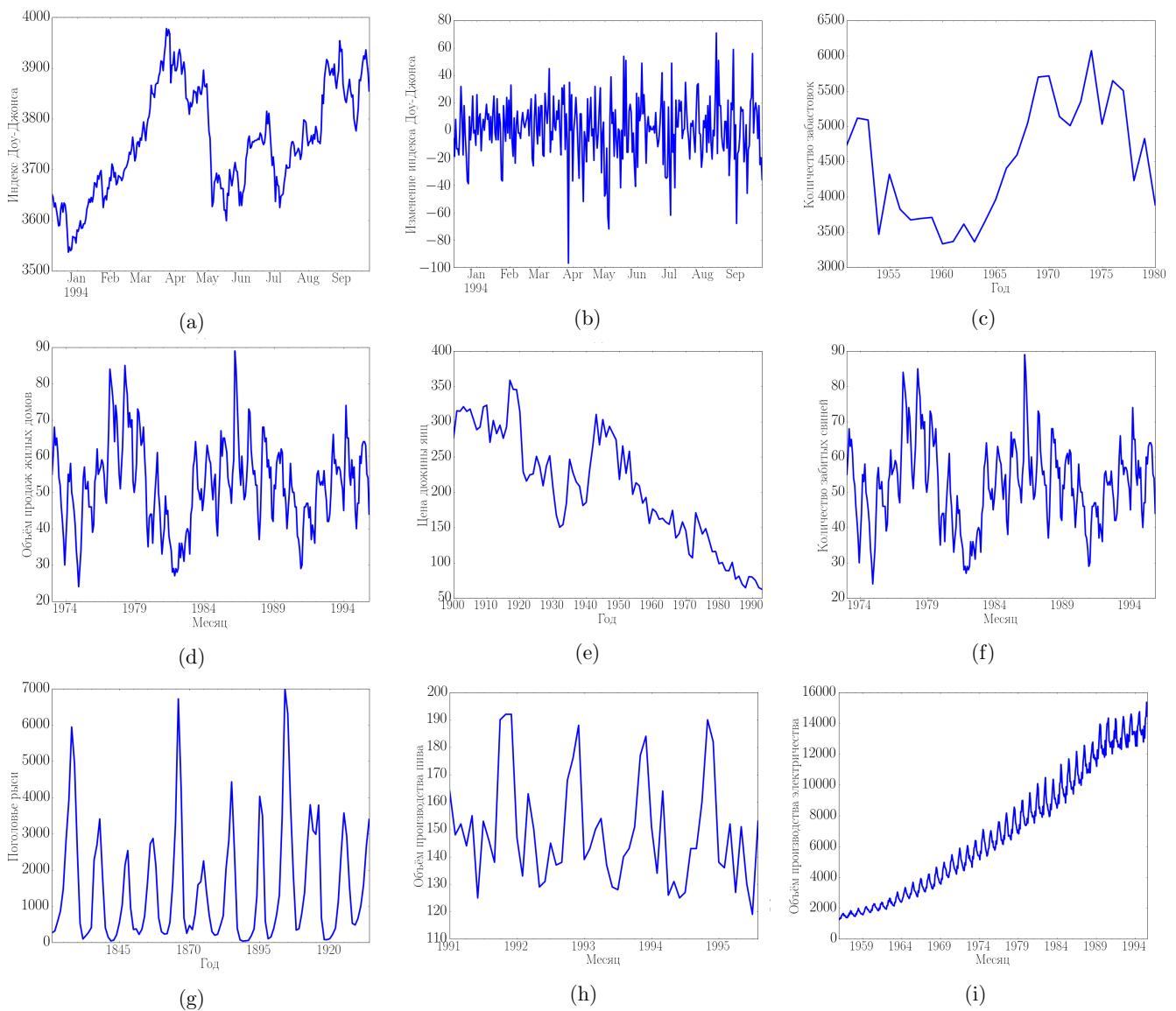


Рис. 1.14: Примеры временных рядов

Первый — это ежедневное изменение индекса Доу-Джонса, этот ряд считается стационарным, а второй — это размер поголовья рыси. Колебания во втором ряду имеют нефиксированный период, то есть это ряд с циклами, значит, он может считаться стационарным.

1.3.2. Критерий Дики-Фуллера

Формально гипотезу о стационарности можно проверить с помощью критерия Дики-Фуллера (таблица 1.2). Статистика данного критерия выглядит достаточно сложно и пока что рассматриваться не будет.

Вообще говоря, существует большое количество критериев для проверки гипотезы о стационарности, на практике можно использовать любой из них. Однако в этом курсе внимание акцентируется на критерии Дики-Фуллера, потому что для него существует реализация в языке Python.

1.3.3. Стабилизация дисперсии

При работе с нестационарными временными рядами используется ряд стандартных трюков, чтобы сделать их стационарными. В случае, если во временном ряде монотонно по времени изменяется дисперсия, применяется специальное преобразование, стабилизирующее дисперсию. Очень часто в качестве такого преобразования

временной ряд:	$y^T = y_1, \dots, y_T$
нулевая гипотеза:	H_0 : ряд нестационарен
альтернатива:	H_1 : ряд стационарен
статистика:	неважно
нулевое распределение:	табличное

Таблица 1.2: Описание статистического критерия Дики-Фуллера

ния выступает логарифмирование. Результат стабилизации дисперсии для ряда производства электричества в Австралии показан на рисунке 1.15. Видно, что после логарифмирования размах колебаний в начале и конце ряда становится очень похожим, и дисперсия примерно стабилизируется.

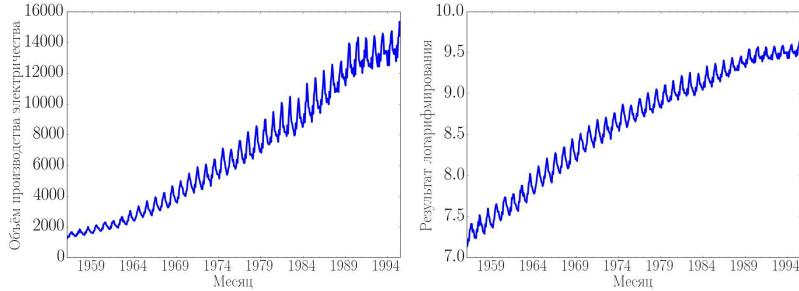


Рис. 1.15: Данные о производстве электричества в Австралии до и после логарифмирования.

Логарифмирование принадлежит к семейству преобразований Бокса-Кокса.

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

Это параметрическое семейство функций, в котором параметр λ определяет, как именно будет преобразован ряд: $\lambda = 0$ — это логарифмирование, $\lambda = 1$ — тождественное преобразование ряда, а при других значениях λ — степенное преобразование. Значение параметра можно подбирать так, чтобы дисперсия была как можно более стабильной во времени. Так, для ряда по данным производства электричества в Австралии оптимальное значение $\lambda = 0.27$, при этом дисперсия немного более стабильна, чем при логарифмировании (рис. 1.16).

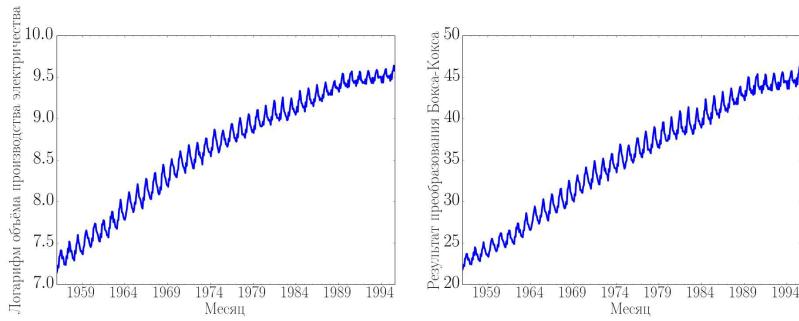


Рис. 1.16: Преобразованный временной ряд по данным о производстве электричества в Австралии. Слева — результат логарифмирования, справа — преобразование Бокса-Кокса с параметром $\lambda = 0.27$.

