Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н. Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н. Э. Баумана)

ФАКУЛЬТЕТ «Информатика, искусственный интеллект и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

# РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

## *К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ*

## *НА ТЕМУ:*

## *«Решения задачи регрессии»*

Студент ИУ7И-71Б (Группа) _____ (Подпись, дата) Й. Н. Везирова (И. О. Фамилия)

Руководитель НИР _____ (Подпись, дата) Ю. В. Строганов (И. О. Фамилия)

*2024 г.*

# РЕФЕРАТ

Расчетно-пояснительная записка 35 с., 4 рис., 5 табл., 27 источн., 1 прил.

Ключевые слова: регрессия, линейная регрессия, логистическая регрессия, адаптивная регрессия, анализ данных, прогнозирование, интерпретация, сравнение методов.

Цель работы — провести анализ методов и подходов к решению задачи регрессии.

В данной работе рассматриваются предметная область и существующие подходы к решению задачи регрессии, формулируются критерии сравнения и проводится классификация методов. По результатам сравнения представлены выводы о рассматриваемых решениях задачи регрессии.

# СОДЕРЖАНИЕ

# ВВЕДЕНИЕ

Регрессионный анализ по праву может быть назван основным методом современной математической статистики. Он стал неотъемлемой частью современных методов анализа данных, находя свое отражение в различных подходах, включая методы усреднения, процедуры сглаживания, алгоритмы согласования противоречивых данных и концепции, основанные на принципах оптимальности. Регрессия — это квинтэссенция понятия целесообразности [1].

Решение задачи регрессии является ключевым этапом в анализе данных и активно применяется в самых разнообразных областях: от анализа экономических процессов и прогнозирования рыночных тенденций до моделирования сложных физических и инженерных систем. Такие методы позволяют учитывать нелинейные зависимости, высокую размерность признаков и наличие выбросов, что делает их универсальным инструментом для обработки и интерпретации сложных данных [2].

Целью работы является проведение анализа методов и подходов к решению задачи регрессии.

Для достижения поставленной цели, необходимо решить следующие задачи:

- провести анализ предметной области;
- провести обзор существующих решений задачи регрессии;
- сформулировать критерии сравнения решений задачи регрессии;
- классифицировать существующие решения задачи регрессии.

# 1 Анализ предметной области

## 1.1 Регрессия

**Определение:** Регрессия — это метод прогнозирования и анализа зависимости целевой переменной от одной или нескольких независимых переменных. Этот подход широко применяется в задачах предсказания, моделирования и объяснения зависимости переменных, позволяя строить аналитические модели, описывающие взаимодействия в сложных системах [3; 4].

Модель регрессии предсказывает числовое значение. Например, модель погоды, которая прогнозирует количество дождя в миллиметрах, является регрессионной моделью.

В таблице 1.1 приведены примеры регрессионных моделей.

Таблица 1.1 – Примеры регрессионных моделей

| Сценарий | Входные данные | Выходные данные (числовой прогноз) |
|---|---|---|
| Будущая цена дома | Площадь участка, количество спален и ванных комнат, размер участка, процентная ставка по ипотеке, ставка налога на недвижимость, затраты на строительство и количество домов, выставленных на продажу в этом районе | Цена дома |
| Будущее время поездки | Исторические условия дорожного движения, расстояние до пункта назначения и погодные условия | Время в минутах и секундах до прибытия в пункт назначения |

Существует множество видов регрессии, и в данной работе будут рассмотрены следующие три вида регрессии: линейная регрессия, логистическая регрессия и адаптивная регрессия.

# 2 Обзор существующих решений задачи регрессии

## 2.1 Линейная регрессия

**Определение:** Линейная регрессия — это статистический метод, используемый для поиска взаимосвязи между переменными. В контексте машинного обучения линейная регрессия находит связь между функциями и меткой [5].

Линейная регрессия является одним из наиболее популярных алгоритмов и чаще всего используется для начала решения любой задачи регрессии, так как считается простейшей моделью машинного обучения [6].

В математической статистике линейная регрессия представляет собой метод аппроксимации зависимостей между входными и выходными переменными на основе линейной модели [7].

Если рассматривается зависимость между одной входной и одной выходной переменными, то имеет место простая линейная регрессия. Уравнение регрессии для этого случая имеет вид:

$$y = ax + b, \tag{2.1}$$

где $a$ — коэффициент наклона (или угловой коэффициент) линии регрессии, $b$ — свободный член (перехват с осью $y$).

Коэффициенты $a$ и $b$, называемые также параметрами модели, определяются таким образом, чтобы сумма квадратов отклонений точек, соответствующих реальным наблюдениям данных, от линии регрессии была бы минимальной [7]. Коэффициенты обычно оцениваются методом наименьших квадратов по следующей формуле:

$$S(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2, \tag{2.2}$$

где $y_i$ — наблюдаемое значение зависимой переменной для i-го наблюдения, $x_i$ — значение независимой переменной для i-го наблюдения, $ax_i+b$ — предсказанное значение зависимой переменной.

