

ЛЕКЦИЯ 6. МОДЕЛИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (МНК)

Составитель:
д.т.н. Колесникова С.И.
skolesnikova@yandex.ru

ВРЕМЕННЫЕ РЯДЫ

Временной ряд (ВР) – последовательность упорядоченных во времени числовых показателей (вообще говоря, случайных), характеризующих состояние изучаемого явления (процесса) в динамике, например, $y_1, y_2, \dots, y_k, \dots$.

Общий вид ВР – случайный процесс (с дискретным временем):

$$Y(k\Delta) = f(k\Delta) + \xi(k\Delta), k = 0, 1, \dots; \Delta > 0$$

где $y_k = Y(k\Delta)$ – временной ряд; $f = f(k\Delta)$ – тренд ВР; $\xi(k\Delta)$ – случайная составляющая ВР (процесс, шум); параметр k принимает дискретные значения, равные номеру отсчета значения ВР.

Месяц	Спрос (тыс. \$)			
янв.00	4039	#Н/Д	#Н/Д	
фев.00	4057	#Н/Д	#Н/Д	
мар.00	4052	4049	#Н/Д	
апр.00	4094	4068	#Н/Д	
май.00	4104	4083	19,39	
июн.00	4110	4103	19,78	
июл.00	4154	4123	22,08	
авг.00	4161	4142	21,67	
сен.00	4186	4167	23,92	
окт.00	4195	4181	17,7	
ноя.00	4229	4203	20,21	
дек.00	4244	4223	20,97	
янв.01	4242	4238	19,38	
фев.01	4283	4256	19,83	
мар.01	4322	4282	27,68	
апр.01	4333	4313	29,99	
май.01	4368	4341	30,09	
июн.01	4389	4363	24,5	



МОДЕЛИ И КРИТЕРИИ при МОДЕЛИРОВАНИИ ВР

Задачи, решаемые при моделировании временных рядов

1. Задать множество аналитически заданных моделей (описаний)

$M = \left\{ f_j^{\text{mod}}(k, \theta_j) \right\}_{j=1, \dots, J}, k = \overline{1, n}$ – претендентов на «лучшую» модель; θ_j - вектор параметров j -й модели.

Примеры моделей	Примеры критериев
$f_1^{\text{mod}}(k, \theta_1) = a \cdot e^{ck\Delta}, t = k\Delta; \theta_1 = \{a, c\}$ $f_2^{\text{mod}}(k, \theta_2) = a_2(k\Delta)^2 + a_1k\Delta + a_0, \theta_2 = \{a_2, a_1, a_0\}$	$\sum_{k=1}^n \varepsilon^2 = \sum_{k=1}^n (y_k^{\text{data}} - f^{\text{mod}}(k, \theta))^2 \rightarrow \min_{\theta}$ $\max_{k=1, \dots, n} y_k^{\text{data}} - f^{\text{mod}}(k, \theta) \rightarrow \min_{\theta}$ $\sum_{k=1}^n y_k^{\text{data}} - f^{\text{mod}}(k, \theta) \rightarrow \min_{\theta}$

2. Определить критерий (возможно множество критериев) согласно которым будет проводиться подгонка модели f_j^{mod} к реальным данным $\{y_k^{\text{data}}\}, k = \overline{1, n}$.

3. Для каждой модели определить «наилучшие» оценки параметров $\hat{\theta}_j$

4. Оценить близость каждой j -й модели к набору исходных данных $\{y_k^{\text{data}}\}, k = \overline{1, n}$ по одному из показателей эффективности, например, **относительная ошибка аппроксимации и средний квадрат ошибки**.

Полагают условно: точность модели хорошая, если среднее значение относительной погрешности не превышает 5%, удовлетворительная, если среднее значение относительной погрешности не превышает 15%, и неудовлетворительная, если среднее значение относительной погрешности больше 15%.

