Введение во временные ряды

Д. А. Ивахненко. Методы прогнозирования, СПбГЭУ 2021 г.

Литература

Rob J Hyndman, George Athanasopoulos. Forecasting: Principles and Practice: https://otexts.-com/fpp2/

---

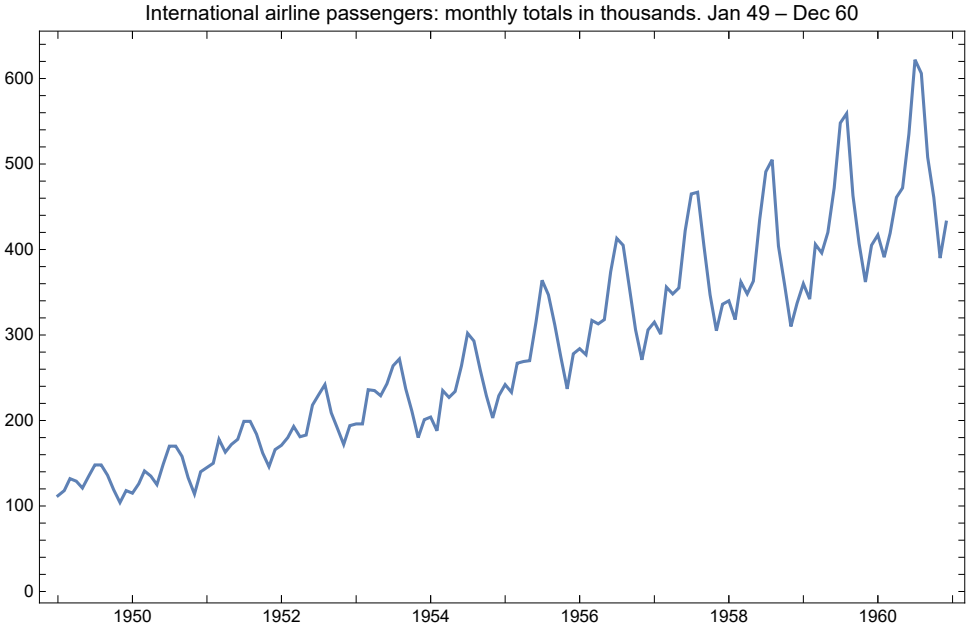
1. Постановка задачи прогнозирования

1.1. Понятия временного ряда и прогнозирования

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

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

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 интернациональных авиакомпаний с января 1949 г. по декабрь 1960 г. (тыс. человек).



Задача прогнозирования состоит в нахождении функции f_T:

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

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

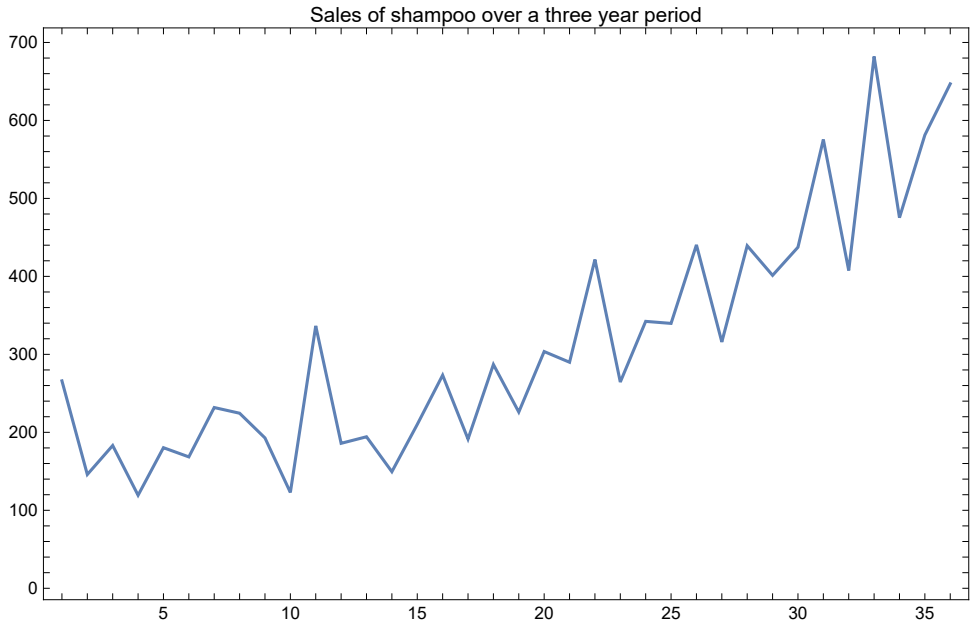
Предсказательный интервал – интервал, в котором предсказываемая величина окажется с вероятностью не меньше заданной.