1.3.4. Дифференцирование

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

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

Для нестационарного ряда часто оказывается, что получаемый после дифференцирования ряд является стационарным. Такая операция позволяет стабилизировать среднее значение ряда и избавиться от тренда, а иногда даже от сезонности. Кроме того, дифференцирование можно применять неоднократно: от ряда первых разностей, продифференцировав его, можно прийти к ряду вторых разностей, и т. д. Длина ряда при этом каждый раз будет немного сокращаться, но при этом он будет стационарным.

Также может применяться сезонное дифференцирование ряда, переход к попарным разностям значений в соседних сезонах. Если длина периода сезона составляет s , то новый ряд задаётся разностями

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

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

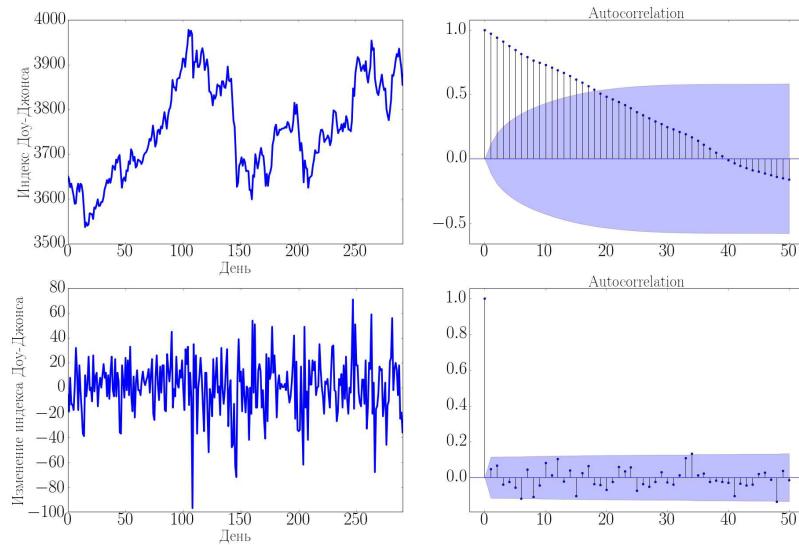


Рис. 1.17: Данные о значении индекса Доу-Джонса и соответствующая коррелограмма. Сверху — исходный ряд, снизу — ряд после дифференцирования.

На верхних графиках на рисунке 1.17 показаны ряд значений индекса Доу-Джонса и его автокорреляционная функция. Видно, что этот ряд достаточно сильно нестационарен — имеется ярко выраженный тренд. От этого тренда удается полностью избавиться, продифференцировав ряд. Снизу на рисунке 1.17 показан ряд после дифференцирования (это ежедневное изменение индекса, и этот ряд уже встречался ранее). Критерий Дики-Фуллера подтверждает, что новый ряд, полученный дифференцированием, является стационарным. Нулевая гипотеза о нестационарности этого ряда отвергается ($p = 5.2 \times 10^{-29}$). Для исходного ряда отвергнуть нулевую гипотезу не удается ($p = 0.3636$).

1.4. ARMA

Первый класс прогнозирующий моделей, который будет разбираться в этом курсе, — это модели ARMA.

1.4.1. Авторегрессия

Ранее была предпринята попытка свести задачу прогнозирования временного ряда к задаче обучения с учителем: предсказывать значения ряда с помощью регрессии, выбирая какие-то признаки, зависящие от времени (например, линейный или квадратичный тренд, рис. 1.2). Результат получился плохим, выбранных признаков явно недостаточно, нужны дополнительные.

Можно перейти к следующей идеи: делать регрессию для ряда не на какие-то внешние признаки, зависящие от времени, а на его собственные значения в прошлом:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t.$$

В этом регрессионном уравнении y_t — это отклик, $y_{t-1}, y_{t-2}, \dots, y_{t-p}$ — признаки, $\alpha, \phi_1, \phi_2, \dots, \phi_p$ — параметры модели, которые необходимо оценить, ε_t — шумовая компонента, описывает отклонения значений ряда от данного уравнения.

Такая модель называется моделью авторегрессии порядка p ($AR(p)$). В этой модели y_t представляет собой линейную комбинацию p предыдущих значений ряда и шумовой компоненты.

1.4.2. Скользящее среднее

Следующий класс моделей — это скользящее среднее. Чтобы лучше понимать, как они устроены, можно начать с рассмотрения независимого, одинаково распределённого во времени шума ε_t (рис. 1.18a).

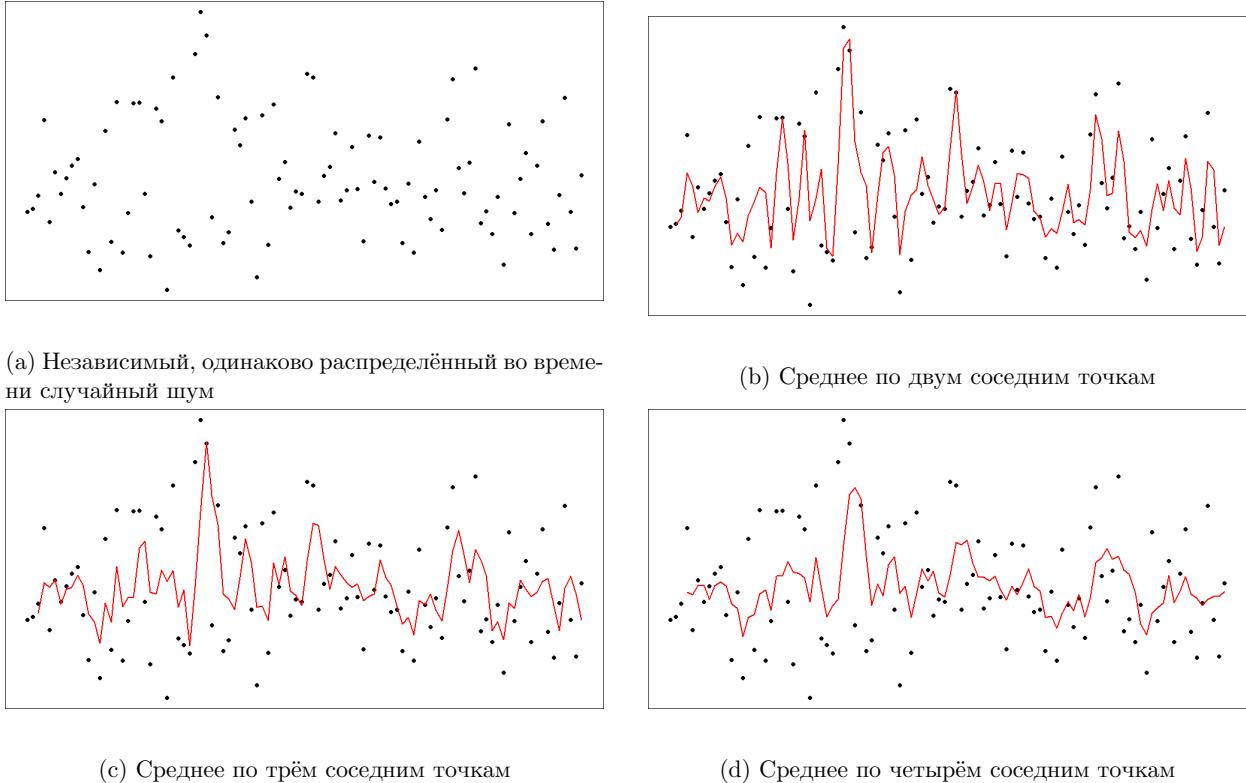


Рис. 1.18

Для каждого значения t можно вычислить среднее арифметическое между точками ε_t и ε_{t-1} (рис. 1.18b). Также можно вычислять среднее не по двум, а по трём (рис. 1.18c) или четырём (рис. 1.18d) точкам. То, что получается в результате такого усреднения, — это уже не простая выборка с независимыми, одинаково распределёнными элементами. Соседние значения на красной линии очень похожи друг на друга, потому что в их вычислении используются одни и те же шумовые компоненты.

