Введение во временные ряды

Литература

1. Rob J Hyndman, George Athanasopoulos. Forecasting: Principles and Practice: https://o-texts.com/fpp2/
2. G. E. P. Box, D. R. Cox. An analysis of transformations, Journal of the Royal Statistical Society, Series B, 26, 211-252 (1964)

---

1. Постановка задачи прогнозирования

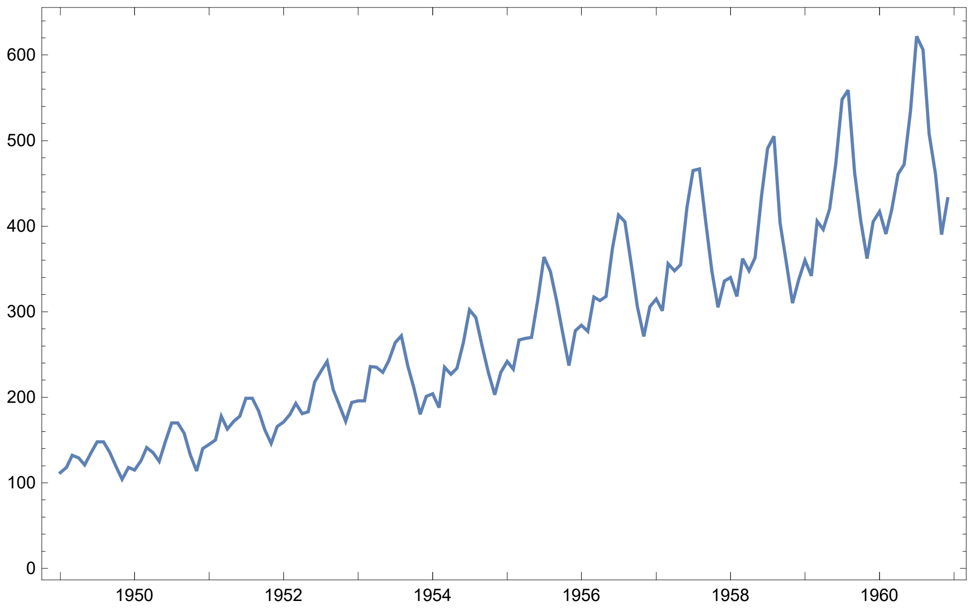
1.1. Понятия временного ряда и прогнозирования

Временной ряд – последовательность значений признака y, измеряемого через **постоянные** временные интервалы:

$$y_1,\ y_2,\ \ldots,\ y_T,\ \ y_t \in \mathbb{R}.$$

Примерами временных рядов могут выступать ряды среднедневных цен на акции определенной компании, рыночные цены, объемы продаж в торговых сетях, объемы потребления и цены электроэнергии, дорожный трафик и т. д. Ещё один пример временного ряда представлен на рисунке – это объемы перевозок международных авиакомпаний с января 1949 г. до декабря 1960 г. в тысячах человек.

International airline passengers: monthly totals in thousands. Jan 49 – Dec 60



Задача прогнозирования состоит в нахождении функции f_T:

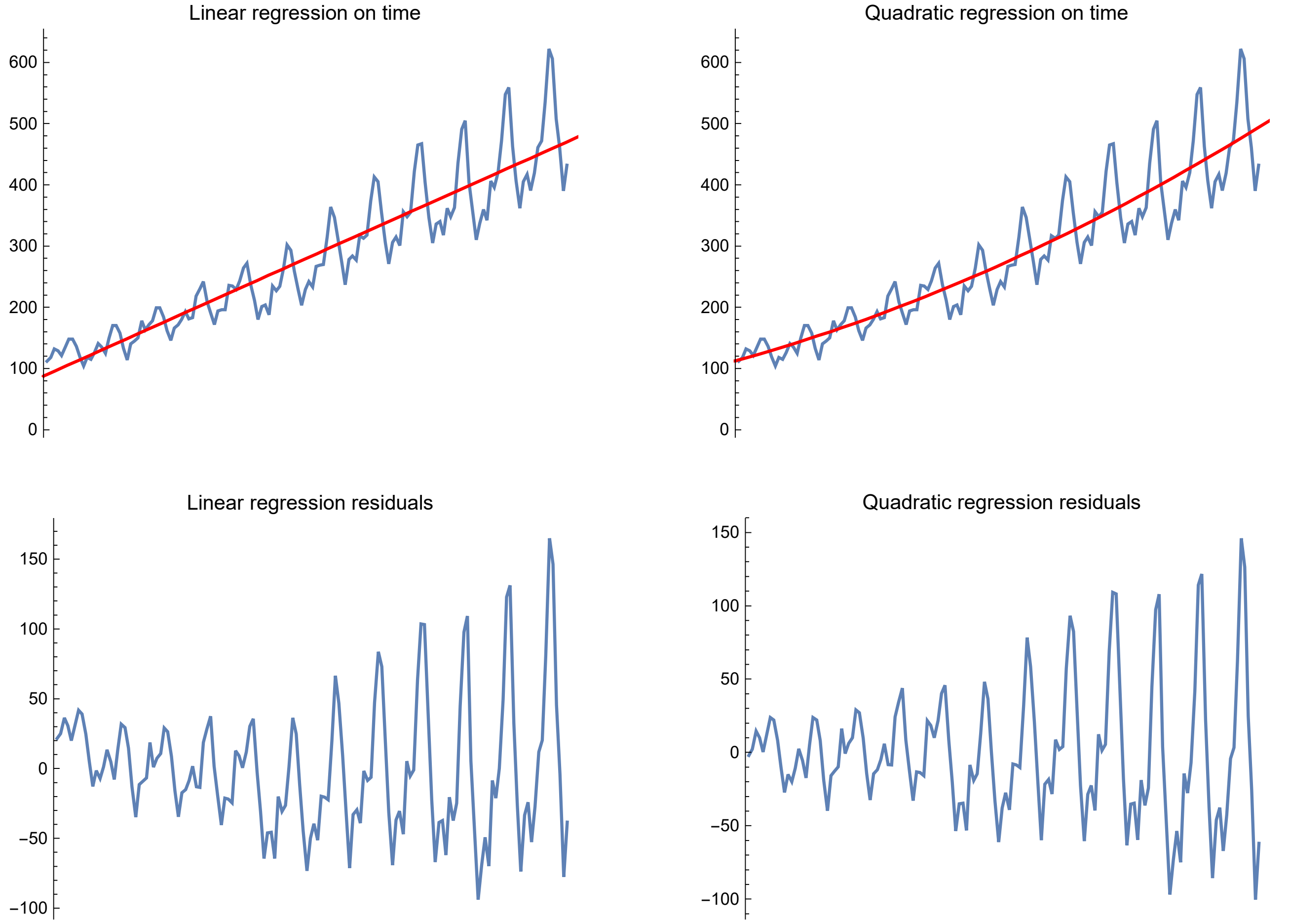
$$y_{T+h} \approx f_T(y_T,\ \ldots,\ y_1,\ h) \equiv \hat{y}_{T+h|T},$$

