Методы прогнозирования. Билеты

Александр Широков ПМ-1701

Преподаватель:

Ивахненко Дарья Александровна

Санкт-Петербург
2020 г., 6 семестр

Список литературы

[1] GLM An Applied Approach, Ulf Olsson

Содержание

1 Понятие прогноза и прогнозирования. Компоненты уровня динамического ряда

1.1 Понятие временного ряда, прогноза и прогнозирования

Определение 1.1.1. Временной ряд - последовательность значений признака y, измеряемого через постоянные временные интервалы:

$$y_1, y_2, \ldots, y_T, \qquad y_t \in R$$

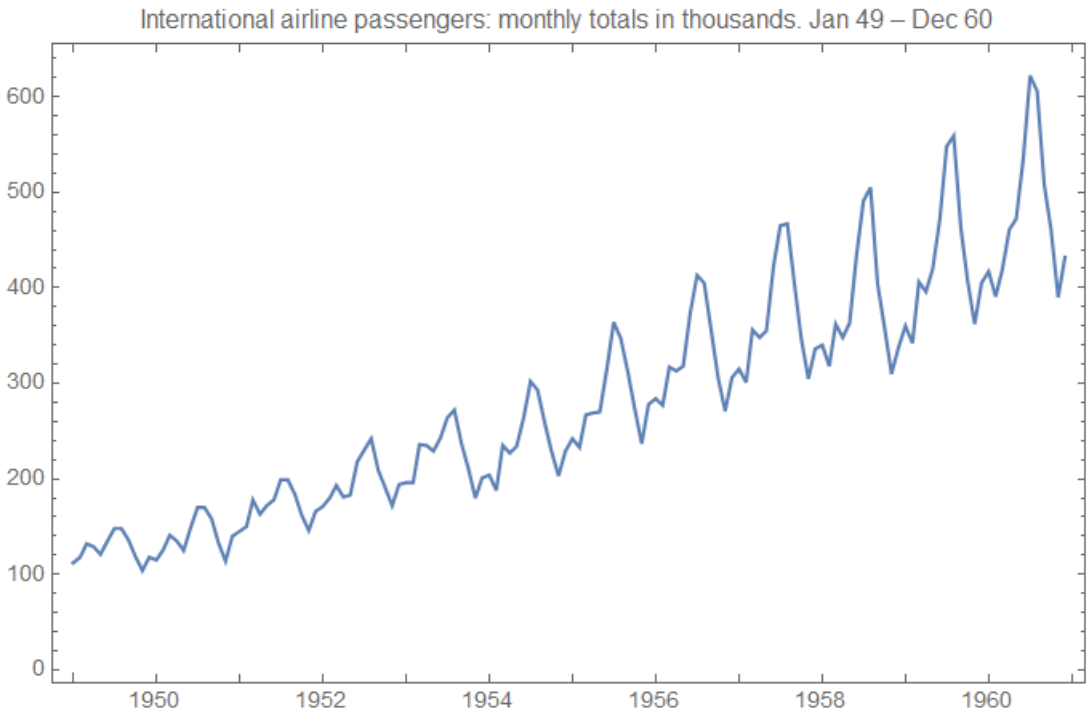

Рис. 1: Временной ряд

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

Определение 1.1.2. Предсказательный интервал - интервал, в котором предсказываемая величина окажется с вероятностью не меньше заданной.

1.2 Компоненты уровня динамического ряда

Поведение временных рядов можно описать следующими характеристиками:

- **Тренд** - плавное долгосрочное изменение уровня ряда.

  Тренд отражает общее направление временного ряда, он может быть повышающимся, понижающимся или постоянным. Тренд может увеличиваться или уменьшаться различными способами (линейно, экспоненциально).

- **Сезонность** - циклические изменения уровня ряда с постоянным периодом.

  Сезонность фиксирует эффекты, которые происходят с определенной частотой. Это можт быть обусловлено большим количеством факторов, например, сезонность, связанная со сменой времен года и.т.д

- **Ошибка** (остатки, случайная компоненты) - непрогнозируемая случайная компонента ряда.

  Остатки - это оставшиеся случайные колебания, после того, как тренд и сезонность удаляются из оригинального временного ряда. В остатках мы *не должны видеть тенденцию или сезонность*, так как остатки представляют собой краткосрочные колеабния. Остатки либо случайные, либо являются частью компонентов тренда и сезонности, которые *были пропущены при разложении* временного ряда.

  Соответственно, вид остатков показывает, содержится ли в остатках большая часть информации, которая не была учтена и можно ли построить более сложную модель, которая будет лучше описывать имеющиеся данные.

- **Цикл** - изменения уровня ряда с переменным периодом (экономические циклы, периоды солнечной активности)

- **Разладка** - смена модели ряда.

Определение 1.2.1. Аддитивная модель - наблюдаемый временной ряд является суммой его компонент - тренда T, сезонности S и ошибки

E:

$$TS = T + S + E$$

Модель является аддитивной, если сезонные колебания и остаточная ошибка колеблятся вместе с изменением тренла и если амплитуда колебаний значений ряда вокргу тренда примерно одинаковая

Определение 1.2.2. Мультипликативная модель - мультипликативные модели предполагают, что наблюдаемый временной ряд является **произведением его компонент**:

$$TS = T \cdot S \cdot E$$

В случае, когда сезонные колебания изменяются пропорционально уровню ряда, говорят о мультипликативной сезонности.

2 Автокорреляция. Коррелограмма. Частичная автокорреляция

2.1 Автокорреляция

Ряд может содержать, помимо случайной составляющей, либо тенденцию, либо только сезонную компоненту, либо все компоненты вместе. Для того, чтобы выявить наличие той или иной неслучайной компоненты, исследуется корреляционная зависимость между последовательными уровнями временного ряда или **автокорреляция уровней ряда**.

Основная идея - при наличии тенденции и циклических колебаний, значения каждого последующего уровня ряда зависят от предыдущих.

Определение 2.1.1. Автокорреляция - количественная характеристика сходства между значениями ряда в соседних точках. Автокорреляционная функция задается следующим соотношением:

$$r_\tau = \frac{E(y_t - Ey)(y_{t+\tau} - Ey)}{Dy}$$

Коэффициент автокорреляции уровней ряда *первого порядка* измеряет зависимость между соседними уровнями ряда t и $t-1$, то есть при *лаге* 1.

Если значение коэффициента автокорреляции близко к единице, это указывает на очень тесную зависимость между соседними уровнями временного ряда и о наличии во временном ряде сильной линейной тенденции, так как автокорреляция - это *корреляци Пирсона - выборочного коэффициента корреляции* между исходным рядом и его версией, сдвинутой на несколько отсчетов.

Аналогично определяются коэффициенты автокорреляции более высоких порядков:

$$r_\tau = \frac{\sum\limits_{t=1}^{T-\tau}(y_t - \bar{y})(y_{\tau+t} - \bar{y})}{\sum\limits_{t=1}^{T}(y_t - \bar{y})^2}$$

$$r_{\tau+1} = \frac{\sum\limits_{t=\tau}^{T}(y_t - \bar{y})(y_{t-\tau} - \bar{y})}{\sum\limits_{t=1}^{T}(y_t - \bar{y})^2}$$

где T - длина временного ряда.

Определение 2.1.2. Количество отсчетов, на которое сдвинут ряд, называется *лагом автокорреляции* τ (число периодов, по которым рассчитывается коэффициент автокорреляции)

Автокорреляция часто приводит к какому-то известному нам виду графика, тогда как временной ряд без автокорреляции будет проявлять случайность.

Свойства автокорреляции

1. Коэффициент автокорреляции строится по аналогии с линейным коэффициентом корреляции, поэтому он характеризует тесноту *только линейной связи* текущего и предыдущего уровня. Для временных рядов с нелинейной тенденции, коэффициент автокорреляции будет близок к 0, хотя тенденция все же в ряде есть.

2. По знаку коэффициента автокорреляции *нельзя* делать выпод о возрастающей или убывающей тенденции в уровнях ряда.

Определение 2.1.3. Последовательность коэффициентов автокорреляции уровней различных порядков, начиная с первого, называется *автокорреляционной функцией временного ряда* (**ACF**).

2.2 Коррелограмма

Определение 2.2.1. График зависимости автокорреляционной функции от величины лага называется *коррелограммой.*

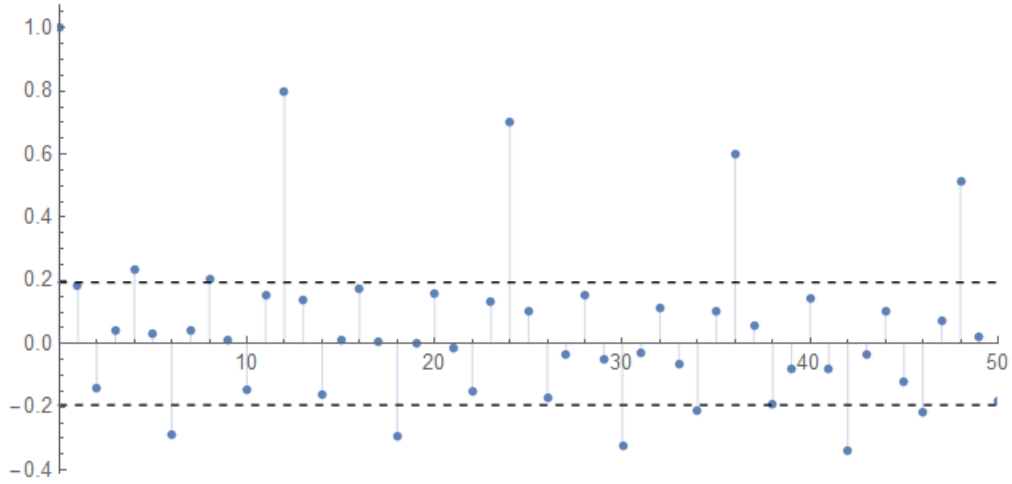

Рис. 2: Коррелограмма

С помощью коррелограммы удобно анализировать при разных значениях лага величину автокорреляции. По оси ординат откладывается автокоореляции, по оси абцисс - размер лага τ.

Анализ автокорреляционной функции позволяет определить лаг, при котором автокорреляция наиболее высокая, лаг, при котором связь между текущим и предыдущим уровнями ряда наиболее тесная.

2.2.1 Значимость автокорреляции

На коррелогрмме помимо значений автокорреляции также изображен коридор вокруг горизонтальной оси.

Это коридор *значимости отличия корреляции от нуля.* Как и для обычной корреляции Пирсона, значимость вычисляется при помощи критерия Стьюдента.

В качестве нулевой гипотезы предполагают, что $H_0 : r_\tau = 0$. Альтернативной гипотезой является двустороння гипотеза о том, что $H_1 : r_\tau <\neq> 0$, потому что при анализе временных рядов крайне редко имеется гипотеза о том, какой по знаку должна быть корреляция - положительной или отрицательной.

Статистическую значимость r_τ можно определить с помощью t-

критерия:

$$T(y) = \frac{r_\tau}{\sqrt{1 - r_\tau^2}} \cdot \sqrt{T - \tau - 2} \sim t(T - \tau - 2)$$

Если $T(y) > t_{table}(T - \tau - 2)$, то принято сичтать, что корреляция статистически значима.

То есть значения автокорреляции, выходящие за пределы коридора вокруг горизонтальной оси, являются статистически значимыми, а значения внутри коридора - *статистически незначимы*

2.2.2 Анализ коррелограммы

1. Если наиболее высоким является коэффициент автокорреляции первого порядка, то исследуемый ряд содержит *только тенденцию.*

2. Есди наиболее высоким оказался коэффициент автокорреляции *второго порядка*, то ряд содержит циклические колебания с циклом, равным двум периодам времени, то есть имеет *пилообразную структуру.*

3. Если наиболее высоким оказался коэффициент автокорреляции порядка τ, то ряд содержит циклические колебания с периодичностью в τ моментов времени.

4. Если ни один из коэффициентов τ не является значимым, то либо ряд не содержит тенденции и циклических колебаний и имеет только случайную составляющую, либо содержит сильную нелинейную связь, для исследования которой нужно провести дополнительный анализ.

2.2.3 Диаграмма рассеяния

Если построить график зависимости y_t от y_{t+1}, то такой график будет называться **диаграммой рассеивания**. Если рассмотреть продажи в одни и те же месяцы соседних лет, то точки на графике стягиваются к главной диагонали. Это значит, что значения в одни и те же месяцы соседних лет сильно похожи:

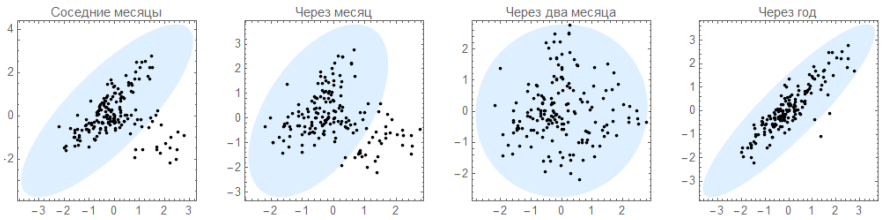

Рис. 5: Различные диаграммы рассеивания, на последнем рисунке видна линейная тесная связь

3 Понятие стационарности

Определение 3.0.1. Стационарный временной ряд - такой временной ряд $y_1, \ldots, y_T$, в которой свойства ряда не зависят от времени, то есть $\forall k$ (ширина окна) распределение $y_t, \ldots, y_{t+k}$ не зависит от времени.