Данную идею можно обобщить и записать следующую модель ряда:

$$y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q},$$

где $\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-q}$ — значения шума в q предыдущих моментов времени, $\alpha, \theta_1, \theta_2, \dots, \theta_q$ — это параметры модели, которые необходимо оценить. Такая модель называется моделью скользящего среднего порядка q ($MA(q)$). В ней предполагается, что значение ряда y_t — это линейная комбинация q последних значений шумовой компоненты.

Данная модель выглядит достаточно странно: шумовая компонента — это что-то, что невозможно наблюдать, и непонятно, как эту модель обучать, и зачем она нужна. Ответы на эти вопросы будут даны далее.

1.4.3. 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:

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

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 . Это прекрасный результат, который означает, что семейство моделей $ARMA(p, q)$ достаточно богато для того, чтобы в нём можно было найти хорошую модель, описывающую любой стационарный ряд.

1.4.4. Пример

Для демонстрации работы модели $ARMA(p, q)$ можно рассмотреть данные о поголовье рыси (рисунок 1.19). Этот пример уже встречался ранее, и было показано, что ряд стационарен, а значит в классе $ARMA(p, q)$ для него можно найти достаточно хорошее описание.

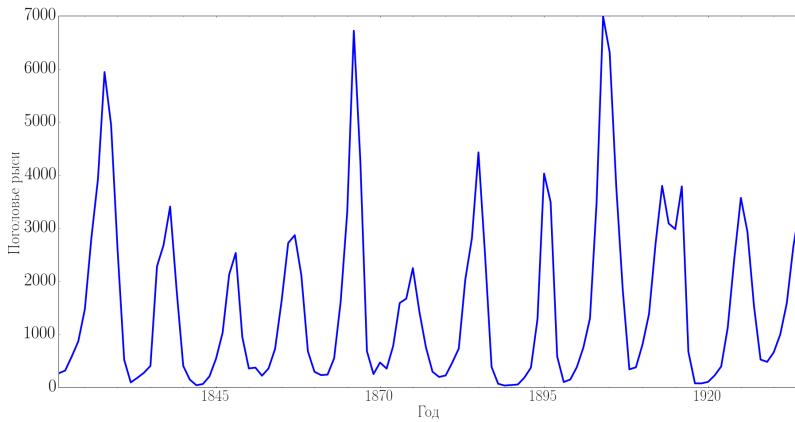


Рис. 1.19: Данные о размере поголовья рыси.

Действительно, модель $ARMA(2, 2)$ даёт результат, который достаточно сильно похож на исходный ряд. Модель не во всех точках близка к истинному значению ряда, однако результат всё равно намного лучше, чем если бы для приближения использовалась регрессия на линейный или квадратичный временной тренд.

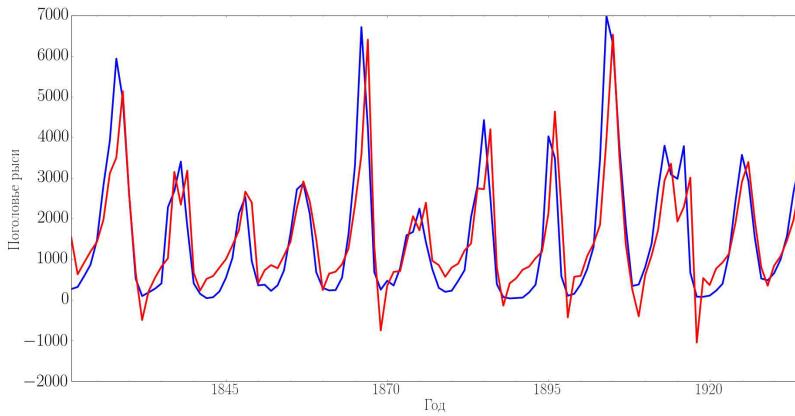


Рис. 1.20: Результат применения модели $ARMA(2, 2)$ к данным о поголовье рыси.

Настроив модель $ARMA(2, 2)$, её можно использовать и для построения прогноза (рис. 1.21), то есть решения той задачи, которая была изначально поставлена. Однако вопрос о том, какую модель использовать для прогнозирования нестационарного временного ряда, остаётся открытым.

1.5. ARIMA

Модели типа ARIMA — это обобщение модели класса ARMA.

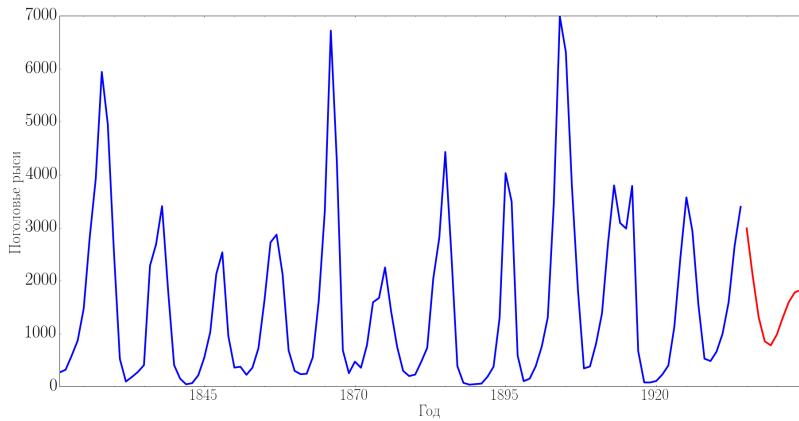


Рис. 1.21: Результат применения модели $ARMA(2, 2)$ для получения прогноза

На данный момент известны два факта:

1. Теорема Вольда: любой стационарный ряд может быть описан моделью $ARMA(p, q)$ с любой наперёд заданной точностью.
2. При помощи дифференцирования нестационарный ряд можно сделать стационарным.

Эти две идеи и лежат в основе моделей класса ARIMA. Модель $ARIMA(p, d, q)$ — это модель $ARMA(p, q)$ для d раз продифференцированного ряда.

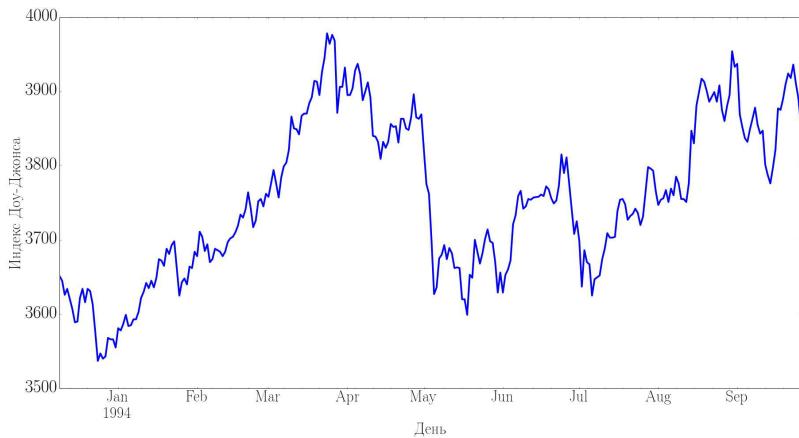


Рис. 1.22: Данные о значении индекса Доу-Джонса

На рисунке 1.22 показаны 300 значений индекса Доу-Джонса. Ранее было показано, что ряд нестационарен (и это видно невооружённым глазом), но зато стационарен ряд его первых разностей. Из этого следует, что для ряда разностей можно подобрать достаточно хорошую модель в классе ARMA. Если сделать это, а затем произвести операцию, обратную дифференцированию, то в результате будет получена модель ARIMA для исходного ряда.

