Добавление экзогенных данных в модель

1. Модель SARIMAX

1.1. Двойная сезонность

Модель SARIMA позволяет строить прогнозные модели, учитывающие сезонность во временных рядах. Однако если сезонностей оказывается несколько, то возникает ряд проблем. Например, если имеется ряд дневных данных по спросу на товары, то могут наблюдаться одновременно:

- недельная сезонность (в выходные покупателей больше);
- месячная сезонность (в конце месяца начисляется зарплата, и люди отправляются за покупками);
- годовая сезонность (летом чаще покупают напитки и мороженое).

В первую очередь возникает проблема, как привести такой ряд к стационарному, ведь необходимо провести сезонное дифференцирование. Для этого требуется определить длину сезонного лага, чего не удастся сделать в случае дневных данных по той причине, что длина текущего года может отличаться от длины предыдущего (366 и 365 дней).

Более того, почему для учета годовой сезонности при прогнозировании на 23 марта 2020 года учитывается именно 23 марта 2019? Вдруг в прошлом году на эту дату выпал выходной в связи с чем наблюдался высокий спрос, а модель этого не заметит? Очевидно, нас будут интересовать не продажи в тот же день прошлого года, а продажи в окрестности этого дня (например, несколько дней до и несколько дней после).

Получается, что для описания поведения такого ряда необходимо выявить те характеристики, которые невозможно описать моделью SARIMA и строить модель с учетом данных особенностей ряда.

1.2. Модель SARIMAX

Модель SARIMAX (Seasonal AutoRegressive Integrated Moving Average model with exogenous variables) является расширением модели SARIMA, которая учитывает дополнительные факторы, которые помогают лучше описать поведение рассматриваемого временного ряда.

Данную модель строят в два этапа:
1. На первом шаге выделяют набор факторов X, которые характеризуют те зависимости в данных, которых не может учесть модель авторегрессии. Это могут быть как факторы, полученные непосредственно из временного ряда, так и сторонние, например, погодные: температура воздуха в рассматриваемый день, скорость ветра и т. д. На полученном наборе факторов строят модель линейной регрессии вида:

$$y_t = \sum_{j=0}^{m} \theta_j x_{tj} + \epsilon_t, \quad (1)$$

где m – количество выделенных прогнозных факторов, ϵ_t – непрогнозируемая ошибка модели. Здесь $x_{t0} = 1$ для учета свободного члена в модели.

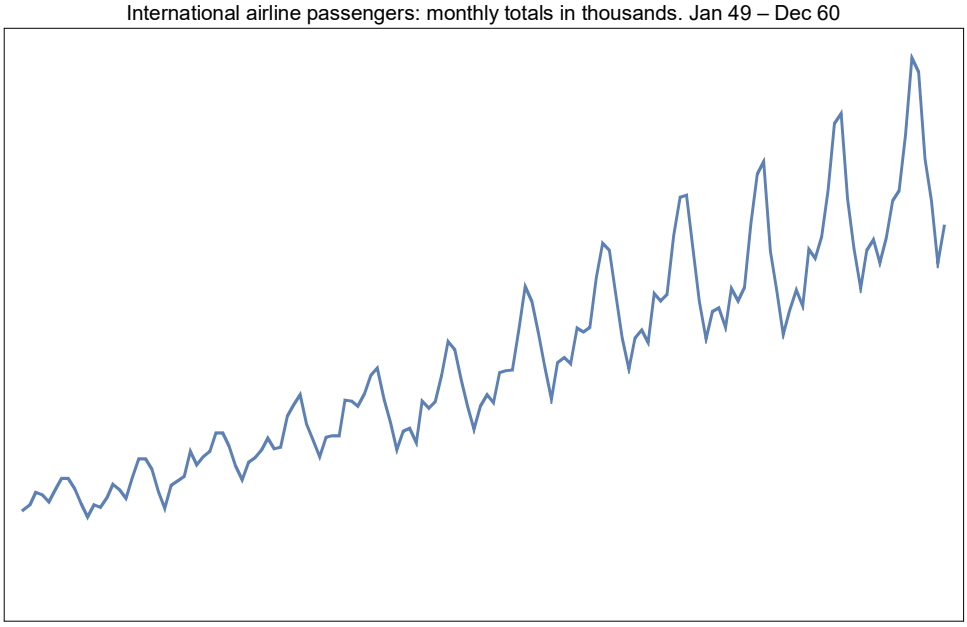
2. Затем находят остатки построенной модели и приближают их моделью SARIMA. Таким образом, если одна из сезонностей была учтена с помощью факторов на первом шаге, то вторую сезонность можно учесть в модели SARIMA.