где $h \in \{1,\ 2,\ \ldots,\ H\}$, H – горизонт прогнозирования.

Предсказательный интервал – интервал, в котором предсказываемая величина окажется с вероятностью не меньше заданной.

1.2. Модель регрессии

Можно свести задачу прогнозирования к задаче обучения с учителем. Процесс разворачивается во времени, поэтому будем строить модель зависимости целевого признака от времени. Регрессия может быть линейной, квадратичной или даже показательной.



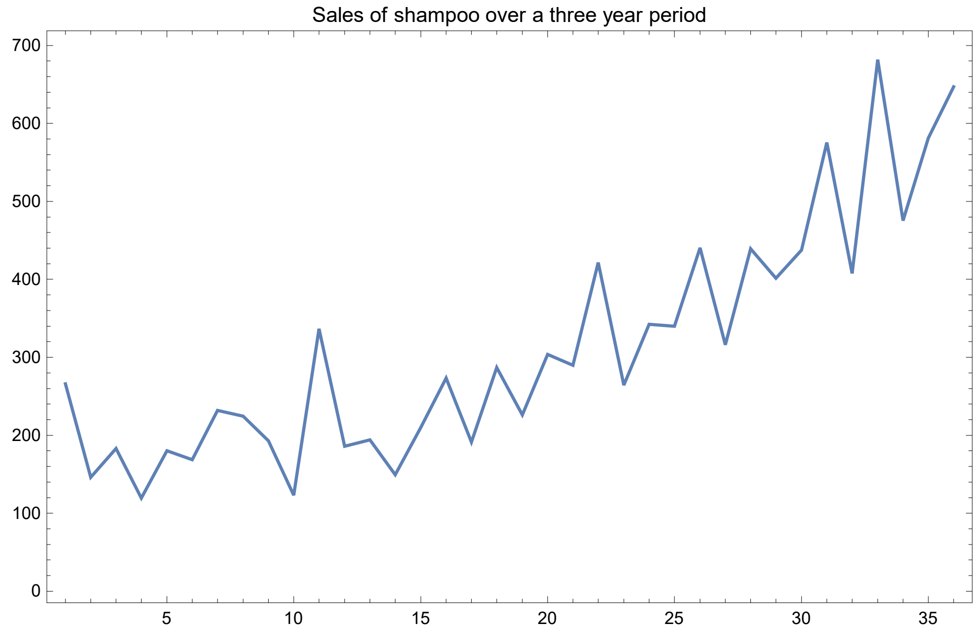
Остатки таких моделей не похожи на случайный шум, в них остается большая часть информации, которая не была учтена. Вид остатков показывает, что можно построить более сложную модель, которая будет лучше описывать имеющиеся данные, а также давать более точные прогнозы в будущем.

1.3. Компоненты временных рядов

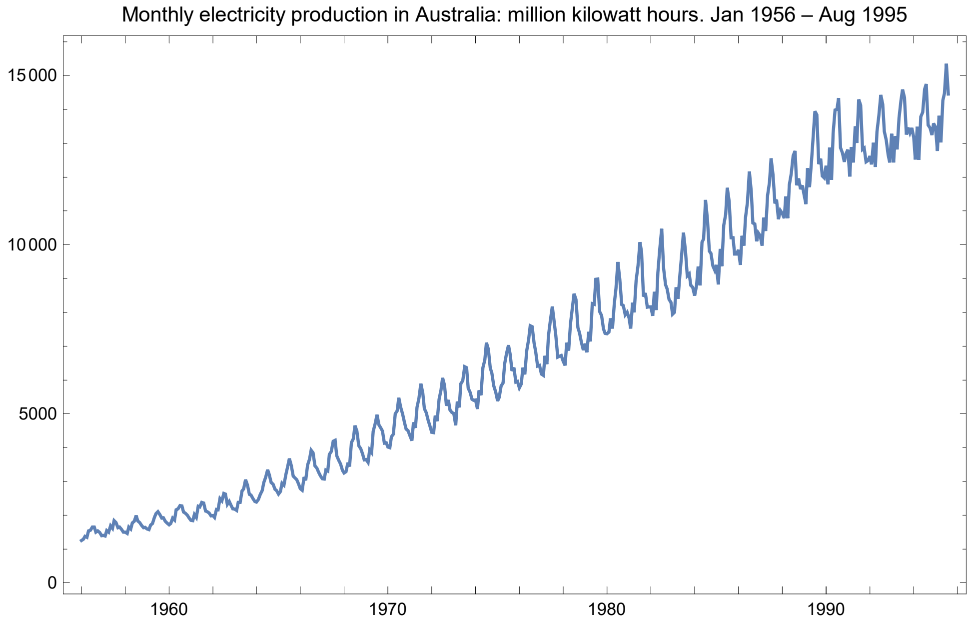
Поведение временных рядов можно описать следующими характеристиками:

- **тренд** – плавное долгосрочное изменение уровня ряда;
- **сезонность** – циклические изменения уровня ряда с постоянным периодом;
- **цикл** – изменения уровня ряда с переменным периодом (экономические циклы, периоды солнечной активности);
- **ошибка** – непрогнозируемая случайная компонента ряда;
- **разладка** – смена модели ряда.

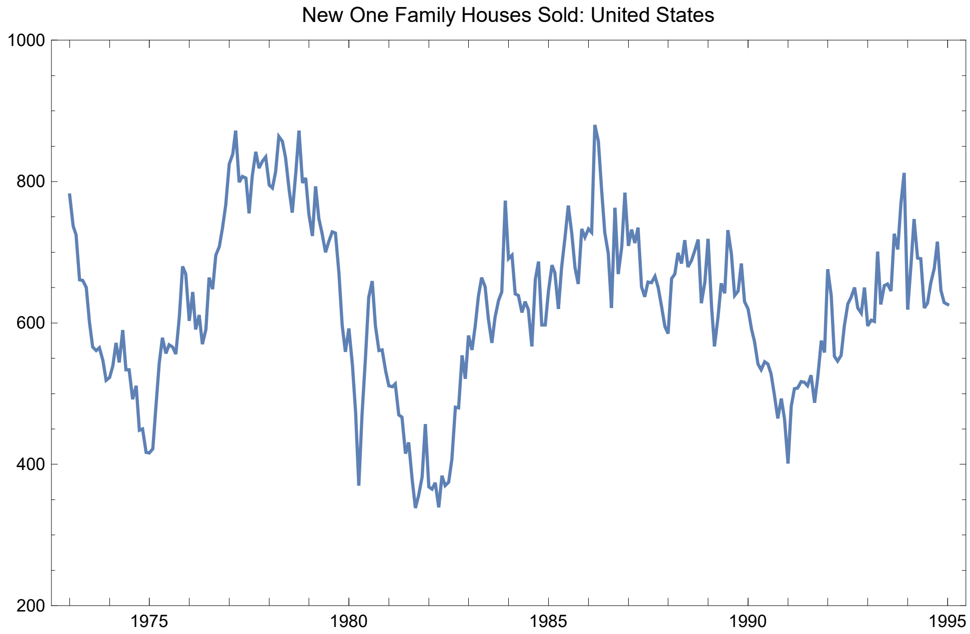
Рассмотрим данные продаж шампуня по месяцам. На графике виден повышающийся тренд, который можно описать линейной или квадратичной функцией. Сложно выделить на этом участке в данных циклы или сезонность.



Теперь рассмотрим данные за несколько лет о суммарном объеме электричества, произведенного за месяц в Австралии. На графике, как и в предыдущем случае, виден повышающийся тренд. 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 теплый сезон в Австралии.



Следующий пример – объем проданной жилой недвижимости в США за месяц (рис. 1.6). 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).



2. Автокорреляция

2.1. Значение автокорреляции

Автокорреляция (автокорреляционная функция, ACF) – количественная характеристика сходства между значениями ряда в соседних точках. Автокорреляционная функция задается следующим соотношением:

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

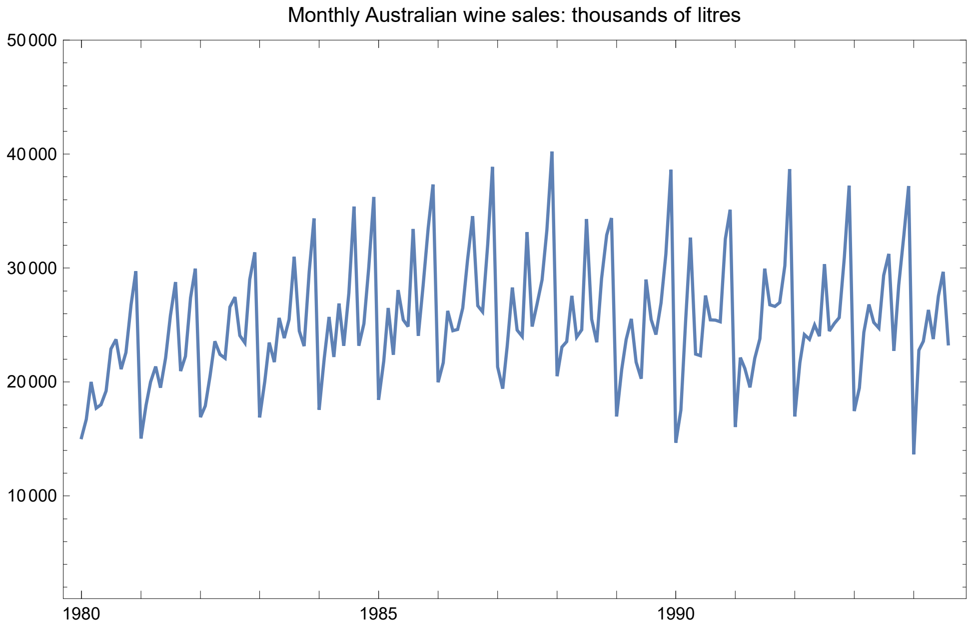
Автокорреляция – это корреляция Пирсона между исходным рядом и его версией, сдвинутой на несколько отсчетов. Количество отсчетов, на которое сдвинут ряд, называется лагом автокорреляции (τ).

