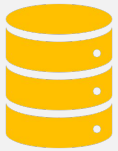
СБОР И ПОДГОТОВКА ДАННЫХ

Лекция 2

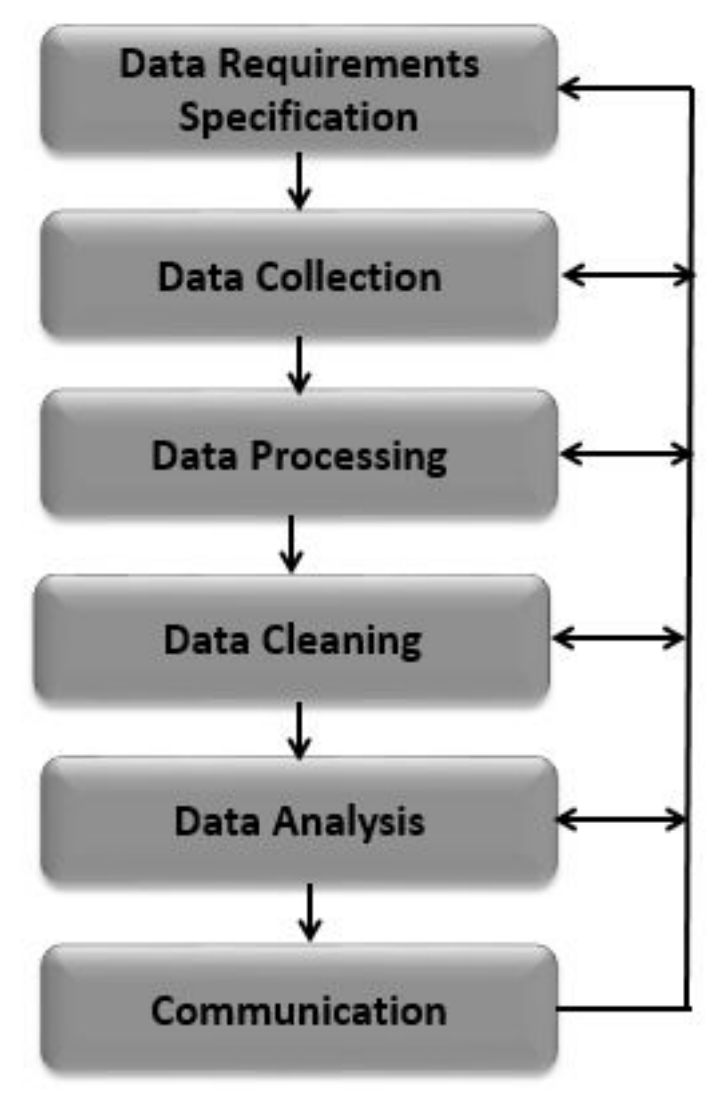
Процесс анализа данных

Сбор данных

Подготовка данных



Обработка данных



Этапы процесса анализа данных , которые носят итеративный характер

- Спецификация требований к данным
- Сбор данных
- Обработка данных
- Очистка данных
- Анализ данных
- Коммуникация

Данные по виду

- Числовые
  - характеризующие состояние какого-либо параметра изучаемого объекта. Наиболее часто такие данные бывают представлены вещественными числами. Примерами числовых данных являются заработная плата, население страны, артериальное давление, температура воздуха
- Категориальные
  - образующие признак принадлежности к какой-либо группе. Примерами категориальных данных являются экзаменационная оценка, цвет автомобиля, уровень образования человека.

Пример

Age	Job	Marital	Education	Balance	Housing
58	management	married	tertiary	2143	yes
44	technician	single	secondary	29	yes
33	entrepreneur	married	secondary	2	yes
47	blue-collar	married	unknown	1506	yes
33	unknown	single	unknown	1	no
35	management	married	tertiary	231	yes
28	management	single	tertiary	447	yes
42	entrepreneur	divorced	tertiary	2	yes
58	retired	married	primary	121	yes
43	technician	single	secondary	593	yes

- В примере поля Age и Balance являются числовыми, а поля Job, Marital, Education и Housing – категориальными

Источники данных

В настоящее время в открытом доступе есть большое количество баз данных, содержащих самые разнообразные сведения.