1.2. Компоненты временных рядов

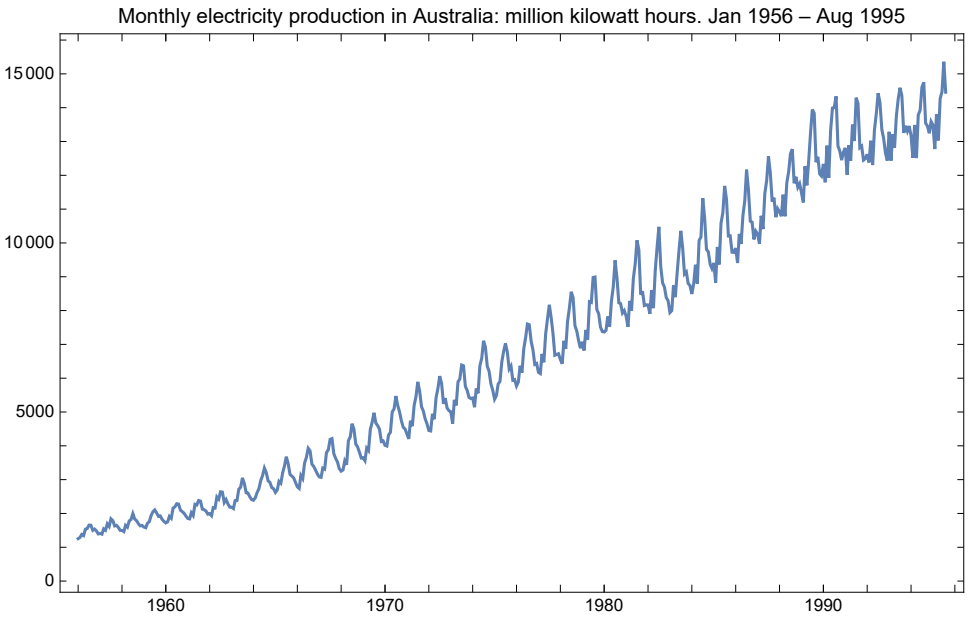
Поведение временных рядов можно описать следующими характеристиками:

- **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.

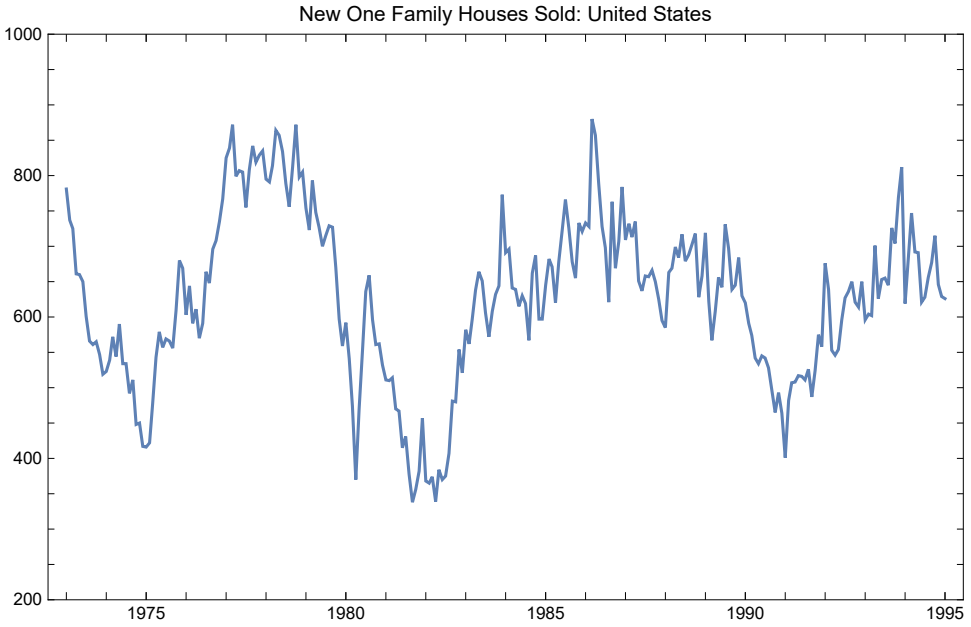
Рассмотрим данные о продажах шампуня по месяцам. На графике виден повышающийся тренд, который можно описать линейной или квадратичной функцией. Сложно выделить на этом участке циклы или сезонность.