2. Извлечение факторов из временного ряда

Как правило, если удачно определить набор факторов, описывающих поведение ряда, то остатки модели, построенной на таком наборе, уже могут оказаться шумом. В качестве факторов могут выступать моменты времени, индикаторы для дня недели, статистики, посчитанные на данных из прошлого. Рассмотрим их подробнее.

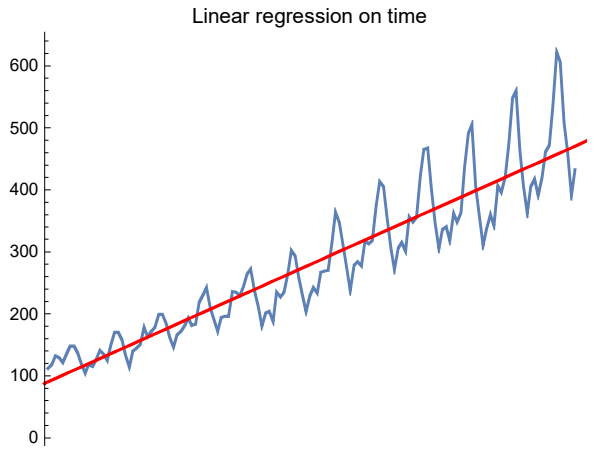
2.1. Зависимость от времени

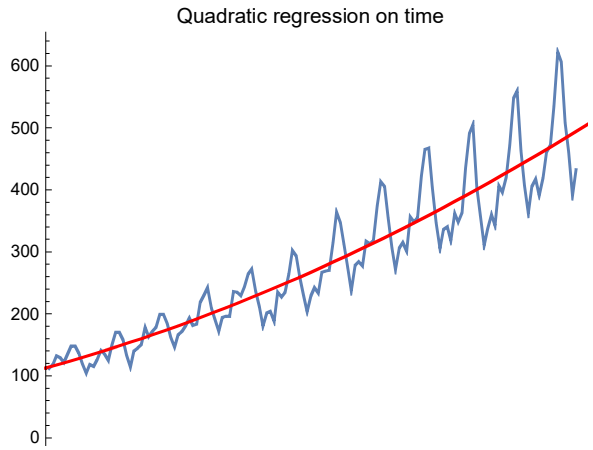
Снова обратимся к данным о пассажирских авиаперевозках. Как было отмечено ранее, со временем наблюдается увеличение объема пассажирских авиаперевозок. Таким образом, можно ожидать, что и дальше будет наблюдаться увеличение пассажиропотока (текущую ситуацию во внимание не принимаем), а значит, описать данную зависимость можно, построив модель линейной регрессии.