ПОКАЗАТЕЛИ КАЧЕСТВА МОДЕЛИРОВАНИИ ВР

<p>Абсолютная относительная ошибка (ARE, AbsoluteRelativeError)</p>	$ARE = \left \frac{Y_t - \hat{Y}_t}{Y_t} \right $
<p>Среднее значение абсолютной относительной ошибки</p>	$MARE = \frac{1}{L} \sum_{t=1}^L \left \frac{Y_t - \hat{Y}_t}{Y_t} \right 100 \%$
<p>Средняя квадратичная ошибка (MSE, MeanSquareError)</p>	$MSE = \frac{1}{L} \sum_{t=1}^L (Y_t - \hat{Y}_t)^2$
<p>Отношение мощности полезного сигнала к мощности шума(ОСШ)</p>	$SER = 10 \lg \left(\frac{\sum_{t=1}^L Y_t^2}{\sum_{t=1}^L (Y_t - \hat{Y}_t)^2} \right)$
<p>Среднее квадратичное отклонение (СКО)</p>	$\gamma_{\text{СКО}} = \sqrt{\frac{1}{L} \sum_{t=1}^L (Y_t - \hat{Y}_t)^2}$

ПРИМЕРЫ МНК-ЗАДАЧ

Задача 1. Чему равна прогнозная численность населения России в начале третьего тысячелетия на основе анализа переписи населения за предыдущие годы? (на примере Excel)

Решение.

A	B	C	D	E	F	G
1	Таблица эксперимента					
2	a	b	Год	Численность стат.	Численность теор.	Отклонение
3			60	117,5		
4			70	130,1		
5			80	137,8		
6			90	147,4		
7			91	148,5		
8			92	147,7		
9			93	148,7		
10			94	148,4		
11			95	148,3		
12				Погрешность		
13						
14						

это модуль разности теоретических и фактических значений функции $f(t)$ (применяется также и квадрат разности)

5. Вычислить погрешность. *Погрешность - это максимальное отклонение.*
6. Подберите значения коэффициентов a , b более точно, используя инструмент <Сервис-Поиск решения> для минимизации погрешности.
7. Определить численность населения России в 2000г.
8. Постройте на одной диаграмме совмещенные графики роста численности населения на основе статистических и теоретических данных:
9. Найти реальные данные и сделать вывод об адекватности математической модели $f(t) = a \cdot e^{kt}$ реальному процессу роста народонаселения.

ПРИМЕРЫ МНК-ЗАДАЧ

Задача 2. Несколько человек решили организовать видеокафе на 6 столиков по 4 места за каждым. С каждого посетителя будет взиматься плата за сеанс видеофильма и ужин (всем посетителям будет предлагаться один и тот же набор блюд). Администрация города постановила, что плата за вход не должна превышать 5\$. Требуется определить такую входную плату, при которой будет получена наибольшая выручка.

Решение. Исследуется на адекватность следующая математическая модель данной задачи, полученная после обобщения опыта работы подобных кафе.

- Пусть X – размер входной платы. $P(X)$ - среднее число посетителей видеосалона как функция от X . Требуется найти такое значение X , при котором выручка $X*P(X)$ достигает максимума. Вид исследуемой модели: $P(X)=ax^2-bx+c$.
- Коэффициенты для каждого кафе свои. Коэффициент c находится из соображений: $P(0)=c$ (если в кафе пускают бесплатно, то свободных мест не будет, c равно числу мест в кафе).

	A	B	C	D	E	F	G	H
1	a	b	Входная плата	Кол-во посетителей	Выручка	Кол-во посетителей	Выручка	Отклонение
2			X	Эксперим. $P(X)$	Эксперимент	Теоретич. $P(X)$	Теоретич.	
4			1,5	17,5				
5			2	16				
6			2,5	14				
7			3	12,5				
8			3,5	11				
9			4	9,2				
10			5	7				
11					Погрешность:			

2. Подобрать а, б. Вычислить теоретическое количество посетителей и теоретическую выручку.
3. Вычислить отклонение между экспериментальной и теоретической выручкой и погрешность.

4. Подобрать а, б, минимизировав погрешность.
5. Построить графики экспериментальной и теоретической зависимости количества посетителей от входной платы.
6. Определить, при какой входной плате выручка будет максимальна. (Каково среднее число посетителей сеанса при найденной оптимальной входной плате).

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (МНК, ORDINARY LEAST SQUARES, OLS)

МНК - базовый метод регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным.