Чтобы найти коэффициенты $a$ и $b$, минимизируем эту сумму по отношению к $a$ и $b$. Обычно, для этого используются аналитические формулы,

полученные путем дифференцирования суммы квадратов остатков:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \tag{2.3}$$

$$b = \frac{1}{n}\sum_{i=1}^{n} y_i - a\frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.4}$$

где $n$ — количество наблюдений.

Для оценки качества модели линейной регрессии часто используется коэффициент детерминации $R^2$, который показывает долю изменчивости зависимой переменной, объясненную моделью. Он рассчитывается как квадрат коэффициента корреляции $r_{xy}$:

$$R^2 = r_{xy}^2, \tag{2.5}$$

где $r_{xy}$ — коэффициент корреляции между $x$ и $y$. Чем ближе $R^2$ к 1, тем лучше модель объясняет зависимость между переменными.

На рисунке 2.1 представлен пример построения линии регрессии.

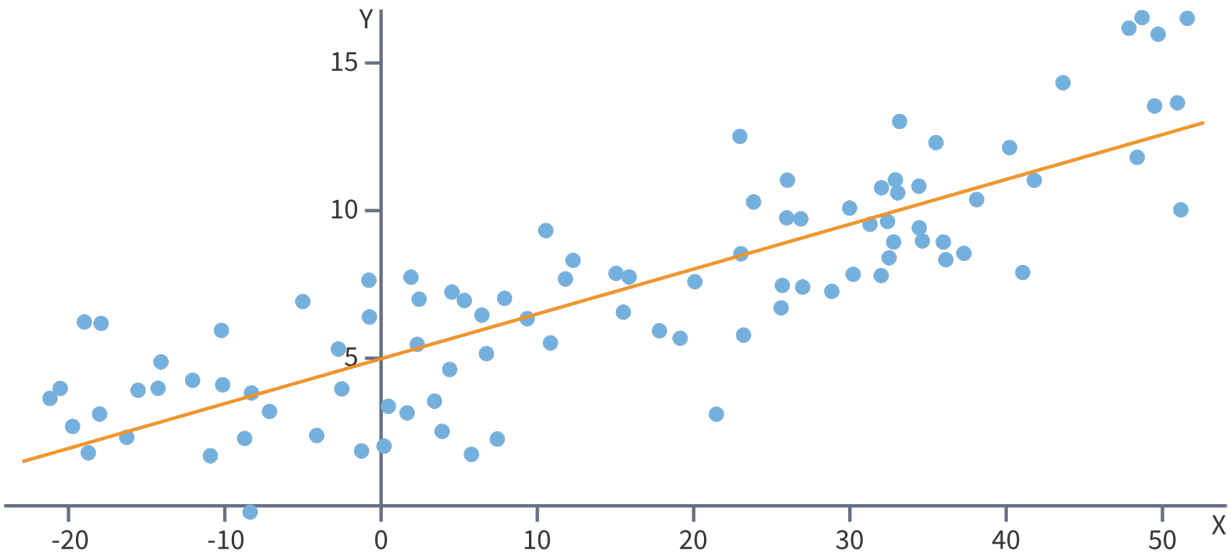



Рисунок 2.1 – Пример построения линии регрессии

Если рассматривается зависимость между несколькими входными и одной выходной переменными, то имеет место множественная линейная ре-

грессия. Соответствующее уравнение имеет вид (2.6):

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n, \tag{2.6}$$

где $n$ — число входных переменных.

В данном случае модель будет описываться не прямой, а гиперплоскостью. Коэффициенты уравнения множественной линейной регрессии подбираются так, чтобы минимизировать сумму квадратов отклонения реальных точек от этой гиперплоскости [7].

## Применение

Линейная регрессия имеет много практических применений, которые можно разделить на две основные категории:

1) прогнозирование — линейную регрессию можно использовать для подгонки модели к наблюдаемому набору данных;

2) объяснение изменчивости — линейный регрессионный анализ применяется для количественной оценки силы взаимосвязи между выходной и входными переменными.

Оценим практическое применение способа построения линейной регрессии в экономике на примере формирования заработной платы, зависящей от показателя среднедушевого прожиточного минимума на человека, на основе представленных способов и формул. В таблице 2.1 представлены данные, на базе которых нужно выявить зависимость показателя заработной платы от фактора $x$ по регионам России за 2016 год:

Таблица 2.1 – Показатели среднемесячной заработной платы под влиянием прожиточного минимума по регионам РФ за 2016 г., тыс. руб.

| № | Среднедушевой прожиточный min на одного работающего (x) | Среднемесячная заработная плата (y) |
|---|---|---|
| 1 | 8.70 | 17.8 |
| 2 | 6.30 | 11.6 |
| 3 | 7.89 | 15.8 |
| 4 | 10.24 | 13.5 |
| 5 | 10.25 | 20.5 |
| 6 | 7.50 | 15.9 |
| 7 | 8.75 | 14.9 |
| 8 | 6.20 | 10.3 |
| 9 | 9.86 | 18.6 |
| 10 | 8.50 | 14.2 |