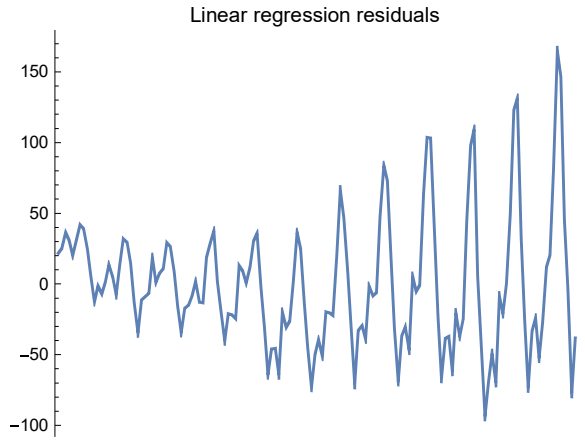


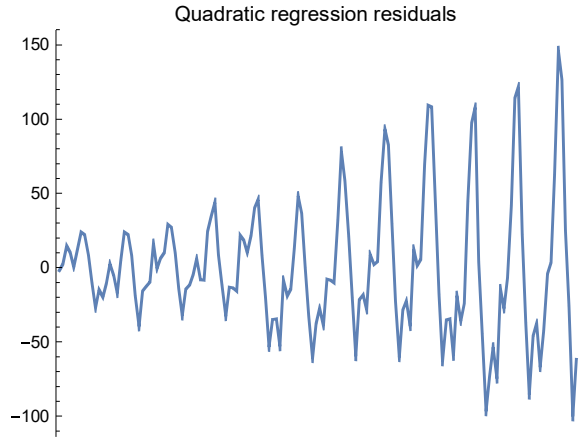
В качестве факторов, описывающих поведение ряда, можно использовать моменты времени t. Поскольку значения целевой переменной y измеряются через равные промежутки времени, то наблюдаемая зависимость:

$$y_t = \sum_{j=0}^{m} \theta_j t^j + \epsilon_t. \tag{2}$$









На левом рисунке целевая переменная описывается моделью

$$y_t = \theta_0 + \theta_1 t + \epsilon_t,$$

а на правом

$$y_t = \theta_0 + \theta_1 t + \theta_2 t^2 + \epsilon_t.$$

Таким образом m определяет степень полинома, которым будет описана целевая переменная.

2.2. Сезонные факторы

В остатках построенной выше модели осталось много информации. Для улучшения качества прогноза необходимо обратить внимание модели на сезонную составляющую. Поскольку целевая переменная измеряется каждый месяц, то в качестве фактора, описывающего сезонность, можно использовать, например, индикатор текущего месяца.

Таким образом, первая запись получит индикатор «1» – январь, вторая «2» – февраль и т. д.

Datetime	NumPassengers	Month
Sat 1 Jan 1949 00:00:00	112	1
Tue 1 Feb 1949 00:00:00	118	2
Tue 1 Mar 1949 00:00:00	132	3
Fri 1 Apr 1949 00:00:00	129	4
Sun 1 May 1949 00:00:00	121	5
Wed 1 Jun 1949 00:00:00	135	6
Fri 1 Jul 1949 00:00:00	148	7
Mon 1 Aug 1949 00:00:00	148	8
Thu 1 Sep 1949 00:00:00	136	9
Sat 1 Oct 1949 00:00:00	119	10
Tue 1 Nov 1949 00:00:00	104	11
Thu 1 Dec 1949 00:00:00	118	12
Sun 1 Jan 1950 00:00:00	115	1
Wed 1 Feb 1950 00:00:00	126	2
Wed 1 Mar 1950 00:00:00	141	3

Если в ряде наблюдается сразу две сезонности, например, годовая и недельная, то можно выделить соответствующие факторы еще и для дня недели. Также важно не забывать о существовании праздников и каникул, ведь праздники также являются выходными и можно в качестве прогнозного фактора модели использовать индикатор «выходной», который принимает значения 1 или 0 (да или нет), куда попадут и праздники или же использовать дополнительный индикатор «праздник».

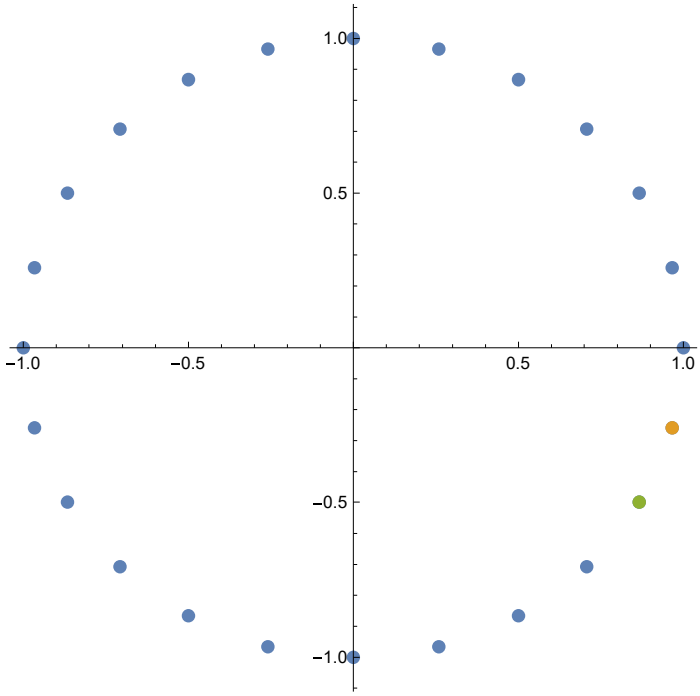
При выделении прогнозных факторов всегда нужно учитывать специфику ряда. Увеличение спроса на цветы наблюдается не только 8 марта, но и несколько дней накануне, поэтому, если решается задача прогнозирования какого-то специфического товара, важно выделить в качестве прогнозных факторов дни особого спроса, характерного только для данного товара.

