ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Лекция 1. Классические линейные модели

Анализ данных с временной структурой

Стационарность

Пусть $\{u_t\}, t \in \mathbb{Z}$ - последовательнсоть случайных величин, определенных на некотором вероятностном пространстве $(\Omega, \mathcal{F}, P)$. Последовательность $\{u_t\}$ - **стационарна в широком смысле** (в смысле Хинчина), если $\forall t$:

$$\mathbb{E}u_t = const, \qquad cov(u_t, u_{t+\tau}) = R(\tau)$$

Последовательность $\{u_t\}$ - **стационарна в узком смысле**, если $\forall (t_1, \ldots, t_k), \tau$:

$$Law(u_{t_1}, ..., u_{t_k}) = Law(u_{t_1+\tau}, ..., u_{t_k+\tau})$$

Белый шум

$\{\varepsilon_t\}$ - **белый шум в широком смысле**, если $\mathbb{E}\varepsilon_t = 0$ и $cov(\varepsilon_t, \varepsilon_s) = \sigma^2\delta_{ts}$. (последовательность некоррелированных случайных величин с нулевым математическим ожиданием).
$\{\varepsilon_t\}$ - **белый шум в узком смысле**, если помимо вышеописанных свойств является гауссовоской последовательностью.

Авторегрессия. AR(p)

Разностное стохастическое уравнение вида

$$u_t = \beta_1 u_{t-1} + ... + \beta_p u_{t-p} + \varepsilon_t, t \in \mathbb{Z}$$

где ε_t - белый шум в широком смысле, называется уравнением авторегрессии порядка p $AR(p)$
Любая последовательность, удовлетворяющая уравнению выше с точностью до почти наверное, называется процессом авторегрессии.
Характеристическим уравнением, присоединенным к уравнению аторегрессии называется уравнение вида

$$x^p = \beta_1 x^{p-1} + ... + \beta_p$$

Интерес для нас представляет стационарность решения уравнения авторегрессии в узком и широком смысле

Авторегрессия. AR(1)

Рассмотрим модель AR(1) поподробнее. Выглядит она следующим образом:

$$u_t = \beta_0 + \beta_1 u_{t-1} + \epsilon_t$$

проследив рекуррентные соотношения до начала временной шкалы, можем получить следующее выражение:

$$u_t = \beta_0(1 + \beta_1 + ... + \beta_1^{t-1}) + \beta_1^t u_0 + (\epsilon_t + \beta_1 \epsilon_{t-1} + ... + \beta_1^{t-1}\epsilon_1)$$

Авторегрессия. AR(1)

$$u_t = \beta_0(1 + \beta_1 + ... + \beta_1^{t-1}) + \beta_1^t u_0 + (\epsilon_t + \beta_1\epsilon_{t-1} + ... + \beta_1^{t-1}\epsilon_1)$$

Из разложения получаем следующие выражения:

$$\mathbb{E}u_t = \frac{\beta_0(1 - \beta_1^t)}{1 - \beta_1} + \beta_1^t \mathbb{E}u_0$$

$$Du_t = \beta_1^{2t} Du_0 + \frac{\sigma^2(1 - \beta_1^{2t})}{1 - \beta_1^2}$$

$$cov(u_t, u_{t-k}) = \beta_1^{2t-k} Du_0 + \frac{\beta_1^k \sigma^2(1 - \beta_1^{2(t-k)})}{1 - \beta_1^2}$$

Авторегрессия. AR(1)

Рассмотрим случай $|\beta_1| < 1$. Предположим также, что начальный элемент ряда имеет конечные матоэидание и дисперсию. В этом случае при переходе в предел по t получим "ассимптотику стационарности":

$$\mathbb{E}u_t \to \frac{\beta_0}{1-\beta_1}$$

$$Du_t \to \frac{\sigma^2}{1-\beta_1^2}$$

$$cov(u_t, u_{t-k}) \to \frac{\sigma^2\beta_1^2}{1-\beta_1^2}$$

Более того, если $Law(u_0) = \mathcal{N}(\frac{\beta_0}{1-\beta_1}, \frac{\sigma^2}{1-\beta_1^2})$, то ряд будет являться стационарным в узком смысле.

Компоненты временных рядов

- Тренд — плавное долгосрочное изменение уровня ряда. Эту характеристику можно получить, наблюдая ряд в течение достаточно долгого времени.
- Сезонность
- Цикл
- Ошибка

Компоненты временных рядов

- Тренд
- Сезонность — циклические изменения уровня ряда с постоянным периодом. В данных о средней зарплате в России очень хорошо видны подобные сезонные колебания: признак всегда принимает максимальное значение в декабре каждого года, а минимальное — в январе следующего года. В целом профиль изменения зарплаты внутри года остаётся более-менее постоянным.
- Цикл
- Ошибка

Компоненты временных рядов

- Тренд
- Сезонность
- Цикл — изменение уровня ряда с переменным периодом. Такое поведение часто встречается в рядах, связанных с продажами, и объясняется циклическими изменениями экономической активности. В экономике выделяют циклы длиной 4 - 5 лет, 7 - 11 лет, 45 - 50 лет и т. д. Другой пример ряда с такой характеристикой — это солнечная активность, которая соответствует, например, количеству солнечных пятен за день. Она плавно меняется с периодом, который составляет несколько лет, причём сам период также меняется во времени.
- Ошибка

Компоненты временных рядов

- Тренд
- Сезонность
- Цикл
- Ошибка — непрогнозируемая случайная компонента ряда. Сюда включены все те характеристики временного ряда, которые сложно измерить (например, слишком слабые).