- открытые данные
  - предоставление свободного доступа к отдельным данным может способствовать повышению качества государственного, регионального и муниципального управления. Принцип открытости получил отдельное название – «открытые данные» (Open Data).
- открытые статистические данные

Сбор данных

- процесс формирования структурированного набора данных в цифровой форме. В некоторых случаях процесс сбора данных может включать также этап оцифровки.

Как правило, оцифрованные данные бывают представлены в виде:

- электронных таблиц в форматах XLS либо ODS;
- текстовых файлов в формате CSV;
- веб-страниц в формате HTML;
- файлов в формате XML;
- базы данных с доступом по технологии JSON либо через специализированный интерфейс (API).

Автоматизированный сбор данных

Особенности набора данных

- Для использования в системах анализа данные должны быть представлены в определенном, как правило, табличном виде.
- Однако зачастую наборы данных имеют следующие особенности:
  - отличную от табличной форму представления;
  - пропуски отдельных данных;
  - некорректные значения;
  - большие числовые значения;
  - текстовые данные.

Подготовка данных

- Для устранения отмеченных несоответствий могут быть применены следующие операции:
- структурирование – приведение данных к табличному (матричному) виду;
- отбор – исключение записей с отсутствующими или некорректными значениями;
- нормализация – приведение числовых значений к определенному диапазону, например к диапазону 0...1;
- кодирование – это представление категориальных данных в числовой форме.
  - Например, при бинарной классификации один из классов можно представить числом «0», а другой класс – числом «1». При множественной классификации система кодирования несколько усложняется: создается несколько числовых полей по количеству классов в выборке данных, каждый класс кодируется проставлением числа «1» в соответствующем поле.

Пример. Анкетные данные клиентов банка

№	Age	Marital	Balance	Housing
1	47	married	1506	yes
2	33	single	1	no
3	35	married	high	yes
4	28	single	447	yes
5	42	divorced	2	yes
6	58		121	yes
7	43	single	593	yes

- Для приведения этой выборки данных в «правильный» формат необходимо выполнить следующие операции:
- 1) исключить записи No3 и No6 как имеющие отсутствующие или некорректные значения;
- 2) нормализовать числовые значения в столбцах *Age* и *Balance*;
- 3) закодировать категориальные данные в столбцах *Marital* и *Housing*.