С помощью представленных данных построим линейное уравнений простой регрессии. Для этого необходимо просчитать коэффициенты $a$ и $b$. Для упрощения расчетов построим вспомогательную таблицу 2.2:

Таблица 2.2 – Расчетная таблица параметров линейного уравнения регрессии

| № | $x$ | $y$ | $xy$ | $x^2$ | $y^2$ | $\hat{y}$ | $y - \hat{y}$ | $(y - \hat{y})^2$ | A (%) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 8.70 | 17.8 | 154.86 | 75.69 | 316.84 | 15.74 | 2.06 | 4.26 | 12% |
| 2 | 6.30 | 11.6 | 73.08 | 39.69 | 134.56 | 12.09 | -0.49 | 0.24 | 4% |
| 3 | 7.89 | 15.8 | 124.66 | 62.25 | 249.64 | 14.51 | 1.29 | 1.67 | 8% |
| 4 | 10.24 | 13.5 | 138.24 | 104.8 | 182.25 | 18.08 | -4.58 | 20.95 | 34% |
| 5 | 10.25 | 20.5 | 210.13 | 105 | 420.25 | 18.09 | 2.41 | 5.80 | 12% |
| 6 | 7.50 | 15.9 | 119.25 | 56.25 | 252.81 | 13.91 | 1.99 | 3.95 | 12% |
| 7 | 8.75 | 14.9 | 130.38 | 76.56 | 222.01 | 15.81 | -0.91 | 0.83 | 6% |
| 8 | 6.20 | 10.3 | 63.86 | 38.44 | 106.09 | 11.94 | -1.64 | 2.68 | 16% |
| 9 | 9.86 | 18.6 | 183.4 | 97.22 | 345.96 | 17.50 | 1.10 | 1.21 | 6% |
| 10 | 8.50 | 14.2 | 120.7 | 72.25 | 201.64 | 15.43 | -1.23 | 1.52 | 9% |
| Сумма | 84.19 | 153.1 | 1318.55 | 728.27 | 2432.05 | 153.10 | 0.00 | 43.11 | 119% |
| ср.знач. | 8.419 | 15.31 | 131.855 | 72.83 | 243.21 | 15.31 | 0.00 | 4.31 | 22% |

Для расчета коэффициентов $a$ и $b$ воспользуемся формулами:

- $b$ (коэффициент наклона), вычисляется по формуле:

$$b = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})}, \tag{2.7}$$

- $a$ (свободный член), вычисляется через средние значения по формуле:

$$a = \overline{y} - b \cdot \overline{x}, \tag{2.8}$$

Подставим значения в формулы:

$$\overline{x} = \frac{\sum x}{n} = \frac{84.19}{10} = 8.42, \tag{2.9}$$

$$\overline{y} = \frac{\sum y}{n} = \frac{153.1}{10} = 15.31, \tag{2.10}$$

$$b = \frac{1318.55}{728.27} = 1.81, \tag{2.11}$$

$$a = 15.31 - 1.51 \cdot 8.42 = 2.52. \tag{2.12}$$

Таким образом, уравнение регрессии имеет вид:

$$y = 2.52 + 1.81x. \tag{2.13}$$

Вычисляем для уравнения коэффициент детерминации:

$$R^2 = \frac{(\sum(x_i - \overline{x})(y_i - \overline{y}))^2}{\sum(x_i - \overline{x})^2(y_i - \overline{y})^2} = \frac{1318.55^2}{728.27 \cdot 2432.05} = 0.51. \tag{2.14}$$

Это означает, что 51% изменчивости зарплаты можно объяснить изменчивостью прожиточного минимума. Остальные 49% объясняются другими факторами [8].

## 2.2 Логистическая регрессия

**Определение:** Логистическая регрессия — это статистическая модель, используемая для предсказания вероятности возникновения некоторого события путем подгонки данных к логистической кривой [9].

Логистическая регрессия применяется для предсказания вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится так называемая зависимая переменная $y$, принимающая лишь одно из двух значений, как правило, это числа: 0 (событие не произошло), и 1 (событие произошло), и множество независимых переменных (также называемых признаками, предикторами или регрессорами) — вещественных $x_1, x_2, \cdots, x_n$, на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной [9].

Стандартная логистическая функция, также известная как сигмовидная функция (*сигмовидная* означает «s–образная»), имеет формулу (2.15):

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{2.15}$$

На рисунке 2.2 показан соответствующий график сигмовидной функции. По мере увеличения входного значения $x$ выходной сигнал сигмовидной функции приближается, но никогда не достигает 1. Точно так же, когда входные данные уменьшаются, выходные данные сигмовидной функции приближаются, но никогда не достигают 0.