Модель $ARIMA(0, 1, 0)$ для значения индекса Доу-Джонса показана на рисунке 1.23. В этой модели происходит одно дифференцирование и не используется ни одной компоненты авторегрессии и скользящего среднего, и это немного странно. Ряд разностей моделируется константой, но после проведения операции, обратной дифференцированию, полученный результат не является константой. В этой модели много странностей, но результат в любом случае лучше, чем то, что можно было бы получить с помощью регрессии ряда на временные признаки.

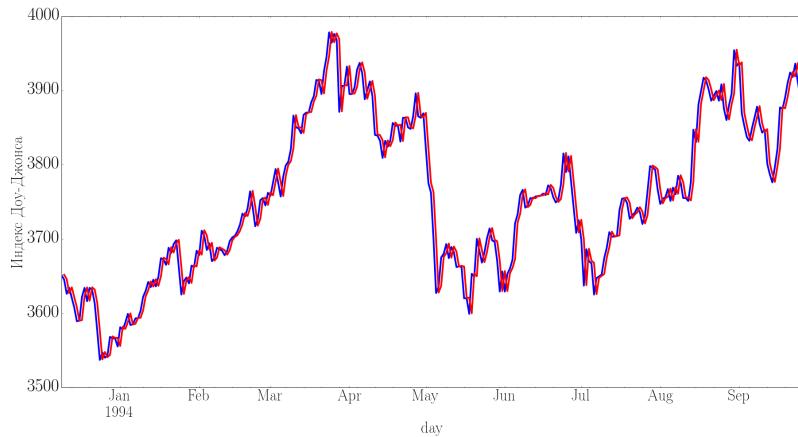


Рис. 1.23: Модель $ARIMA(0, 1, 0)$, применённая к ряду значений индекса Доу-Джонса.

1.5.1. 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:

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

Результат — это модель $SARMA(p, q) \times (P, Q)$.

1.5.2. SARIMA

Модель $SARIMA(p, d, q) \times (P, D, Q)$ — модель $SARMA(p, q) \times (P, Q)$ для ряда, к которому d раз было применено обычное дифференцирование и D раз — сезонное. Такую модель часто называют просто ARIMA: первая буква не пишется, но подразумевается, что сезонная компонента тоже может быть (такая терминология будет встречаться и дальше в курсе).

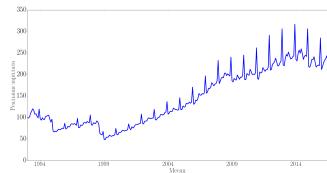
1.5.3. Пример

Для демонстрации рассмотренных моделей будет использоваться временной ряд реальной заработной платы в России (рис. 1.24a). Критерий Дики-Фуллера не отвергает гипотезу о том, что этот ряд нестационарный ($p = 0.2265$). Это неудивительно, потому что видно, что многие параметры этого ряда меняются во времени. Во-первых, меняется дисперсия: разброс скачков в начале совсем не такой, как ближе к концу.

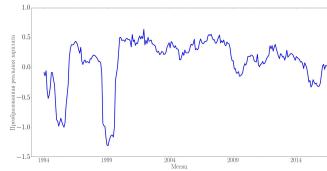
Ряд после применения преобразования Бокса-Кокса с параметром $\lambda = 0.22$ показан на рисунке 1.24b. Критерий Дики-Фуллера всё ещё не отвергает для этого ряда гипотезу о нестационарности ($p = 0.1661$). Это можно объяснить наличием в ряду сезонности и тренда.

После применения к ряду сезонного дифференцирования (рис. 1.24c) критерий Дики-Фуллера отвергает гипотезу о нестационарности ($p = 0.01$). Относительно этого ряда можно говорить, что он стационарный, а значит, можно попытаться подобрать для него модель в классе ARMA или даже сезонную модель. Если после этого провести обратные преобразования к преобразованию Бокса-Кокса и сезонному дифференцированию, модель может выглядеть, например, как на рисунке 1.24d. Красная линия на графике — это предсказание модели, видно, что она достаточно хорошо описывает исходные данные, а значит, можно надеяться, что и прогнозы она будет давать хорошие.

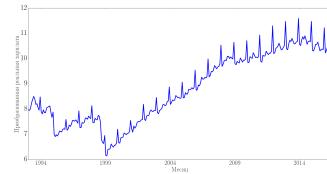
При применении регрессии с линейным или квадратичным трендом по времени в остатках этой модели было видно достаточно много структуры (рис. 1.3), а значит, что в данных оставалось много информации, которую не учитывает модель.



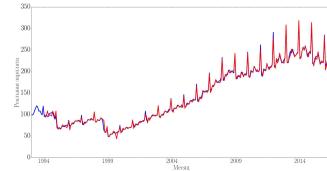
(a) Реальная месячная заработная плата в России.



(c) Ряд реальной месячной заработной платы в России после применения сезонного дифференцирования.



(b) Ряд реальной месячной заработной платы в России после применения преобразования Бокса-Кокса.



(d) Синим — ряд реальной месячной заработной платы, красным — модель SARIMA(2, 0, 2) x (0, 1, 0) с преобразованием Бокса-Кокса.

Рис. 1.24

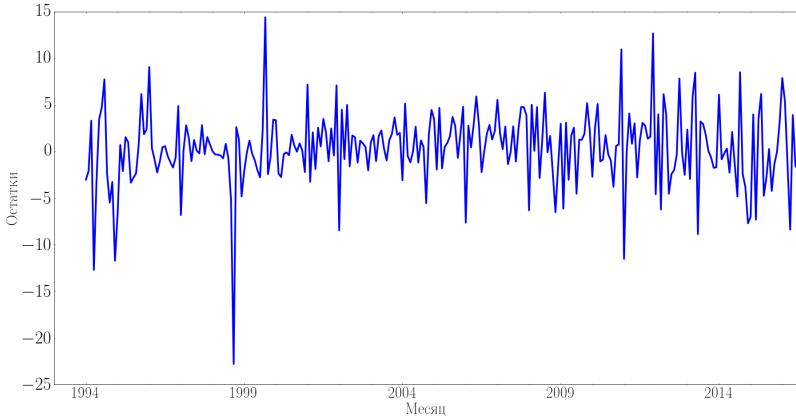


Рис. 1.25: Остатки построенной модели SARIMA.

Остатки для построенной модели SARIMA показаны на рисунке 1.25. Они уже гораздо больше похожи на белый шум. Выброс в остатках — это кризис 1998 года, который плохо описывается построенной моделью. Тем не менее, в этих остатках уже практически не имеется структуры, а значит, полученный результат лучше, чем при использовании линейной регрессии.

1.6. Выбор ARIMA и прогнозирование

После изучения устройства моделей класса ARIMA настало время разобраться, как настраивать эти модели и получать с их помощью прогнозы.

1.6.1. Подбор параметров

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 на-

страивание привычной регрессии методом минимизации квадратичной ошибки.

Единственный трюк заключается в определении коэффициентов θ , которые стоят при шумовых компонентах из прошлого. Наблюдать шумовые компоненты невозможно, поэтому, чтобы подставить их в регрессионное уравнение, их нужно предварительно оценить. Обычно оценка производится с помощью остатков от авторегрессии, которая предварительно строится по исследуемым данным.

Если шум, который стоит в модели ARIMA, является белым (независимый, одинаково распределённый, гауссовский), то метод наименьших квадратов даёт оценки максимального правдоподобия для параметров α, ϕ, θ , то есть заведомо известно, что эти оценки хороши.

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. Дело в том, что выгодно дифференцировать ряд как можно меньше раз, потому что с увеличением количества дифференцирований растёт дисперсия итогового прогноза.

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. Чем больше значения гиперпараметров, тем больше параметров в модели и тем она сложнее. Таким образом, с увеличением значения этих гиперпараметров значение правдоподобия может только увеличиваться. Поэтому для сравнения моделей с разным количеством параметров необходим другой критерий.

В качестве искомого критерия можно использовать, например, критерий 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.