Данный ряд обладает следующими особенностями, следующими из определения:

- **Постоянное среднее значение**, то есть у стационарного ряда *нет тренда*
- **Постоянная дисперсия** (гомоскедастичность), то есть дисперсия не зависит от другой случайной величины. Если дисперсия не постоянна и зависит от предыдущих наблюдений, то ряд нестационарен
- **Постоянная автокорреляционная структура**, то есть значения автокорреляции незначимы - отсутствует линейная связь с предыдущими значениями.
- **Отсутствие сезонности**, ведь если ширина окна меньше сезонного периода, то распределение будет разным, в зависимости от положения окна.

Замечание: ряды в которых есть непериодические циклы, не обязательно являются нестационарными.

Почему же стационарность настолько важна? Стационарность является фундаментальным предположением в большинстве случаев прогнозирования моделей временных рядов:

- Без стационарности множество базисных моделей временных рядов попросту не работали бы
- Преобразования могут применяться для преобразования нестационарного временного ряда в стационарный перед моделированием, например, поведения ряда или прогноза.

По стационарному ряду просто строить прогноз, так как предполагается, что его будущие характеристики не будут отличаться от наблюдаемых текущих.

3.1 Способы определения стационарности ряда

Для того, чтобы определить, является ли ряд стационарным, можно использовать несколько способов:

- Построение графика

  С самого начала необходимо построить **график** временного ряда. Обычно, данного шага достаточно для определения, является ли ряд нестационарным, потому что видно либо наличие сезонности, либо тренда, либо автокорреляции, либо непостоянство дисперсии, которую необходимо стабилизировать.

- Оценка с помощью статистик

  Вычисления **математических ожиданий и дисперсий** на протяжении всего временного ряда является важным способом определения, в каких моментах ряд является стационарным. Легким способом является **разделение временного ряда** на *несколько временных периодов* и *подсчет* статистических значений на данных промежутках.

  Большие отклонения в среднем или дисперсии между различными промежутками означает, что данные *нестационарны*.

  Если мы хотим еще более уточнить, то можем воспользоваться тестами, чтобы определить, является ли разница в среднем или разница в отклонениях статистически значимой.

- Гистограмма

  Гистограмма дает хорошие предположения о виде ряда. Если распределение похоже на *нормальное*, то ряд практически наверняка является *стационарным*.

- Тест Дики-Фуллера

  Гипотезу о стационаности можно проверить с помощью критерия Дики-Фуллера. Статистику данного критерия рассмотрим чуть позже.

3.2 Преобразование ряда в стационарный

Существует несколько способов преобразования нестационарного ряда в стационарный:

- Убрать тренд
- Привести ряд к ряду с постоянной дисперсией путем логарифмирования
- Убрать автокорреляцию путем диффенцирования ряда
- Убрать сезонность

Зачастую приходится делать несколько из данных преобразований в одном наборе данных. Рассмотрим последовательно каждое из преобразований.

3.2.1 Стабилизация дисперсии. Преобразование Бокса-Кокса

Если во временном ряде **монотонно по времени изменяется дисперсия**, то применяется специальное преобразование, стабилизирующее дисперсию. Часто в качестве такого преобразования модно использовать простое логарифмирование.

В результате логарифмирования размер колебаний в начале и конце ряда становится очень похожим и дисперсия стабилизируется.

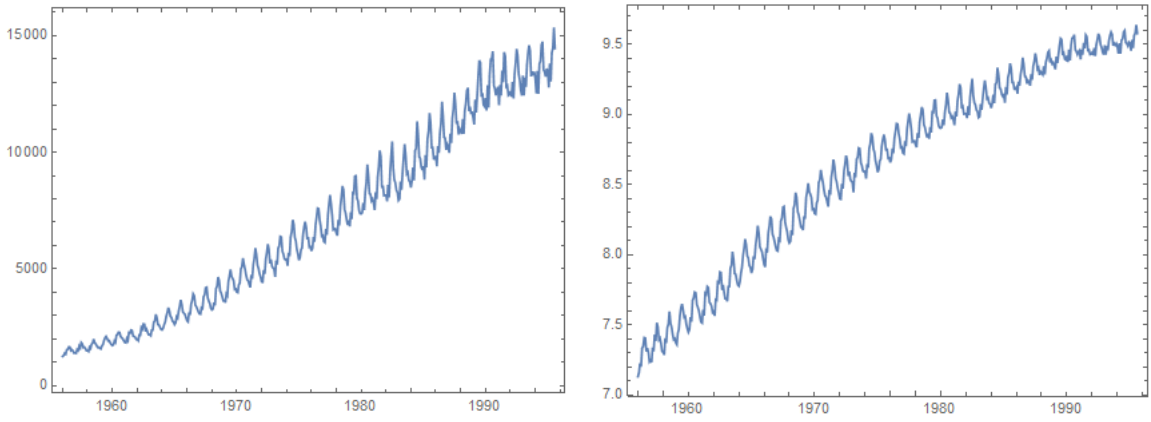

Рис. 4: Слева - исходный ряд, справа - логарифмированный

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

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

Заметим, что экспонента может быть разложена в ряд Тэйлора:

$$e^x = 1 + \sum_{n=1}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \dots$$

поэтому:

$$y^{(\lambda)} = e^{\lambda \ln y} = 1 + \lambda \ln y + O((\lambda \ln y)^2)$$

и получаем, что $y^{(\lambda)} = \ln y$ для $\lambda \to 0$.

Параметр λ определяет, как будет преобразован ряд. Если $\lambda = 0$, то будет произведено логарифмирование, если $\lambda = 1$ - тождественное преобразование со смещением на единицу, при других λ - степенное преобразование.

Значение параметра λ применяется так, чтобы дисперсия была как можно более стабильной во времени. Так, параметр λ выбирается методом максимального правдоподобия.

Найдем функцию правдоподобия:

Доказательство.

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

Прологорифмируем:

$$\ln L = \sum_{i=1}^{T} \left(\ln \frac{1}{\sqrt{2\pi} \cdot \sigma} - \frac{(y_i^{(\lambda)} - \overline{y}^{(\lambda)})^2}{2\sigma^2} + \ln J(\lambda, y) \right)$$

Воспользуемся оценкой выборочной дисперсии:

$$\hat{\sigma}^2 = \sum_{i=1}^{T} \frac{(y_i^{(\lambda)} - \overline{y}^{(\lambda)})^2}{T}$$

$$\ln \frac{1}{\sqrt{2\pi\sigma^2}} = \ln \frac{1}{\sqrt{2\pi}} + \ln \frac{1}{\sqrt{\sigma^2}} = \ln \frac{1}{\sqrt{2\pi}} - \ln \frac{(y_i^{(\lambda)} - \overline{y}^{(\lambda)})^2}{T}$$

Тогда:

$$\ln L = \sum_{i=1}^{T} \ln \frac{1}{\sqrt{2\pi}} - \ln \frac{(y_i^{(\lambda)} - \overline{y}^{(\lambda)})^2}{T} - \frac{T}{2} + \log \prod_{i=1}^{T} y_i^{\lambda-1} =$$

$$= \sum_{i=1}^{T} \ln \frac{1}{\sqrt{2\pi}} - \sum_{i=1}^{T} \ln \frac{(y_i^{(\lambda)} - \overline{y}^{(\lambda)})^2}{T} - \frac{T}{2} + (\lambda - 1) \sum_{i=1} \ln y_i =$$

$$= -\frac{T}{2} \cdot \ln \sum_{i=1}^{T} \frac{(y_i^{(\lambda)} - \overline{y}^{(\lambda)})^2}{T} + (\lambda - 1) \sum_{i=1} \ln y_i$$

Итого: логафрим функции правдоподобия Бокса-Кокса:

$$\ln L = -\frac{T}{2} \cdot \ln \sum_{i=1}^{T} \frac{(y_i^{(\lambda)} - \overline{y}^{(\lambda)})^2}{T} + (\lambda - 1) \sum_{i=1} \ln y_i$$

где

$$\overline{y}^{(\lambda)} = \frac{1}{T} \sum y_i^{(\lambda)}$$

T - количество элементов в выборке.

Взяв производную по параметру λ и прировняв к нулю, найдем решение, получим оценку методом максимального правдоподобия. □

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

$$y^{(\lambda)} = \begin{cases} \frac{(y+\lambda_2)^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ \ln(y + \lambda_2), \lambda = 0 \end{cases}, \qquad y > \lambda_2$$

При возвращении к исходному ряду важно и проделывать *обратное преобразование Бокса-Кокса:*

$$y^{(\lambda)} = \begin{cases} e^y \cdot \frac{\ln(\lambda \cdot y)+1}{\lambda}, \lambda \neq 0 \\ e^y, \lambda = 0 \end{cases}$$

3.2.2 Дифференцирование

Как было сказано ранее, дифференцирование - один из способов привести ряд к стационарному, убрав из него автокорреляцию.

Определение 3.2.1. Дифференцирование - переход к попарным разностям соседних значений:

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

Мы вычитаем прошлую стоимость, отсоящую от нынешней на, допустим, 1 день, как в формуле выше.

Данная операция позволяет стабилизировать среднее значение ряда и избавиться от тренда, а иногда даже от сезонности. Дифференцирование можно применять неоднократно: от ряда первых разностей, продифференцировав его, можно прийти к ряду вторых разностей и.т.д.

Длина ряда, соответственно, буедт сокращаться на величину порядка разностей, но при этом мы добьемся стационарности.

Определение 3.2.2. Сезонное дифференцирование ряда - переход к попарным разностям в соседних сезонах. Если длина периода сезонности составляет s, то новый ряд задается разностями:

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

Можно проводить сезонное и обычное дифференцирование в любом порядке, однако, если у ряда есть *ярковыраженный сезонный профиль*, то рекомендуется начинать с сезонного дифференцирования.

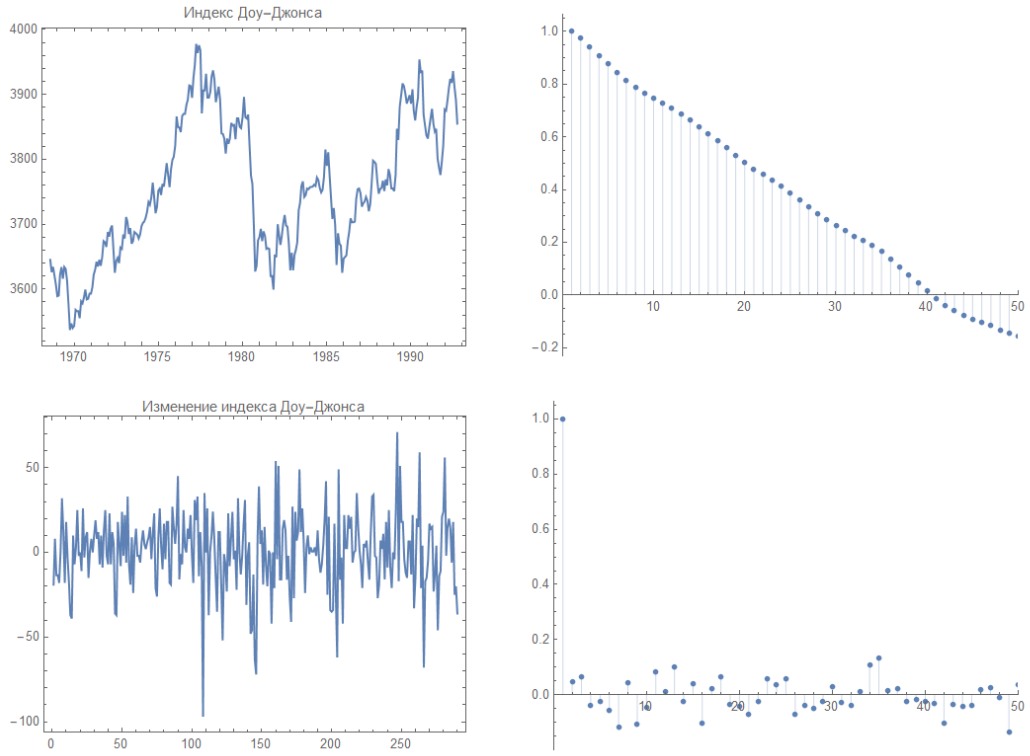

Рис. 5: Вверху - нестационарный ряд, имеется ярко выраженный тренд, внизу - стационарный после сезонного дифференцирования

Обратное преобразование

Исходя из правил дифференцирования, перевод к исходному временному ряду может быть получен по следующему правилу:

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

где y_t - исходный ряд, а k - лаг дифферницирования.

4 Наивные методы прогнозирования

4.1 Прогноз средним (Simple Average)

Наиболее простой способ построить прогноз на h точек вперед - посмотреть, какие значения име 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 позволет 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.

Недостатки прогноза средним: если в данных наблюдается *тренд*, то среднее значение плохо описывает динамику поведения ряда, очень плохо..