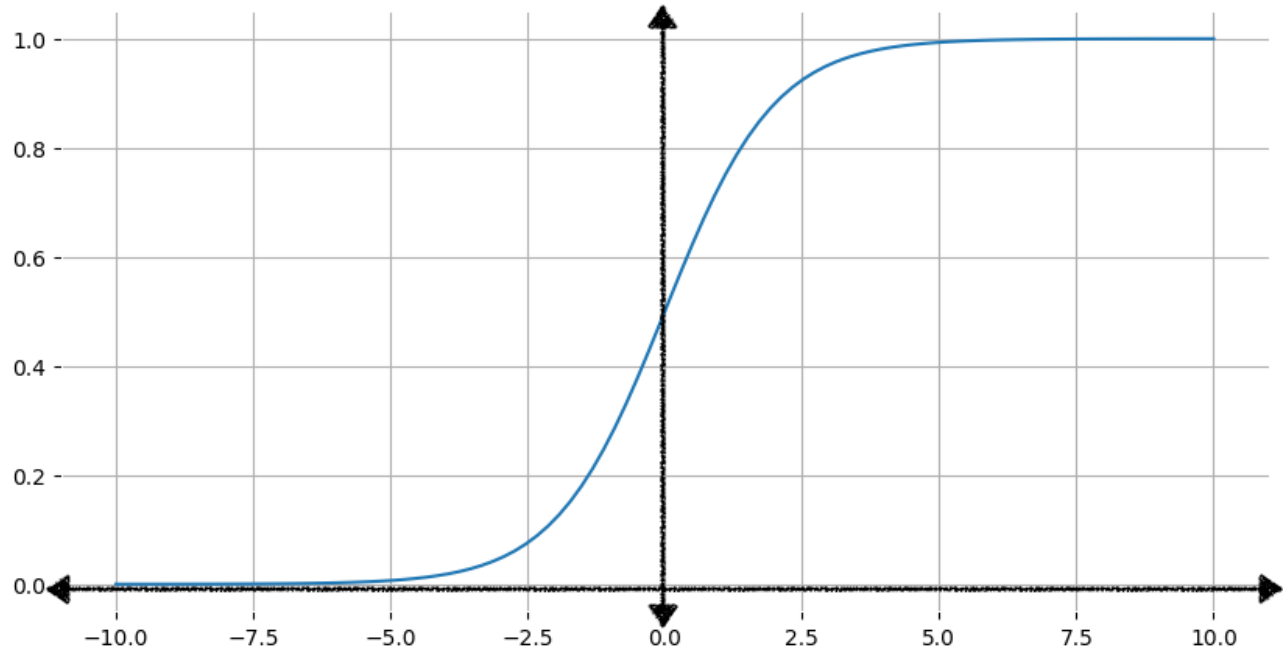



Рисунок 2.2 – График сигмовидной функции

Линейный компонент модели логистической регрессии описывается следующим уравнением (2.16):

$$z = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n, \tag{2.16}$$

где $n$ — число входных переменных, $z$ — результат линейного уравнения (логарифм шансов), $b_i$ — коэффициент регрессии для $i$–го признака, $x_i$ — значения признаков.

Чтобы получить прогноз логистической регрессии, значение $z$ затем передается сигмовидной функции, что дает значение (вероятность) от 0 до 1 (формула 2.17):

$$y' = \frac{1}{1 + e^{-z}}, \tag{2.17}$$

где $y'$ — результат модели логистической регрессии, $z$ — линейный выход (рассчитанный в уравнении 2.16).

На рисунке 2.3 показано как линейный результат преобразуется в результат логистической регрессии.

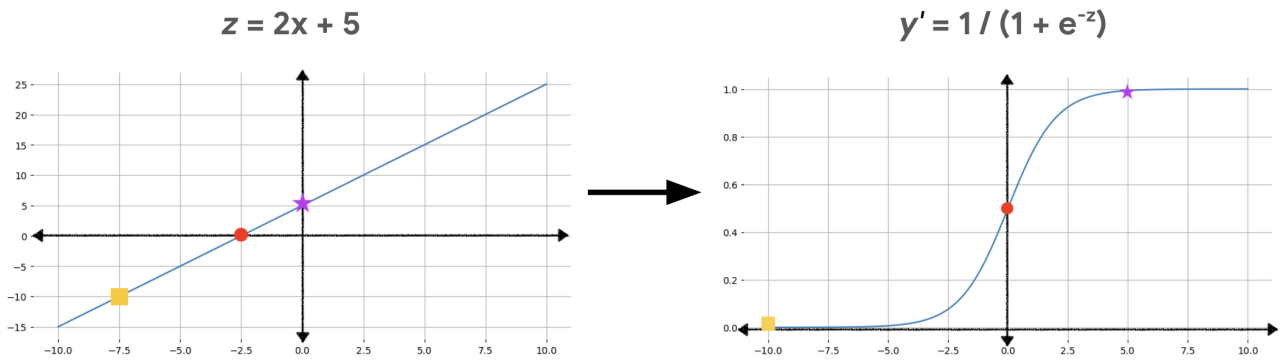



Рисунок 2.3 – Слева: график линейной функции $z = 2x + 5$, выделены три точки. Справа: сигмовидная кривая с теми же тремя точками, выделенными после преобразования сигмовидной функцией

В качестве функции потерь в линейной регрессии используется метод наименьших квадратов (квадрат потерь). Этот метод подходит для линейной модели, где скорость изменения выходных значений постоянна. Например, для линейной модели $y' = b + 3x_1$ каждый раз, когда увеличивается входное значение $x_1$ на 1, выходное значение $y'$ увеличивается на 3 [10].

