МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования «Ульяновский государственный университет»

Факультет математики, информационных и авиационных технологий

«Прогнозирование. Временные ряды. Регрессия»

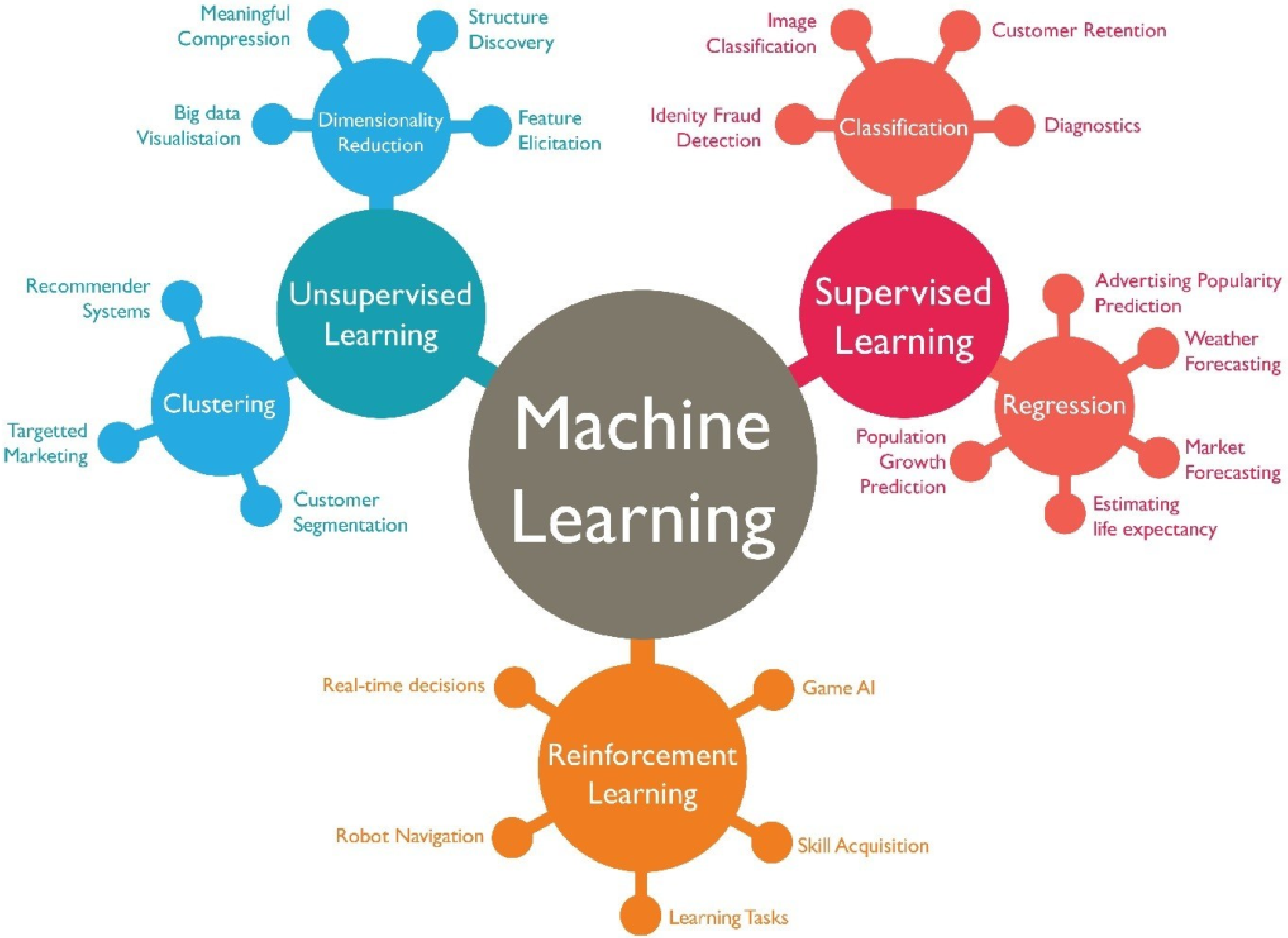
Павлов Павел Юрьевич

Кандидат технических наук

Ульяновск, 2023

Agenda

1. Прогнозирование
2. Методы прогнозирования временных рядов
3. Регрессионные модели
4. Линейная регрессия
5. Логистическая регрессия


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Прогнозирование

Прогнозирование (Forecasting) — это разработка прогноза. В узком значении — специальное научное исследование конкретных перспектив дальнейшего развития какого-либо процесса.

Точность любого прогноза обусловлена:

- объёмом истинных (верифицированных) исходных данных и периодом их сбора;
- объёмом неверифицированных исходных данных, периодом их сбора;
- свойствами системы, объекта, подвергающихся прогнозированию;
- методиками и подходами прогнозирования.

Прогнозирование

Прогнозы делятся (условно):

- по срокам: краткосрочные, среднесрочные, долгосрочные, дальнесрочные;
- по масштабу: частные, местные, региональные, отраслевые, страновые, мировые (глобальные).
- по ответственности (авторству): личные, на уровне предприятия (организации), на уровне государственных органов.

Методы прогнозирования

Статистические методы:

- Экстраполяционные методы (Extrapolative Methods)
  - Скользящее среднее, простое скользящее среднее (Moving Average, Simple Moving Average)
  - Экспоненциальное сглаживание (Exponential Smoothing)
  - Авторегрессионное скользящее среднее (Auto Regression Moving Average (ARMA))
- Объясняющие переменные методы (Explanatory Variable Methods)
  - Регрессионный анализ (Regression Analysis)
  - Прогнозное моделирование (Predictive Modeling)
  - Искусственные нейронные сети (Artificial Neural Networks)
  - Эконометрическое моделирование (Econometric Modeling)

Методы прогнозирования

Экспертные оценки (Judgmental Methods)

- Составные прогнозы (Composite forecasts)
- Статистическое обследование (Statistical surveys)
- Метод Дельфи (Delphi Method)
- Построение сценариев (Scenario building)
- Прогнозирование технологий (Technology forecasting)
- Прогноз по аналогии (Forecast by analogy)

Методы прогнозирования

Методы моделирования (Simulation Modeling Methods)

- Моделирование на основе клеток (Cell-based Modeling)
- Моделирование динамики систем (Systems Dynamics Simulation)
- Мультиагентное моделирование (Multi-agent Simulation)

Интуитивные (то есть выполненные без применения технических средств, экспромтом, «в уме» специалистом, имеющим опыт ранее применяемых научных методов в данном типе прогнозов).