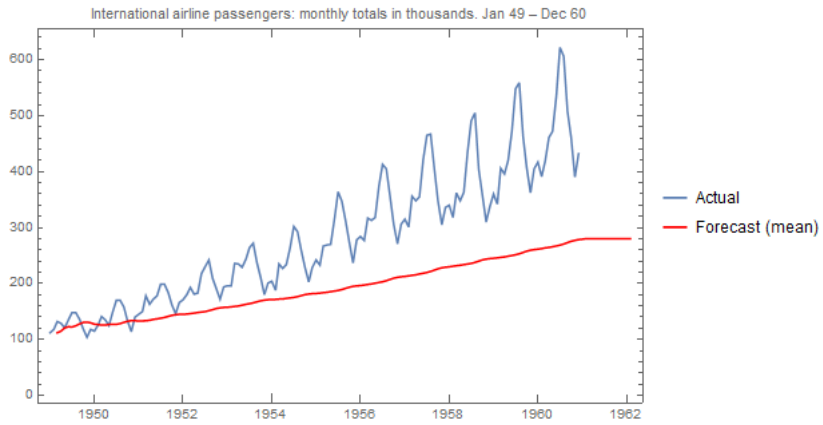

Рис. 6 Прогнозирование средним, прогнозируется плохо в рядах с трендом

Видно, что прогноз сильно остает от тренда из-за того, что в начале ряда значения малые и перетягивают прогноз на себя.

Попытаемся улучшить метод прогнозирования для учитывания тренда.

4.2 Скользящее среднее (Moving Average)

Чтобы сгладить недостатки предыдущего метода прибегают к методу *скользящего среднего*. Прогноз будущего значения признака зависит от среднего по k последних наблюдений, где k - *ширина окна*:

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

Само же сглаживание исходного ряда может быть записано следующим образом:

$$\hat{y}_i = \frac{1}{k} \sum_{l=i}^{i+k-1} y_l$$

Действительно, допустим, что у нас есть ряд $y_i = [1.5, 2.0, 2.5, 3.0, 3.5, 4.0], i = 1, \ldots, 6$. Зададим ширину обзора нашего ряда: $k = 3$. Выполнения сглаживания скользящим средним происходит следующим образом, как показано на рисунке:

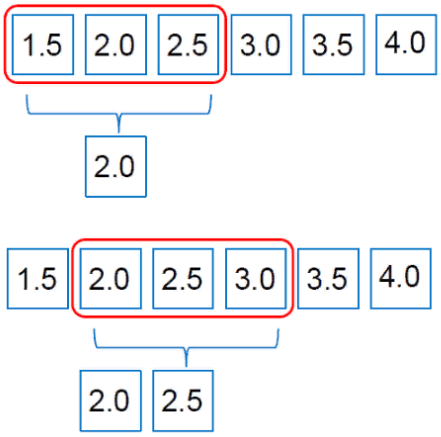

Рис. 7 Принцип работы скользяшего среднего для ширины окна $k = 3$

Для данного ряда прогноз на 1 день вперед будет равен:

$$y_7 = \frac{3.0 + 3.5 + 4.0}{3} = 3.5$$

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.

При применении метода меньших порядков становятся более очевидными сезонные колебания.

В общем, мы видим, что данная техника извлекает основные структуры из временных рядов. Но для лелинейной структуры тренда, например, скользящее среднее будет оставать от тренда и это становится более серьезной проблемой.

Значит, можно рассмотреть модифицированный способ сглаживания

4.2.1 Взвешенное скользящее среднее (Weighted Moving Average)

Вместо того, чтоы присваивать одинаковые веса всем наблюдениям, присвоем веса экспоненциально. Существует много способов генерации весов. Один из них:

$$\sum_{i=1}^{k} w_{t-i}^{k} = 1$$

где k - ширина обзора нашего ряда. Мы видим, что каждому значению присваивается определенный вес, причем значению, стоящему ближе к значению временного ряда, дается больший вес.

Веса можно раздавать и другими способами, тогда модель будет называться просто - *взвешенное скользящее среднее*.

4.3 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 данного подхода - в простоте реализации и в отсутствии необходимости в большом количестве исторических данных. Недостаток - низкое качество прогнозирования.

Если развить идею, то в качестве прогноза можно брать предыдущие значения с сезонным лагом, если в данных наблюдается сезонность:

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

где S - период сезонности, а $\lfloor \_ \rfloor$ - частное от деления

4.4 Экстраполяция тренда

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

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

Второе слагаемоме - уравнение прямой через две заданные точки, а именно через первую и последнюю точку заданного временного ряда. Таким образом можно построить долгосрочный прогноз, однако представленная модель не учитывает сезонные колебания.

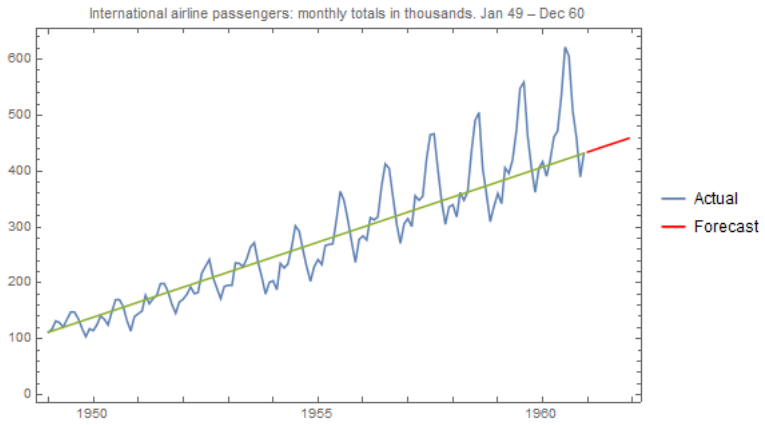

Рис. 8 Экстраполяция тренда

5 Адаптивные методы краткосрочного прогнозирования

5.1 Простое экспоненциальное сглаживание (Модель Брауна)

Экспоненциальнео сглаживание использует идею метода скользящего среднего с той лишь разницей, что каждое предшествующее наблюдение имеет свой вес, экспоненциально убывающий по мере углубления в историю.

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \dots$$

где $\alpha \in [0;1]$ - коэффициент варьирования горизонта прошлого (smoothing constant, параметр сглаживания).

Вообще почему экспоненциальное сглаживание наывается экспоненциальным сглаживаением?) Разберем следующее доказательство:

Доказательство. Запишем формулу члена y_{t+1}:

$$y_{T+1} = \alpha y_T + (1-\alpha)y_{T-1}$$

теперь запишем формулу для y_T:

$$y_T = \alpha y_{T-1} + (1-\alpha)y_{T-2}$$

Подставим y_T в первую формулу:

$$y_{T+1} = \alpha(\alpha y_{T-1} + (1-\alpha)y_{T-2}) + (1-\alpha)y_{T-1}$$

Мы видим, что y_{T+1} зависит от y_{t-1} как α^2, если раскрыть скобки. Если мы будем дальше разворачивать, то на каждом шаге сам на себя умножается коэффициент. Любое наблюдение имеет экспоненциально малый вклад, потому что на каждом шаге убывает в $1-\alpha$ раз. Поэтому данный метод и называется методом экспоненциального сглаживания. □

Экспоненциальное сглаживание по данному методу также называют *методом Брауна экспоненциального сглаживания.*

Экспоненциальное сглаживание учитывает, какое окно учитывают наблюдения. Чем больше параметр сглаживания α, тем больший вес имеют последние наблюдения, тем больше они влияют на будущее предсказание. Поэтому α и называется *коэффициетом варьирования горизонта прошлого.*

Модлеь экспоненциального сглаживания можно выразить через рекуррентную формулу:

$$\hat{y_{t+1|t}} = l_t, \qquad l_t = \alpha y_t + (1-\alpha) l_{t-1}$$

где l_t - это сглаженное значения ряда на момент времени t, значение уровня ряда.

Таким образом, прогноз на один шаг вперед можно записать в виде:

$$\hat{y}_{T+1|T} = \left(\sum_{i=0}^{T-1} \alpha(1-\alpha)^i \cdot y_{T-i} \right) + (1-\alpha)^T \cdot l_0$$

Данный метод подходит для *краткосрочного прогнозирования* рядов *без тренда и сезонности.* В качестве l_0 можно использовать арифметическую среднюю всех имеющихся данных или какой-то их части. Также можно взять в качестве l_0 первое значение временного ряда y_1.

Выбор оптимального значения α происходит таким образом, чтобы минимизировать среднюю квадратичную ошибку.

5.2 Метод Хольта

5.2.1 Метод Хольта с аддитивным трендом

У экспоненциального сглаживания были недостатки - при прогнозе на долгосрочный срок метод эскпоненциального сглаживания предсказывал постоянное значение, не учитыва ни тренд, ни сезонность.

Учесть тренд позволяет **метод Хольта**, который предполагает разбиение временного ряда на две составляющие: уровень l_t и тренд b_t. Экспоненциальное сглаживание применяется также к тренду в предположении, что будущее направление изменения ряда зависит от взвешенных предыдущих изменений.

В случае линейного тренда:

$$\hat{y}_{t+h|t} = l_t + h \cdot b_t$$

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$

где $\alpha, \beta \in [0; 1]$

Теперь компонента, описывающая уровень, зависит от текущего значения ряда, а второе слагаемое разбивается на предыдущее значение уровня и тренда.

Компонента, отвечающая за тренд, зависит от изменения уровня на текущем шаге и от предыдущего значения тренда. Результирующее значение прогноза - сумма модельных значений уровня и тренда.

Данный метод также называют методом *двойного экспоненциального сглаживания*. В качестве начального значения тренда можно брать разность между вторым и первым значением, что логично.

5.2.2 Метод Хольта с мультипликативным трендом

В случае мультипликативного тренда с затуханием

$$\hat{y}_{t+h|t} = l_t \cdot b_t^h$$

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} \cdot b_{t-1}^{\varphi})$$

$$b_t = \beta \left(\frac{l_t}{l_{t-1}} \right) + (1-\beta) b_{t-1}^{\varphi}$$

где $\alpha, \beta \in [0; 1]$ и $\varphi \in [0, 1]$ - параметр затухания.

Чтобы использовать обычный мультипликативный тренд, просто уберем коэффициент затухания.

5.3 Метод Хольта - Уинтерса

В модели Хольта-Уинтерса возможны два подхода в зависимости от характера поведения сезонной составляющей. Рассмотренный ранее аддитивный метод применяется, когда сезонные колебания примерно постоянны по всему ряду.

5.3.1 Метод Хольта - Уинтерса с аддитивной сезонностью

Модель Хольта-Уинтерса является модификацией модели Хольта. Данная модификация позволяет учесть не только последние наблюдаемые значения и тренд, но и сезонность. Дополнительная сезонная компонента в модели объясняет повторяющиеся колебания вокруг уровня и тренда и характеризуется длиной сезона m:

Для реализации метода Хольта-Уинтерса с аддитивной сезонностью возьмем за основу реализацию метода Хольта и добавим сезонную составляющую:

$$\hat{y}_{t+h\,|\,t} = l_t + h\,b_t + s_{t-m+(h \bmod m)}\,,$$

$$l_t = \alpha\,(y_t - s_{t-m}) + (1-\alpha)\,(l_{t-1} + b_{t-1}),$$

$$b_t = \beta\,(l_t - l_{t-1}) + (1-\beta)\,b_{t-1},$$

$$s_t = \gamma\,(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)\,s_{t-m}\,.$$

Уровень теперь зависит от текущего значения ряда без учета соответствующей сезонной компоненты, а сезонная компонента зависит от текущего значения ряда за вычетом уровня и от предыдущего значения компоненты.

5.3.2 Метод Хольта - Уинтерса с мультипликативной сезонностью

В случае, когда сезонные колебания изменяются пропорционально уровню ряда, говорят о мультипликативной сезонности. В данном случае мультипликативный метод Хольта-Уинтерса является предпочительным:

$$\hat{y}_{t+h\,|\,t} = (l_t + hb_t)\,s_{t-m+(h \bmod m)}\,,$$

$$l_t = \alpha\,\frac{y_t}{s_{t-m}} + (1-\alpha)\,(l_{t-1} + b_{t-1}),$$

$$b_t = \beta\,(l_t - l_{t-1}) + (1-\beta)\,b_{t-1},$$

$$s_t = \gamma \frac{y_t}{l_{t-1} + b_{t-1}} + (1 - \gamma)\, s_{t-m}\,.$$

При аддитивном методе сезонная компонента s_t выражается в абсолютных величинах в масштабе наблюдаемого ряда, а в уравнении уровня ряд сезонно корректируется путем вычитания из него сезонной компоненты. При использовании же мультипликативного метода сезонная компонента выражается в относительных единицах, а ряд корректируется путем деления значений на сезонную компоненту.

Метод Хольта-Уинтерса также часто называется *тройным экспоненциальным сглаживанием* из-за сглаживания компонент уровня, тренда и сезонной составляющей.

6 Авторегрессионные методы прогнозирования

6.1 Модель авторегрессии AR

Определение 6.1.1. Авторегрессия порядка p (AR(p)) - регрессия для ряда на его собственные значения в прошлом:

$$y_t = \alpha + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \varepsilon_t$$

$$y_t = \alpha + \sum_{i=1}^{p} \alpha_p \cdot y_{t-i} + \varepsilon_t$$

где y_t - стационарный ряд, ε_t - гауссов белый шум с нулевым средним и постоянной дисперсией.

y_t в данной модели представляет собой линейную комбинацию p предыдущих значений ряда и шумовой компоненты.