Однако скорость изменения модели логистической регрессии не является постоянной. Когда значение логарифма шансов ($z$) ближе к 0, небольшое увеличение $z$ приводит к гораздо большим изменениям $y$, чем когда $z$ является большим положительным или отрицательным числом.

В таблице 2.3 показаны выходные данные сигмовидной функции для входных значений от 5 до 10, а также соответствующая точность, необходимая для учета различий в результатах.

Таблица 2.3 – Выходные данные сигмовидной функции

| | | |
|---|---|---|
| 6 | 0.997 | 3 |
| 7 | 0.999 | 3 |
| 8 | 0.9997 | 4 |
| 9 | 0.9999 | 4 |
| 10 | 0.99998 | 5 |

В данном случае нельзя успешно применить метод наименьших квадратов для оценки параметров $b$ и построения прогнозов, так как в этом случае прогнозные значения вероятности могут принимать как отрицательные значения, так и значения больше единицы [11]. Поэтому для оценки коэффициентов модели используют метод максимального правдоподобия, который заключается в оценивании параметров путем максимизации функции правдоподобия.

Положительный коэффициент говорит о том, что данный фактор увеличивает общий риск, то есть повышает вероятность анализируемого исхода. Отрицательный коэффициент означает, что данный фактор уменьшает риск, то есть понижает вероятность наступления исхода [11].

**Определение:** Метод максимального правдоподобия — еще один способ построения оценки неизвестного параметра. Состоит он в том, что в качестве «наиболее правдоподобного» значения берут значение $\Theta$, максимизирующее вероятность получить при $n$ опытах данную выборку $X = (X_1, X_2, \cdots, X_n)$ [11]. Это значение параметра $\Theta$ зависит от выборки и является искомой оценкой.

Формула функции правдоподобия имеет вид:

$$f(X, \Theta) = f_\Theta(X_1) \cdot f_\Theta(X_2) \cdot \cdots \cdot f_\Theta(X_n) = \prod_{i=1}^{n} f_\Theta(X_i). \tag{2.18}$$

Формула логарифма правдоподобия имеет вид:

$$L(X, \Theta) = \ln f(X, \Theta) = \sum_{i=1}^{n} \ln f_\Theta(X_i). \tag{2.19}$$

## Применение

Согласно проведенному анализу современных исследований, связанных с использованием логистической регрессии, было выявлено несколько особенностей. Во–первых, применение этой модели наиболее распространено в социально–экономических и медицинских исследованиях, хотя есть опыт применения в работах технического характера. Во–вторых, с помощью этой модели решают три типа задач: прогнозирование, классификация и оценка влияния факторов на исход [12].

При использовании логистической регрессии возникают следующие трудности:

1) ошибка соотнесения с классом значения — возникает при классификации объектов, которые близки к границе класса (вероятность близка к 0.5). Например, в [13] модель логистической регрессии используется для решения задачи кредитного скоринга. Задача заключается в классификации клиентов банка на два класса: надежные и ненадежные. Автор отмечает, что клиентов, для которых вероятность

возвращения кредита близка к 0.5, невозможно классифицировать однозначно;

2) мультиколлинеарность — возникает, когда два или более предиктора в модели линейно зависимы. Например, в [14] отмечается, что пренебрежение зависимостями между независимыми переменными ведет к построению ошибочных моделей. Автор предлагает проводить предварительное статистическое исследование и исключать такие переменные из анализа;

3) несбалансированные данные — возникает, когда один из классов в обучающей выборке представлен значительно меньшим количеством объектов. Например, в работе [15] автор отмечает, что при обучении модели на выборке с неравномерным распределением классов значений зависимой переменной была получена низкая точность прогноза. После «выравнивания» выборки и повторного обучения модель показала 99% точности;

4) проблема «границ чувствительности» — возникает, когда модель логистической регрессии не может корректно оценить вероятность принадлежности объекта к классу. Например, в кредитном скоринге клиенты в возрасте 30 и 31 год практически одинаковые группы, а клиенты с возрастом 60 и 61 год — весьма разные группы заемщиков [16].

Тем не менее большинство исследований показывают эффективность модели логистической регрессии. Кроме высокой точности прогнозирования, стоит отметить ее достоинство решать задачи различного масштаба: от 3 независимых [17] переменных до 230 [18], от 40 записей [19] до 20 миллионов [15], а построенные модели понятны для интерпретации.

Несмотря на существующие ограничения, модель логистической регрессии показывает высокую точность прогнозирования и широкий спектр применения в различных предметных областях.

## 2.3 Адаптивная регрессия

**Определение:** Адаптивная регрессия — это метод статистического моделирования, который использует функции для представления нелинейных зависимостей между переменными. В отличие от линейных моделей, адаптивная регрессия может автоматически подстраиваться под данные, включая нелинейные связи и взаимодействия между предикторами, что позволяет улучшить точность предсказаний [20].