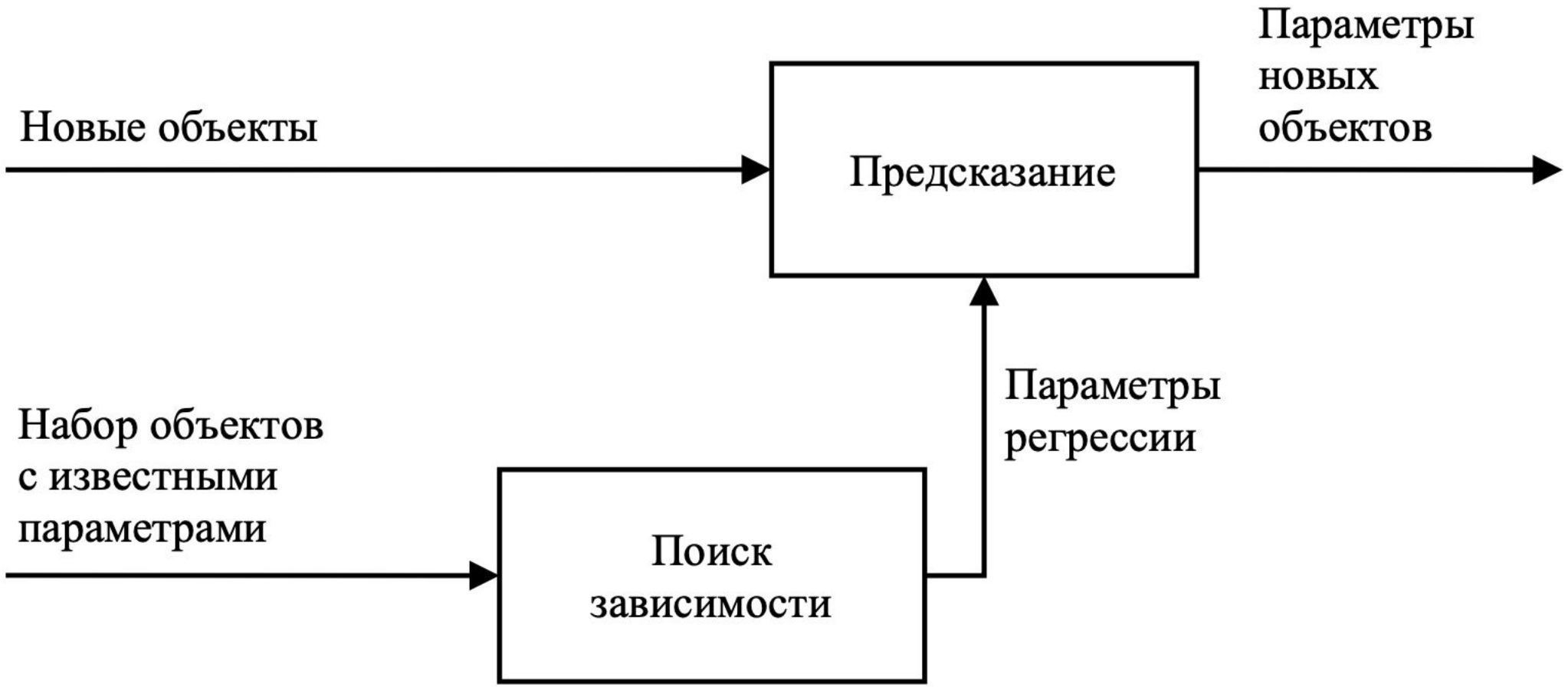
Пример. Обработанная выборка данных

№	Age	Marital1	Marital2	Marital3	Balance	Housing
1	1,000	1	0	0	1,000	1
2	0,263	0	1	0	0,000	0
4	0,000	0	1	0	0,296	1
5	0,737	0	0	1	0,001	1
7	0,789	0	1	0	0,393	1

РЕГРЕССИОННЫЙ АНАЛИЗ

- Предсказание значения зависимой переменной с помощью независимой переменной (независимых переменных) является задачей регрессионного анализа.
- Регрессия относится к типу задач обучения с учителем (Supervised Learning в терминах Machine Learning). Предполагается, что имеется некоторая выборка данных, в которой представлены несколько объектов с известными свойствами.
- Решение задачи предсказания включает два этапа:
  - поиск характера зависимости
  - предсказание

Схема применения регрессии



линейная функция гипотезы

$$h(x) = \theta_0 \cdot x_0 + \theta_1 \cdot x_1 + \ldots + \theta_m \cdot x_m = \sum_{j=0}^{m} \theta_j \cdot x_j \,. \quad (1)$$

- С учетом того, что наборы значений θ и *x* по сути являются векторами, выражение (1) для удобства записывают в виде произведения векторов:

$h(x) = x^* \boldsymbol{\theta}$ (2)