6.2 Скользящее среднее MA

Сразу отметим, что это не то, что мы разбирали раньше, то есть не простое, не взвешенное, не экспоненциальное скользящее среднее.

Будем считать, что будущее значение переменной зависит от среднего ее предыдущих значений. Это приводит к идее модели **скользящего среднего**:

$$y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

где $\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ - значения шума в q предыдущих моментах времени, $\alpha, \theta_1, \ldots, \theta_q$ - параметры модели, которые необходимо оценить. Эта модель называется моделью скользящего среднего порядка q - MA(q).

6.3 ARMA

Чтобы улучшить результаты предсказания модели, построим модель скользящего среднего на остатках, полученных после вычитания из исходных данных значений, предсказанных при помощи модели AP(p) $\varepsilon_i = y_i - \hat{y}_i$

Определение 6.3.1. Модель ARMA - сумма авторегрессионной модели порядка p AR(p) и модели скользящего среднего порядка q MA(q):

$$y_t = \alpha + \sum_{i=1}^{p} \alpha_i \cdot y_{t-i} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

Теорема 6.1. ***Теорема Вольда***

Любой стационарный ряд может быт описан моделью ARMA(p,q) с любой точностью.

6.4 ARIMA

В основе моделей класс ARIMA лежат идеи о том, что нестационарный ряд можно сделать стационарным при помощи дифференцирования, а любой стационарный ряд может быть описан моделью ARMA(p,q).

Определение 6.4.1. ARIMA(p,d,q) - это модель авторегресии интегрированного скользящего среднего - модель ARMA(p,q) для d раз продифференцируемого ряда.

6.5 Тест Дики-Фуллера на стационарность

Нестационарный временной ряд описывается моделью авторегресии интегрированного скользящего среднего ARIMA(p,d,q), если временной ряд y_t является интегрированным порядка $d, d \geq 1$, а случайный процесс $\Delta^d y_t$ является стационарным и описывается моделью ARMA(p,q).

Рассмотрим модель авторегрессии первого порядка $AR(1)$:

$$y_t = \alpha + \alpha_1 y_{t-1} + \varepsilon_1$$

Временной ряд y_t является стационарным при $0 < \alpha_1 < 1$ и нестационарным интегрированным процессом при $\alpha_1 = 1$. Во втором случае модель можно записать в следующем виде:

$$y_t = \alpha + y_{t-1} + \varepsilon_t$$

Такая модель называется **моделью случайного блуждания** или моделью единичного корня. Таким образом, значения параметра α_1 влияет на тип модели временного ряда:

- Если $\alpha_1 = 1$, то временной ряд нестационарный и описывается моделью случайного блуждания
- Если $\alpha_1 < 1$, то временной ряд стационарный и описывается моделью $AR(1)$.

В первом случае будущие значения рассматриваего ряда непредсказуемы, а во втором возникает возможность прогнозирования. Для финансово-экономических процессов значение $|\alpha_1| > 1$ не свойственно, так как в этом случае процесс является взрывным

Процесс проверки гипотезы теста Дики-Фуллера

Тест Дики-Фуллера проверяет нулевую гипотезу, что ряд является *нестационарным* и описывается моделью единичного корня при альтернативе H_1, что ряд является *стационарным.* Проще говоря, для модели $AR(1)$ гипотезы принимают вид:

$$\begin{cases} H_0 : \alpha_1 = 1 - \text{нестационарный ряд} \\ H_1 : \alpha_1 < 1 - \text{стационарный ряд} \end{cases}$$

Процедура тестирования включает оценивание параметров модели $AR(1)$ с помощью МНК и проверку гипотез о значимости коэффициентов модели. Далее рассматриваются пороговые значения t-статистики Дики-Фуллера относительно коэффициента α_1, которые представлены в виде таблиц для различных уровней значимости и различных значений длины ряда.

В качестве проверки гипотезы о нестационарности ряда, используется функция $statsmodels.tsa.stattools.adfuller$, возвращающее $pvalue$ для оцениваемого ряда. Если pvalue $< T_{0.05}$, то нулевая гипотеза отвергается и ряд является стационарным. Однако важно, чтобы $pvalue$ был очень и очень маленьки, чтобы с большей степенью уверенности утверждать, что ряд является стационарным.

Даже если тест Дики-Фуллера дал $pvalue$, для которого ряд стационарен, лучше проверить данную гипотезу еще раз теми метода, что были получены ранее.

6.6 SARIMA

Пусть ряд имеет сезонный период длины S. Возьмем $ARMA(p, q)$:

$$y_t = \alpha + \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

и добавим P сезонных авторегрессионных компонент и Q компонент скользящего среднего:

$$y_t = \alpha + \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{r=1}^{P} \alpha_{rS} \cdot y_{t-r\cdot S} + \sum_{l=1}^{Q} \theta_{lS} \cdot \varepsilon_{t-l\cdot S}$$

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

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.

Параметры d, D необходимо подобрать так, чтобы ряд стал стационарным. Дифференцировать нужно как можно меньше, потому что с увеличением дифференцирования растет дисперсия итогового прогноза.

6.6.1 Частичная автокорреляция

Определение 6.6.1. *Частичная автокорреляция* - это автокорреляция после снятия авторегрессии предыдущего порядка.

Для определения значения частичной автокорреляции с лагом 2, например, необхоидмо построить авторегрессию порядка $1 AR(1)$, вычесть эту авторегрессию из ряда и вычислить автокорреляцию на полученных остатках, то есть:

$$pacf(h) = \begin{cases} \rho(y_{t+h}, y_t), & h = 1 \\ \rho(y_{t+h} - AR(t, h-1), y_t - AR(t+h, h-1)), & h \geq 1 \end{cases}$$

где $AR(t, h-1)$ - авторегрессия следующего вида:

$$AR(t, h-1) = y_t^{h-1} = \alpha_1 y_{t+1} + \alpha_2 y_{t+2} + \ldots + \alpha_{h-1} y_{t+h-1}$$

$$AR(t+h, h-1) = y_{t+h}^{h-1} = \alpha_1 y_{t+h+1} + \alpha_2 y_{t+h+2} + \ldots + \alpha_{h-1} y_{t+1}$$

Частичная автокорреляция используется для нахождения периодичностей во временных рядах и для определени порядка авторегрессионной модели ряда, чем мы и займемся ниже.

6.6.2 Определение начального порядка авторегрессии и скользящего среднего (гиперпараметров)

Определение 6.6.2. *Гиперпараметрами модели называют те параметры, которые задаются пользователем.* Параметрами же называют подобранные путем минимизации функционала качества моделью коэффициенты.

Для выбора порядка авторегрессии обращаются к частичной автокорреляции.

Для выбора порядка авторегрессии p_0 мы ищем номер последнего значимого лага перед незначимым, для выборка сезонного порядка авторегрессии P_0 мы определяем величину сезонности, а после ищем номер последнего значимого сезонного лага перед незначимым.

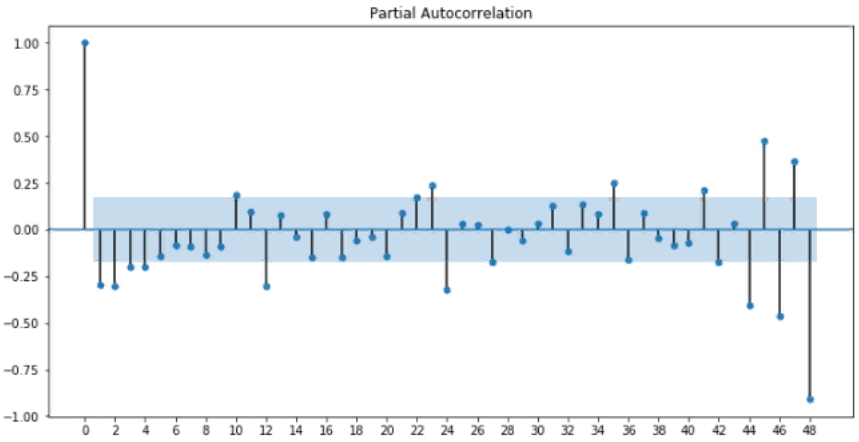

Рис. 9 Определение начального порядка авторегрессии по частичной автокорреляции (PACF)

Так, в данном примере, последним значимый лаг оказался равным 4, поэтому $p_0 = 4$. Период сезонности равен 12, при лаге, равном 12, частичная автокорреляци значима, при 24 значима, а вот при 36 уже нет. Следовательно, $P_0 = 2$. Теперь обратимся к определению порядка скользяшего среднего.

Для выборка порядка скользящего среднего обратимся к коррелограмме или автокорреляции.

Для начального значения q_0 выбираем последний значимый лаг перед следующим незначимым, для начального значения Q_0 определяем величину сезонности, а после выбираем последний значимый сезонный лаг перед следующим незначимым.

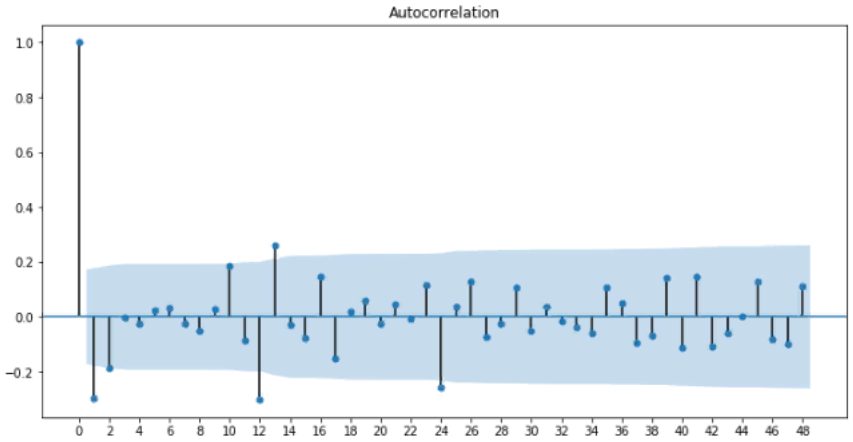

Рис. 10 Определение начального порядка скользящего среднего по коррелограмме или автокорреляции (ACF)

Последний значимый лаг - 2. А последний значимый сезонный лаг равен 2 (при лаге $12, 24$ автокорреляция значима, при 36 - уже нет). Таким образом, $q_0 = 2, Q_0 = 2$.

Заметим, что гиперпараметры модели нельзя выбирать методом максимального правдоподобия, поскольку с увеличением количества парамтров значение функции правдоподобия L растет.

Поэтому для сравнения различных моделей применяется информационный критерий Акаике:

$$AIC = 2K - 2\ln L$$

где $K = P + Q + p + q + 1$ - число параметров в статистической модели, а L - максимизированное значение функции правдоподобия модели.

Оптимальный по критерию Акаике будет модель с наименьшим значением этого критерия. Такая модель будет хорошо описывать данные, а с другой - содержать не слишком большое количество параметров.

Параметры p, P, q, Q определяются перебором и выбирается та комбинация, у которой Акаике меньший.

7 Методы оценки качества модели

Чтобы оценить, насколько хорошо модель справляется с задачей прогнозирования, сравнить между собой различные модели или выбрать набор признаков, позволяющий налучшим образом описать данные, используются различные методы оценки сравнения качества.

7.1 Отложенная выборка

Для того чтобы оценка была достоверной, неправильно сравнивать качество модели на тех же данных, на которых она обучалась, то есть на которых были оценены параметры модели. Очевидным способом оценки качества является разбиение всей выборки на две части: *обучающую* (**train**) длины l и *контрольную* (**test**) длины k.

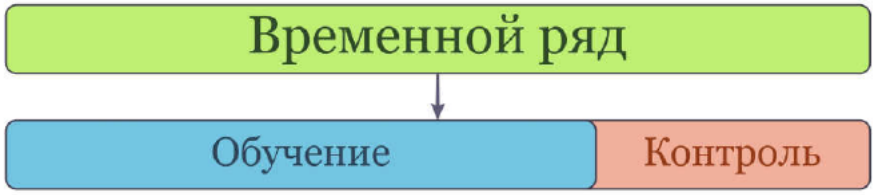

Рис. 11 Разделение временного ряда на **train** и **test**

На тренировочном множестве модель обучается, а на *тестовом* множестве происходит валидация результата.

Данный метод логичен, но если разделить множество один раз на тестовое и тренировочное, то это может значительно повлиять на результат. Вдруг в тестовое множество случайно попали такие точки, в которых предсказать значение тяжелее или легче, чем обычно?

Тогда применяют следующий способ.

7.2 Перекрестная проверка (кросс-валидация) на временных рядах

Давайте для сглаживания недостатков предыдущего метода применять так называемую **кросс-валидацию** (или перекрестную проверку) на временных рядах. Особенность задачи прогнозирования временных рядов состоит в том, что оценка качества модели может производиться только последовательно, так как временной ряд имеет временную структуру и порядок данных важен и перемешивать их нельзя.