Комбинированные

- Байесовское прогнозирование (Bayesian Forecasting)
- другие

Применение

Прогнозирование может применяться в разных сферах деятельности:

- Управление цепочками поставок - для того, чтобы убедиться, что продукт находится в нужном месте в нужное время, сократить издержки за счет оптимизации поставок товара в соответствии с прогнозируемой потребностью.
- Управление технологиями - технологический форсайт
- Маркетинг - прогнозирование успешности выхода нового продукта на рынок (Product forecasting)
- Телекоммуникации - прогнозирование спроса на телекоммуникационные услуги для планирования развития сетей (Telecommunications forecasting)
- Продажи - прогнозирование спроса (Sales Forecasting)

Понятие временного ряда

Динамика финансово-экономических показателей обычно отражается динамическими и временными рядами.

Динамические ряды — упорядоченная совокупность последовательных наблюдений одного показателя y в зависимости от последовательно возрастающих или убывающих значений другого показателя x.

Временные ряды — динамические ряды, у которых в качестве признака упорядочения выбрано время t.

Временной ряд — это последовательность значений, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени времени, обычно через равные промежутки промежутки.

Пример временного ряда

Уровни временного ряда

Уровни ряда могут измеряться в различных величинах:

- абсолютных (размер прибыли, издержек, …);
- относительных (объем производства с/х продукции на душу населения);
- средних за некоторый период времени (среднесуточная выработка продукции,…);
- индексных (индексы роста накопленного дохода,…).

Уровни временного ряда могут принимать:

- детерминированные значения – не представляют интереса (например, число дней месяце);
- случайные значения – подвергаются научному анализу, при этом они могу быть:
  - дискретными и непрерывными.

Длина временного ряда определяется количеством наблюдений n.

Модель временного ряда

Тренд (T) – плавно изменяющаяся компонента, описывающая чистое влияние долговременных факторов (рост населения, изменение структуры возрастного состава и т.д.);

- **циклическая компонента (C)** – плавно изменяющаяся компонента, описывающая длительные периоды относительного подъема и спада, состоит из циклов, меняющихся по амплитуде и протяженности (в экономике бывает связана со взаимодействием спроса и предложения, ростом и истощением ресурсов, изменением в финансовой и налоговой политике и т.п.);
- **сезонная компонента (S)** – состоит из последовательности почти повторяющихся циклов (объем продаж накануне Нового Года, объем перевозок пассажиров городским транспортом);
- **случайная компонента (e)** – остается после полного вычленения закономерных компонент.

Аддитивная и мультипликативная модели ВР

ВР представляет собой

- либо сумму этих компонент X = T+C+S+e в аддитивной модели,
- либо произведение X = T*C*S*e в мультипликативной модели.

Второй вариант более распространен в экономических приложениях и сводится к первому логарифмированием

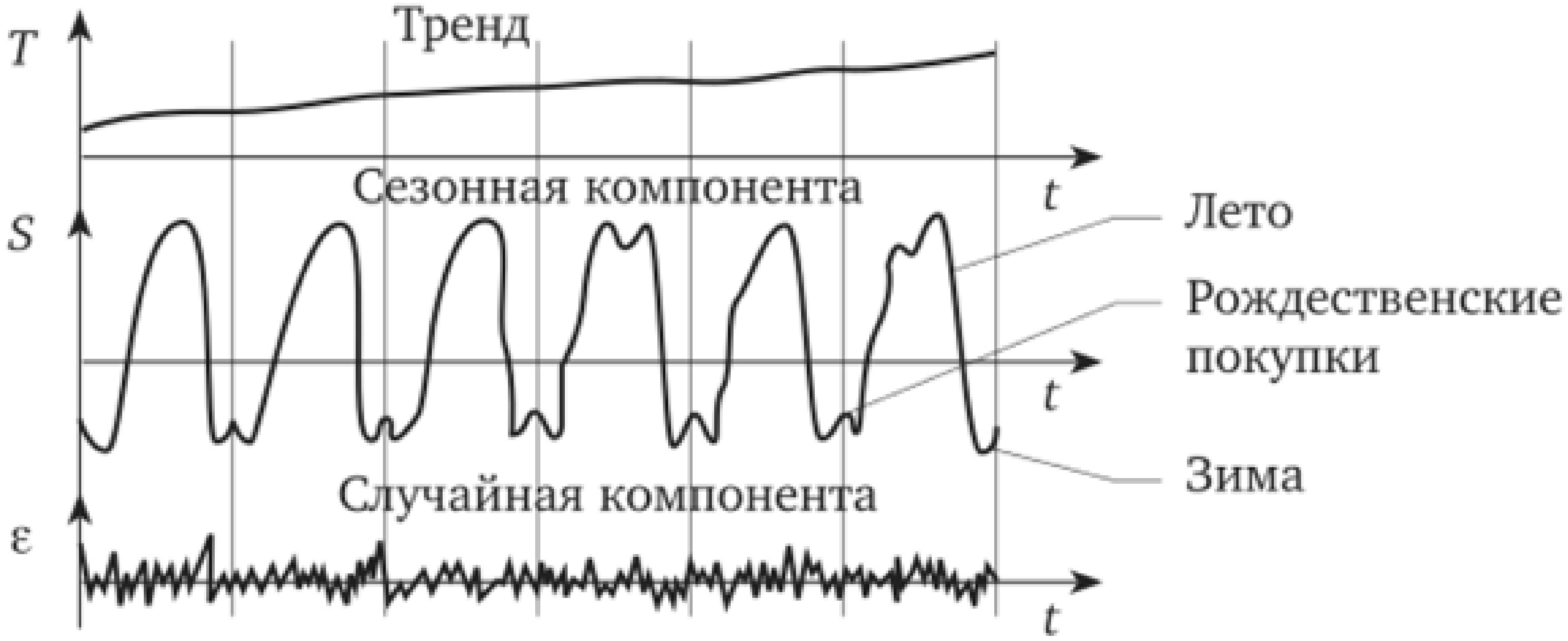
Выделение тренда

Наиболее распространены следующие модели трендов:

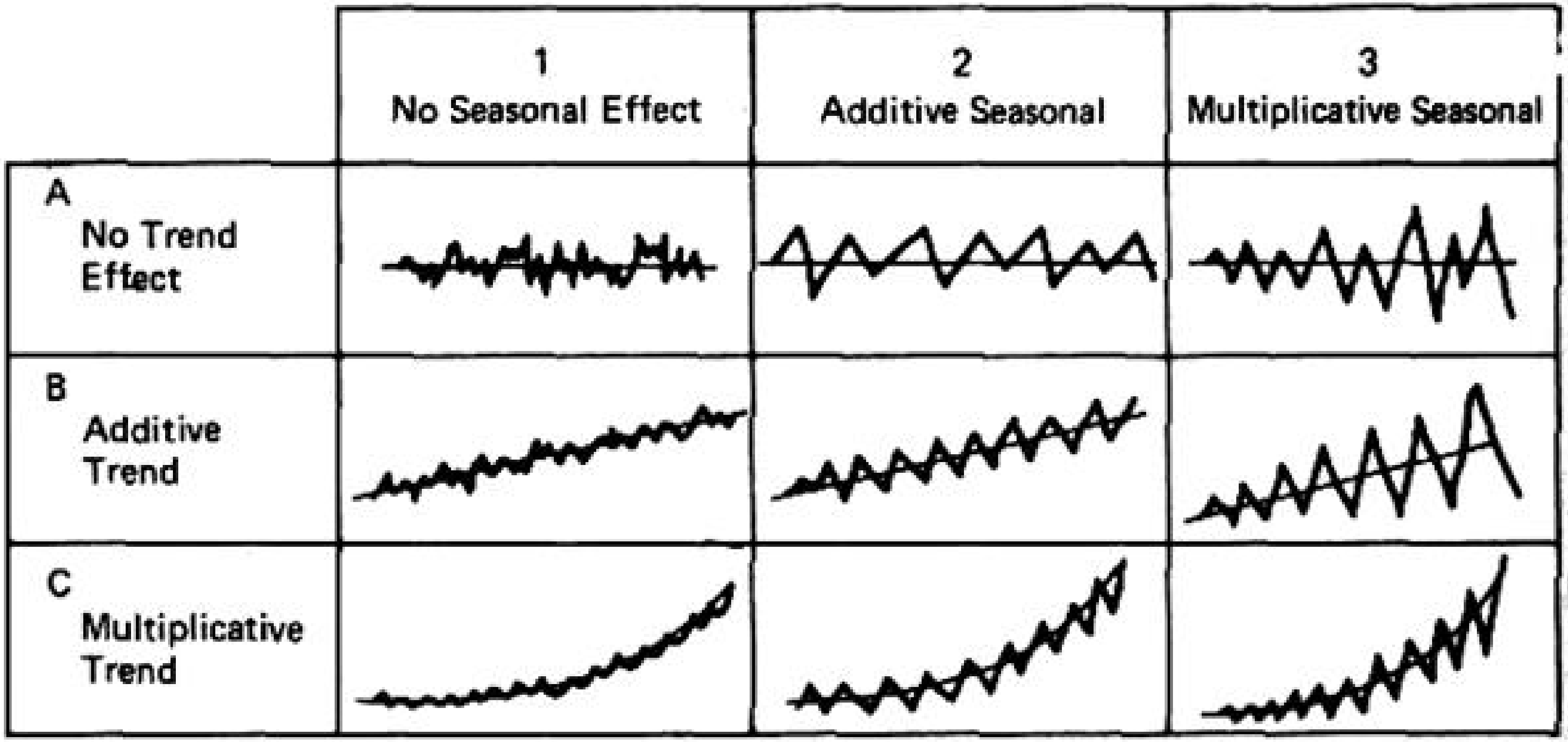
- линейная - $T_t = \beta_0 + \beta_1 t$
- полиномиальная - $T_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_k t^k$
- логарифмическая - $T_t = \exp(\beta_0 + \beta_1 t)$

Представление о характере тренда можно получить из графика ВР.

Выделение тренда



Выделение тренда



Выделение циклической составляющей

Чтобы выделить циклическую компоненту, работают с ВР, не содержащими сезонности (например, годовыми ВР).

Из мультипликативной модели X=T*C*S*e

C_t выделяется следующим образом:

- Удаляют тренд из модели $(C \cdot \varepsilon)_t = \dfrac{X_t}{T_t}$
- Вычисляют процентную долю, приходящуюся на $(C \cdot \varepsilon)_t$: $(C \cdot \varepsilon)_t$*100%.
- Вычисляют скользящее среднее для процентной доли: MA[$(C \cdot \varepsilon)_t$*100%]

  эту величину и считают циклической компонентой.
- Для построения прогноза бывает необходимо выделять и исследовать случайную компоненту :

[$(C \cdot \varepsilon)_t$ *100%] / MA[$(C \cdot \varepsilon)_t$ *100%].

По полученному графику можно определить точки поворота цикла.

Выделение циклической составляющей

Для выделения сезонной компоненты поступают следующим образом:

- сглаживают ВР, т.е. вычисляют

$$\text{MA}[X_t] = (T \cdot C)_t$$

- вычисляют компоненту

$$(S \cdot \varepsilon)_t = \frac{X_t}{(T \cdot C)_t} = \frac{X_t}{MA(X_t)}$$

- изолируют сезонную компоненту усреднением $(S \cdot \varepsilon)_t$

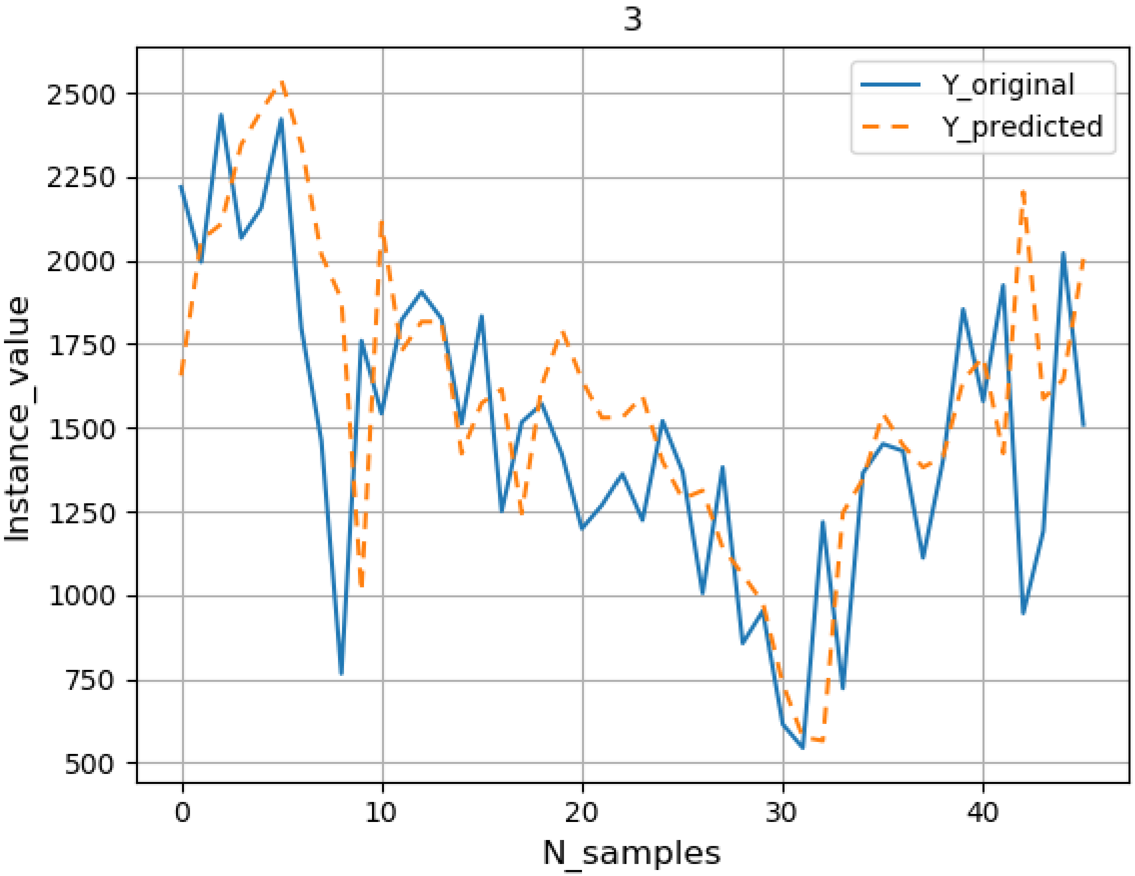
Задача прогнозирования

Задача прогнозирования состоит в том, чтобы по имеющимся наблюдениям ВР предсказать неизвестные будущие значения.

Прогнозирование в бизнесе играет очень большую роль, поскольку оно является рациональной основой для принятия большую роль, поскольку оно является рациональной основой для принятия решений.

Например, предсказание ежемесячных объемов продаж товара – это основа политики контролирования запасов, предсказание будущих доходов корпорации – основа для принятия решений в инвестиционной политике.

Задача прогнозирования



Задача прогнозирования

Библиотеки Python:

- Pandas
- NumPy
- Matplotlib
- Statsmodels
- Scikit-learn

Ссылки:

https://habr.com/ru/company/ods/blog/327242/

http://distrland.blogspot.com/2019/09/7-python.html

https://www.machinelearningmastery.ru/time-series-forecast-case-study-python-monthly-armed-robberies-boston/

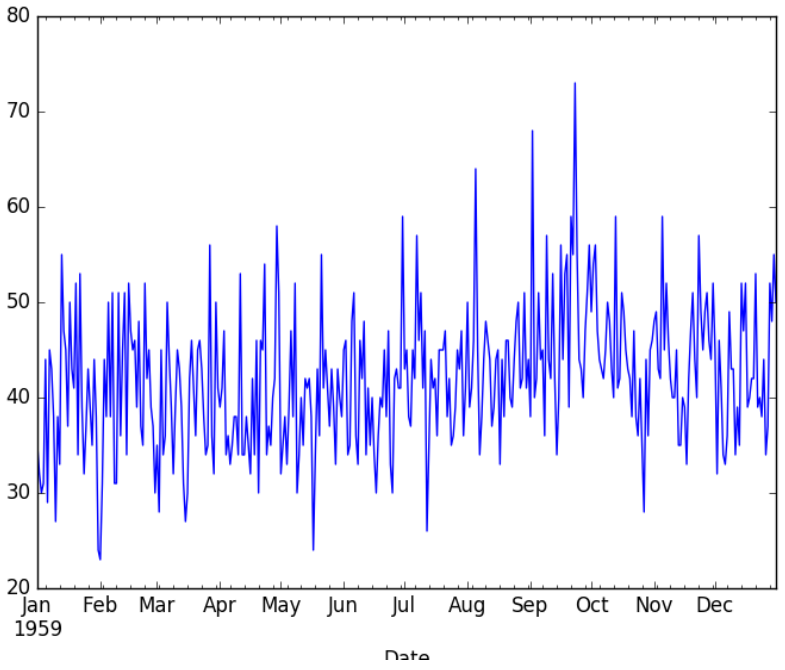
https://coderlessons.com/tutorials/python-technologies/izuchit-iskusstvennyi-intellekt-s-python/ai-s-python-analiz-dannykh-vremennykh-riadov

https://habr.com/ru/post/207160/

Понятие стационарности ряда

Ряд является стационарным, если он совершает колебания вокруг своего математического ожидания.

Сами значения ряда не являются, как правило, независимыми, но корреляция между членами ряда зависит только от расстояния S между ними.

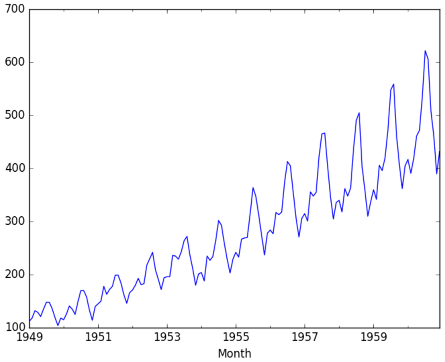


Нестационарный временной ряд

Наблюдения из нестационарного временного ряда показывают сезонные эффекты, тренды и другие структуры, которые зависят от временного индекса.

Сводные статистические данные, такие как среднее значение и дисперсия, меняются с течением времени, предоставляя дрейф в концепциях, которые модель может попытаться отразить.