В конечном итоге значения параметров q, Q, p, P определяются перебором: из разных значений гиперпараметров выбираются те, у которых значение критерия Акаике будет минимальным. Начальные приближения для этого перебора можно выбрать с помощью автокорреляционной функции.

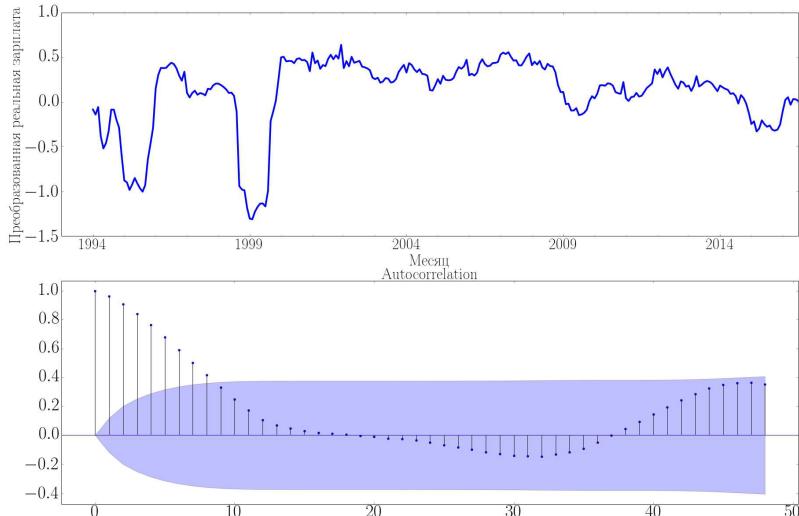


Рис. 1.26: Сверху — ряд реальной заработной платы в России после преобразования Бокса-Кокса и сезонного дифференцирования, снизу — автокорреляционная функция этого ряда.

Принцип подбора параметров можно продемонстрировать на данных о реальной заработной плате в России (1.26). Начальное значение для параметра $Q * S$ даёт номер последнего сезонного лага, при котором автокорреляция значима. В рассматриваемом примере сезонных лагов со значимой корреляцией нет, значит, начальное приближение $Q = 0$. Параметр q задаётся номером последнего несезонного лага, при котором автокорреляция значима. В данном случае можно взять начальное значение $q = 8$.

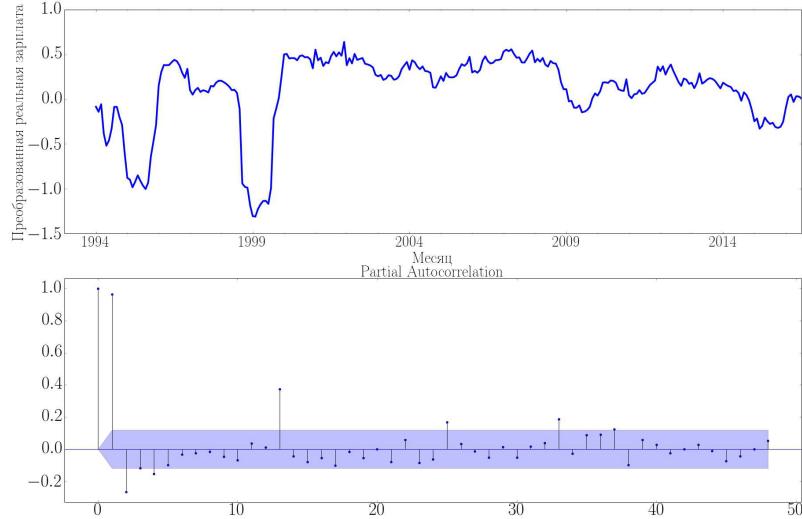


Рис. 1.27: Сверху — ряд реальной заработной платы в России после преобразования Бокса-Кокса и сезонного дифференцирования, снизу — частичная автокорреляционная функция этого ряда.

Значения параметров p, P подбираются с использованием не автокорреляционной функции, а частичной автокорреляционной функции (рис. 1.27). Частичная автокорреляция — это автокорреляция после снятия авторегрессии предыдущего порядка. Например, чтобы подсчитать частичную автокорреляцию с лагом $\tau = 2$, требуется построить авторегрессию порядка 1, вычесть эту авторегрессию из ряда и подсчитать автокорреляцию на полученных остатках.

Начальное приближение для параметра $P * S$ задаёт номер последнего сезонного лага, при котором частичная автокорреляция значима. В данных из примера это лаг под номером 24, значит начальное приближение $P = 2$, поскольку длина сезонного периода $S = 12$. Аналогично, p задаётся как номер последнего несезонного лага, при котором частичная автокорреляция значима. В данном случае можно взять начальное приближение $p = 2$.

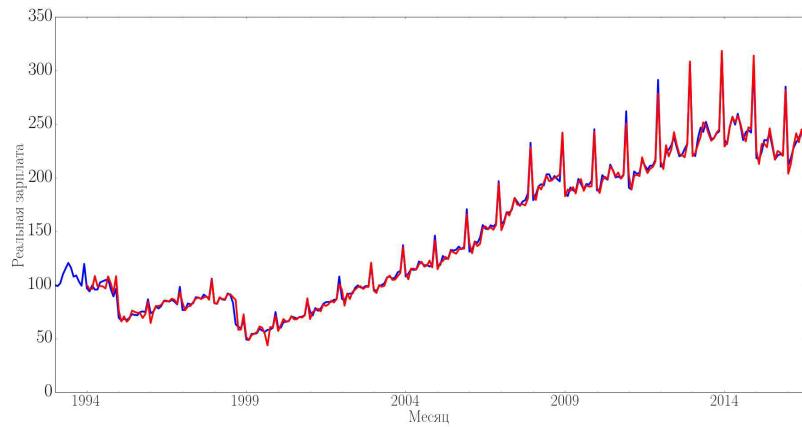


Рис. 1.28: Модель $ARIMA(2, 0, 1) \times (2, 1, 2)$, которая по критерию Акаике оптимально описывает данные о реальной заработной плате в России

Обобщая всю вышеизложенную информацию о прогнозировании ряда реальной заработной платы в России, далее необходимо перебрать разные модели в классе $ARIMA$ со значениями параметров $D = 1, d = 0$

и преобразованием Бокса-Кокса, начиная с начальных приближений, которые были получены из автокорреляционной функции. Сравнение моделей будет производиться по информационному критерию Акаике. В результате, самая лучшая по этому критерию модель, — это $ARIMA(2, 0, 1) \times (2, 1, 2)$ (рис. 1.28).

1.6.2. Подбор ARIMA

Итоговый алгоритм подбора модели в классе ARIMA состоит в следующем:

- В первую очередь необходимо построить график ряда и посмотреть на него. Уже из визуального анализа можно сделать определённые 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.
- Следующий шаг — это стабилизация дисперсии при необходимости. Стабилизация производится с помощью метода Бокса-Кокса или логарифмированием, что является частным случаем того же метода.
- Если исследуемый ряд нестационарен, необходимо подобрать порядок дифференцирования, при котором он становится стационарным. Таким образом фиксируются параметры d, D модели ARIMA.
- Далее необходимо построить графики автокорреляционной функции (ACF) и частичной автокорреляционной функции (PACF) и из этих графиков определить примерные значения параметров p, q, P, Q . Фактически эти значения — начальные приближения, с которых начинается перебор разных моделей.
- Полученные модели необходимо обучить, сравнить их по информационному критерию Акаике и выбрать ту, которая его минимизирует.
- Необходимо посмотреть на остатки получившейся модели, чтобы понять, насколько хорошей она получилась, можно ли, теоретически, её улучшить, нет ли в ней каких-то видимых недостатков. Подробнее об анализе остатков будет рассказано позднее.