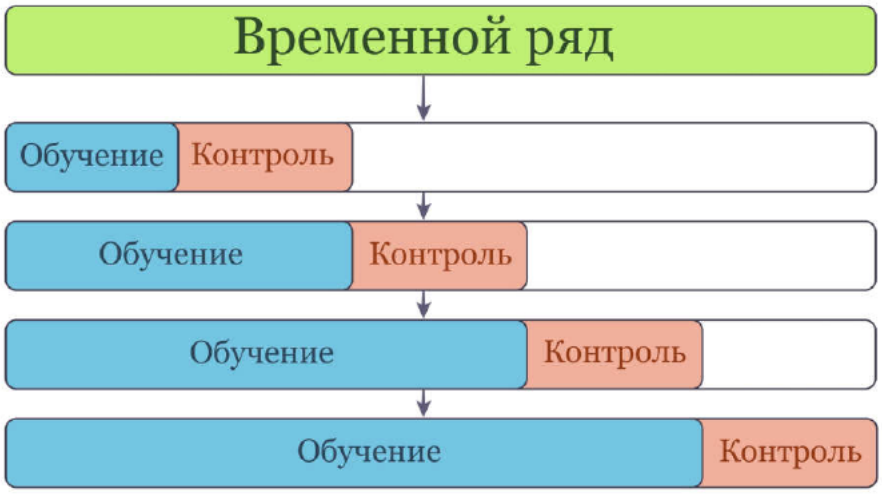

Рис. 12 Кросс-валидация

Частным случаем метода кросс-валидации является контрольный по отдельным объектам (leave-one-out).

7.3 Метрики оценки качества

7.3.1 Коэффициент детерминации

Цель регрессии - объяснение поведения Y. В любой выборке Y оказывается низким, а в других - высоким. Разброс значений Y можно описать с помощью суммы квадратов отклонений от выборочного среднего.

$$\sum (Y - \overline{Y})^2$$

Все показатели корреляции основаны на правиле сложения дисперсий $\Rightarrow$ можно разложить **общую сумму квадратов отклонений** переменной Y от среднего значения $\overline{Y}$ на две части - **"объясненную"** сумму квадратов и **"необъясненную"**.

$$\sum (Y - \overline{Y})^2 = \sum (\widehat{Y} - \overline{Y})^2 + \sum (Y - \widehat{Y})^2 \tag{12}$$

Данное равенство можно переписать как:

$$SS_T = SS_R + SS_E \tag{13}$$

где:

$SS_T = \sum (Y - \overline{Y})^2$ - общая сумма квадратов отклонений *(total sum of squares)*

$SS_R = \sum (\widehat{Y} - \overline{Y})^2$ - **сумма квадратов отклонений, объясненная** регрессией, **факторная сумма** *(sum of square due to regression)*

$SS_E = \sum (Y - \widehat{Y})^2 = \sum e_i^2$ - **остаточная сумма** квадратов отклонений, *(sum of square due to error)*.

Введем **коэффициент детерминации**:

$$R^2 = r^2 = \frac{\sigma^2_{y,obyasn}}{\sigma^2_{y,obch}} = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = 1 - \frac{\sum (Y - \widehat{Y})^2}{\sum (Y - \overline{Y})^2}$$

$$R^2 = \frac{\sum (\widehat{Y} - \overline{Y})^2}{\sum (Y - \overline{Y})^2} = 1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \overline{Y})^2}$$

Коэффициент детерминации- обобщающий показатель оценки качества построенного уравнения регрессии.

Если фактор x не влияет на результат, то линия регрессии параллельна оси ox и $\bar{y} = \hat{y}$. Тогда вся дисперсия результативного признака

обусловлена воздействием прочих факторов и общая сумма квадратов отклонений совпадет с остаточной.

Если же прочие факторы не влияют на результат, то Y связан с X функционально и остаточная сумма квадратов $SS_E = \sum e_i^2 = 0$. В этом случае сумма квадратов отклонений равна объясненной сумме квадратов:

$$SS_T = SS_R$$

Поскольку не все точки поля корреляции лежат на линии регрессии, то всегда имеет место их разброс как обусловленный влиянием фактора X, т.е регрессией Y по X, так и вызванный действием прочих причин(необъясненная вариация).

Так как

$$R^2 = 1 - \frac{\sum(Y - \widehat{Y})^2}{\sum(Y - \overline{Y})^2} = 1 - \frac{SS_E}{SS_T},$$

то если SS_T будет больше остаточной суммы квадратов SS_E, то уравнение регрессии статистически значимо и фактор X оказывает существенное воздействие на результат Y. Это равносильно тому, что коэффициент детерминации R^2 будет приближаться к единице.

Числитель показывает дисперсию ошибки модели, а знаменатель – дисперсию рассматриваемого ряда. Таким образом, второе слагаемое показывает долю необъясненной моделью дисперсии ряда.

Коэффициент детерминации принимает значения от 0 до 1. Чем ближе значение к 1, тем сильнее зависимость.

При расчете коэффициента множественной корреляции используется остаточная дисперсия, которая имеет систематическую ошибку в сторону преуменьшения, чем больше параметров определяется в модели при заданном объеме n, то есть его значение увеличивается с увеличением параметров модели, что не всегда является показателем того, что модель стала предсказывать лучше и обладает обобщающей способностью. Значение R^2 увеличивается от добавления в модель новых переменных, даже если эти переменные никакого отношения к объясняемой переменной не имеют.

Поэтому принятно считать *скорректированный* (нормированный) коэффициент множественной корреляции, учитывающий число степеней

свободы, чтобы не допустить возможного преувеличения тесноты связи.

$$R = \sqrt{1 - \frac{MS_E}{MS_T}}$$

$$\hat{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1}$$

где n - число наблюдений, а m - число параметров.

Данный коэффициент зависит от объема наблюдения и числа параметров расчитываемой модели. Чем больше m, тем больше коэффициент детерминации и скорректированный коэффициент детерминации различаются.

Низкое значение скорректированного коэффициента детерминации означает, что в гегрессионную модель не включены существенные факторы, с другой стороны, рассматриваемая форма связи не отражает реальные соотношения между переменными, включенными в модель.

7.3.2 Среднеквадратичная ошибка MSE

В качестве метрики качества берут квадрат отклонения фактических значений от прогнозных (Mean Squared Error), которая выступает в качестве оптимизируемой функции при оценке параметров модели:

$$MSE = \frac{1}{T-k+1}\sum_{t=k}^{T}(y_t - \hat{y}_t)^2$$

MSE достаточно удобная функция, ведь она выпуклая, дифференцируемая функция, но есть и недостатки - при больших значениях временного ряда MSE выдает большие значения, завышая их. Возможно и занижение ошибки, если $-1 \leq y \leq 1$.

7.3.3 Средняя абсолютная ошибка MAE

Наиболее интерпретируемой метрикой выступает средняя абсолютная ошибка (Mean Squared Error - MAE), которая показывает среднее отклонение прогнозных значений от фактических:

$$MAE = \frac{1}{T-k+1}\sum_{t=k}^{T}|y_t - \hat{y}_t|$$

Недостатком данной метрики является невозможность сравнения качества, если решается задача прогнозирования нескольких временных рядов (разные величины имеют разные единицы измерения).

7.3.4 Средняя абсолютная процентная ошибка

Для сравнения качества для нескольких рядов можно сравнивать не абсолютную ошибку, а процентную (Mean Absolute Percentage Error):

$$MAPE = \frac{1}{T-k+1}\sum_{t=k}^{T}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

У MAPE есть недостаток - деление на ноль. И с этим нужно как-то жить и что-то придумывать. Некоторые программы отбрасывают периоды с нулевыми фактически значениями, но это не очень хорошая идея, потому что это фактически означает, что нам все равно, что мы прогнозировали, если значение нулевое.

7.3.5 Средняя симметричная абсолютная ошибка

Идея симметричной абсолютной процентной ошибки: в знаменателе указать среднее между фактическим и прогнозным значениями:

$$SMAPE = \frac{1}{T-k+1}\sum_{t=k}^{T}\frac{2|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|}$$

8 Регрессионные модели прогнозирования

8.1 Понятие обучающей выборки. Предсказательная модель. Функция потерь и функционал качества

Определение 8.1.1. Обучающая выборка - набор пар $X^l = (x_i, y_i), i = 1, \ldots, l$, в котором для каждого объекта $x_i \in X^l$, характеризующегося рядом признаков $(x_{i1}, x_{i2}, \ldots, x_{im})$, известно значение целевой переменной y_i.

Определение 8.1.2. Признаки - числовые характеристики рассматриваемых объектов $X \to D_j, j = 1, \ldots, M$. В зависимости от принимаемых значений можно выделить следующие типы признаков:

- $D_j \in \{0, 1\}$ - бинарный
- $|D_j| < \infty$ - номинальный
- $|D_j| < \infty, D_j$ - упорядоченный порядковый
- $D_j \in R$ - количественный (вещественный)

Задача состоит в том, чтобы по обучающей выборке определить неизвестную зависимость $a : X \to Y$.

Определение 8.1.3. Предсказательная модель - параметрическое семейство функций:

$$A = \{a(x) = g(x, \theta) \mid , \theta \in \Theta\},$$

где $gLX \times \Theta \to Y$ - фиксированная функция, а Θ - множество параметров модели.

Определение 8.1.4. $Q(a, X^l)$ называется **функционалом качества** алгоритма a на обучающей выборке X^l. Величина ошибки накаждом объекте $x \in X^l$ называется **функцией потерь** $L(a, x)$:

$$Q(a, X^l) = \frac{1}{l} \sum_{i=1}^{l} L(a, x_i)$$

Определение 8.1.5. Обучение модели - процесс настройки параметров модели $\theta \in \Theta$ путем решения задачи оптимизации:

$$Q(a, X^l) \to \min_{a \in A}$$

Определение 8.1.6. Функция потерь (Loss function) - это неотрицательная функция $L(a, x)$, характеризующая величину ошибки алгоиртма a на объекте x.

8.2 Общий вид модели линейной регрессии. Предпосылки метода наименьших квадратов

В соответствии с введенными ранее обозначениями, модель линейной регрессии может быть представлена в виде:

$$a(x) = g(x, \theta) = \theta_0 + \sum_{j=1}^{m} \theta_j x_j$$

где θ_0 - свободный член.

Если добавить фиктивный признак $x_0 = 1$ для каждого объекта, то модель линейной регрессии запишется так:

$$a(x) = g(x, \theta) = \sum_{j=0}^{m} \theta_j x_j = (\theta, x) = \theta^T x$$

где θ, x - вектора параметров коэффициентом и признаков.

Таким образом, наблюдаемые значения y описываются следующим выражением:

$$y = X\theta + \varepsilon$$

где X - матрица "объекты-признаки ε - случайная непрогнозируемая ошибка.

8.2.1 Предпосылки метода наименьших квадратов

Условия, необходимые для получения несмещеных, состоятельных и эффективных оценок, представляют собой предпосылки МНК, соблюдение которых желательно для получения достоверности результатов.

Делаются предположения относительно поведения остатков ξ_i.

1. Модель линейна по параметрам

2. $\mathbb{E}\xi_i = 0 \forall i$, т.е ожидание значения случайного члена должно быть равно нулю в каждом наблюдении из-за того, что каждое наблюдение не должно включать в себя смещения ни в каком из направлений.

3. Гомоскедастичность - $\mathbb{D}\xi_i = Const$, т.е его значение в каждом наблюдении получено из распределения с постоянной теоретической дисперсией. Также не должно быть причин , делающих его больше подверженным ошибке в одних наблюдениях по сравнению с другим. Заметим, что

$$\mathbb{E}\xi_i^2 = \mathbb{D}\xi_i = \mathbb{D}\sigma_{\xi_i}^2 | \forall i$$

4. Отсутствие автокорреляции остатков - значения случайного члена имеют взаимно независимые распределения. Случайный член не подвержен автокорреляции, т.е отсутствует систематическая связь между его значениями в любых двух наблюдениях. Ковариация равна нулю:

$$\sigma_{\xi_i \xi_j} = \mathbb{E}(\xi_i \xi_j) = \mathbb{E}\xi_i \cdot \mathbb{E}\xi_j = 0 | \forall i \neq j$$

5. $\xi_i \sim \mathbb{N}(0, \sigma^2)$: если случайный член нормально распределен, то распределены нормально и коэффициенты регрессии.

После построения уравнения регрессиии проводится проверка наличия у оценок ξ_i тех свойств, которые предполагались. Связано это с тем, что оценки параметров регресии должны отвечать определенным критериям: быть несмещенными, состоятельными эффективными.

Определение 8.2.1. Оценка является *несмещенной*, если математическое ожидание остатков равно нулю.

Следовательно, остатки не будут накапливаться и найденный параметр регрессии b_i можно рассматривать как среднее значение из возможного большого количества несмещенных оценок.

Определение 8.2.2. Оценки называются *эффективными*, если они характеризуются наименьшей дисперсией.

Это означает возможность перехода от точечного оценивания к интервальному.

Определение 8.2.3. Состоятельность оценок характеризует увеличение их точности с увеличением объема выборки.

При соблюдении введенных предпосылок, оценки, полученные методом наименьших квадратов, обладают данными важнейшими свойствами, на основании которых можно быть уверенным в достоверности результатов модели.

Теорема 8.1. *Теорема Гаусса-Маркова*

Пусть рассматривается модель линейной регрессии:

$$y_i = X\theta + \varepsilon$$

и выполнены предполсыки МНК.