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

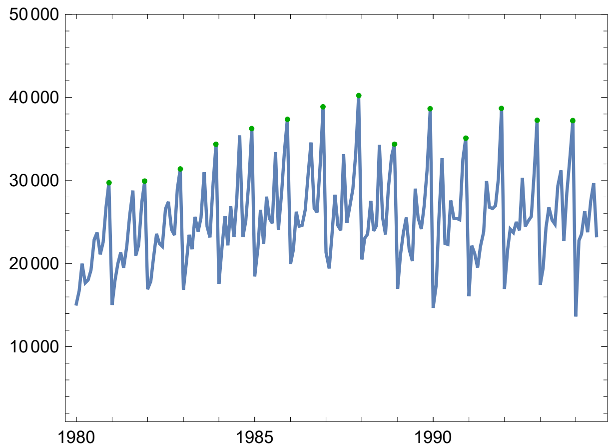
$$r_\tau = \frac{\sum_{t=1}^{T-\tau}(y_t - \overline{y})(y_{t+\tau} - \overline{y})}{\sum_{t=1}^{T}(y_t - \overline{y})^2}.$$

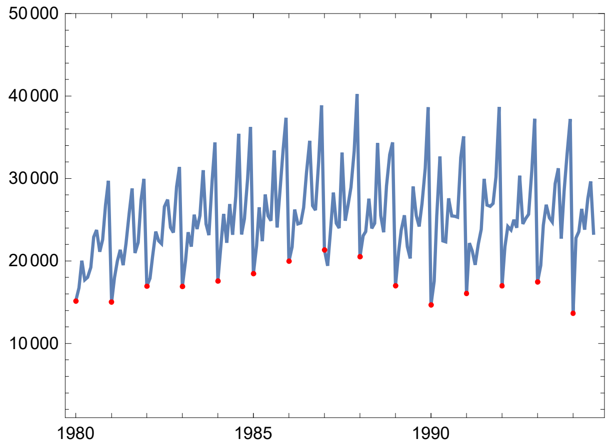
2.2. Диаграмма рассеяния

Рассмотрим данные о суммарном объеме продаж вина в Австралии за месяц.

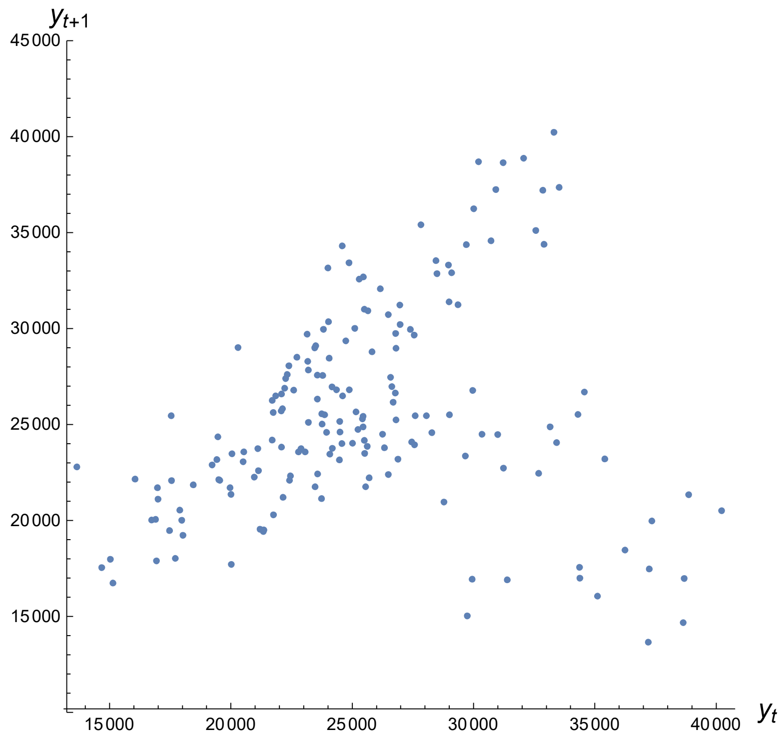


Заметим, что в декабре продажи вина больше, а в январе продажи падают. Значит, ряд обладает ярко выраженной годовой сезонностью.

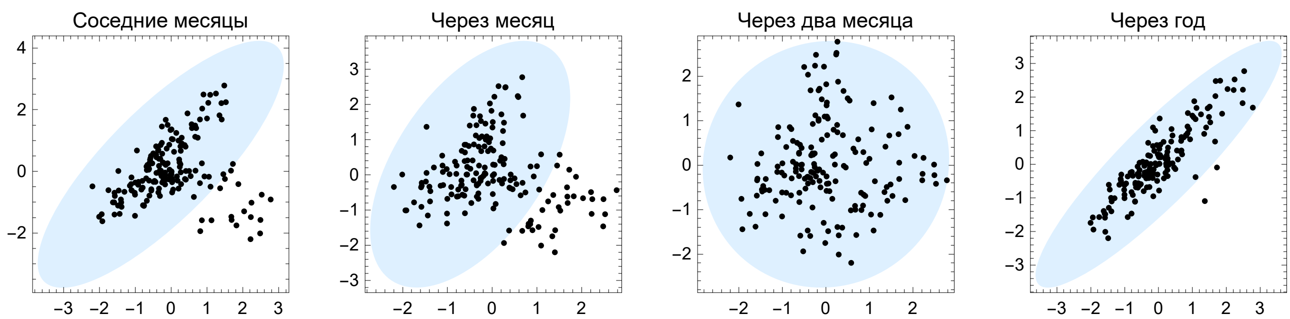




Если построить график зависимости объемов продаж вина в соседние месяцы, то будет видно, что большая часть точек **диаграммы рассеяния** группируется вокруг главной диагонали. Это говорит о том, что в основном значения продаж в соседние месяцы похожи. Еще одно подмножество точек выделяется в правом нижнем углу, оно связано с падением продаж от декабря к январю, которое было видно на предыдущем графике.