Один из популярных методов адаптивной регрессии для выявления нелинейных связей в данных — это многомерные адаптивные регрессионные сплайны (MARC) [20].

MARC — это статистическая процедура, позволяющая решить классическую задачу регрессии: установить вид и параметры аппроксимирующей функции, описывающую функциональную зависимость отдельных наблюдений (исходные данные) с указанной точностью [21]. Пространство значений входных переменных разбивается на области со своими собственными уравнениями базисных функций. Это позволяет использовать MARC даже в случае задач с «проклятием размерности», когда высокая размерность пространства значений входных переменных ограничивает применимость иных статистических процедур.

Метод MAP–сплайнов не имеет ограничения, характерного для иных статистических методов, в части наличия исходных предположений о типе зависимостей (линейных, степенных, экспоненциальных) между предикторными и выходными переменными [21]. Кроме того, метод MARC чувствителен к изменению вида связи между предиктором и откликом, будь то: изменение формы связи (например, от линейной к степенной), добавление или вычитание некоторой константы для прогноза отклика справа от узловой точки предиктора, изменение наклона регрессионной функции [21].

Подобные особенности сплайнов достигаются за счет использования следующих базисных функций особого вида:

$$(x-t)_+ = \begin{cases} x-t, & x > t, \\ 0, & \text{иначе}, \end{cases} \tag{2.20}$$

$$(t-x)_- = \begin{cases} t-x, & x < t, \\ 0, & \text{иначе}, \end{cases} \tag{2.21}$$

где $t$ — точка разрыва (узел). Этот метод оценивает каждую точку данных для каждого предиктора в качестве узла и создаёт линейную регрессионную модель с выбранной(ыми) переменной(ыми).

Формулы вида (2.20 — 2.21) можно представить в следующем виде соответственно:

$$(x-t)_+ = max(0, x-t), \tag{2.22}$$

$$(t-x)_+ = max(0, t-x). \tag{2.23}$$

В многомерном случае для каждой компоненты $x_j$ вектора предикторов $x = (x_1, x_2, \cdots, x_n)$ строятся базисные функции вида (2.20 — 2.21) с узлами в каждой наблюдаемой переменной $x_i j$, где $i = \overline{1, 2, \cdots, n}$ и $j = \overline{1, 2, \cdots, m}$.

Общее уравнение MAP–сплайнов для модели из $M$ членов, отличных от константы, представляет собой взвешенную сумму базисных функций и их произведений и записывается в виде (2.20):

$$y = f(x) = b_0 + \sum_{m=1}^{M} b_m h_m(x), \tag{2.24}$$

где $b_0$ — свободный член, $b_m$ — коэффициенты регрессии, определяемые методом наименьших квадратов, $h_m(x)$ — базисная функция, $M$ — число базисных функций.

Основной принцип работы модели состоит в выборе нужной взвешенной суммы базисных функций из общего набора базисных функций, покрывающих все значения каждого предиктора (т.е. набор будет состоять из одной базисной функции и параметра $t$ для каждого отдельного значения каждой предикторной переменной). Алгоритм MAP–сплайнов отыскивает в пространстве всех входных и предикторных переменных расположение узловых точек, а также взаимосвязи между переменными [21]. В процессе поиска число добавленных к модели базисных функций из общего набора возрастает до тех пор, пока не будет максимизирован общий критерий качества модели — обобщенное

скользящее среднее, который имеет следующий вид:

$$GCV(M) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{(1 - \frac{C}{n})^2}, \tag{2.25}$$

где $GCV(M)$ — критерий точности модели, отражающий рост дисперсии с ростом числа базисных функций, $C$ — эффективное число параметров модели, которое в общем случае определяется как $C = 2K - 1$, где $K$ — общее число параметров модели.

Пошаговый алгоритм построения MAP–сплайна схож с алгоритмом линейной регрессии, только вместо регрессионных функций используются базисные функции. Например, рассмотрим нелинейные, немонотонные данные, где $y = f(x)$. Процедура MARC сначала ищет одну точку в диапазоне значений $x$, где две различные линейные зависимости между $y$ и $x$ минимизируют ошибку методом наименьших квадратов [22].

Результатом становится функция $h(x - a)$, где $a$ — это значение точки разрыва. Для одного узла функция $h(x - a)$ выглядит как $h(x - 1.183606)$, следовательно, две линейные модели для $y$ имеют вид:

$$\mathrm{y} = \begin{cases} \beta_0 + \beta_1(1.183606 - \mathrm{x}) & \mathrm{x} < 1.183606, \\ \beta_0 + \beta_1(\mathrm{x} - 1.183606) & \mathrm{x} > 1.183606 \end{cases} \tag{2.26}$$

После нахождения первого узла поиск продолжается для второго узла, который обнаруживается при $x = 4.898114$. Это приводит к созданию трёх линейных моделей для $y$:

$$\mathrm{y} = \begin{cases} \beta_0 + \beta_1(1.183606 - \mathrm{x}) & \mathrm{x} < 1.183606, \\ \beta_0 + \beta_1(\mathrm{x} - 1.183606) & \mathrm{x} > 1.183606 \quad \& \quad \mathrm{x} < 4.898114, \\ \beta_0 + \beta_1(4.898114 - \mathrm{x}) & \mathrm{x} > 4.898114 \end{cases} \tag{2.27}$$

На рисунке 2.4 показан пример адаптивного регрессионного сплайна с одним, двумя, тремя и четырьмя узлами соответственно.