Теперь рассмотрим данные 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. 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,

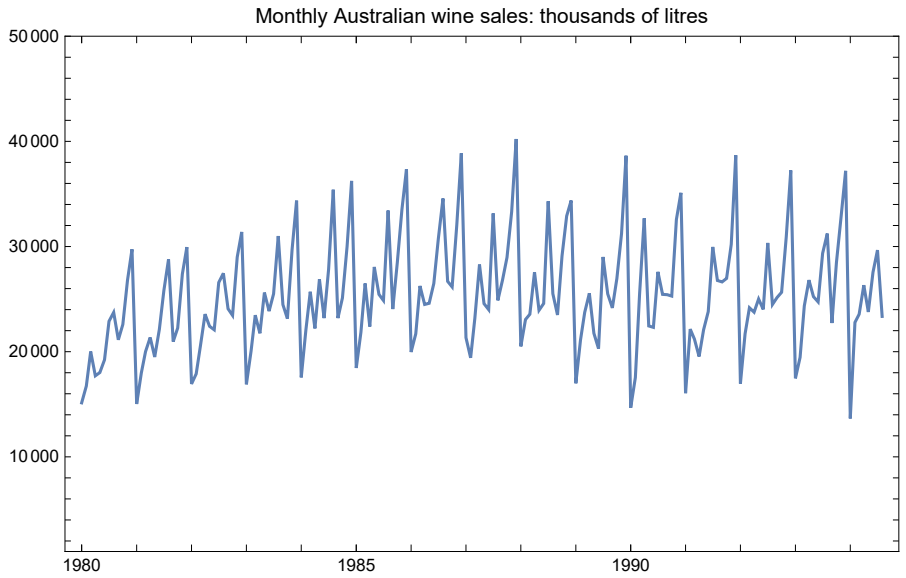
называется лагом автокорреляции (τ).

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

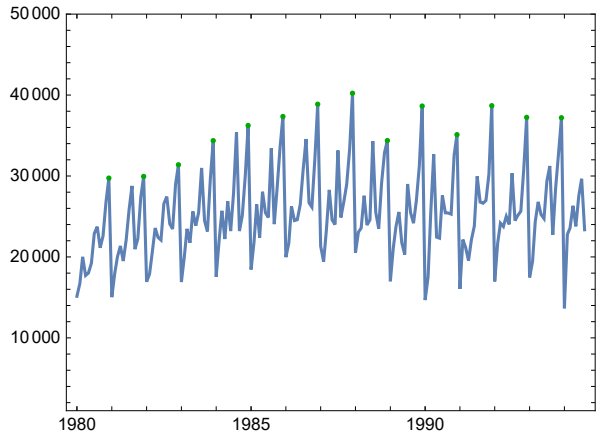
$$r_\tau = \frac{\sum_{t=1}^{T-\tau} (y_t - \overline{y})\,(y_{t+\tau} - \overline{y})}{\sum_{t=1}^{T} (y_t - \overline{y})^2}.$$