1.6.3. Прогнозирование

Теперь необходимо разобраться, как на основании настроенной модели ARIMA правильно строить прогноз. Пусть модель построена, определены значения всех неизвестных параметров, получены их оценки $\hat{\alpha}, \hat{\phi}, \hat{\theta}$, которые записаны в этом уравнении:

$$y_t = \hat{\alpha} + \hat{\phi}_1 y_{t-1} + \cdots + \hat{\phi}_p y_{t-p} + \varepsilon_t + \hat{\theta}_1 \varepsilon_{t-1} + \cdots + \hat{\theta}_q \varepsilon_{t-q}.$$

Чтобы построить прогноз на момент времени $T + 1$, нужно в этом уравнении заменить все индексы t на $T + 1$:

$$\hat{y}_{T+1|T} = \hat{\alpha} + \hat{\phi}_1 y_T + \cdots + \hat{\phi}_p y_{T+1-p} + \varepsilon_{T+1} + \hat{\theta}_1 \varepsilon_T + \cdots + \hat{\theta}_q \varepsilon_{T+1-q}.$$

В этом уравнении присутствует значение ошибки из будущего ε_{T+1} . Неизвестно, какой будет наблюдаться шум в будущем, однако можно предполагать, что в среднем он будет равен 0. Поэтому значения будущих ошибок можно безболезненно заменить на 0. Фактически из уравнения просто удаляются все члены, которые связаны с ошибками из будущего:

$$\hat{y}_{T+1|T} = \hat{\alpha} + \hat{\phi}_1 y_T + \cdots + \hat{\phi}_p y_{T+1-p} + \hat{\theta}_1 \varepsilon_T + \cdots + \hat{\theta}_q \varepsilon_{T+1-q}.$$

В уравнении также присутствуют ошибки из прошлого. Их необходимо заменить на остатки модели в этих точках, потому они являются самыми лучшими оценками ошибок из имеющихся:

$$\hat{y}_{T+1|T} = \hat{\alpha} + \hat{\phi}_1 y_T + \cdots + \hat{\phi}_p y_{T+1-p} + \hat{\theta}_1 \hat{\varepsilon}_T + \cdots + \hat{\theta}_q \hat{\varepsilon}_{T+1-q}.$$

Если прогноз необходимо построить не на одну точку вперёд, а, например, на две, то в формуле появляется значение ряда из будущего y_{T+1} :

$$\hat{y}_{T+2|T} = \hat{\alpha} + \hat{\phi}_1 y_T + \cdots + \hat{\phi}_p y_{T+1-p} + \hat{\theta}_1 \hat{\varepsilon}_T + \cdots + \hat{\theta}_q \hat{\varepsilon}_{T+2-q}.$$

Его необходимо заменить на прогноз $\hat{y}_{T+1|T}$.

Прогноз реальной заработной платы в России на два года вперёд показан на рисунке 1.29. Использовалась модель $ARIMA(2, 0, 1) \times (2, 1, 2)$ и преобразование Бокса-Кокса. Полученный прогноз не выглядит тривиальным: он не является константой, не ведёт себя странно, не падает до нуля. Визуально этот прогноз можно одобрить.

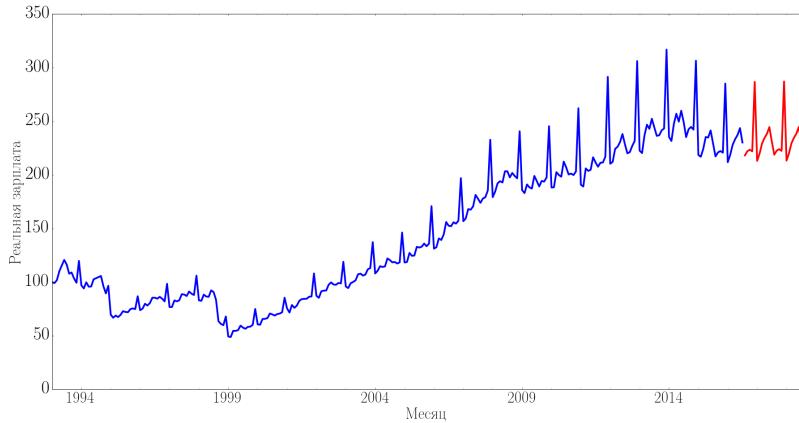


Рис. 1.29: Прогноз реальной заработной платы в России на два года вперёд (красным).

1.7. Анализ остатков

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

Остатки — это разность между фактом и прогнозом:

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

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:

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

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

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

полученные прогнозы сравниваются с истинными значениями ряда, и с их помощью вычисляются остатки.

В зависимости от задачи могут использоваться разные определения остатков, однако чаще используется первое. Остатки оценивают ошибку, то есть шумовую компоненту, которую наблюдать невозможно. При построении модели делаются предположения об этой шумовой компоненте, и логично, что свойства остатков должны согласовываться с выдвинутыми предположениями.

С предположениями о шумовой компоненте необходимо разобраться подробнее.

1.7.1. Несмешённость

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

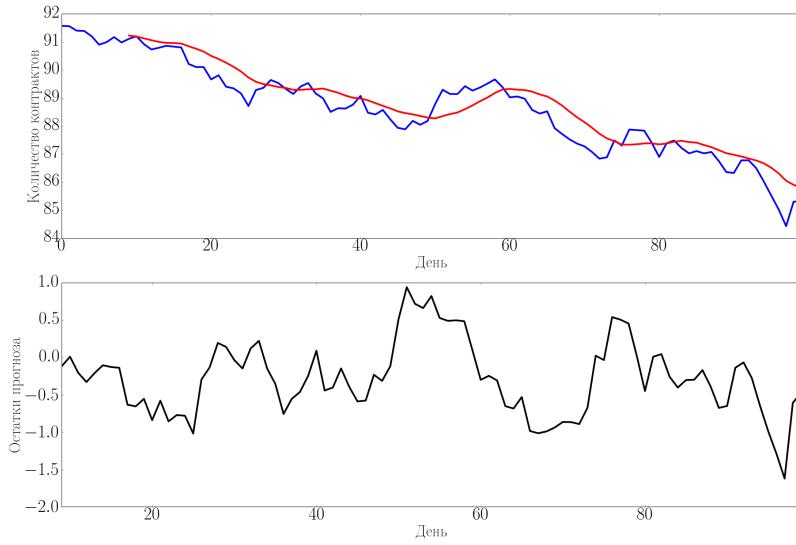


Рис. 1.30: Сверху — количество контрактов в сокровищнице США и прогноз, выполненный методом скользящего среднего, снизу — остатки модели.

На рисунке 1.30 показан прогноз, полученный методом скользящего среднего (по 10 предыдущим значениям). Это достаточно простой и грубый метод прогнозирования, и для рядов с ярко выраженным трендом он даёт плохие результаты. В данном случае видно, что прогноз завышает значения ряда, то есть отстаёт от понижающегося тренда. По этой причине остатки в среднем получаются отрицательными, что видно на нижнем графике на рисунке 1.30.

Гипотезу о несмещённости остатков $H_0: \varepsilon = 0$ 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. В этом случае рекомендуется провести визуальный анализ, чтобы посмотреть, почему прогнозы систематически завышаются или занижаются.

На самом деле, модель очень легко скорректировать в случае, если остатки имеют смещение. Достаточно вычислить среднее значение остатков, это и будет константой, на которую необходимо скорректировать все прогнозы, чтобы остатки стали несмещёнными. После этого преобразования прогнозирующая модель улучшится.

1.7.2. Стационарность

Ещё одно свойство, наличие которого предполагается у ошибок, — это стационарность, то есть отсутствие зависимости от времени. Таким образом, остатки во времени должны быть распределены примерно одинаково.

На рисунке 1.31 показан прогноз, построенный так называемым наивным сезонным методом. То есть каждый январь значение ряда предсказывается равным предыдущему январю и т.д. Поскольку в ряде присутствует ярко выраженный тренд, то такой прогноз будет некачественным, а остатки — смещёнными. В таком случае необходимо провести корректировку на смещение. По остаткам скорректированного несмещённого прогноза (рис. 1.31, снизу) видно, что в модели всё ещё есть проблема. Остатки ведут себя систематически: сначала они в основном отрицательные, затем в течение какого-то времени — положительные, в конце ряда — снова преимущественно отрицательные.