Тогда оценки параметров θ_j, полученные методом наименьших квадратов, будут ***несмещенным и состоятельными****, а также оценка методов наимеших квадратов является оптимальной в классе линейных несмешенных моделей.*

Доказательство. $\varepsilon_i = y_i - x_i\theta \sim N(0, \sigma^2)$, можно записать следующее:

$$y_i \sim N(\theta^T \cdot x, \sigma^2)$$

Тогда плотность нормального распределения выглядит следующим образом:

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \theta^T x_i)}{2\sigma^2}}$$

Составим функцию правдоподобия:

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \theta^T x_i)}{2\sigma^2}} \to \max$$

Прологарифмируем:

$$\ln L(\mu, \sigma) = \ln\left(\frac{1}{(\sqrt{2\pi}\cdot\sigma)^n}\right) + \left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \theta^T x_i)^2}{\sigma^2}\right) \to \max$$

$$\sum_{i=1}^{n}(y_i - \theta^T x_i)^2 \to \min$$

А это и есть метод наименьших квадратов! □

8.2.2 Применение модели линейной регрессии для прогнозирования

Найдем решение МНК в общем виде. Запише выражение, которое надо минимизировать в матричной форме $L = (\theta X - y)^T (X\theta y) \to \min$, где X - матрица объектов-признаков, а θ-вектор искомых коэффициентов модели.

Продифференцируемым по всем переменным θ_k и приравняем производные к нулю:

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial \theta}(\theta^T X^T X\theta - 2y^T X\theta + y^y) = 0$$

$$2X^T X\theta = 2(y^T X)^T$$

$$\theta = (X^T X)^{-1} X^T y$$

матрица $(X^T X)^{-1} X^T$ называется псевдообратной к матрице X. Это понятие является естественным обобщением понятия обратной матрицы на случай неквадратных матриц.

Матрица x выглядит следующим образом:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,m-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2,m-1} \\ \dots & \dots & \dots & \dots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,m-1} \end{bmatrix}$$

На место элементов матрицы можно добавлять различные базисные функции, например, в качестве одного из столбцов добавить квадрат признака x_1, тогда мы будем строить квадратичную регрессию, то есть мы можем нагенерировать поведение признаков такое, какое захотим. Такой метод называется *Feature Engineering*.

Подобным образом мы можем, например, использовать линейную регрессию для матрицы признаков, составленных из соответствующих гиперпараметрам сдвигам временного ряда.

8.3 Извлечение признаков из временного ряда. SARIMAX

8.3.1 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, то возникает недельная, месячная и годовая сезонность.

В первую очередь нельзя нормально привести ряд к стационарному, ведь если проводить сезонное дифферцнирование, то для этого требуется определить длину сезонного лага, что не удастся сделать, если мы рассматриваем, например, високосный и невисокосный год 366 и 365 дней соответственно.

Более того, почему именно дата год назад учитывается в прогнозе? Ведь сегодня может быть будний день, а год назад мог быть выходной и спрос мог быть совершенно другим, а модель может этого не заметить.

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

8.3.2 Модель SARIMAX

Определение 8.3.1. Модель **SARIMAX** - расширение 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, такие как погода, температура воздуха и.т.д.

   На полученном наборе факторов строят модель линейной регрессии:

$$y_t = \sum_{j=0}^{m} \theta_j x_{tj} + \varepsilon_t$$

где m - количество выделенных прогнозных факторов, ε_t - непрогнозируемая ошибка модели. $x_{t0} = 1$, так как необходимо учесть свободный член в модели.

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.

8.3.3 Извлечение факторов из временного ряда

Если удачно определить набор факторов, описывающих поведение ряда, то остатки модели могут оказаться шумом. В качестве факторов могт выступать моменты времени, индикаторы для дня недели, статистики, посчитанные на данных из прошлого. Рассмотрим подробнее.

По сути, мы ищем способы, как добавить в матрицу признаков X новые столбцы, чтобы модель оставляла в остатках как можно меньше информации.

Зависимость от времени

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

$$y_t = \sum_{j=0}^{m} \theta_j t^j + \varepsilon_t$$

Таким образом, m может быть степенью полинома, которым будет описана целевая переменная.

Сезонные факторы

Допустим, в остатках осталось много информации. Для улучшения качества модели посмотрим на сезонную составляющую. В качестве фактора можно использовать, например, индикатор текущего месяца. Первая запись получит 1 - январь, 12 - декабрь.

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

Рис. 13 Сезонные факторы, нумерация месяцев

Также стоит не забывать о существовании праздников и каникул, будний день можно выделить 0, а выходной - 1. можно добавить и годовую, и недельную сезонность, если это важно. Необходимо также учитывать спефицику ряда, а именно учитывать предпраздничные дни и их динаминку, например. То есть необходимо тщательно продумать выбор факторов при добавлении сезонности.

Тригонометрический ряд Фурье

Хорошим способом описать сезонные колебания является применение Фурье-преобразований. Недостаток прошлого способа кодирования - расстояние от явнаря до декабря оказалось больше, чем от января до октября, что мешает выявить соответствующие закономерности. Решить данную задачу позволяет следующее преобразование:

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

где i - соответствующее наблюдению значение - месяц, например, отсчет ведем с нуля, а s - длина сезонного колебания.

Дневная сезонность, состоящая из 24 часов описывается следующим образом:

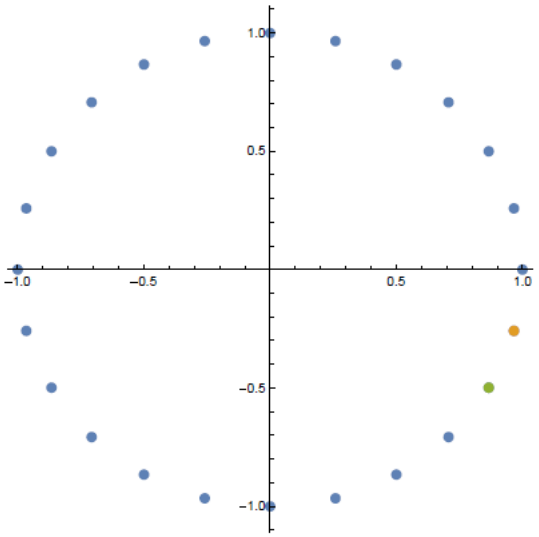

Рис. 14 Дневная сезонность

Выделенные точки соответсттвуют 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**, то есть для аддитивной сезонности. Добавим факторы в регрессионную модель, описываемую полиномом mой степени и получим:

$$y_t = \sum_{j=0}^{m} \theta_j t^j + \beta_1 \varphi_{t \bmod s} + \beta_2 \psi_{t \bmod s} + \varepsilon_t$$

Коэффициенты β_1 и β_2 находятся с помощью МНК.

На случай мультипликативной сезонности логично предположить, что с ростом ряда необходимо увеличиение амплитуды колебаний, поэтому домножим на t или возведем в степень:

$$y_t = \sum_{j=0}^{m} \theta_j t^j + \beta_1 \varphi_{t \bmod s} \cdot t + \beta_2 \psi_{t \bmod s} \cdot t + \varepsilon_t$$

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 неко-

торые идентификаторы. Допустим, у нас есть точки в Москве, Санкт-Петербурге и Сочи. Тогда:

$$\begin{cases} \text{Москва} \to 1 \\ \text{Санкт-Петербург} \to 2 \\ \text{Сочи} \to 3 \end{cases}$$

Оказалось, что Сочи больше Москвы и такой способ кодирования не несет никакой информации. Можно отталкиваться от населения и присвоить номер больший самому большому городу.

Способ преобработки данных 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

Рис. 15 One-Hot-Encoding

При применение такого способа для описания всех 12 месяцев достаточно 11 колонок, иначе возникнет проблема мультиколлинеарности - зависимости признаков друг от друга.

Недостаток - хранение большой разряженной матрицы переменных.

Другим способом является кодирование с помощью двоичного представления. Для 8 значений - 3 колонки, для 30 - 5 и.т.д.

Представленные способы хоть и продемонстрированы на примере кодирования, но они не решают проблему расстоний между месяцами.

Дополнительная информация статистик

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;

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

8.4 Проблема переобучения. Регуляризация

8.4.1 Понятие недообучения и переобучения

Определение 8.4.1. Недообучение - называют нежелательное явление, когда алгоритм обучения не обеспечивает достаточно малой величины средней ошибки на обучающей выборке. Недообучение возникает при использовании недостаточно сложных моделей.

Определение 8.4.2. Переобучение (overfitting) - нежелательное явление, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке. Переобучение возникает при использовании избыточно сложных моделей.

То есть, если мы используем достаточно сложную модель, например, многочлен высокой степени, у которой много параметров, то у модели появляется возможность выучить все точки, которые она видела.

Таким образом нужно следить, чтобы модель хорошо описывала данные, но при этом не слишком настраивалась под них.

Определение 8.4.3. Говорят, что алгоритм обучения обладает **обобщающей способностью**, если вероятность ошибки на тестовой выборке достаточно мала или хотя бы предсказуема, то есть не сильно отличается от ошибки на обучающей выборке.

8.4.2 Регуляризация

Если используется слишком сложная модель, а данных недостаточно, чтобы точно определить ее параметры, эта модель легко может получиться переобученной. Бороться с этим можно различными способами:

- взять больше данных (но, понятное дело, что такое не всегда доступно).
- упростить модель, исключи некоторые признаки, потому что модельможет быть на них заточена.
- использовать регуляризацию. У переобученной линейной модели значения параметров в модели становятся огромными и разными по знаку. Если ограничить значения весов модели, то с переобучением можно до какой-то степени бороться.

Рассмотрим методы регуляризации. Основнымиявляются L_2 - (*ridge-регрессия* или гребневая регрессия) и добавление L_1-регуляризатора *lasso-регрессия*.

8.4.3 Гребневая регрессия

Метод наименьших квадратов состоит в минимизации функционала качества:

$$Q(a, X^l) = \sum_{i=1}^{l} \left(\sum_{j=0}^{m} \theta_j x_{ij} - y_i \right)^2 \to \min_{\theta}$$

Большие значения параметров θ приводят к переобучению, добавим в функционал качества штраф на слишком большие значения θ:

$$Q(a, X^l) = \sum_{i=1}^{l} \left(\sum_{j=0}^{m} \theta_j x_{ij} - y_i \right)^2 + \lambda \sum_{j=1}^{m} \theta_j^2 \to \min_{\theta}$$

Такая модель называется гребневой регрессией.

8.4.4 Лассо регрессия

Отличие лассо-регрессии лишь в том, что штрафующим слагаемым выступает модуль:

$$Q(a, X^l) = \sum_{i=1}^{l} \left(\sum_{j=0}^{m} \theta_j x_{ij} - y_i \right)^2 + \lambda \sum_{j=1}^{m} |\theta_j| \to \min_{\theta}$$

8.4.5 Особенности регуляризаторов

Рассмотрим особенности регуляризаторов

Пусть матрица "объекты-признаки"X является единичной матрицей размера $l \times l$:

$$X = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Тогда при решении задачи линейной регрессии использование МНК без регуляризации:

$$Q(a, X^l) = \sum_{i=1}^{l} (\theta_i - y_i)^2 \to \min_{\theta}$$

дает следующий вектор θ:

$$\theta_i^* = y_i$$

При добавлении L_2 регуляризации получаем:

$$\sum_{i=1}^{l} (\theta_i - y_i)^2 + \lambda \sum_{j=1}^{l} \theta_j^2 \to \min$$

$$2(\theta_i - y_i) + 2\lambda\theta_i = 0$$

$$\theta_i = \frac{y_i}{1 - \lambda}$$

А при добавлении L_2 регуляризации получаем:

$$\sum_{i=1}^{l} (\theta_i - y_i)^2 + \lambda \sum_{j=1}^{l} |\theta_j| \to \min$$

Если $\theta_i > 0$, то $2(\theta_i - y_i) + \lambda = 0 \Longrightarrow \theta_i^* = y_i - \frac{\lambda}{2}$.

Если $\theta_i < 0$, то $2(\theta_i - y_i) - \lambda = 0 \Longrightarrow \theta_i^* = y_i + \frac{\lambda}{2}$.
Если $\theta_i = 0$, то $\theta_i^* = 0$:

$$\theta_i^* = \begin{cases} y_i - \frac{\lambda}{2}, \theta_i > 0 \\ y_i + \frac{\lambda}{2}, \theta_i < 0 \\ 0, \theta_i = 0 \end{cases}$$

При использовании L_2-регуляризации зависимость θ_i^* от y_i все еще линейная, компоненты вектора весов ближе расположены к нулю.

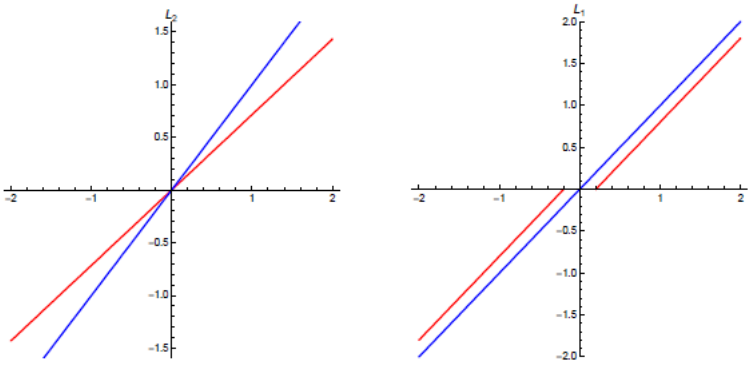

Рис. 15 Регуляризация

В случае L_1-регуляризации график выглядит иначе: существует область размера λ значений y_i, для которых $\theta_i = 0$. То есть L_1 позволяет отбирать признаки, а именно: параметры веса признаков, обладающих низкой предсказательной способностью, оказываются равными нулю.