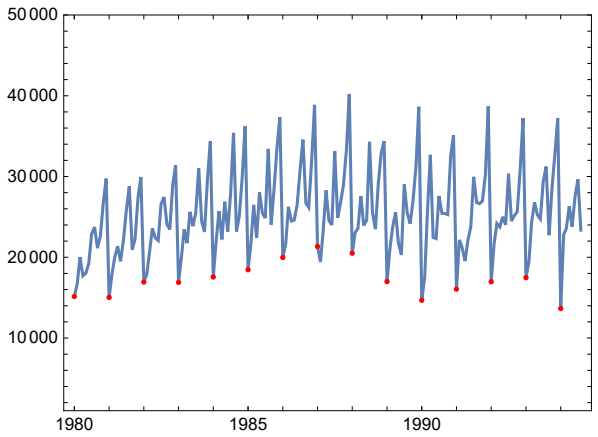
2.2. Диаграмма рассеяния

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

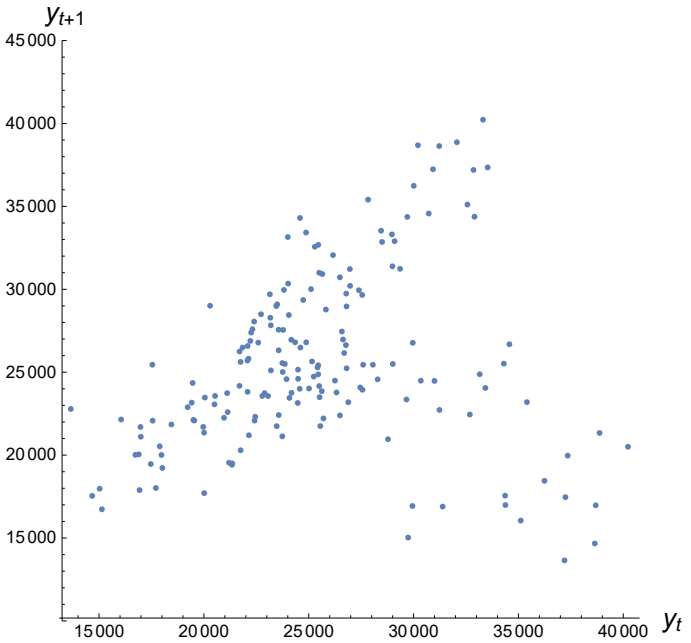


Заметим, что в декабре продажи вина больше, а в январе продажи падают. Значит, ряд обладает ярко выраженной годовой сезонностью.

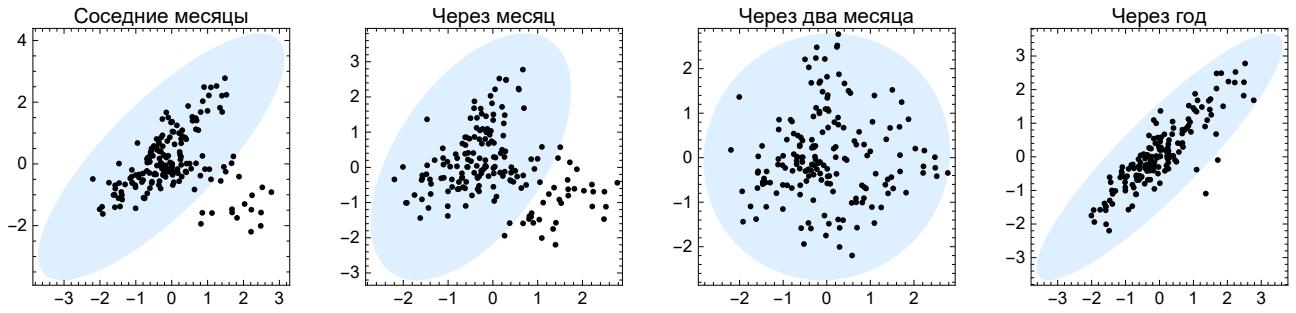




Если построить график зависимости объемов продаж вина в соседние месяцы, то будет видно, что большая часть точек **диаграммы рассеяния** группируется вокруг главной диагонали. Это говорит о том, что в основном значения продаж в соседние месяцы похожи. Еще одно подмножество точек выделяется в правом нижнем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