Виды регрессии

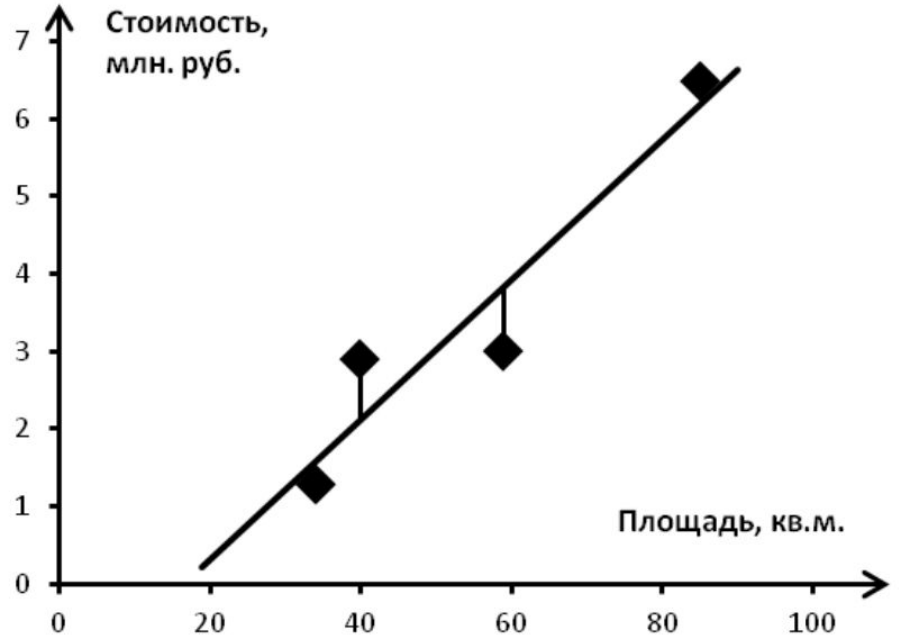
- В зависимости от характера функции гипотезы регрессию подразделяют на линейную и нелинейную.
- В зависимости от числа независимых переменных регрессию подразделяют на парную и множественную.

Примером парной линейной регрессии является задача выявления зависимости стоимости квартир от их площади

Пример регрессии с помощью линейной функции. Характеристики квартир

- Регрессия с помощью линейной функции

Площадь, кв. м	Стоимость, млн. руб.
34	1,3
40	2,9
59	3,0
85	6,5



Функция штрафа

- Подбор параметров регрессионной функции обычно осуществляется по критерию минимума суммы квадратов отклонений:

$$CF = \sum_{i=1}^{n} \left[h(x_i) - y_i \right]^2 \rightarrow \min . \qquad (3)$$

- При этом выражение [*h*(*xi*) − *yi*]2 называется функцией штрафа

(cost function, CF; либо loss function, LF).

Оптимизационная задача

- В формулировке (3) задача нахождения параметров регрессионной функции является оптимизационной.
- Существует два основных подхода к решению задачи регрессии в постановке (1): аналитический и численный.
- Следует отметить, что решения регрессионной задачи, полученные разными методами, могут различаться.

Аналитическое решение

- Известно аналитическое решение задачи линейной регрессии в постановке (1):
- $\boldsymbol{\theta} = (X^T X)^{-1} X^T y$, (4)
- где X – матрица, содержащая значения независимых переменных,
- y – вектор, содержащий значений зависимых переменных.

- Для набора данных о характеристиках квартиры матрица X и вектор y примут вид

$$X = \begin{bmatrix} 1 & 34 \\ 1 & 40 \\ 1 & 59 \\ 1 & 85 \end{bmatrix}, \; y = \begin{bmatrix} 1{,}3 \\ 2{,}9 \\ 3{,}0 \\ 6{,}5 \end{bmatrix}. \qquad (5)$$

- При исходных данных (3) выражение (2) дает результат
- $\theta \approx \begin{bmatrix} -1{,}506 \\ 0{,}090 \end{bmatrix}$

Вычисления в Microsoft Excel

- для умножения матриц используется функция МУМНОЖ, для транспонирования матриц – функция ТРАНСП, а для нахождения обратной матрицы – МОБР