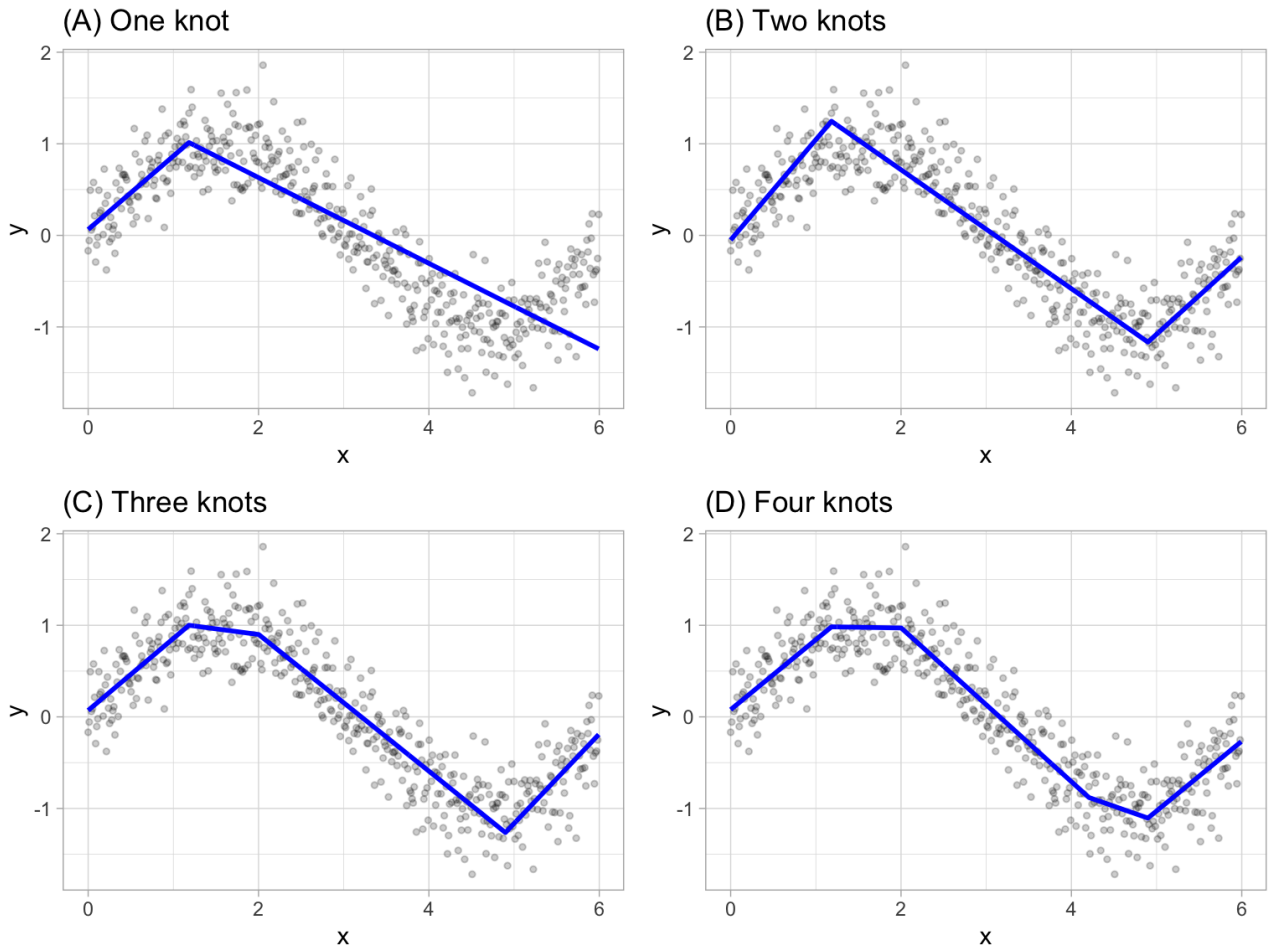



Рисунок 2.4 – Пример адаптивного регрессионного сплайна

Этот процесс продолжается до тех пор, пока не будет найдено множество узлов, что приводит к созданию точного нелинейного уравнения предсказания. Добавление большого количества узлов может позволить модели полностью соответствовать обучающим данным, но привести к недостаточной обобщаемости на новые, ранее не виденные данные.

Поэтому, после того как полный набор узлов найден, следует поочередно удалять узлы, которые не вносят значительного вклада в точность предсказания [22].