2.3. Тригонометрический ряд Фурье

Хорошим способом описать сезонные колебания является применение Фурье-преобразований. Недостатком предыдущего способа «кодирования» месяца является то, что расстояние от января до декабря оказалось больше, чем от января до октября, что мешает модели выявить соответствующие закономерности. Решить данную проблему позволяет применение следующих преобразований:

$$\varphi_i = \sin\left(\frac{2\pi(i-1)}{s}\right), \quad \psi_i = \cos\left(\frac{2\pi(i-1)}{s}\right).$$

где i – соответствующее наблюдению значение (текущий час / день недели / месяц, отсчет при этом ведется **с нуля**), s – длина сезонного периода (24 / 7 / 12). Тригонометрические функции описывают те колебания, которые наблюдаются при сезонной цикличности. Посмотрим, как можно описать дневную сезонность. В данном случае длина сезонного периода составляет 24 часа. Применим указанные выше функции к значениям от 0 до 23 и отобразим на графике пары точек $\{\varphi_i, \psi_i\}$.



Выделенные точки соответствуют значениям 0 и 23, а значит 0 находится от 22 на таком же расстоянии, как и от 2, что соответствует реальной разнице во времени. Таким образом, значения φ_i, ψ_i позволяют моделировать сезонные колебания с **постоянной** амплитудой. Добавим данные факторы в модель (2), получим:

$$y_t = \sum_{j=0}^{m} \theta_j t^j + \beta_1 \varphi_{t \bmod s} + \beta_2 \psi_{t \bmod s} + \epsilon_t. \quad (3)$$

Коэффициенты β_1 и β_2 могут быть найдены методом наименьших квадратов.

2.4. Обработка категориальных переменных

При решении практических задач возникает необходимость также кодировать категориальные переменные. К категориальным переменным относятся, например, город, в котором совершалась покупка или тип осадков, если для прогнозирования используются погодные факторы. Можно закодировать все города, где есть торговые точки сети через некоторые идентификаторы. Допустим, у нас есть точки в Москве, Санкт-Петербурге и Сочи. Тогда

Москва → 1,
Санкт-Петербург → 2,
Сочи → 3.

Оказалось, что Санкт-Петербург и Сочи больше Москвы и такой способ кодирования не несет никакой информации. В этом примере можно отталкиваться от населения города. Поскольку Москва – самый большой город в списке, присвоим ей идентификатор 3. Возможно, в Сочи больше торговых точек нашей сети, нежели в Петербурге и логично было бы присвоить городу Сочи идентификатор 2. Таким образом, способ предобработки данных всегда определяется из специфики задачи, однако есть и универсальные методы.

Одним из них является **One-Hot-Encoding** или кодирование в унитарный код, который каждому значению сопоставляет двоичный код фиксированной длины, содержащий только одну 1 на позиции, соответствующей данному значению.