9 Обобщенные линейные модели

9.1 Понятие функции связи. Роль функции связи в прогнозировании.

Обобщенные линейные модели обеспечивают единый подход к моделированию всех типов целевых переменных. Рассмотрим общий вид линейной модели:

$$y = X\theta + \varepsilon$$

Введем следующее обозначение: $\eta = X\theta$ как часть линейного предиктора.

Обобщение линейных моделей происходит следующими способами:

- В модели линейной регрессии делается предположение о том, что компоненты y независимые нормально распределённые случайные величины с постоянной дисперсией. Мы можем расширить это предположение чтобы использовать любое распределение из тех, что принадлежат экспоненциальному семейству распределений: нормальное, пуассоновское, гамма и биномиальное.

- Вместо того, чтобы оценивать целевую переменную $\mu = E(y)$ как функцию линейного предиктора $X\theta$, ,будем оценивать некоторую функцию $g(\mu)$ от параметра μ. Таким образом параметр распределения связан с линейной комбинацией следующей формулой:

$$g(\mu) = \eta = X\theta$$

В обобщенных линейных моделях ожидаемые значения отклика представляют собой линейную комбинацию предикторов, которые связаны с зависимой переменной через функцию связи g:

Функция $g(\cdot)$ в формуле называется *функцией связи*.

Определение 9.1.1. Функция связи - это функция, связывающая ожидаемое значение переменной Y с предикторами $X_0, \ldots, X_m$. Функция $g(\cdot)$ должна быть монотонной и дифференцируемой, ведь для монотонной функции можно обратить функцию:

$$g^{-1}(g(\mu)) = g^{-1}(X\theta) = \mu$$

Выбор функции связи зависит от типа данных

Функция связи используется в откликах модели, когда предполагается **нелинейная связь зависимой переменной с предикторами**.

Таким образом, спецификация обобщенной линейной модели включает в себя:

- Выбор распределения
- Определение функции связи $g(\cdot)$
- Спецификацию линейного предиктора $X\theta$

9.1.1 Оценка распределения целевой переменной

Линейные модели крайне полезны для анализа данных. Тем не менее, линейные модели во многих отношениях являются ограниченными. Формально, классические приложения линейных моделей опираются на предположения о нормальности, линейности и гомоскедастичности.

Обобщение линейной модели, которое будет представлено в этой главе, позволит моделировать данные используя не только нормальное распределение.

В линейных моделях целевую переменную Y чаще всего считают количественной, нормально распределенной величиной. Но это не единственный тип переменных, который встречается на практике. Вот некоторые примеры типов целевых переменных:

- вещественные
- счетные
- бинарные
- переменные в форме пропорций
- переменные в форме долей
- упорядоченные

9.1.2 Экспоненциальное семейство распределений. Экспоненциальное семейство распределений

Экспоненциальнальное семейство является классом распределений, который включает в себя много известных нам прежде распределений в специальной форме. Общий вид экспоненциального семейства распределений будет выглядеть следующим образом:

$$f(y, \alpha, \varphi) = exp\left[\frac{y \cdot \alpha - c(\alpha)}{\varphi} + h(y, \varphi)\right]$$

где $h(\cdot)$ и $c(\cdot)$ - некоторые функции, а α называется *каноническим параметров*, φ-параметр дисперсии.

Есть еще следующие выражение для статистик μ:

$$Ey = \mu = c'(\alpha) \qquad\qquad Dy = \varphi c''(\alpha)$$

Некоторые распределения являются частными случаями экспоненциального семейства распределений, которые мы будем получать в дальнейшем.

9.2 Функционал качества и функция связи в предположении нормального распределения целевой переменной.

Обобщенные линейные модели расширяют модели линейной регрессии: предсказываемые значения связаны линейной комбинацией входных переменных через обратную функцию связи g:

$$\hat{y}(\theta, X) = g(\theta X)$$

Пусть случайная величина y распределена по следующему закону:

$$y_i = \sum_{j=0}^{m} \theta_j x_{ij} + N(0, \sigma^2) = N(\sum_{j=0}^{m} \theta_j x_{ij}, \sigma^2)$$

То есть

$$\mu = \sum_{j=0}^{m} \theta_j x_{ij} = \bar{\theta}^T \cdot \bar{x}_i$$

Найдем оценку параметра θ методом максимального правдоподобия.

Функция правдоподобия записывается следующим образом:

$$L(\mu, \sigma) = \prod_{i=1}^{n} p(y_i, \mu, \sigma^2) = \frac{1}{(\sqrt{2\pi} \cdot \sigma)^n} \cdot e^{-\frac{1}{2} \sum\limits_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2}} \to \max_{\theta}$$

Прологарифмируем:

$$\ln L(\mu, \sigma) = \ln \left(\frac{1}{(\sqrt{2\pi} \cdot \sigma)^n} \right) + \left(-\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2} \right) = \ln \left((\sqrt{2\pi} \cdot \sigma)^{-n} \right) -$$

$$-\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2} = -n \ln \sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 =$$

$$= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \bar{\theta}^T \cdot \bar{x}_i)^2 \to \max_{\theta}$$

$$Q = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \bar{\theta}^T \cdot \bar{x}_i)^2 \to \max_{\theta}$$

Q является функционалом качества для нормального распределе-

ния.

В обобщенных линейных моделях ожидаемые значения отклика представляют собой линейную комбинацию предикторов, которые связаны с зависимой переменной через функцию связи g:

$$\mu = g^{-1}(\bar{\theta}^T \cdot \bar{x}_i)$$

$$g(\mu) = \bar{\theta}^T \cdot \bar{x}_i$$

Для нормального распределения, изначально было положено:

$$g(\mu) = \bar{\theta}^T \cdot \bar{x}_i = \mu$$

Утверждение: Нормальное распределение находится в классе экспоненциального семейства. Покажем это:

Доказательство.

$$f(y, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = exp\left[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] =$$

$$= exp\left[\frac{\mu y - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]$$

Это эксопненциальное распределение с:

$$\alpha = \mu, c(\alpha) = \frac{\alpha^2}{2}$$

$$E(y) = \mu = c'(\alpha)$$

$$D(y) = \varphi \cdot c''(\alpha) = \sigma^2 \cdot 1 = \sigma^2$$

$$g(\mu) = [c']^{-1}(\mu)$$

$$c(\mu)' = \mu = \bar{\theta}^T \cdot \bar{x}$$

Итого, для нормального распределения N:

$$\bar{\theta}^T \cdot \bar{x} = g(\mu) \qquad g(\mu) = \mu \qquad \mu = g^{-1}(\bar{\theta}^T \cdot \bar{x})$$

Данная функция связи называется *тождественной функцией связи.* □

9.3 Прогнозирование счетной переменной

Существуют задачи, где на y наложены ограничения: например, y должен быть целым и неотрицательным. В таком случае надо пользоваться не МНК, а другими способами.

Если целевые переменные являются вещественными или счетными, а также если предположить, что y неотрицательный, то можно использовать Пуассоновское распределение с логистической функцией связи, которую мы сейчас и получим.

Пусть величина $y \in \{0, 1, 2, \dots, \}$, тогда для нее применимо распределение Пуассона:

$$y \sim Poiss(y = k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

$$Ey = \lambda, Dy = \lambda$$

Так как λ должно быть больше нуля, то хотелось бы линейную комбинацию $(-\infty, \infty)$ переделать в промежуток $(0, \infty)$. Это можно сделать с помощью потенцирования.

$$\log y = \bar{\theta}^T \cdot \bar{x} \Rightarrow \lambda = e^{\bar{\theta}^T \cdot \bar{x}}$$

Опять же найдем с помощью ммп логарифм функции правдоподобия

$$P(\xi = y) = \frac{e^{-\lambda} \cdot \lambda^y}{y!}, y \in \{0, 1, 2, \dots, \}$$

$$L(\lambda) = \prod_{i=1}^{n} p(y_i, \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!}$$

$$\ln L(\lambda) = \sum_{i=1}^{n} \ln \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} = \sum_{i=1}^{n} (-\lambda + y_i \ln \lambda - \ln(y_i!))$$

Последнее слагаемое не сожержит оцениваемого параметра λ, поэтому:

$$Q = \sum_{i=1}^{n} (-\lambda + y_i \ln \lambda) \to \max_{\theta}$$

$$Q = \sum_{i=1}^{n} (e^{\bar{\theta}^T \cdot \bar{x}} - y_i \ln e^{\bar{\theta}^T \cdot \bar{x}}) \to \min_{\theta}$$

$$Q = \sum_{i=1}^{n} (e^{\bar{\theta}^T \cdot \bar{x}} - y_i \bar{\theta}^T \cdot \bar{x}) \to \min_{\theta}$$

Для пуассоновского распределения:

$$g(\mu) = \log \lambda$$

$$\mu = g^{-1}(\log \mu) = e^{\log \mu} = \mu$$

Утверждение: Распределение Пуассона лежит в классе экспоненциального семейства, покажем это:

Доказательство.

$$f(y, \lambda) = exp\left(\frac{y \cdot \log \lambda - \lambda}{1} - log(y!)\right)$$

$$\alpha = \log(\lambda), \lambda = exp(\alpha), c(\alpha) = \lambda, c(\lambda) = e^{\lambda}$$

$$g(\lambda) = \log(\lambda) = \bar{\theta}^T \cdot \bar{x}$$

И это верно, так как:

$$\lambda = g^{-1}((g(\bar{\theta}^T \cdot \bar{x}))) = e^{\log \lambda} = \lambda$$

Данная функция связи называется *логистической функцией связи.*

□

9.4 Прогнозирование времени наступления события. Анализ выживаемости

Задача анализа выживаемости состоит в том, чтобы предсказать, сколько проживет человек, поступивший в больницу или в предсказании времени, которое пройдет до следующего события.

Экспоненциальное распределение моделирует время между двумя последовательными свершениями одного и того же события и очень хорошо подходит для моделирования задачи анализа выживаемости.

Сначала покажем, что экспоненциальное распределение принадлежит экспоненциальному семейству распределения (было бы забавно, если бы не принадлежало):

Доказательство.

$$f(y) = \mu e^{-\mu y} = \exp(-\mu y + \log \mu) = \exp\left(\frac{y(-\mu) + \log \mu}{1} - 0\right) =$$

$$= \exp\left(\frac{y(-\mu) - (-\log \mu)}{1} - 0\right)$$

$$\alpha = -\mu \qquad c(\alpha) = -\log \mu$$

$$g(\mu) = \bar{\theta}^T \cdot \bar{x} = -\frac{1}{\mu}$$

И это верно, так как:

$$g^{-1} = \frac{-1}{\mu} \Leftrightarrow \mu = g^{-1}(g(\mu)) = -\frac{1}{\left(-\frac{1}{\mu}\right)} = \mu$$

□

А теперь выведем функционал качества для задач анализа выживаемости используя метод максимального правдоподобия и функцию правдоподобия соответственно. Пусть y подчиняется экспоненциальному распределению, которое является непрерывным:

$$P(y) = \mu e^{-\mu y}$$

В качестве функции связи возьмем следующую величину:

$$\lambda = -\frac{1}{\sum\limits_{j=0}^{m} \theta_j x_{ij}} = -\frac{1}{\bar{\theta}^T \cdot \bar{x}}$$

Запишем функцию правдоподобия:

$$L = \prod_{i=1}^{n} -\frac{1}{\bar{\theta}^T \cdot \bar{x}} \exp\left(\frac{y}{\bar{\theta}^T \cdot \bar{x}}\right) \to \max$$

$$\ln L = \sum_{i=1}^{n} \ln\left(-\frac{1}{\bar{\theta}^T \cdot \bar{x}}\right) + \left(\frac{y}{\bar{\theta}^T \cdot \bar{x}}\right) \to \max$$

$$Q = \sum_{i=1}^{n} \ln\left(-\bar{\theta}^T \cdot \bar{x}\right) - \left(\frac{y}{\bar{\theta}^T \cdot \bar{x}}\right) \to \min$$

Это и есть функционал качества для экспоненциального распределения, с помощью которого можно решать задачу анализа выживаемости.

Вкратце рассмотрим два распределения, также имеюищих важное значение.

Отрицательное биномиальное распределение - также называемое распределением Паскаля — это распределение дискретной случайной величины равной количеству произошедших неудач в последовательности испытаний Бернулли с вероятностью успеха p, проводимой до n-го успеха.

$$P(y = k) = C_{n+k-1}^{k} p^n (1-p)^k$$

Гамма распределение широко применяется для моделирования сложных потоков событий, сумм временных интервалов между событиями, в экономике, теории массового обслуживания, в логистике, описывает продолжительность жизни в медицине. Является своеобразным аналогом дискретного отрицательного биномиального распределения.