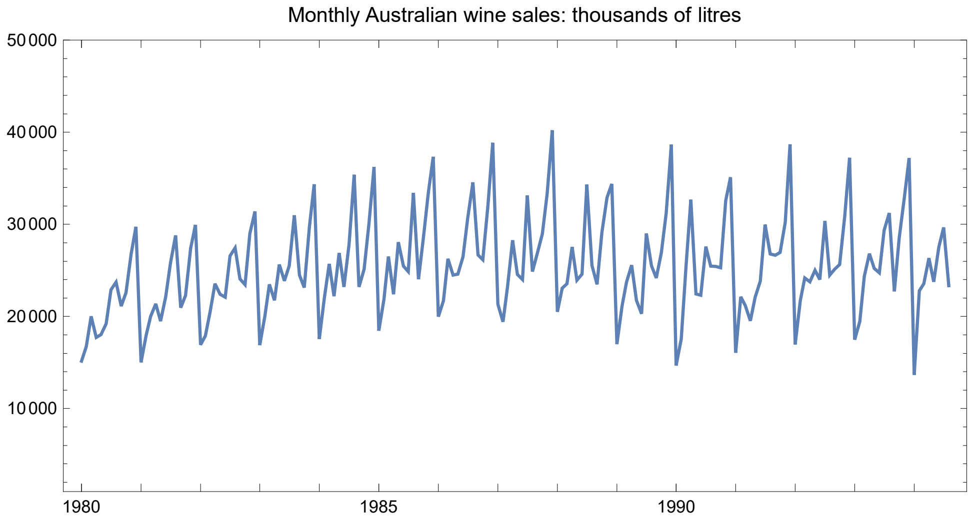


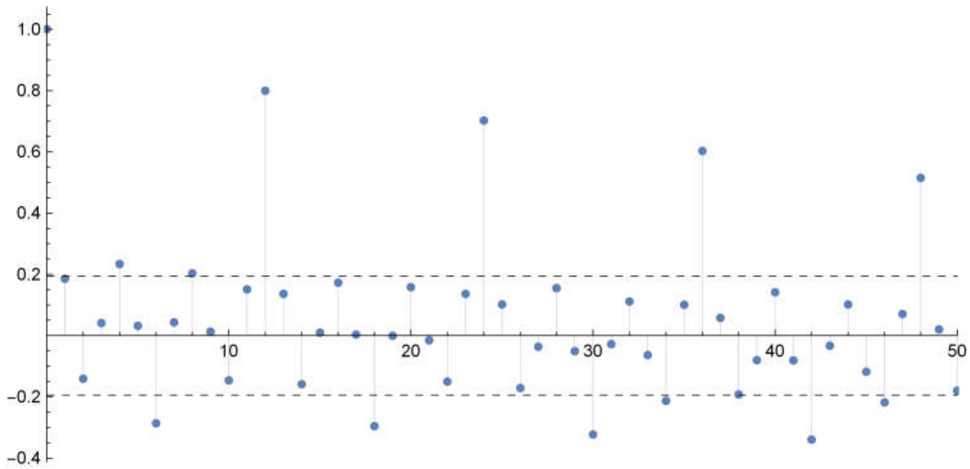
Если построить аналогичный график, но по вертикальной оси отложить y_{t+2}, то видно, что точки в основном облаке начинают «расплываться» вокруг главной диагонали, то есть сходство между продажами через месяц уменьшается по сравнению с соседними месяцами. Если посмотреть связь между продажами через два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



2.3. Коррелограмма

Анализировать величину автокорреляции при разных значениях лагов удобно с помощью графика, который называется **коррелограммой**. По оси ординат на нем откладывается автокорреляция, а по оси абсцисс – размер лага τ. На графике для продаж вина в Австралии видно, что автокорреляция принимает большие значения в лагах, кратных сезонному периоду.