Классические методы анализа и прогнозирования временных рядов направлены на то, чтобы сделать данные нестационарных временных рядов стационарными путем выявления и устранения тенденций и устранения сезонных эффектов.

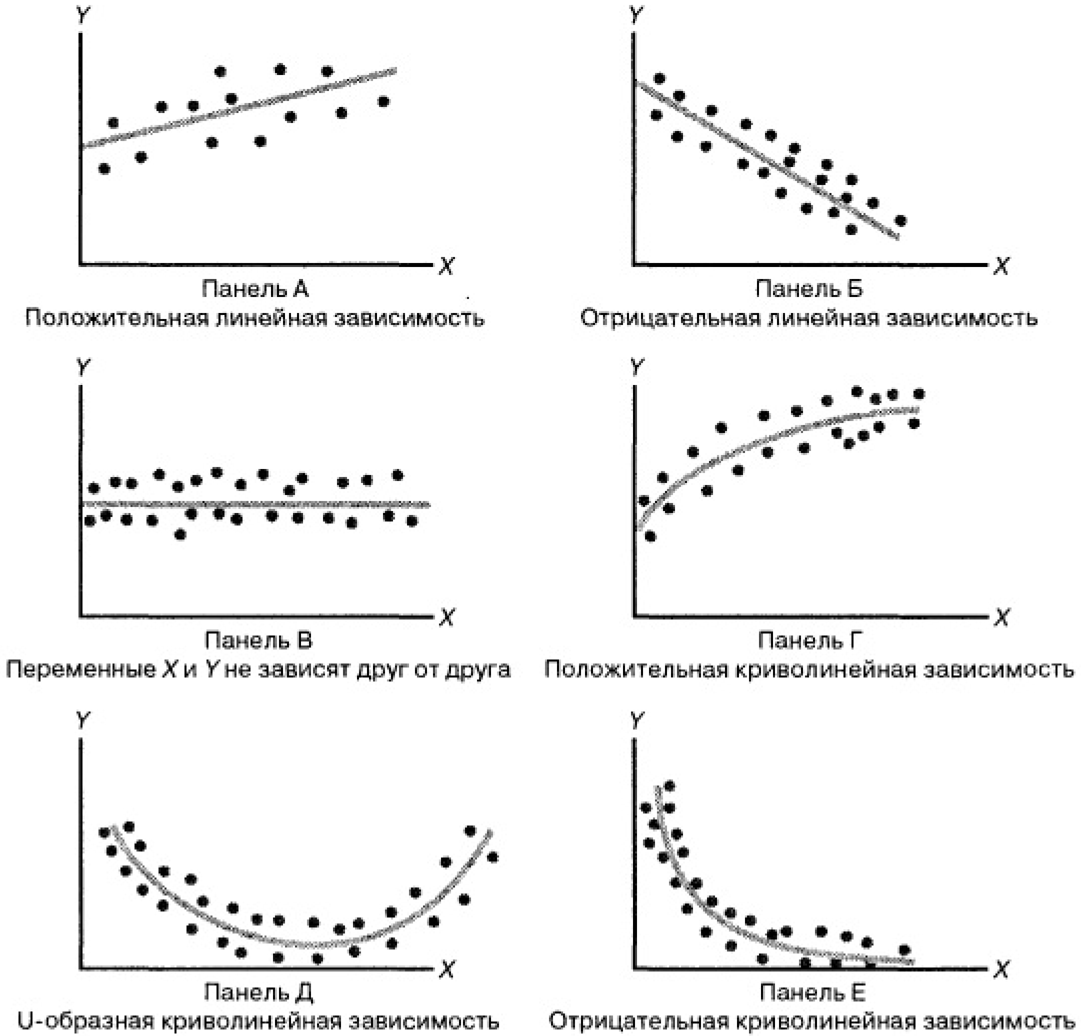


Регрессионный анализ

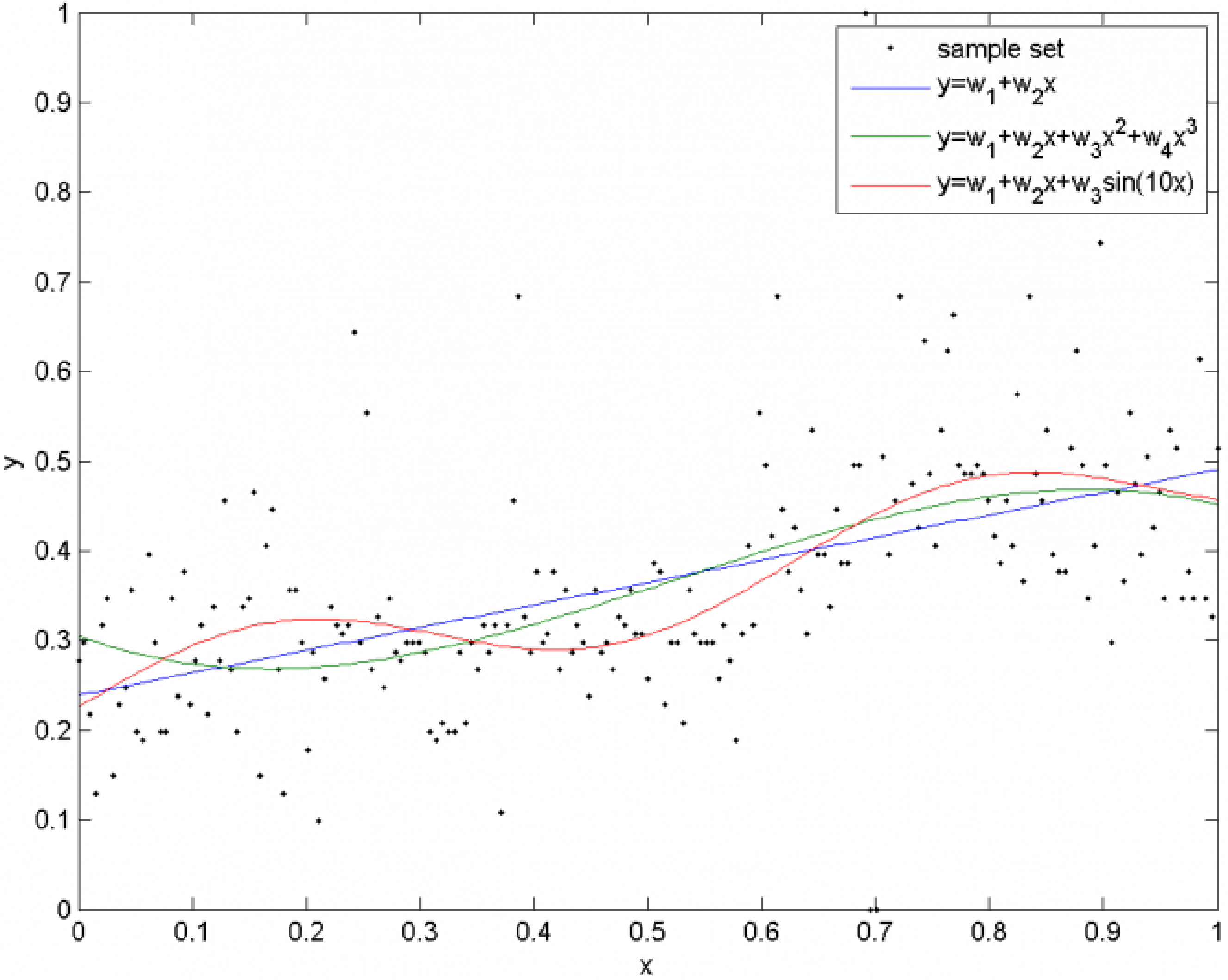
Регрессионный анализ — статистический метод исследования влияния одной или нескольких независимых переменных $X_1,X_2,\ldots,X_p$ на зависимую переменную Y. Уравнение линейной регрессии $Y_x = a+b*X$, где a и b оцененные коэффициенты регрессии.