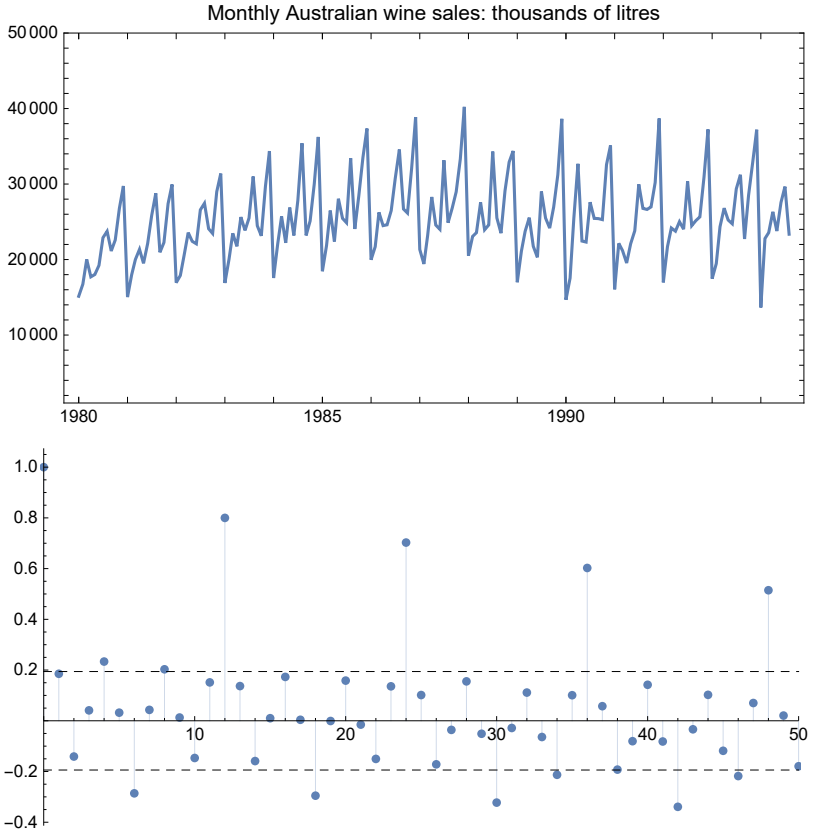


Если построить аналогичный график, но по вертикальной оси отложить y_{t+2}, то видно, что точки в основном облаке начинают «расплываться» вокруг главной диагонали, то есть сходство между продажами через месяц уменьшается по сравнению с соседними месяцами. Если посмотреть связь между продажами через два 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

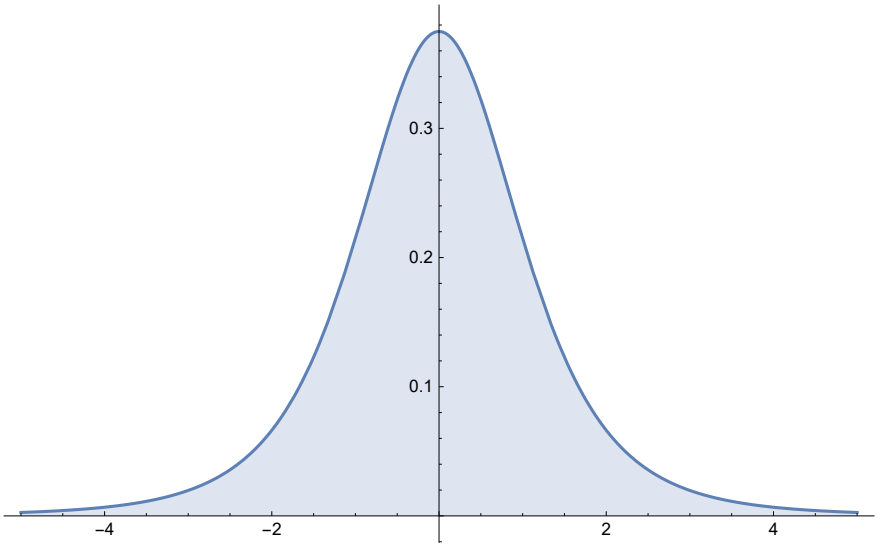
Анализировать величину автокорреляции при разных значениях лагов удобно с помощью графика, который называется **коррелограммой**. По оси ординат на нем откладывается автокорреляция, а по оси абсцисс – размер лага τ. На графике для продаж вина в Австралии видно, что автокорреляция принимает большие значения 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: $r_\tau = 0$
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:	$T(y^T) \sim St(T-\tau-2)$