Datetime	NumPassengers	Jan	Feb	Mar	Apr	May	Jun	Jul
Sat 1 Jan 1949 00:00:00	112	1	0	0	0	0	0	0
Tue 1 Feb 1949 00:00:00	118	0	1	0	0	0	0	0
Tue 1 Mar 1949 00:00:00	132	0	0	1	0	0	0	0
Fri 1 Apr 1949 00:00:00	129	0	0	0	1	0	0	0
Sun 1 May 1949 00:00:00	121	0	0	0	0	1	0	0
Wed 1 Jun 1949 00:00:00	135	0	0	0	0	0	1	0
Fri 1 Jul 1949 00:00:00	148	0	0	0	0	0	0	1
Mon 1 Aug 1949 00:00:00	148	0	0	0	0	0	0	0
Thu 1 Sep 1949 00:00:00	136	0	0	0	0	0	0	0
Sat 1 Oct 1949 00:00:00	119	0	0	0	0	0	0	0
Tue 1 Nov 1949 00:00:00	104	0	0	0	0	0	0	0
Thu 1 Dec 1949 00:00:00	118	0	0	0	0	0	0	0
Sun 1 Jan 1950 00:00:00	115	1	0	0	0	0	0	0
Wed 1 Feb 1950 00:00:00	126	0	1	0	0	0	0	0
Wed 1 Mar 1950 00:00:00	141	0	0	1	0	0	0	0

Важно отметить, что при применении такого способа для описания всех 12 месяцев достаточно 11 колонок, поскольку последняя выражается через 11 других. При попытке построить модель на всех 12 факторах возникнет проблема мультиколлинеарности.

Недостатком такого метода является возникновение большой разреженной матрицы переменных, которую приходится хранить в памяти. Другим, менее требовательным к объему памяти способом является кодирование переменных двоичным представлением. То есть, для кодирования 8 различных значений достаточно трех колонок ($2^3 = 8$). Для кодирования 12 – 4 колонки соответственно. В результате каждому месяцу соответствует уникальный набор значений e1-e4:

Datetime	NumPassengers	e1	e2	e3	e4
Sat 1 Jan 1949 00:00:00	112	0	0	0	0
Tue 1 Feb 1949 00:00:00	118	0	0	0	1
Tue 1 Mar 1949 00:00:00	132	0	0	1	0
Fri 1 Apr 1949 00:00:00	129	0	0	1	1
Sun 1 May 1949 00:00:00	121	0	1	0	0
Wed 1 Jun 1949 00:00:00	135	0	1	0	1
Fri 1 Jul 1949 00:00:00	148	0	1	1	0
Mon 1 Aug 1949 00:00:00	148	0	1	1	1
Thu 1 Sep 1949 00:00:00	136	1	0	0	0
Sat 1 Oct 1949 00:00:00	119	1	0	0	1
Tue 1 Nov 1949 00:00:00	104	1	0	1	0
Thu 1 Dec 1949 00:00:00	118	1	0	1	1
Sun 1 Jan 1950 00:00:00	115	0	0	0	0
Wed 1 Feb 1950 00:00:00	126	0	0	0	1
Wed 1 Mar 1950 00:00:00	141	0	0	1	0

Важное замечание: представленные способы хоть и продемонстрированы на примере кодирования месяца, но они не решают проблему расстояний между месяцами!

2.5. Дополнительная информация

Основные закономерности в данных были учтены при использовании методов 2.1-2.3. Чтобы учесть отклонения от основных тенденций в качестве дополнительных факторов могут рассматриваться различные статистики, например:

- средний показатель за тот же месяц в предыдущие годы;
- среднеквадратичное отклонение показателя за тот же месяц в предыдущие годы;
- отклонение показателей на прошлой неделе по сравнению с позапрошлой при прогнозировании на неделю вперед;
- и т. д.

Таким образом, задача сводится к выявлению набора факторов (переменных, признаков), наилучшим образом описывающих наблюдаемые закономерности.

Задание

Для ряда пассажирских авиаперевозок постройте прогнозную модель вида (3). Подумайте, как можно модифицировать модель на случай мультипликативной сезонности. Постройте вашу собственную модель, учитывающую мультипликативную сезонность. Отобразите результат на графике.

Найдите остатки построенной модели. Является ли ряд остатков стационарным? Для проверки отобразите ряд остатков и воспользуйтесь критерием Дики-Фуллера (UnitRootTest в математике).

Постройте авторегрессионную модель прогнозирования ряда остатков. Итоговую модель SARIMAX отобразите на графике.

Указания

Выполнять желательно в математике, но можно и в питоне. Для построения модели ARIMA можно пользоваться встроенными функциями.

```
passengers = ExampleData[{"Statistics", "InternationalAirlinePassengers"}, "TimeSeries"];
```