Регрессия — функция, позволяющая по средней величине одного признака определить среднюю величину другого признака, корреляционно связанного с первым.

Регрессионный анализ



Регрессионный анализ



Проблема мультиколлинеарности

Мультиколлинеарность — тесная корреляционная взаимосвязь между отбираемыми для анализа факторами, совместно воздействующими на общий результат, которая затрудняет оценивание регрессионных параметров. Среди последствий выделить следующие:

- увеличение дисперсий оценок параметров
- уменьшение значений t-статистик для параметров, что приводит к неправильному
- выводу об их статистической значимости
- получение неустойчивых оценок параметров модели и их дисперсий
- возможность получения неверного с точки зрения теории знака у оценки параметра

Дисперсия в регрессионной модели

Общее отклонение есть сумма объяснимой и необъяснимой вариации:

$$\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$$

Общая вариация	Объяснимая вариация	Необъяснимая вариация
Общая сумма квадратов	**Объяснимая моделью вариация**	**Сумма квадратов остатков**
TSS Total Sum of Squares	**ESS** Explained Sum of Squares	**RSS** Residual Sum of Squares

Виды регрессионных моделей

Множественная линейная регрессия

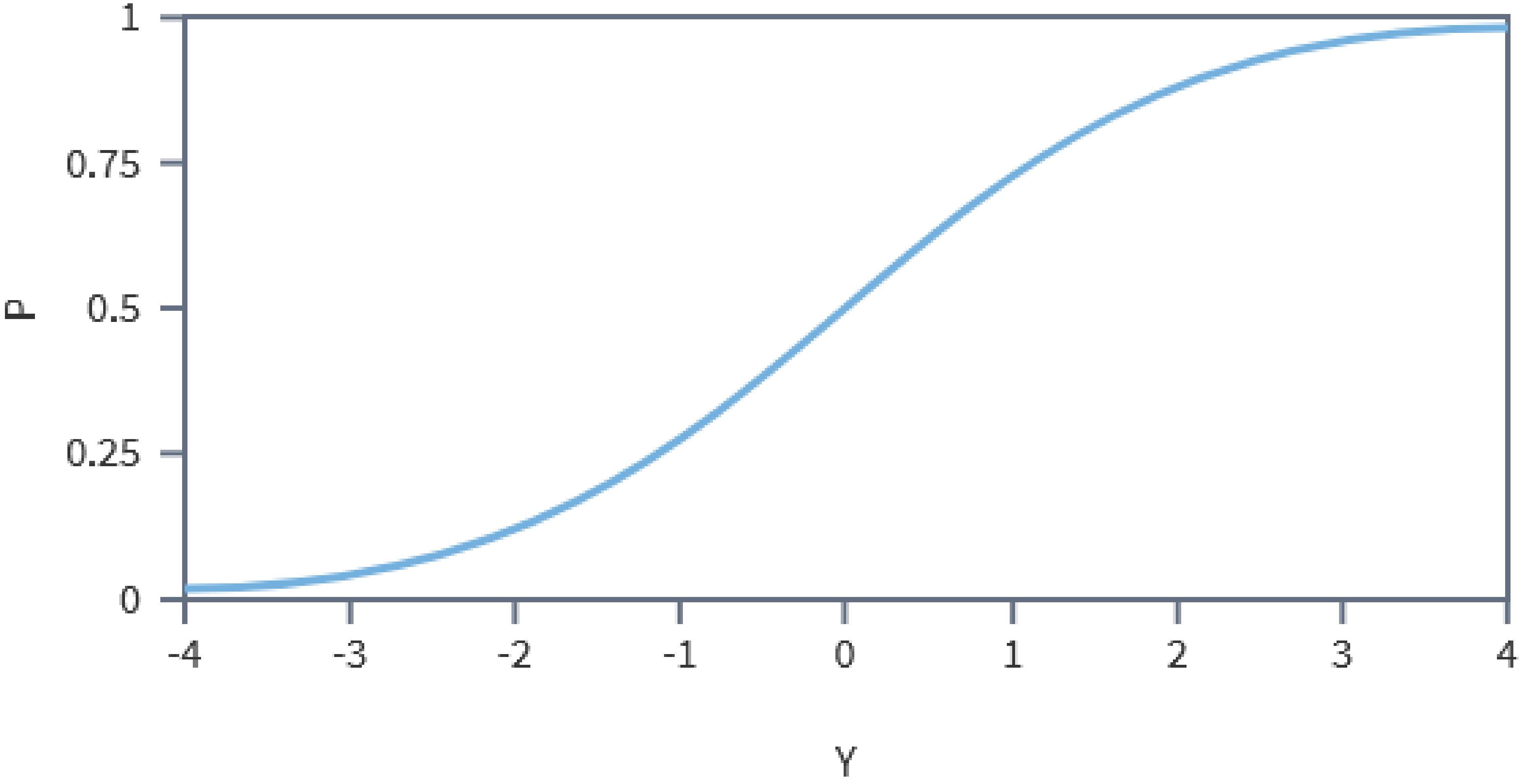
Существует прогнозируемая переменная Y (зависимая переменная) и отобранный заранее комплект переменных, от которых она зависит - X_1, X_2, ..., X_N (независимые переменные). Природа независимых переменных может быть различной. Например, если предположить, что Y - уровень спроса на некоторый продукт в следующем месяце, то независимыми переменными могут быть уровень спроса на этот же продукт в прошлый и позапрошлый месяцы, затраты на рекламу, уровень платежеспособности населения, экономическая обстановка, деятельность конкурентов и многое другое. Модель множественной регрессии в общем случае описывается выражением: $Y = \mathcal{F}(X_1,\ X_2,\ \ldots,\ X_N) + \varepsilon$