Функция связи у гамма-распределения такая же, как у экспоненциального: $g(\mu) = \bar{\theta}^T \cdot \bar{x} = -\frac{1}{\mu}$. И гамма-распределение, и отрицательное биномиальное распределение входит в состав экспоненциального семейства.

9.5 Применение биномиального распределения

Биномиальное распределение - дискретное распределение вероятностей случайной величины y, принимающей целочисленные значения $k = 0, 1, \ldots, n$ и применяется для ситуаций, когда необходимо высчитать вероятность определенного количества усехов в последовательности из n независимых экспериментов, таких, что вероятность успеха в каждом из них постоянна и равна p:

$$F(Y = y) = C_n^y \cdot p^y \cdot (1-p)^{n-y} \qquad Ey = np \qquad Dy = np(1-p)$$

Пусть $y \sim Bin(n, p)$. Запишем функцию вероятности и попытаемся вывести функцию связи для биномиального распределения:

$$F(Y = y) = C_n^y \cdot p^y \cdot (1-p)^{n-y} = \exp\left[\ln C_n^y + \ln p + (n-y)\ln(1-p)\right] =$$

$$= \exp\left[y \ln\left(\frac{p}{1-p}\right) + n\ln(1-p) + \ln C_n^y\right]$$

В качестве $\alpha = \ln\left(\frac{p}{1-p}\right)$, то есть $p = \frac{e^\alpha}{1+e^\alpha}$, тогда подставив это в выражение, получим:

$$f(y, p) = \exp\left[y \cdot \alpha + n \ln\left(\frac{1}{1+e^\alpha}\right) + \ln C_n^y\right]$$

Биномиальное распределение содержится в экспоненциальном распределении со следующими параметрами:

$$\alpha = \ln\left(\frac{p}{1-p}\right) \qquad c(\alpha) = n\ln(1+e^\alpha)$$

Соответственно функция связи имеет вид:

$$g(p) = \ln\frac{p}{1-p} = \bar{\theta}^T \cdot \bar{x}$$

Действительно, ведь:

$$p = g^{-1((g(p)))} = \frac{e^{g(p)}}{1+e^{g(p)}} = \frac{e^{\ln\frac{p}{1-p}}}{1+e^{\ln\frac{p}{1-p}}} = p$$

Такая функция связи называется *logit-функцией связи.*

Теперь с помощью метода максимального правдоподобия выведем

функционал качества для биномиального распределения:

$$p = g^{-1((g(p)))} = \frac{e^{g(p)}}{1 + e^{g(p)}} = \frac{e^{\bar{\theta}^T \cdot \bar{x}}}{1 + e^{\bar{\theta}^T \cdot \bar{x}}}$$

Запишем функцию правдоподобия:

$$L = \prod_{i=1}^{l} C_n^{y_i} p_i^{y_i} (1 - p_i)^{n - y_i} \to \max_{\theta}$$

$$\ln L = \sum_{i=1}^{n} \log C_n^{y_i} + y_i \log p_i + (n - y_i) \log(1 - p_i) \to \max_{\theta}$$

Биномиальный коэффициент не содержит оцениваемого параметра p, поэтому функционалом качества биномиального распределения будет:

$$Q = \sum_{i=1}^{n} y_i \log \frac{e^{\bar{\theta}^T \cdot \bar{x}}}{1 + e^{\bar{\theta}^T \cdot \bar{x}}} + (n - y_i) \log(1 - \frac{e^{\bar{\theta}^T \cdot \bar{x}}}{1 + e^{\bar{\theta}^T \cdot \bar{x}}}) \to \max_{\theta}$$

9.6 GARMA

GARMA переводится как Generalized Autoregressive Moving Average Models - модели скользящего среднего для обобщенных линейных моделей, у которых распределение целевой переменной не является нормальным гауссовским. Оценка модели выполняется с использованием итеративно взвешенного алгоритма наименьших квадратов.

Модели GARMA может быть исоплзована для распределения Пуассона, отрицательного биномиального распределения или для непрерывного *gamma*-распределения, называемое $GARCH$ моделью.

9.6.1 Определение модели

В GARMA моделе, условное распределение *каждого наблюдения* $y_t, t = 1, \ldots, n$ вычисляется с учетом предыдущей информации

$$H_t = \{x_r, \ldots, x_1;\ y_{t-1}, \ldots, y_1;\ \ , \mu_{t-1}, \ldots, \mu_1\}$$

в предположении о том, что данные наблюдения принадлежат одному и тому экспоненциальному семейству:

$$f(y_t|H_t) = exp\left[\frac{y_t \cdot \alpha - c(\alpha)}{\varphi} + h(y, \varphi)\right]$$

где α и φ - канонический и параметр масштаба, $c(\cdot)$ и $h(\cdot)$ функции, определяющие экспоненциальное семейство функция и x, являющийся вектором из r объясняющих переменных.

$$E(y_t|H_t) = \mu = c'(\alpha) \qquad\qquad D(y_t|H_t) = \varphi c''(\alpha)$$

выражают условную вероятность каждого значения y_t по H_t.

Как видно, обозначения здесь очень похожи на те, что и для независимых наблюдений GLM, но в данном случае моделируются условные распределения, что является важным уточнением модели.

Как и в стандартном GLM, параметр μ_t связан с предиктором η_t с помощью дважды дифференцируемой мотонной функции g, которая называется *функцией связи*.

По сравнению с GLM, где предиктор $\eta = \bar{x_t}^T \cdot \bar{\theta}$, где $\theta = (\theta_1, \theta_2, \ldots, \theta r)$, в данном случае добавляется компонента τ_t, которая позволяет добавить автогресионную компоненту скользяшего среднего к предиктору. Про-

шлые значения, зависящие от вермени, также включены в предиктор.

Модель для μ_t в GARMA $g(\mu_t) = \eta_t = \bar{x_t}^T \cdot \bar{\theta} + \tau_t$ где:

$$\tau_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

функции, которые представляют собой авторегрессионную AR компоненту и компоненту скользящего среднего MA.

Данное выражение записывается по-другому:

$$g(\mu_t) = \eta_t = \bar{x_t}^T \cdot \bar{\theta} + \sum_{i=1}^{p} \left(g(y_{t-i}) - \bar{x}_{t-j}^T \cdot \bar{\theta} \right) + \sum_{j=1}^{q} \phi_j \left(g(y_{t-i}) - \eta_{t-j} \right)$$

Данная модель в экспоненциальном семействе распределения называется моделью *GARMA(p,q)*. Параметры θ, α и ϕ определяются с помощью метода максимального правдоподобия.

9.6.2 Poisson Garma(p,q)

В предположении, что g является логистической функции связи $g(\lambda_t) = \log(\lambda_t)$:

$$\log(\lambda_t) = \bar{x_t}^T \cdot \bar{\theta} + \sum_{i=1}^{p} \left(\log(y_{t-i}^*) - \bar{x}_{t-j}^T \cdot \bar{\theta} \right) + \sum_{j=1}^{q} \phi_j \left(\log(y_{t-i}^*) - \lambda_{t-j} \right)$$

Так как $\alpha_t = \log(\lambda_t) = \bar{x_t}^T \cdot \bar{\theta}$, то:

$$\log(\lambda_t) = \alpha_t + \sum_{i=1}^{p} \left(\log(y_{t-i}^*) - \alpha_{t-i} \right) + \sum_{j=1}^{q} \phi_j \left(\log(y_{t-i}^*) - \lambda_{t-j} \right)$$

10 Дополнительные главы

Мы обучались на многих библиотеках и многие функции мы использовали очень часто. Разберем некоторые из них.

Базовые библиотеки:

1. **IMPORT NUMPY AS NP** - библиотека для ускоренных вычислений *Numpy*
2. **IMPORT PANDAS AS PD** - для обработки данных *Pandas*
3. **IMPORT MATPLOTLIB.PYPLOT AS PLT** - для визуализации данных *Matplotlib*
4. **IMPORT SEABORN AS SNS** - библиотека *Seaborn*, более качественная визуализация

Документация Numpy, Pandas, Matplotlib и Seaborn

10.1 Операции с временными рядами

1. **scipy.stats.boxcox** - преобразование Бокса-Коса, принимает временной ряд, в возвращает преобразованный ряд и значение λ, при котором достигается максимум правдоподобия

2. **scipy.stats.boxcox_llf** - возвращает значение функции правдоподобия

3. **pd.Series.diff(lag)** - дифференцирование временного ряда с лагом τ

4. **pd.Series.shift(n)** - сдвиг на n периодов ряда.

5. **from scipy.optimize import minimize** - минимизация функции от параметра, параметр передается через *lambda*-функцию.

   Документация Scipy

10.2 Statsmodels

1. **import statsmodels.api as sm** - вся библиотека *Statsmodels* для прогнозирования временных рядов
2. **from statsmodels.tsa.holtwinters import**
   (a) **SimpleExpSmoothing** - простое экспоненциальное сглаживание
   (b) **Holt** - методы Хольта, двойное экспоненциальное сглаживание
   (c) **Exponential Smoothing** - методы Хольта-Уинтерса, тройное экспоненциальное сглаживание
3. **from statsmodels.graphics.tsaplots import plot_acf** - коррелограмма, автокорреляционная функция (ACF)
4. **from statsmodels.graphics.tsaplots import plot_pacf** - график частичной автокорреляции (PACF)
5. **from statsmodels.tsa.arima_model import ARIMA** - построение модели ARIMA
6. **from statsmodels.tsa.statespace.sarimax import SARIMAX** - модель SARIMA
7. **from statsmodels.tsa.stattools import adfuller** - тест на стационарность Дики-Фуллера, возвращает pvalue вторым выходным аргументом
8. **from statsmodels.tsa.seasonal import seasonal_decompose** - разложение ряда на тренд, сезонную компоненту и случайную компоненту

Документация Statsmodels

10.3 Prophet

Prophet для визуализации результатоп прогнозирования использует библиотеку **Plotly**, которая позволяет строить интерактивные графики.

Модель прогнозирования в библиотеке от Facebook имеет название **Prophet.** Основной идеей построения прогноза стало разложение временного ряда на основные составляющие:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

где $g(t)$ – функция, описывающая тренд временного ряда, $s(t)$ – компонента, описывающая сезонные колебания, $h(t)$ – компонента, отвечающая за различные праздники и события, которые могут оказывать влияние на целевую переменную $y(t)$, а ϵ_t представляет собой непрогнозируемую перечисленными компонентами ошибку.

from fbprophet import Prophet - помощью данного метода строятся модели

Документация Prophet

10.4 Sklearn

Механизм создания модели - сначало необходимо инициализировать модель в отдельную переменную, затем обучить на выборке.

1. **from sklearn.linear_model import LinearRegression** - модель линейной регрессии, на вход подается матрица объекты-признаки X и целевая переменная y. Обучение происходит при помощи следующих команды - model.fit(X,y)

2. **from sklearn.preprocessing import PolynomialFeatures** - служит для построения полиномиальной регрессии, создает матрицы объекты-признаки $1, x, x^2, \ldots$. По умолчанию, кроме степеней будут созданы столбцы с всевозможными комбинациями этих признаков при имеюищихся признаках a, b. Чтобы получить результат преобразования, нужно применить fit_transform

3. **from sklearn.metrics import mean_squared_error, mean_absolute_error, explained_variance_score** - метрики качеств MSE, MAE, R^2.

4. **from sklearn.model_selection import train_test_split** - разделение данных на тренировочную и тестовую выборку, передается матрица объекты-признаки X, y и доля тестовой выборки.

   Если нет привязки ко времени, то надо перемешать выборку, указав shuffle=True и задав random_state=42, чтобы зафикисровать разбиение

5. **from sklearn.model_selection import cross_val_score** - кросс-валидация, на вход дается модель, обучающую выборку, метрику качеству и число разбиений cv.

6. **from sklearn.model_selection impor KFold** - генератор разбиения KFold, в котором задается число блоков, а также фиксируется разбиение. Чем больше значение метрики качества, тем луче

7. **sklearn.model_selection.TimeSeriesSplit** - кросс-валидация на временных рядах, передается в cv к cross_val_score.

8. **from sklearn.pipeline import Pipeline** - Pipeline позволяет последовательно выполнять преобразование над данными. Нет необходимости сначала создавать полиномиальные признаки, затем строить модель, достаточно передавать в готовый Pipeline исходные данные.

9. **from sklearn.linear_model import Ridge, Lasso** - L_2 и L_1 регрессии

10. **from sklearn.model_selection import GridSearchCV** - позволяет подобрать параметр регуляризации λ, на вход подается словарь параметров, которые нужно перебрать, их названия и значения. Выбор параметров реализован через fit.

    Документация Sklearn

Заключение

Написание теоретических материалов и конспектов подходит к концу, но многое еще не сказано и ожидает нас впереди.

Я благодарю Дарью Александровну за честный труд и живую увлеченность при подаче материала, а так же за практические навыки, которые мы приобрели на занятиях в такое непростое время. Надеюсь, что дальше будет светлее, хоть ночи и будут длиннее.

Благодарю немногочисленных моих одногруппников за мотивацию и вдохновление, благодаря Вам я стараюсь описывать материал доступнее и понятнее.

Практические материалы можно посмотреть здесь

Мы все переживем, ведь Нева впадает в море Балтийское.

Александр Широков