X

1	34
1	40
1	59
1	85

y

1,3
2,9
3,0
6,5

X^T

1	1	1	1
34	40	59	85

$X^T \cdot X$

4	218
218	13462

$(X^T \cdot X)^{-1} \cdot X^T$

0,957	0,750	0,095	-0,801
-0,013	-0,009	0,003	0,019

$(X^T \cdot X)^{-1}$

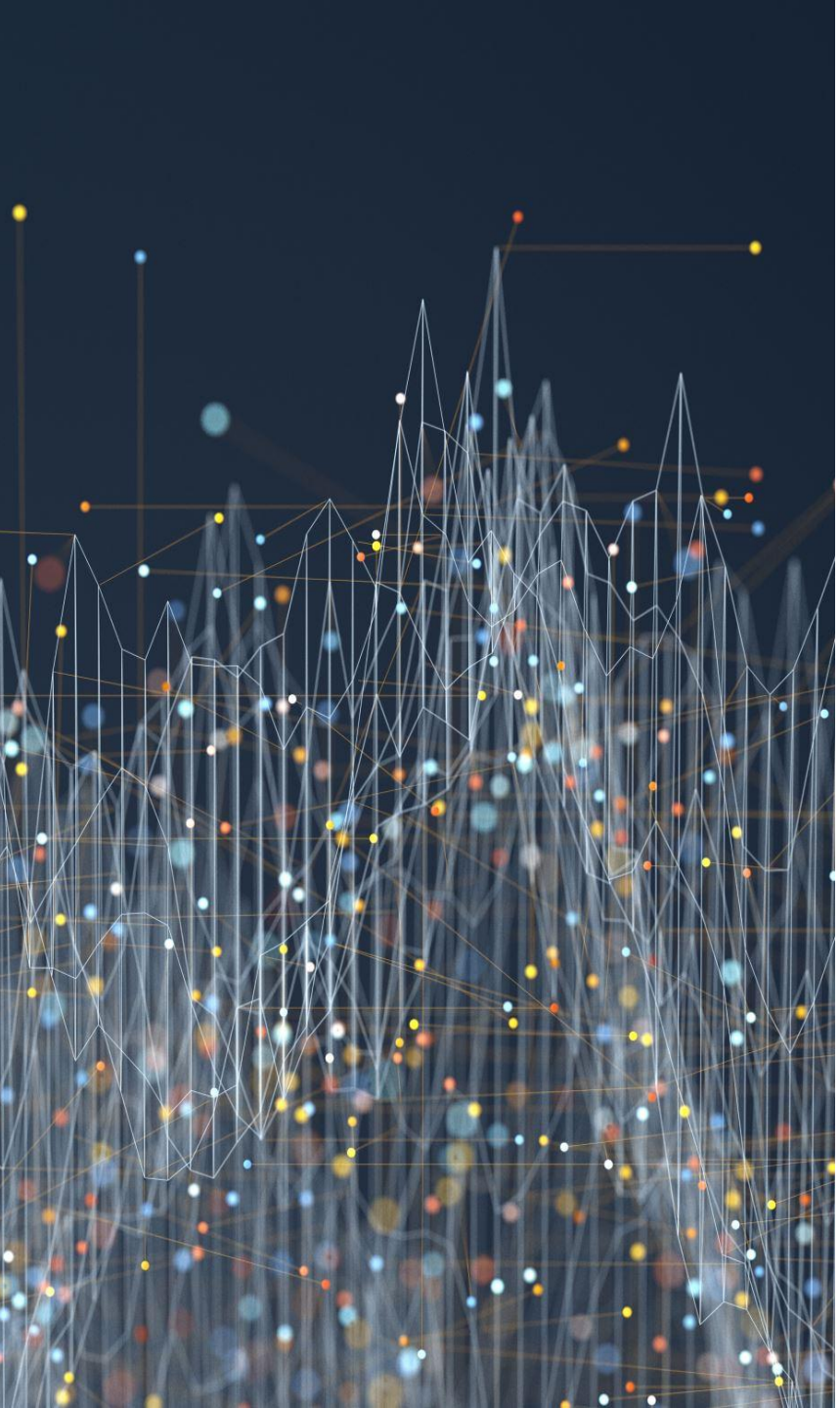
2,128716	-0,0344719
-0,03447	0,0006325

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

-1,506
0,090

Пример в Excel

	A	B	C	D	E	F	G	H	I
1									
2	**Прощадь**	**Стоимость**							
3	34	1,3							
4	40	2,9							
5	59	3							
6	85	6,5							
7									
8		X		y		X trans			
9	1	34		1,3		1	1	1	1
10	1	40		2,9		34	40	59	85
11	1	59		3					
12	1	85		6,5					
13									
14		X trans * X				(X trans * X)^(-1)*X trans			
15	=MMULT(F9:I10;A9:B12)					=MMULT(A19:B20;F9:I10)			
16									
17									
18		(X trans * X)^(-1)				(X trans * X)^(-1)*X trans * y			
19	=MINVERSE(A15:B16)					=MMULT(F15:I16;D9:D12)			



Особенности

- Относительно низкая устойчивость к отдельным сочетаниям данных. Так, дублирование какой-либо строки в наборе данных приведет к сбою в вычислениях при операции нахождения обратной матрицы.
- Большая вычислительная сложность. Относительно большие наборы данных, содержащие порядка тысячи и более строк, будут обрабатываться относительно медленно.