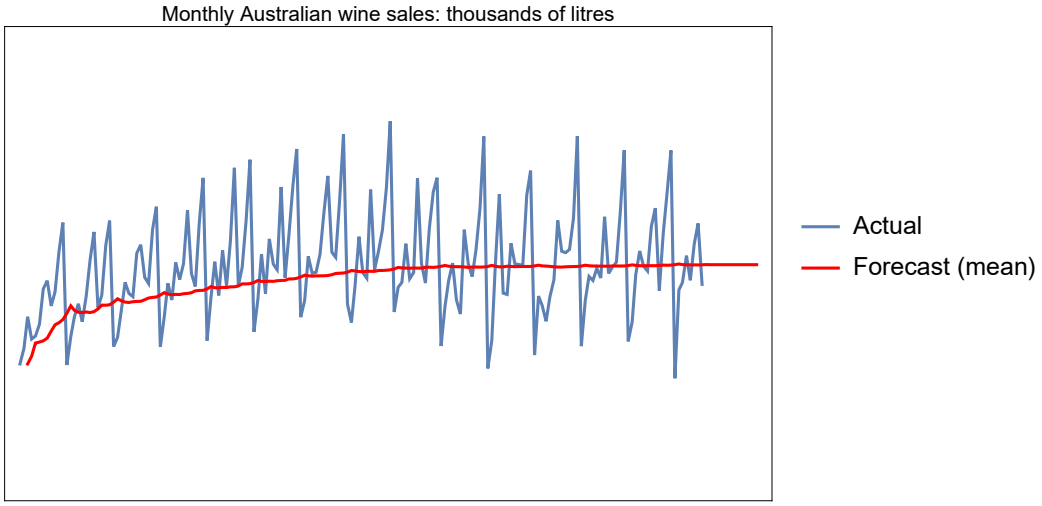
3. Простейшие модели прогнозирования

3.1. Прогноз средним

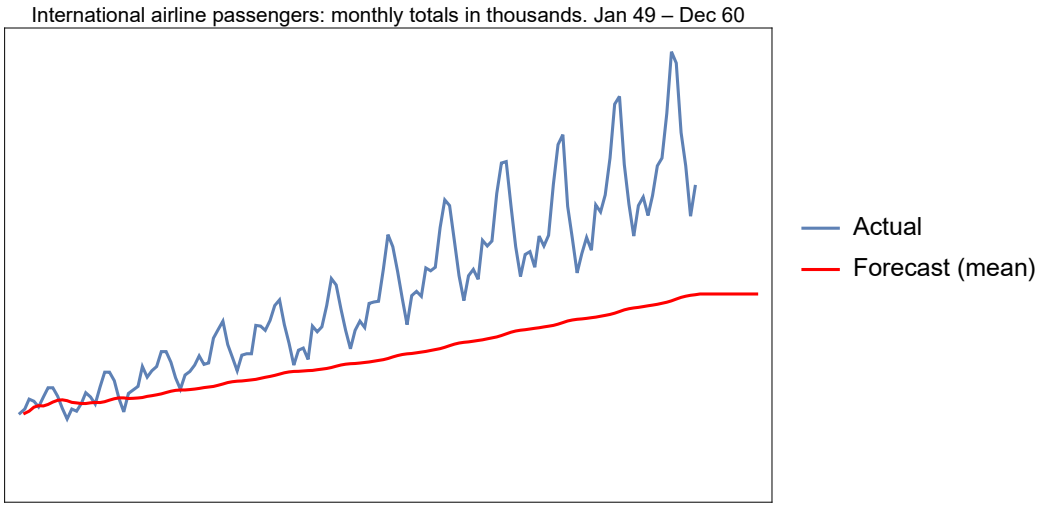
Наиболее простой способ построить прогноз на h точек вперед – посмотреть, какие значения им предшествовали и в качестве прогноза взять среднее значение по всему имеющемуся временному ряду:

$$\hat{y}_{T+h} = \frac{1}{T}\sum_{t=1}^{T} y_t.$$

Такой способ позволяет получить прогноз на сколь угодно длительный период времени и в задачах, где нет явных закономерностей во временном ряду, наивный прогноз средним значением является не худшим вариантом. Рассмотрим данные по продажам вина в Австралии. Построим по имеющимся данным прогноз на 12 месяцев вперед:



Очевидны недостатки такого способа прогнозирования. Если в данных наблюдается тренд, то среднее значение плохо описывает динамику поведения ряда. Это хорошо видно на примере задачи прогнозирования объемов пассажирских авиаперевозок.



Как видно на графике, прогноз начинает сильно отставать от тренда из-за того, что значения признака в начале ряда довольно маленькие по сравнению со значениями в конце ряда и «перетягивают» на себя прогноз.

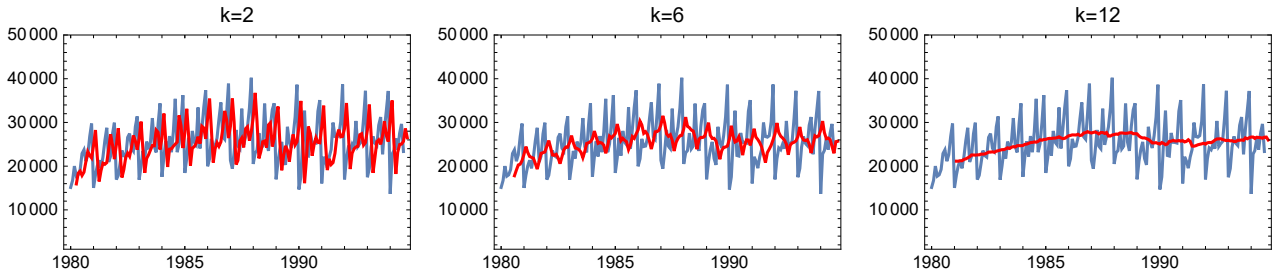
3.2. Скользящее среднее

Чтобы сгладить недостатки предыдущего метода прибегают к методу скользящего среднего. В данном методе прогноз будущего значения признака зависит от среднего от k последних наблюдений:

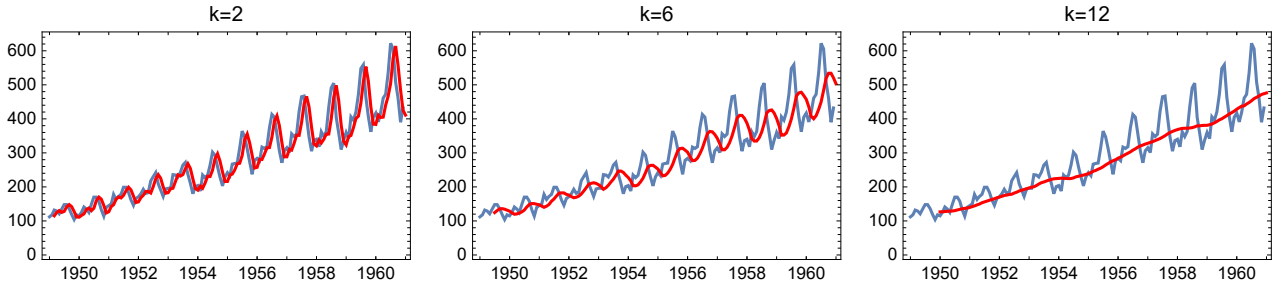
$$\hat{y}_{T+h} = \frac{1}{k} \sum_{t=T-k}^{T} y_t.$$

Долгосрочный прогноз таким способом построить не удастся, поскольку каждое следующее значение зависит от фактически наблюдаемых величин. Однако скользящее среднее позволяет сгладить исходный ряд, выявив таким образом тренд и предоставляя возможность обнаружить закономерности в данных.

Построим модель скользящего среднего на примере продаж вина в Австралии, чтобы убедиться в отсутствии в нем тренда.