При использовании МНК минимизируется следующая функция:

$$Q = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left(y_i - y_i^{\text{model}} \right)^2,$$

где $y_i, y_i^{\text{model}} = f(x_i)$ - фактические данные ВР в зависимости от некоторого аргумента x_i и модельные данные, соответственно; величина e_i есть невязка.

СВОЙСТВА ОЦЕНОК, ПОЛУЧЕННЫХ НА ОСНОВЕ МНК.

- Математическое ожидание случайного отклонения $M e_i = 0 \quad \forall i = \overline{1, n}$.
- Случайные отклонения $e_i, e_j = 0 \quad \forall i \neq j, i, j = \overline{1, n}$ являются независимыми (отсутствие автокорреляции).
- Факторы x_i и случайные ошибки e_i — независимые случайные величины.
- Постоянная (одинаковая) дисперсия случайных ошибок во всех наблюдениях (отсутствие гетероскедастичности называется гомоскедастичностью).
- **МНК-оценки для линейной регрессии:** несмешённые, состоятельные и эффективные (в классе всех линейных несмешённых оценок) (Best Linear Unbiased Estimator, BLUE).

МНК. ПРИМЕР. ПОСТАНОВКА ЗАДАЧИ (Указания к ЛР-4)

Исходные данные: экспериментальные данные о значениях показателей X и Y

X, Y	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
x_i	0	1	2	4	5
y_i	2,1	2,4	2,6	2,8	3,0

Задание: В результате предварительной аппроксимации данных получена зависимость (тренд) – функциональная модель $g(x) = \sqrt[3]{x+1} + 1$.

- 1) Используя метод наименьших квадратов, аппроксимировать эти данные линейной моделью $y=ax+b$ (найти параметры a и b).
- 2) Выяснить, какая из двух моделей лучше (адекватность в смысле метода наименьших квадратов) моделирует экспериментальные данные.
- 3) Сделать чертеж.
- 4) Сделать прогноз в момент $i=6$ по «лучшей» модели.

Суть МНК применительно к поставленной задаче:

найти коэффициенты линейной зависимости, при которых функция двух переменных a и b $F(a,b) = \sum_{i=1}^n (y_i - (ax_i + b))^2 \xrightarrow{a,b} \min$ (принимает наименьшее значение).

Решение задачи сводится к нахождению экстремума функции двух переменных.

МНК. ВЫВОД ФОРМУЛ ДЛЯ НАХОЖДЕНИЯ КОЭФФИЦИЕНТОВ a и b (Указания к ЛР-4)

$$F(a, b) = \sum_{i=1}^n (y_i - (ax_i + b))^2 \xrightarrow[a, b]{} \min$$

Находим частные производные функции $F(a, b)$ по переменным a и b , приравниваем эти производные к нулю:

$$\frac{\partial F}{\partial a} = -2 \sum_{i=1}^n (y_i - (ax_i + b))x_i = 0,$$

$$\frac{\partial F}{\partial b} = -2 \sum_{i=1}^n (y_i - (ax_i + b)) = 0.$$

Получаем соотношения для a и b в виде:

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}, \quad b = \frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n}.$$

ПРОВЕРИТЬ САМОСТОЯТЕЛЬНО!

МНК. ВЫВОД ФОРМУЛ ДЛЯ НАХОЖДЕНИЯ КОЭФФИЦИЕНТОВ a и b (Указания к ЛР-4)

Утверждение 1 (см. математический анализ). В стационарной точке (a,b) функция $F(a,b)$ принимает минимум, если дифференциал второго порядка положительно определен, или матрица квадратичной формы дифференциала второго порядка для функции $F(a,b)$, в точке (a,b) она должна быть положительно определенной (по критерию Сильвестра).
Дифференциал второго порядка имеет вид:

$$d^2F = \frac{\partial^2 F}{\partial a^2} d^2a + 2 \frac{\partial^2 F}{\partial a \partial b} dadb + \frac{\partial^2 F}{\partial b^2} d^2b.$$