- Чувствительность к большим значениям. Для наборов данных, в отдельных столбцах которых содержатся большие значения, может потребоваться предварительная нормализация.

Численное решение

- Для линейной регрессии задача в формулировке (1) имеет единственное решение, что позволяет без каких-либо оговорок применять численные методы.
- можно использовать
  - метод Ньютона
  - либо метод сопряженных градиентов.

Оба этих метода представлены в инструменте «Поиск решения» ПО *Microsoft Excel*.

Шаги яисленного решения регрессионной задачи

1) подготовку данных;

2) задание функции гипотезы, в том числе начальных значений её параметров;

3) задание целевой функции;

4) решение оптимизационной задачи каким-либо численным методом.

Пример на основе данных о стоимости квартир

Для удобства запишем выражение для функции гипотезы в следующей форме:

$$h(x) = a_0 + a_1 \cdot x \,. \qquad (6)$$

запишем формулировку оптимизационной задачи:

$$CF = \sum_{i=1}^{4} \left[(a_0 + a_1 \cdot x_1) - y_i \right]^2 \to \min \,. \qquad (7)$$

Пример на основе данных о стоимости квартир см 16 слайд

Подготовка к численному решению

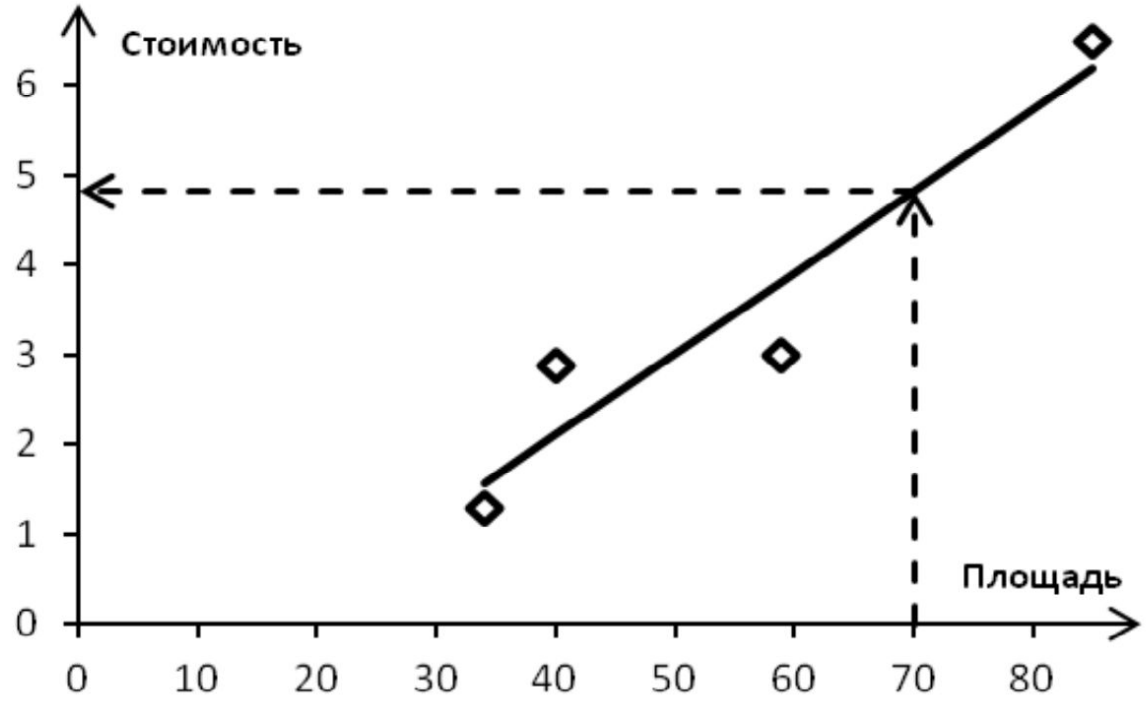
	A	B	C	D	E
1					
2	**X**	**Y**	**h(x)**	**h(x)-y**	**(h(x)-y)^2**
3	34	1,3	4	2,7	7,3
4	40	2,9	4	1,1	1,2
5	59	3	4	1	1,0
6	85	6,5	4	-2,5	6,3
7				∑=	15,8
8					
9					
10	a0=	4			
11	a1=	0			
12					