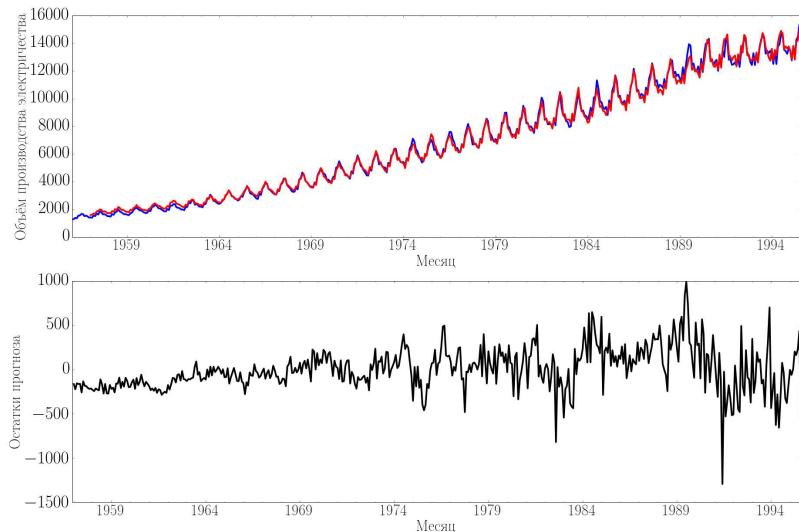


Рис. 1.31: Сверху — количество произведённого в Австралии электричества и прогноз, выполненный "наивным сезонным методом" с корректировкой на смещение, снизу — остатки модели.

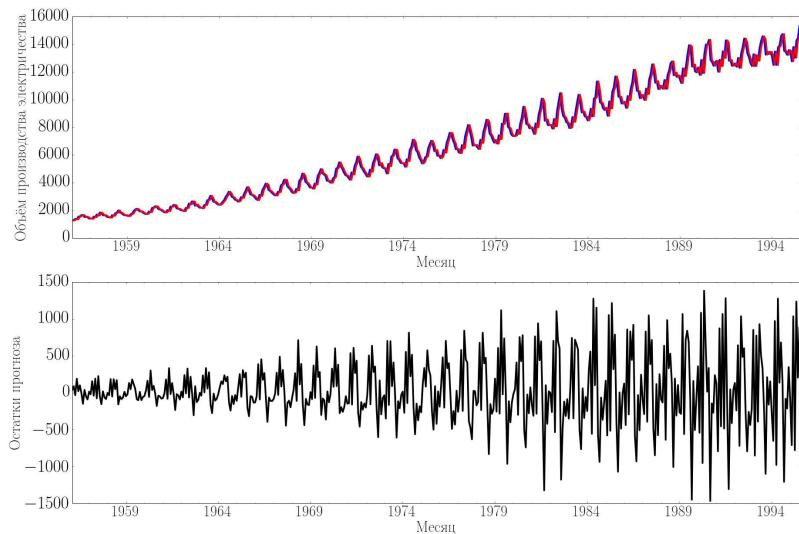


Рис. 1.32: Сверху — количество произведённого в Австралии электричества и наивный прогноз, снизу — остатки модели.

Результат применения наивного прогноза (в каждой точке значение прогнозируется предыдущим значением этого же ряда) показан на рисунке 1.32. Как и в предыдущем случае, остатки модели нестационарны: дисперсия меняется во времени.

Формально гипотезу о стационарности можно проверить с помощью критерия Дики-Фуллера. Если стационарность отсутствует, то модель неодинаково точна в разные периоды времени. Необходимо провести визуальный анализ, чтобы понять, что с моделью не так, и почему прогнозы в разные периоды времени систематически имеют разную ошибку.

1.7.3. Неавтокоррелированность

Ещё одно желаемое свойство остатков — это неавтокоррелированность, то есть отсутствие зависимости от предыдущих наблюдений.

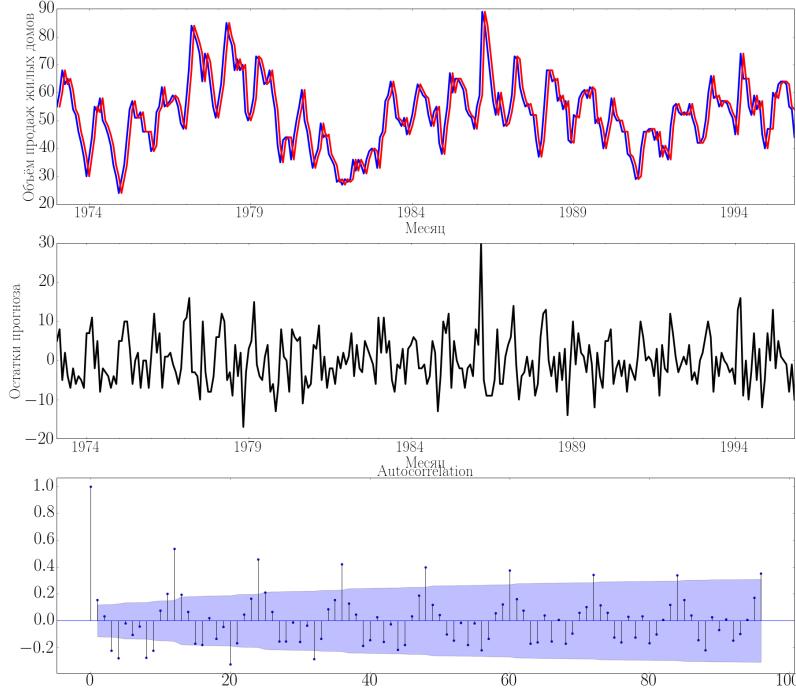


Рис. 1.33: Сверху — продажи недвижимости в США и прогноз, выполненный наивным методом, по центру — остатки модели, снизу — автокорреляционная функция остатков.

На рисунке 1.33 показан результат прогнозирования наивным методом продаж недвижимости в США. Наивный метод работает очень плохо при работе с рядами с ярко выраженной сезонностью. Это хорошо видно на графике автокорреляционной функции остатков прогноза. Она имеет очень значимые и ярко выраженные пики на лагах, кратных сезонному периоду.

Гипотезу о неавтокоррелированности можно проверить по коррелограмме, а также с помощью Q-критерия Льюнга-Бокса (таблица 1.3). Этот критерий позволяет проверить гипотезу о равенстве нулю одновременно нескольких автокорреляций при разных лагах (с лага 1 по лаг Q). Параметр Q можно выбирать, например, перебором, а можно пользоваться значением по умолчанию, использующемся в функции, которая производит оценку модели ARIMA.

ряд ошибок прогноза:	$\varepsilon^T = \varepsilon_1, \dots, \varepsilon_T$
нулевая гипотеза:	$H_0: r_1 = \dots = r_Q = 0$
альтернатива:	$H_1: H_0$ неверна
статистика:	$Q(\varepsilon^T) = T(T+2) \sum_{\tau=1}^Q \frac{r_\tau^2}{T-\tau}$
нулевое распределение:	$Q(\varepsilon^T) \sim \chi_{Q-K}^2$ при H_0
K —	число настраиваемых параметров модели

Таблица 1.3: Описание Q-критерия Льюнга-Бокса

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. Скорректированная модель будет лучше, а её остатки будут больше похожи на белый шум. Однако это можно сделать далеко не всегда — возможности модели класса ARIMA не безграничны, и с помощью таких моделей нельзя учесть всю структуру ряда. Таким образом, автокоррелированность остатков только указывает на потенциальную возможность улучшить модель, и не факт, что улучшения можно добиться на практике с помощью рассматриваемого класса моделей.