Находим соответствующие величины: **ПРОВЕРИТЬ САМОСТОЯТЕЛЬНО!**

$$\frac{\partial^2 F}{\partial a^2} = 2 \sum_{i=1}^n x_i^2, \quad \frac{\partial^2 F}{\partial b^2} = 2n, \quad \frac{\partial^2 F}{\partial a \partial b} = 2 \sum_{i=1}^n x_i.$$

Матрица квадратичной формы имеет вид:

$$D = \begin{bmatrix} \frac{\partial^2 F}{\partial a^2} & \frac{\partial^2 F}{\partial a \partial b} \\ \frac{\partial^2 F}{\partial a \partial b} & \frac{\partial^2 F}{\partial b^2} \end{bmatrix} = \begin{bmatrix} 2 \sum_{i=1}^n x_i^2 & 2 \sum_{i=1}^n x_i \\ 2 \sum_{i=1}^n x_i & 2n \end{bmatrix}.$$

Утверждение 2. Значения элементов не зависят от a и b , главные миноры положительны, следовательно, минимум $F(a,b)$ в точке (a,b) имеет место по критерию Сильвестра.

МНК. ВЫВОД ФОРМУЛ ДЛЯ НАХОЖДЕНИЯ КОЭФФИЦИЕНТОВ a и b (Указания к ЛР-4)

Заполняем таблицу для удобства нахождения коэффициентов a и b .

X, Y	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$\sum_{i=1}^5 ()$
x_i	0	1	2	4	5	$\sum_{i=1}^n x_i = 12$
y_i	2,1	2,4	2,6	2,8	3,0	$\sum_{i=1}^n y_i = 12,9$
$x_i y_i$	0	2,4	5,2	11,2	15	$\sum_{i=1}^n x_i y_i = 33,8$
x_i^2	0	1	4	16	25	$\sum_{i=1}^n x_i^2 = 46$

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \approx 0,165, \quad b = \frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n} \approx 2,184.$$

Далее

1. Записываем итоговое выражение для линейной модели

$$y(x) = ax + b = 0,165x + 2,184.$$

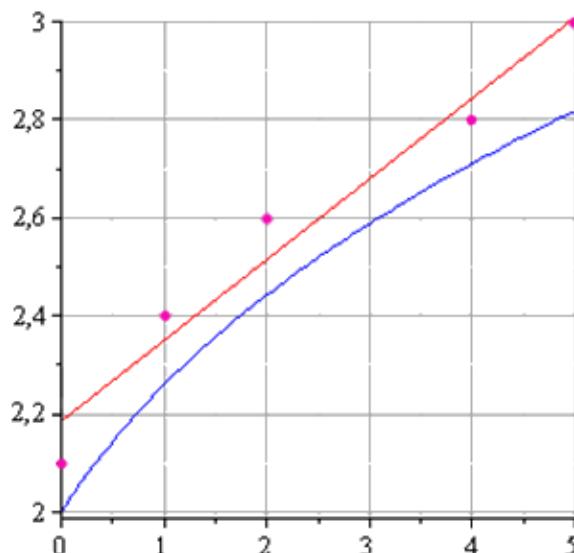
2. Выбираем «лучшую» модель из анализируемых

$$y(x) = ax + b = 0,165x + 2,184 \text{ или } g(x) = \sqrt[3]{x+1} + 1$$

Оцениваем погрешности аппроксимаций:

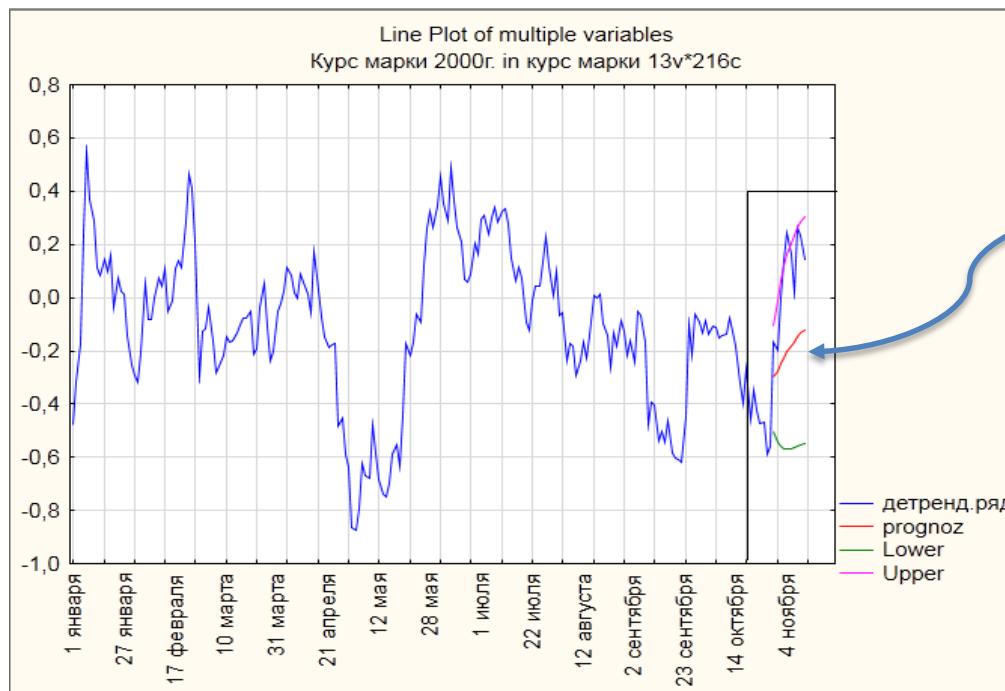
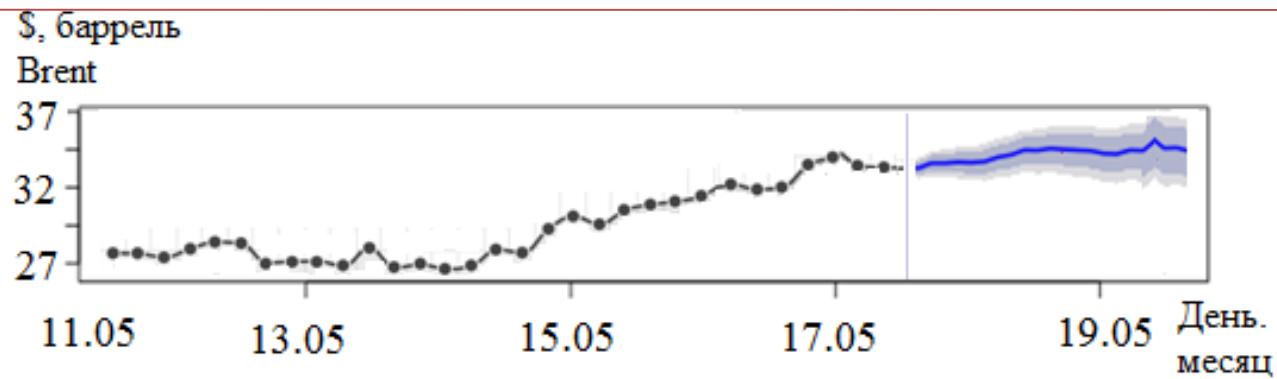
$$\sigma_1 = \sum_{i=1}^n (y_i - (ax_i + b))^2 \approx 0,019, \quad \sigma_2 = \sum_{i=1}^n (y_i - g(x_i))^2 \approx 0,096.$$

Утверждение 3. Прямая $y=0,165x+2,184$ является лучшим приближением исходных данных по сравнению с кривой $g(x) = \sqrt[3]{x+1} + 1$, так как $\sigma_1 < \sigma_2$.



3. *Графическая интерпретация* степени близости кривых к исходным данным:
 кривая $y(x) = 0,165x + 2,184$ (красное начертание),
 $g(x) = \sqrt[3]{x+1} + 1$ (синее начертание),
 исходные данные – набор точек $(x_i, y_i), i = \overline{1, 5}$
 (сиреневый цвет начертания).

ПРИМЕРЫ ВИЗУАЛЬНОГО ОТРАЖЕНИЯ КАЧЕСТВА ПРОГНОЗА



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

Открытый курс машинного обучения. Тема 9. Анализ временных рядов с помощью Python