	A	B	C	D	E
1					
2	**X**	**Y**	**h(x)**	**h(x)-y**	**(h(x)-y)^2**
3	34	1,3	=$B$10+$B$11*A3	=C3-B3	=D3^2
4	40	2,9	=$B$10+$B$11*A4	=C4-B4	=D4^2
5	59	3	=$B$10+$B$11*A5	=C5-B5	=D5^2
6	85	6,5	=$B$10+$B$11*A6	=C6-B6	=D6^2
7				∑=	=SUM(E3:E6)
8					
9					
10	a0=	4			
11	a1=	0			
12					

Зададим функцию гипотезы и начальные значения коэффициентов функции гипотезы, зададим функцию штрафа

Поиск решения

- В настройках инструмента «Поиск решения» (MS Excel) зададим целевую ячейку, содержащую выражение для функции штрафа, и изменяемые ячейки, содержащие значения коэффициентов функции гипотезы $a_0 \approx -1{,}5062$, $a_1 \approx 0{,}0905$.
- График функции гипотезы представляет собой прямую линию
- Прогнозирование стоимости квартиры осуществляется с помощью подстановки площади квартиры и найденных коэффициентов в выражение (6).
- Например, для квартиры площадью 70 кв. м прогнозная стоимость составит $-1{,}5062+0{,}0905 \cdot 70 \approx 4{,}83$ млн. тенге.



Выбор функции гипотезы

- В случае парной регрессии выбор функции гипотезы можно осуществлять визуально по соответствующему графику.
- В случае множественной регрессии этот подход неприменим.
- Предп