В более простом варианте линейной регрессионной модели зависимость зависимой переменной от независимых имеет вид: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N + \varepsilon$

Логистическая регрессия

Логистическая регрессия — это разновидность множественной регрессии, общее назначение которой состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной. Бинарная логистическая регрессия применяется в случае, когда зависимая переменная является бинарной (т.е. может принимать только два значения). С помощью логистической регрессии можно оценивать вероятность того, что событие наступит для конкретного испытуемого (больной/здоровый, возврат кредита/дефолт и т.д.).

Логистическая регрессия



Достоинства регрессионных моделей

- простота
- гибкость
- единообразие их анализа и проектирования
- при использовании линейных регрессионных моделей результат прогнозирования может быть получен быстрее, чем при использовании остальных моделей.
- прозрачность моделирования, т. е. доступность для анализа всех промежуточных вычислений.

Недостатки регрессионных моделей

- сложность определения вида функциональной зависимости
- трудоемкость определение параметров модели
- низкая адаптивность
- отсутствие способности моделирования нелинейных процессов

Авторегрессионные модели

Авторегрессия - это модель временного ряда, в которой наблюдения из предыдущих временных шагов используются в качестве входных данных для уравнения регрессии для прогнозирования значения на следующем временном шаге.

Модель авторегрессии предполагает, что наблюдения на предыдущих временных шагах полезны для прогнозирования значения на следующем временном шаге.

Авторегрессионные модели. Модель ARIMA

В середине 90-х годов прошлого века был разработан принципиально новый и достаточно мощный класс алгоритмов для прогнозирования временных рядов. Большую часть работы по исследованию методологии и проверке моделей была проведена двумя статистиками, Г.Е.П. Боксом (G.E.P. Box) и Г.М. Дженкинсом (G.M. Jenkins). С тех пор построение подобных моделей и получение на их основе прогнозов иногда называться методами Бокса-Дженкинса.

В классическом варианте интегрированной модели авторегрессии — скользящего среднего или ARIMA (autoregressive integrated moving average) не используются независимые переменные. Модели опираются только на информацию, содержащуюся в предыстории прогнозируемых рядов, что ограничивает возможности алгоритма. В настоящее время в научной литературе часто упоминаются варианты моделей ARIMA, позволяющие учитывать независимые переменные.

В методологии ARIMA не предполагается какой-либо четкой модели для прогнозирования данной временной серии. Задается лишь общий класс моделей, описывающих временной ряд и позволяющих как-то выражать текущее значение переменной через ее предыдущие значения. Затем алгоритм, подстраивая внутренние параметры, сам выбирает наиболее подходящую модель прогнозирования.

Авторегрессионные модели. Модель ARIMA

- Предположение о том, что существует связь между соседними значениями ВР и составляет основу методов ARMA.
- Именно эта гипотеза позволяет предсказать значения $\varepsilon_{t+1}, \varepsilon_{t+2}, \ldots$ на основании известных значений $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_t$
- затем на основании регрессионной модели $X_t = \beta_0 + \beta_1 t + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \varepsilon_t$
- строятся будущие значения $X_{t+1}, X_{t+2}, \ldots$

Модель ARIMA. Иерархия

AR(p)+MA(q) -> ARMA(p,q)-> ARMA(p,q)(P,Q) -> ARIMA(p,q,r)(P,Q,R)->...

AR(p) - авторегрессионая модель порядка p.

Модель имеет вид:

Y(t)=f_0+f_1*Y(t-1)+f_2*Y(t-2)+...+f_p*Y(t-p)+E(t), где

Y(t) - зависимая переменная в момент времени t. f_0, f_1, f_2, ..., f_p - оцениваемые параметры. E(t) - ошибка от влияния переменных, которые не учитываются в данной модели. Задача заключается в том, чтобы определить f_0, f_1, f_2, ..., f_p. Их можно оценить различными способами. Правильнее всего искать их через систему уравнений Юла-Уолкера, для составления этой системы потребуется расчет значений автокорреляционной функции. Можно поступить более простым способом - посчитать их методом наименьших квадратов.

MA(q) -модель со скользящим средним порядка q.

Модель имеет вид:Y(t)=m+e(t)-w_1*e(t-1)-w_2*e(t-2)-...-w_p*e(t-p), где Y(t)-зависимая переменная в момент времени t. w_0, w_1, w_2, ..., w_p - оцениваемые параметры.

Модель ARIMA. Этапы

1. Тестирование на стационарность (тест Unit Root). Если результат положительный, то пункт 3.
2. Приведение к стационарному виду взятием 1-ой или 2-ой разности и снова пункт 1.
3. Идентификация параметров p и q процесса ARMA(p,q) по коррелограммам AC и PAC.
4. Оценивание параметров методом максимального правдоподобия и выбор наилучшей модели (**критерии Акаике, Шварца**).
5. Диагностическая проверка (анализ коррелограмм AC и PAC).
6. Прогнозирование.

Модель ARIMA

Частной автокорреляционной функцией называют серию частных коэффициентов автокорреляции r, измеряющих связь между текущим лагом временного ряда Y_t и предыдущими лагами временного ряда Y_{t-1}, Y_{t-2}….,Y_{t-k-1} с устранением влияния других промежуточных временных лагов. Вполне естественно, что при нулевом лаге коэффициент частной корреляции $\rho_0 = 1$, а при лаге $k = 1$, $\rho_1 = r_1$, т. е. коэффициент частной корреляции равен коэффициенту автокорреляции.