Additive and Multiplicative TS

Аддитивно:

$$Value = BaseLevel + Trend + Seasonality + Error$$

Мультиаликативно:

$$Value = BaseLevel \times Trend \times Seasonality \times Error$$

Заметим, что $y_t = S_t \times T_t \times R_t$ эквивалентно
$\log y_t = \log S_t + \log T_t + \log R_t$.

Стационарность модели

- Взять разность
- Взять логарифм
- Взять n-ый корень
- Комбинация указанного выше

Проверка стационарности

- Пристальное вглядывание в график временного ряда
- Разбиение на части и подсчет статистик
- Тесты на единичные корни
- Dickey Fuller test (ADH Test)
- Kwiatkowski-Phillips-Schmidt-Shin – KPSS test (trend stationary)
- Philips Perron test (PP Test)

Чаще всего используют тест Дикки Фуллера, нулевая гипотеза о нестационарности ряда. Если p значение меньше уровня значимости(0.05) - гипотеза отклоняется в пользу альтернативы.

Взятие разности

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

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

Для нестационарного ряда часто оказывается, что получаемый после дифференцирования ряд является стационарным. Дифференцирование можно применять неоднократно: от ряда первых разностей, продифференцировав его, можно прийти к ряду вторых разностей, и т. д.

Взятие разности

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

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

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

Стабилизация дисперсии

В случае, если во временном ряде монотонно по времени изменяется дисперсия, применяется специальное преобразование, стабилизирующее дисперсию.

Очень часто в качестве такого преобразования выступает логарифмирование. Логарифмирование принадлежит к семейству преобразований Бокса-Кокса.

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

Это параметрическое семейство функций, в котором параметр определяет, как именно будет преобразован ряд:

$\lambda = 0$ — это логарифмирование,

$\lambda = 1$ — тождественное преобразование ряда,

а при других значениях λ — степенное преобразование.

Как избавиться от тренда

- Воспользоваться линейной регрессией на шаг временного ряда (используя также степенные компоненты)
- Вычесть среднее
- Baxter-King filter(statsmodels.tsa.filters.bkfilter) или Hodrick-Prescott Filter (statsmodels.tsa.filters.hpfilter), чтобы убрать среднее или циклическую компоненту

Как учесть сезонность?

1. Скользящее среднее с сезонным окном
2. Сезонные разности
3. Разбить ряд по временным индексами

ACF

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

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

Автокорреляция — это уже встречавшаяся ранее корреляция Пирсона между исходным рядом и его версией, сдвинутой на несколько отсчётов. Количество отсчётов, на которое сдвинут ряд, называется лагом автокорреляции (τ).

ACF

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

Анализировать величину автокорреляции при разных значениях лагов удобно с помощью графика, который называется коррелограммой. По оси ординат на нём откладывается автокорреляция, а по оси абсцисс — размер лага τ.

ACF

Как и для обычной корреляции Пирсона, значимость вычисляется с помощью критерия Стьюдента. Альтернатива чаще всего двусторонняя, потому что при анализе временных рядов крайне редко имеется гипотеза о том, какой должна быть корреляция, положительной или отрицательной.

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

Авторегрессия

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

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

В этом регрессионном уравнении y_t — это отклик, $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ — признаки, $\alpha, \phi_1, \phi_2, \ldots, \phi_p$ — параметры модели, которые необходимо оценить, ε_t — шумовая компонента, описывает отклонения значений ряда от данного уравнения. Такая модель называется моделью авторегрессии порядка p (AR(p)). В этой модели y_t представляет собой линейную комбинацию p предыдущих значений ряда и шумовой компоненты.

PACF

Частичная автокорреляция - это сводка взаимосвязи между наблюдением во временном ряду с наблюдениями на предыдущих временных этапах с удалением взаимосвязей между промежуточными наблюдениями.

Частичная автокорреляция при лаге k - это корреляция, возникающая после устранения влияния любых корреляций, связанных с членами с более короткими лагами.

PACF

Поскольку автоковариация стационарного ряда зависит только от сдвига, то и автокорреляция – это функция только сдвига k,

$$(k) = \frac{cov(y_t, y_{t-k})}{\sigma_y^2} = \frac{\gamma(k)}{\gamma(0)}, k \geq 0.$$

По сути, вместо того, чтобы находить корреляции настоящего с лагами, такими как ACF, он находит корреляцию остатков (которая сохраняется после устранения эффектов, которые уже были объяснены более ранним лагом (ами)) со следующим значением лага, следовательно, «частичным», а не «полным».

Скользящее среднее

Рассмотрим независимый, одинаково распределённый во времени шум ε_t.
Для каждого значения t можно вычислить среднее арифметическое между точками ε_t и ε_{t1}. Также можно вычислять среднее не по двум, а по трём или четырём точкам и т.д. Данную идею можно обобщить и записать следующую модель ряда:

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

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

Использованные источники

1. Jonathan D. Cryer , Kung-Sik Chan. Time Series Analysis With Applications in R. – 2008. – 491 c.

2. Kane M.J, Price N., Scotch M., Rabinowitz P. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. – 2014. – 9 c.

3. Robert Nau. Statistical forecasting: notes on regression and time series analysis.: Fuqua School of Business, Duke University. – 2020.

4. Brian Christopher. Time Series Analysis (TSA) in Python - Linear Models to GARCH - applying the ARIMA models family to the task of modeling financial indicators.