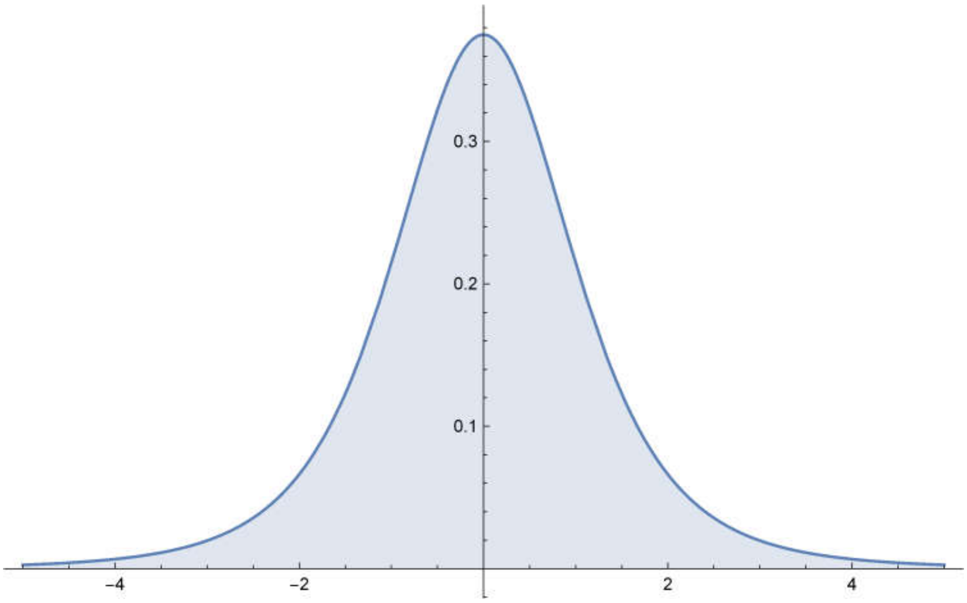




2.4. Значимость автокорреляции

На первой коррелограмме помимо значений автокорреляции также изображен коридор вокруг горизонтальной оси. Это коридор значимости отличия корреляции от нуля. Как и для обычной корреляции Пирсона, значимость вычисляется с помощью критерия Стьюдента. 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:	$y^T = y_1, \ldots, y_t$
нулевая гипотеза:	$H_0\colon \mathrm{r}_\tau = 0$
альтернатива:	$H_1\colon \mathrm{r}_\tau < \neq > 0$
статистика:	$T(y^T) = \frac{\mathrm{r}_\tau \sqrt{T-\tau-2}}{\sqrt{1-\mathrm{r}_\tau{}^2}}$
нулевое распределение:	$T(y^T) \sim \mathrm{St}(\mathrm{T}-\tau-2)$



3. Стационарность

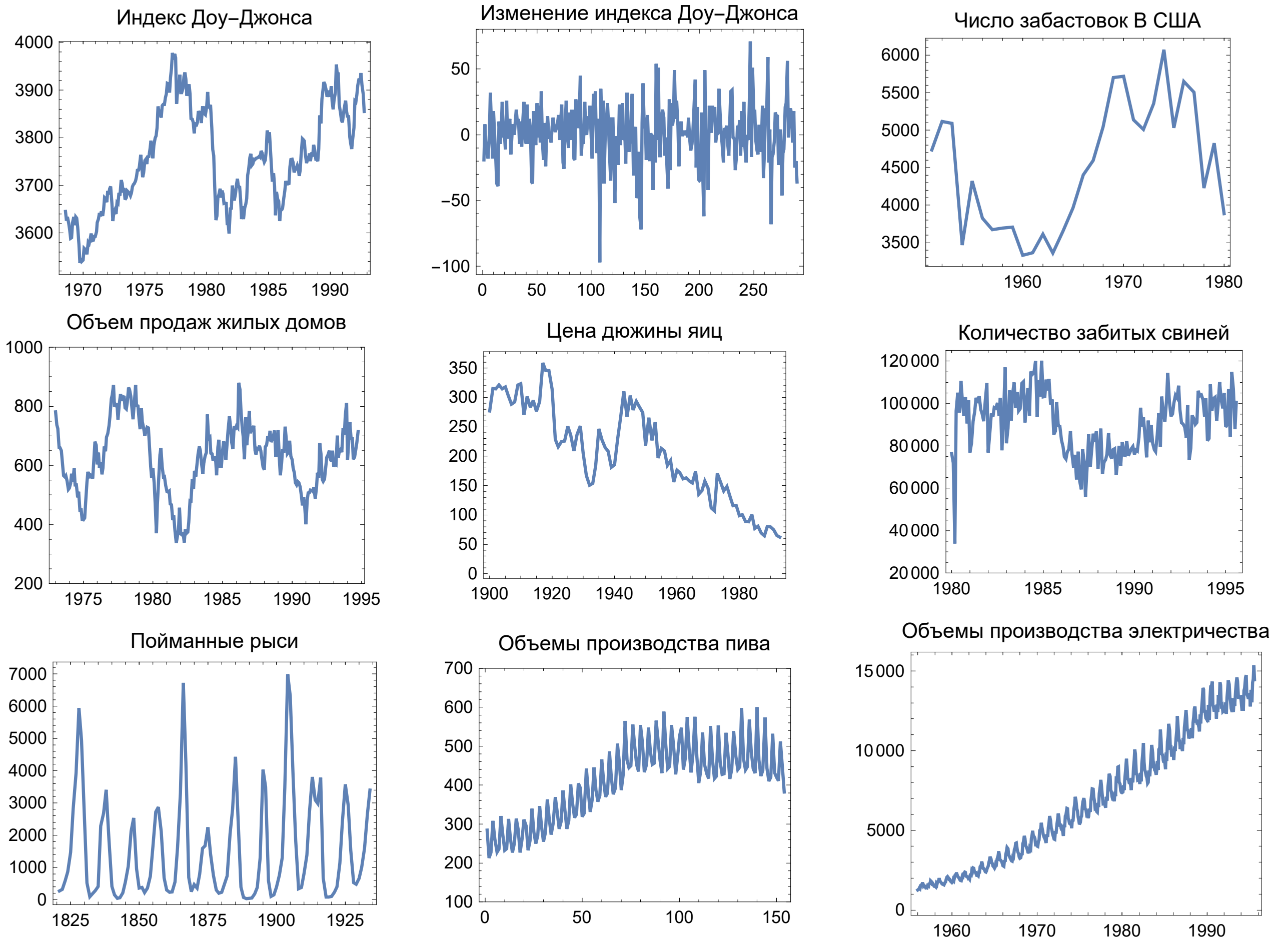
3.1. Понятие стационарности

Ряд $y_1, \ldots, y_T$ **стационарен**, если $\forall\, k$ (ширина окна) распределение $y_t, \ldots, y_{t+k}$ 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.