1.8. Пример

1.9. Регрессионный подход к прогнозированию

1.9.1. Праздники

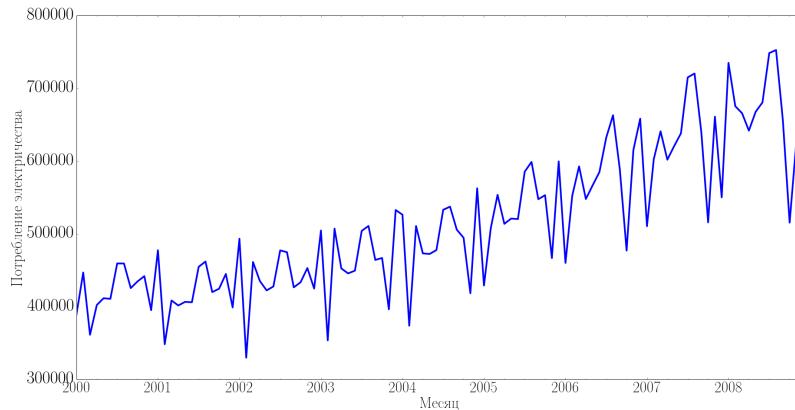


Рис. 1.34: Ежемесячное потребление электричества в Турции

На рисунке 1.34 представлены ежемесячные данные о потреблении электричества в Турции. На графике очень хорошо видна годовая сезонность, а также странные падения. Эти падения соответствуют месяцам, на которые выпадают праздники по исламскому календарю. Этот календарь примерно на 11 дней короче, чем григорианский. По этой причине праздники, которые определяются по исламскому календарю всё время попадают в разные места стандартного григорианского года, и их не получается учесть с помощью стандартной сезонности с периодом 12. Это плохая новость.

Есть и хорошая новость, которая заключается в том, что для каждого года точно известно, на какой месяц приходится такой праздник. На основании этого можно сформировать бинарный признак, в котором 1 будет соответствовать месяцам с праздником, 0 — остальным. Такие признаки можно попытаться встроить в модель типа ARIMA.

1.9.2. ARIMAX

Пусть считается, что значение ряда y в момент времени t задаётся следующим образом:

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

то есть фактически y задаётся линейной регрессией, но остатки этой регрессии прогнозируются при помощи модели ARIMA:

$$\begin{aligned} z_t = & \alpha + \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \\ & + \phi_S z_{t-S} + \cdots + \phi_{PS} z_{t-PS} + \theta_S \varepsilon_{t-S} + \cdots + \theta_{PS} \varepsilon_{t-PS} + \varepsilon_t. \end{aligned}$$

Такая модель, которая комбинирует линейную регрессию и ARIMA называется regARIMA или ARIMAX.

1.9.3. Сложная сезонность

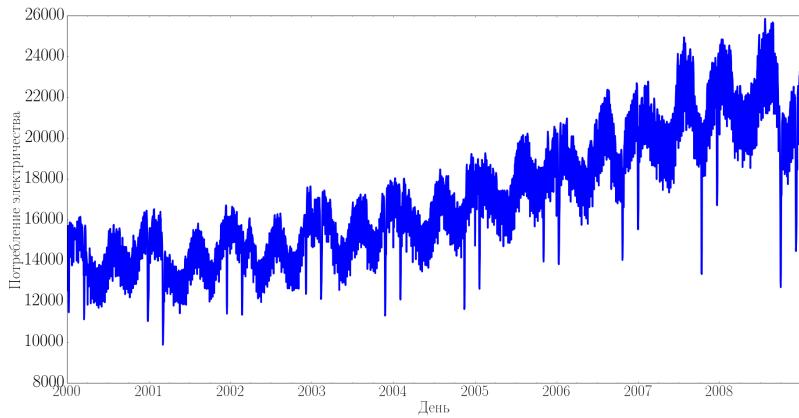


Рис. 1.35: Ежедневное потребление электричества в Турции

Описанная выше модель очень полезна при работе с рядами со сложной сезонностью. Например, если рассмотреть потребление электричества не по месяцам, а по дням (рис. 1.35), то в этом ряду будет недельная сезонность, годовая и, опять же, праздники по исламскому календарю.

К сожалению, модели класса ARIMA дают плохие результаты при работе с рядами со сложной сезонностью. Возникает несколько проблем. Во-первых, при длинных сезонных периодах в модели ARIMA становится слишком много параметров, и их невозможно оценить по ряду.

Во-вторых, модель ARIMA явно задаёт значение ряда как функцию от его значения, например, один сезонный период назад. Странно ожидать, что при работе с подневным рядом его значение будет определяться значением в эту же дату прошлого года. Скорее всего, значения ряда в некой окрестности текущей даты похоже на его значения в этой же самой окрестности этой же даты год назад.

Наконец, довольно большую проблему представляет то, что длина года нецелая. Она равна не в точности 365 дням, а 365.25. Если в ряду собраны недельные данные, то длина года в неделях также не 52 недели ровно, а 52.18. С этой проблемой ARIMA не может работать никак, в ряде нет измерения с индексом 52.18, которое можно было бы подставить в модель.

Решить эту проблему можно в рамках регрессионного подхода. Для модели ARIMA можно оставить самый короткий из имеющихся сезонных периодов. Все остальные периоды будут учитываться с помощью регрессии на специально построенные признаки. В качестве признаков будут выступать фурье-гармоники с периодами, пропорциональными длине сезонных периодов (например, 365.25, 365.25/2, 365.25/3 и т.д.). Какое-то количество таких гармоник необходимо подставить как регрессионную компоненту в регрессионную ARIMA.

1.9.4. Регрессионные признаки

Вообще говоря, можно придумать много вещей, которые можно подставлять в регрессионную компоненту:

- гармоники по длинным периодам сезонности;
- для коротких периодов сезонности можно использовать индикаторы (например, дни недели можно явно задать как индикатор понедельника, индикатор вторника, и т. д.) и в явном виде подставить их в регрессионную ARIMA;
- индикаторы праздников;
- полезными часто оказываются индикаторы пред- и постпраздничных дней;
- тренды (линейный, квадратичный и т. д.);
- скользящие средние ряда за предыдущие периоды (например, в каждой точке вычисляется среднее за прошлый месяц или прошлую неделю, и такой признак подставляется в регрессионную ARIMA).

Часто оказывается, что если хорошо подобрать признаки, и построить на них регрессию, то добавлять поверх неё ARIMA уже не нужно. Выигрыш в качестве при добавлении авторегрессионной добавки оказывается практически незначимым.

1.9.5. Массовое прогнозирование

Специфические методы прогнозирования рядов хороши, когда работа производится не более чем с десятками временных рядов, которые постоянно прогнозируются. На каждый из этих рядов можно посмотреть глазами, тщательно подобрать хорошую модель, проанализировать остатки, при необходимости перестроить модель и т.д. Это ручной труд, и часто нет возможности для такого ручного труда. Нет такой возможности в задачах массового прогнозирования.

Пусть необходимо спрогнозировать дневные продажи товаров в магазинах. В распоряжении имеются данные о продажах товаров, их остатках в магазинах, ценах, скидках, промоакциях, иерархии товаров, иерархии и расположение торговых точек. Стоит задача построить прогнозы продаж всех товаров во всех магазинах. Однако временных рядов слишком много, чтобы их можно было спрогнозировать вручную.

На помощь может прийти регрессионный подход. Если хорошо сконструировать признаки, которые будут содержать всю имеющуюся информацию, и использовать на них какую-то регрессионную модель (не обязательно линейную), то уже таким способом можно получить достаточно хорошее решение.

1.9.6. Резюме

Оказалось, что задачу прогнозирования временных рядов можно свести к задаче обучения с учителем, не применяя специфические методы. Тем не менее, важно помнить две вещи. Во-первых, для того, чтобы регрессионный подход работал, необходимо явно учитывать временную природу данных при конструировании признаков. Во-вторых, даже если получилось построить хорошую регрессионную модель, стоит попробовать поверх неё наложить авторегрессию или какую-либо ещё модель временных рядов, о которых не упоминалось в этом курсе. Возможно, это поможет улучшить качество прогнозов.