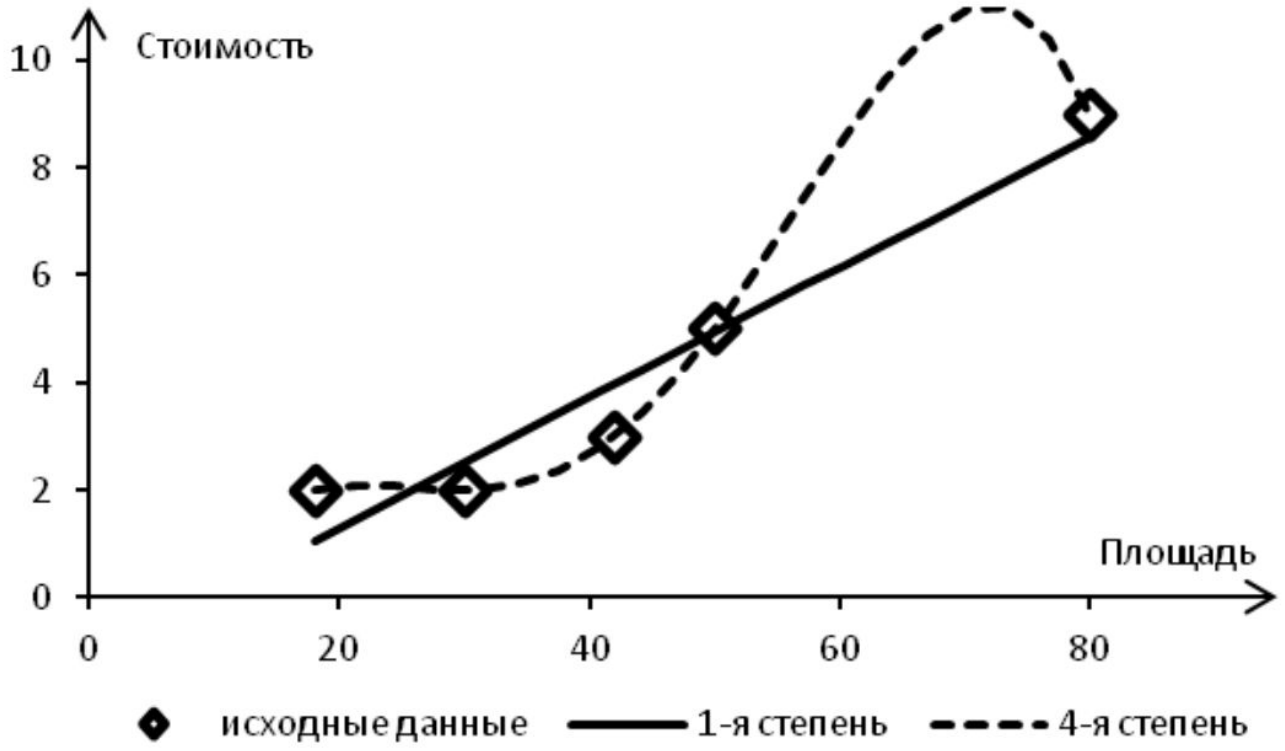
Площадь, кв. м.	Цена, млн.тенге
18	2,0
30	2,0
42	3,0
50	5,0
80	9,0

Решение с применением линейной функции гипотезы и функции гипотезы

Функция гипотезы	R^2	Функция штрафа
Линейная	0,935	2,271
Полином 4-й степени	1,000	0

Регрессия при разных функциях гипотезы

- В терминологии Machine Learning ситуация, иллюстрируемая сплошной линией, соответствующей линейной функции гипотезы, обозначается термином underfitting (недообученность).
- Ситуация, иллюстрируемая пунктирной линией, соответствующей полиномиальной функции регрессии, обозначается термином «переобученность» (overfitting).



Выбор функции регрессии

1 Разделение случайным образом исходной выборки данных на две части: обучающую, содержащую от 70 до 80% исходных данных, и проверочную, содержащую от 20 до 30% исходных данных.

2 Задание нескольких функций гипотезы.

3 Выполнение для каждой из функций гипотезы подбора параметров функции по обучающей выборке (минимизация функции штрафа по обучающей выборке) и вычисления функции штрафа по тестовой выборке.

4 Выбор функции гипотезы по критерию минимальной функции штрафа по тестовой выборке.

Заключение

- понятие регрессионного анализа, парной регрессии, множественной регрессии
- способы решения задачи регрессии.
- особенности решения регрессионной задачи аналитическим методом
- особенности решения регрессионной задачи численными методами
- эффекты недообученности и переобученности
- алгоритм подбора функции регрессии

Источники

- Поручиков, М. А., Анализ данных: учеб. пособие / М.А. Поручиков. – Самара: Изд-во Самарского университета, 2016. – 88 с.
- Data analysis with Excel. https://www.tutorialspoint.com/excel_data_analysis/excel_data_analysis_tutorial.pdf
- Guerrero H. (2019) Modeling and Simulation: Part 1. In: Excel Data Analysis. Springer, Cham. https://doi.org/10.1007/978-3-030-01279-3_7