Корреляционную зависимость между последовательными уровнями временного ряда называют автокорреляцией уровней ряда. Так, коэффициент автокорреляции уровней первого порядка измеряет корреляционную зависимость между динамикой курса доллара временного ряда t и динамикой курса доллара временного ряда t-1.

Модель ARIMA

Если мы хотим построить модель авторегрессионного процесса AR(p), для определения оптимального числа p мы должны использовать частную автокорреляционную функцию. При этом следует исходить из следующего критерия: оптимальное число p в уравнении авторегрессии должно быть меньше лага, в котором частная автокорреляционная функция начинает стремиться к нулю.

В свою очередь при идентификации модели ARMA(p,q) в качестве лага p выбирается лаг, после которого начинает убывать частная автокорреляционная функция, а в качестве лага q — лаг, после которого начинает убывать автокорреляционная функция.

Авторегрессионные модели. Модель ARCH

Из эмпирических наблюдений над поведением ВР процентных ставок, валютных курсов и т. п. было замечено, что наблюдения с большими отклонениями от среднего и с малыми отклонениями склонны к образованию кластеров. В 1982 г. Р. Энгль предложил модель, которая определяет зависимость дисперсии от других величин. Данная модель получила название — **Авторегрессионная условно-гетероскедастическая модель (Autoregressive Conditional Heteroscedastic model (ARCH))** модель, в которой используется условная, зависимая от времени дисперсия, которая может быть выражена через квадрат значений показателей прошлых периодов, где a – коэффициент задержки (лага) или базовая волатильность.

$$\sigma^2(t) = a + \sum_{i=1}^{q} b_i r_{t-1}^2$$

Авторегрессионные модели. Модель GARCH

В 1986 г. Т. Боллерслев предложил **GARCH-модель (Generalized Autoregressive Conditional Heteroscedastic model)** – обобщенную авторегрессионную модель гетероскедастичности, которая предполагает, что на текущую изменчивость дисперсии влияют как предыдущие изменения показателей, так и предыдущие оценки дисперсии (т.н. «старые новости»). Согласно данной модели (GARCH(p,q)) расчет дисперсии производится по следующей формуле: $\sigma^2(t) = a + \sum_{i=1}^{q} b_i r_{t-1}^2 + \sum_{i=1}^{p} c_i \sigma_{t-1}^2$

Авторегрессионные модели. Преимущества GARCH

- Метод позволяет оценивать регрессии

$$y_t = x_t' \beta + u_t$$

- с не гауссовскими (не нормальными) распределениями ошибок при наличии тяжелых хвостов,
- успешно справляется с сериальной корреляцией квадратов ошибок
- несложно приспосабливаются для моделирования финансовых ВР

Экспоненциальное сглаживание. Модель Хольта. Модель Брауна

В середине прошлого века Хольт предложил усовершенствованный метод экспоненциального сглаживания, впоследствии названный его именем. В предложенном алгоритме значения уровня и тренда сглаживаются с помощью экспоненциального сглаживания. Причем параметры сглаживания у них различны.

$$\begin{cases} \Omega_t = \alpha Y_t + (1-\alpha)(\Omega_{t-1} - T_{t-1}), \\ T_t = \beta(\Omega_t - \Omega_{t-1}) + (1-\beta)T_{t-1}, \\ \hat{Y}_{t+p} = \Omega_t + pT_t \end{cases}$$

Здесь первое уравнение описывает сглаженный ряд общего уровня. Второе уравнение служит для оценки тренда. Третье уравнение определяет прогноз на p отсчетов по времени вперед.

Постоянные сглаживания в методе Хольта идеологически играют ту же роль, что и постоянная в простом экспоненциальном сглаживании. Подбираются они, например, путем перебора по этим параметрам с каким-то шагом. Можно использовать и менее сложные в смысле количества вычислений алгоритмы. Главное, что всегда можно подобрать такую пару параметров, которая дает большую точность модели на тестовом наборе и затем использовать эту пару параметров при реальном прогнозировании. Частным случаем метода Хольта является метод Брауна, когда a=ß.

Экспоненциальное сглаживание. Модель Хольта-Винтерса

Хотя описанный выше метод Хольта (метод двухпараметрического экспоненциального сглаживания) и не является совсем простым (относительно «наивных» моделей и моделей, основанных на усреднении), он не позволяет учитывать сезонные колебания при прогнозировании. Говоря более аккуратно, этот метод не может их «видеть» в предыстории. Существует расширение метода Хольта до трехпараметрического экспоненциального сглаживания. Этот алгоритм называется методом Винтерса. При этом делается попытка учесть сезонные составляющие в данных. Дробь в первом уравнении служит для исключения сезонности из Y(t). После исключения сезонности алгоритм работает с "чистыми" данными, в которых нет сезонных колебаний. Появляются они уже в самом финальном прогнозе, когда "чистый" прогноз, посчитанный почти по методу Хольта умножается на сезонный коэффициент.

$$\begin{cases} \Omega_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(\Omega_{t-1} - T_{t-1}), \\ T_t = \beta(\Omega_t - \Omega_{t-1}) + (1-\beta)T_{t-1}, \\ S_t = \gamma \frac{Y_t}{\Omega_t} + (1-\gamma)S_{t-s}, \\ \hat{Y}_{t+p} = (\Omega_t + pT_t)S_{t-s+p} \end{cases}$$

Авторегрессионные модели и методы. Достоинства

- простота и прозрачность моделирования
- единообразие анализа и проектирования, заложенное в работе
- большое количество примеров применения в открытом доступе для решения задач прогнозирования временных рядов различных предметных областей.

Авторегрессионные модели и методы. Недостатки

- большое число параметров модели, идентификация которых неоднозначна и ресурсоемка
- низкая адаптивность моделей, а также линейность и, как следствие, отсутствие способности моделирования нелинейных процессов, часто встречающихся на практике

Дополнительные ссылки

https://proglib.io/p/linear-regression

https://coderlessons.com/tutorials/python-technologies/uznaite-mashinnoe-obuchenie-s-python/regressionnye-algoritmy-obzor

https://www.machinelearningmastery.ru/simple-and-multiple-linear-regression-in-python-c928425168f9/

https://habr.com/ru/post/206306/

https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

Спасибо за внимание!