Упражнение

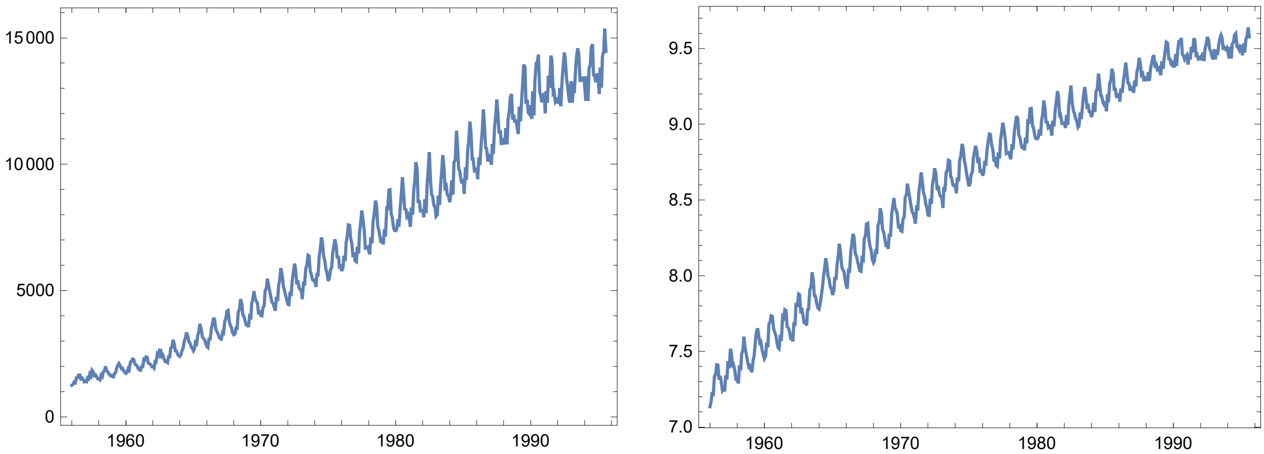
Какие из представленных ниже рядов являются стационарными?



Гипотезу о стационарности можно проверить с помощью критерия Дики-Фуллера. Статистику данного критерия будем рассматривать позже.

3.2. Стабилизация дисперсии

Если во временном ряде монотонно по времени изменяется дисперсия, применяется специальное преобразование, стабилизирующее дисперсию. Часто в качестве такого преобразования выступает логарифмирование. В результате логарифмирования ряда производства электричества в Австралии размах колебаний в начале и конце ряда становится очень похожим, и дисперсия примерно стабилизируется.

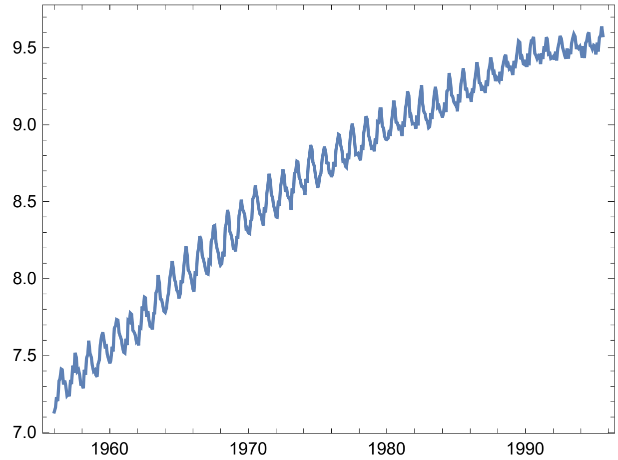


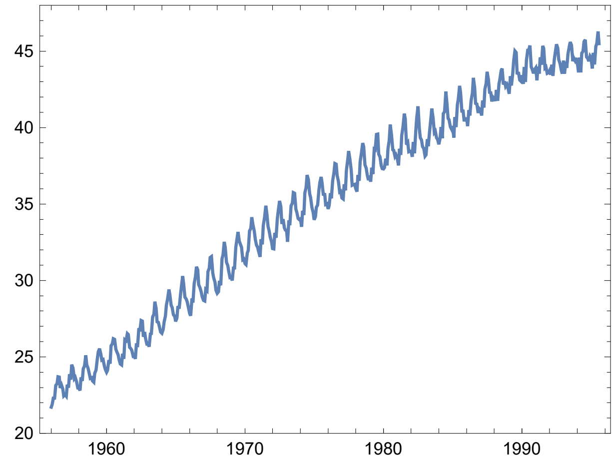
Логарифмирование принадлежит к параметрическому семейству преобразований Бокса-Кокса. В случае, когда значения ряда $y > 0$, преобразование Бокса-Кокса имеет вид:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log y, & \lambda = 0. \end{cases}$$

Заметим, что $y^\lambda = e^{\lambda \log(y)} = 1 + \lambda \log(y) + O((\lambda \log(y))^2)$. Тогда $y^{(\lambda)} = \log(y)$ в случае, когда λ бесконечно мало.

Параметр λ определяет, как именно будет преобразован ряд: $\lambda = 0$ – логарифмирование, $\lambda = 1$ – тождественное преобразование ряда, 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.





Параметр λ выбирается методом максимального правдоподобия. Преобразование Бокса-Кокса относится к семейству степенных преобразований вида:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda (\mathrm{gm}(y))^{\lambda - 1}}, & \lambda \neq 0, \\ \mathrm{gm}(y) \log y, & \lambda = 0, \end{cases}$$

где $\mathrm{gm}(y) = \left(\prod_{i=1}^{T} y_i\right)^{\frac{1}{T}} = \sqrt[T]{y_1 y_2 \ldots y_T}$ – среднее геометрическое ряда.