Напротив, на примере данных об объемах пассажирских авиаперевозок при применении метода скользящего среднего четко виден повышающийся тренд. Кроме того, при применении метода меньших порядков становятся более очевидными сезонные колебания.



3.3. Наивный прогноз

Еще более простой метод краткосрочного прогнозирования основан на предположении, что будущие значения переменной зависят от последнего наблюдаемого значения:

$$\hat{y}_{T+h} = y_T.$$

Преимущество такого подхода в простоте реализации, а также отсутствии необходимости в большом количестве исторических данных. Однако очевидным серьезным недостатком такого подхода является низкое качество результатов прогнозирования. Если развивать эту идею, то в качестве прогноза можно брать предыдущие значения с сезонным лагом, если в данных наблюдается сезонность:

$$\hat{y}_{T+h} = y_{T+h-kS}, \; k = \left\lfloor \frac{(h-1)}{S} \right\rfloor + 1,$$

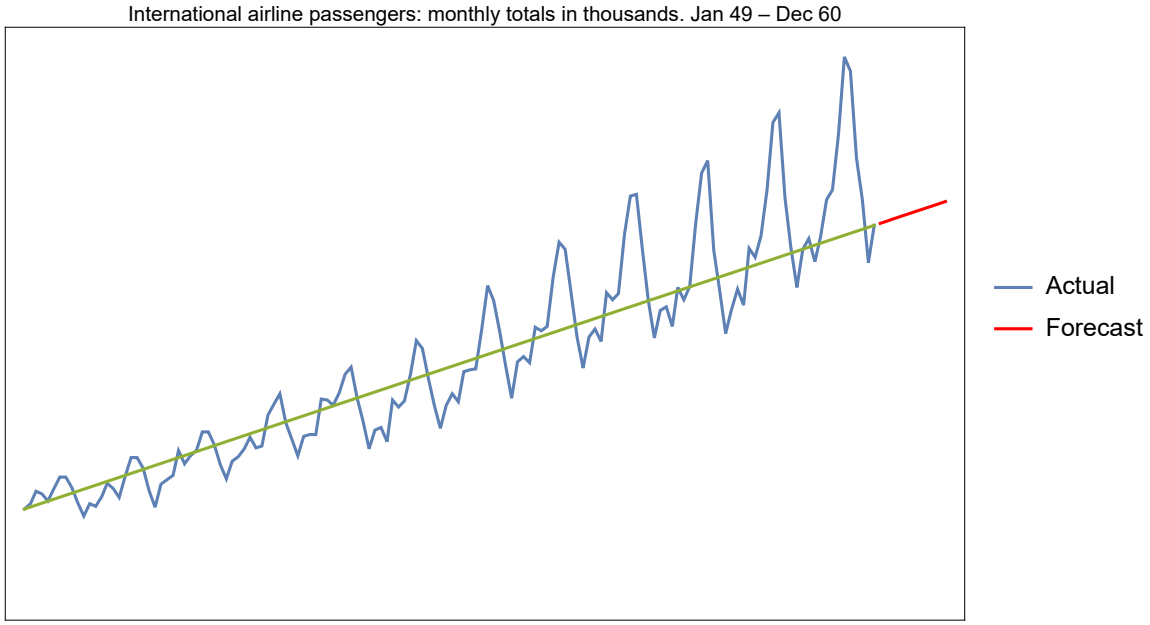
где S – период сезонности, $\lfloor \_ \rfloor$ – частное от деления.

3.4. Экстраполяция тренда

При наличии во временном ряду тренда, можно усложнить наивную модель с помощью экстраполяции тренда:

$$\hat{y}_{T+h} = y_T + h\,\frac{y_T - y_1}{T-1}.$$

Второе слагаемое в данной модели представляет собой ничто иное как уравнение прямой, проходящей через две заданные точки, а именно через первую и последнюю точки временного ряда. Таким образом можно построить долгосрочный прогноз, однако представленная модель не учитывает сезонные колебания.



Упражнение

Какая комбинация простейших методов может улучшить прогноз объемов пассажирских авиаперевозок?