Бокс и Кокс в своей статье включили среднее геометрическое в преобразование, связав плотность распределения исходного ряда с плотностью преобразованного следующим соотношением:

$$J(\lambda; y_1, y_2, \ldots, y_T) = \prod_{i=1}^{T} \left| \frac{\partial y_i^{(\lambda)}}{\partial y} \right| = \prod_{i=1}^{T} y_i^{\lambda - 1} = \mathrm{gm}(y)^{T(\lambda - 1)},$$

$$f(y_1, \ldots, y_T) = f_{(\lambda)}\left(y_1^\lambda, \ldots, y_T^\lambda\right) J(\lambda; y_1, y_2, \ldots, y_T).$$

Из предположения, что значения ряда $y_i^{(\lambda)}$ $(i = 1, \ldots, T)$ распределены нормально с математическим ожиданием $\overline{y}^{(\lambda)}$ и постоянной дисперсией σ^2, оценка параметра λ может быть получена путем максимизации логарифма правдоподобия:

$$L_{\max}(\lambda) = \prod_{i=1}^{T} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(- \frac{\left(y_i^{(\lambda)} - \overline{y}^{(\lambda)}\right)^2}{2\sigma^2} \right) J(\lambda; y),$$

$$\log(L_{\max}(\lambda)) = -\frac{T}{2} \log\left(\frac{\sum_i \left(y_i^{(\lambda)} - \overline{y}^{(\lambda)}\right)^2}{T} \right) + (\lambda - 1) \sum_i \log(y_i).$$

Если ряд содержит отрицательные значения, то можно переписать правила преобразования следующим образом:

$$y^{(\lambda)} = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \lambda_1 \neq 0, \\ \log(y+\lambda_2), & \lambda_1 = 0, \end{cases}$$

где $y > -\lambda_2$.

3.3. Дифференцирование

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

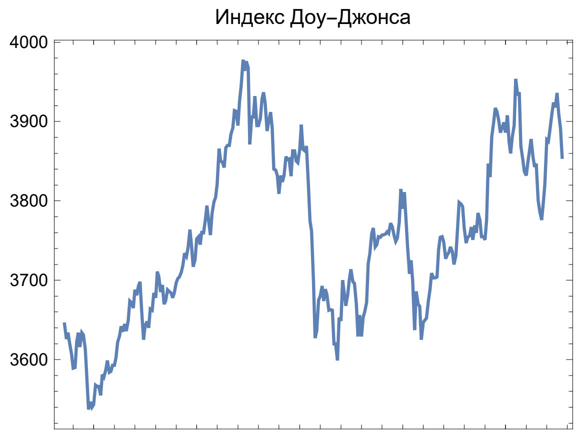
$$y'_t = y_t - y_{t-1}.$$

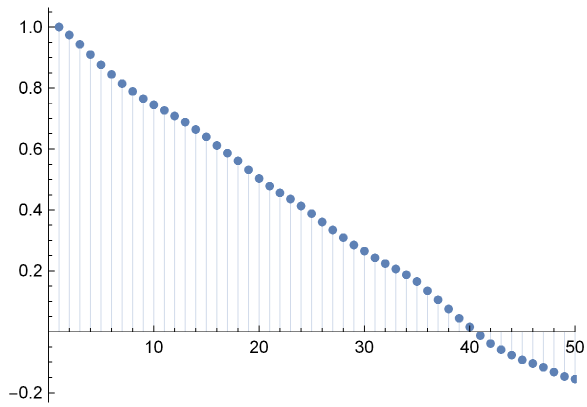
Для нестационарного ряда часто оказывается, что получаемый после дифференцирования ряд является стационарным. Такая операция позволяет стабилизировать среднее значение ряда и избавиться от тренда, а иногда даже от сезонности. Кроме того, дифференцирование можно применять неоднократно: от ряда первых 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.

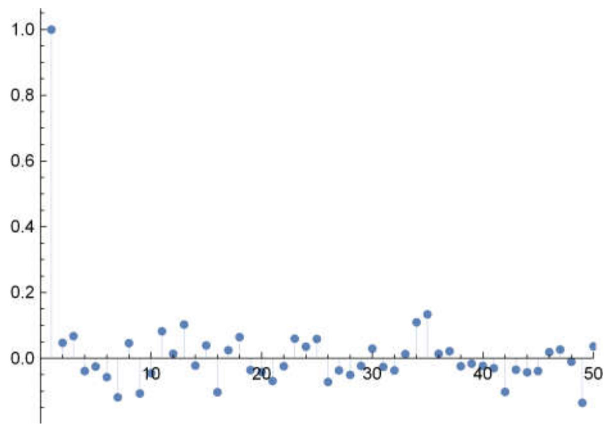
Также может применяться сезонное дифференцирование ряда, переход к попарным разностям значений в соседних сезонах. Если длина периода сезона составляет s, то новый ряд задается разностями:

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

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.







На верхних графиках показаны ряд значений индекса Доу-Джонса и его автокорреляционная функция. Видно, что этот ряд нестационарен – имеется ярко выраженный тренд. От тренда удается полностью избавиться, продифференцировав ряд.

Таким образом, для приведения временного ряда к стационарному первым делом необходимо стабилизировать дисперсию, то есть применить преобразование Бокса-Кокса, затем, при наличии ярко выраженной сезонности провести сезонное дифференцирование с лагом, равным сезонному периоду. При необходимости провести обычное дифференцирование.

3.4. Обратное преобразование

Исходя из правил дифференцирования, переход к исходному временному ряду может быть выполнен следующему правилу:

$$y'_t = y_t - y_{t-k},$$
$$y_t = y'_t + y_{t-k},$$

где y_t – исходный ряд, а k – лаг дифференцирования.