## Применение

MAP–сплайны находят свое применение во многих сферах науки и технологий, например, в предсказании видов распределений по имеющимся данным [23], кишечного поглощения лекарств [24], а также в воспроизведении речи [25] и поиске глобального оптимума в проектировании конструкций [26].

Многомерные адаптивные регрессионные сплайны обладают рядом пре-

имуществ перед другими регрессионными методами [27]:

1) модели, построенные с использованием MAP–сплайнов, обладают большей гибкостью, чем модели, построенные при помощи линейной регрессии;

2) MAP–сплайны могут автоматически находить нелинейные зависимости между переменными, что позволяет улучшить точность прогнозов;

3) MAP–сплайны позволяют работать с численными и категориальными признаками;

4) благодаря разделению исходных данных на области базисными функциями, MAP–сплайны позволяют определять выбросы;

5) MAP–сплайны не требуют значительных мер по подготовке входных данных;

6) метод демонстрирует высокую устойчивость к многоколлинеарности, что особенно важно при работе с большими наборами данных;

Благодаря своим преимуществам, MAP–сплайны нашли применение в биоинформатике для анализа геномных данных, включая прогнозирование экспрессии генов и идентификацию биомаркеров. Метод активно используется в экологии для моделирования и прогнозирования ареалов обитания видов с учетом множества факторов окружающей среды, в экономике и финансах — для прогнозирования цен, анализа временных рядов и моделирования волатильности; В задачах управления производственными процессами MAP–сплайны помогают оптимизировать технологические параметры и контролировать качество продукции.

# 3 Сравнение существующих решений

В качестве критериев сравнения решений задачи регрессии можно выделить следующие характеристики:

1) возможность реализации без специальных требований;

2) способность учитывать нелинейные зависимости;

3) интерпретируемость параметров;

4) возможность анализа как числовых, так и категориальных данных;

5) возможность использования для больших объемов данных;

6) устойчивость к переобучению;

7) корректная обработка выбросов в данных.

В таблице 3.1 приведено сравнение существующих решений задачи регрессии по выделенным критериям.

Таблица 3.1 – Сравнение существующих решений задачи регрессии

| Критерий | Линейная регрессия | Логистическая регрессия | Адаптивная регрессия |
|---|---|---|---|
| 1 | Да | Да | Нет |
| 2 | Нет | Нет | Да |
| 3 | Да | Да | Нет |
| 4 | Нет | Нет | Да |
| 5 | Да | Да | Нет |
| 6 | Да (при регуляризации) | Нет | Нет |
| 7 | Нет | Частично | Да |

Линейная регрессия отличается высокой интерпретируемостью, однако она ограничена в способности моделировать нелинейные зависимости. Логистическая регрессия применима в задачах бинарной классификации, но имеет ограничения в условиях высокой размерности данных или сложных

нелинейных связей. Адаптивная регрессия на примере метода MARC, продемонстрировала высокую гибкость и способность моделировать сложные и нелинейные зависимости, что делает её подходящей для задач с большим количеством переменных и сложными взаимодействиями, однако она требует тщательной настройки для предотвращения переобучения и обеспечения интерпретируемости.

# ЗАКЛЮЧЕНИЕ

В данной работе были рассмотрены различные методы регрессии, включая линейную регрессию, логистическую регрессию и адаптивную регрессию.

Цель работы достигнута.

В ходе выполнения работы были решены следующие задачи:

- был проведен анализ предметной области;
- был проведен обзор существующих решений задачи регрессии;
- были сформулированы критерии сравнения решений задачи регрессии;
- были классифицированы существующие решения задачи регрессии;

В ходе работы было выяснено, что каждый метод регрессии имеет свои преимущества и ограничения, которые делают его более подходящим для конкретных типов задач.

Таким образом, выбор подходящего метода зависит от особенностей данных и задачи. Линейная регрессия и логистическая регрессия применимы в случае простых линейных зависимостей и бинарных задач, в то время как адаптивные методы регрессии обеспечивают более высокую точность в сложных случаях, требующих гибкости и учёта нелинейных взаимодействий между переменными.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. *Н. Д., Г. С.* Прикладной регрессионный анализ. — —М.: Издательский дом «Вильямс» (дата обращения: 26.11.2024), 2012. — С. 350.
2. *К. Б.* Распознавание образов и машинное обучение. — —М.: Издательский дом «Вильямс» (дата обращения: 26.11.2024), 2006. — С. 700.
3. *Сибер Г. А. Ф., Ли А. Д.* Анализ линейной регрессии. — Уайли (дата обращения: 26.11.2024), 2012. — С. 456.
4. *Монтгомери Д. К., Пек Э. А., Вининг Г. Г.* Введение в анализ линейной регрессии. — Wiley (дата обращения: 26.11.2024), 2021. — С. 640.
5. Словарь машинного обучения [Электронный ресурс]. — — Режим доступа: (дата обращения: 16.11.2024) `https://developers.google.com/machine-learning/glossary`.
6. Линейная регрессия [Электронный ресурс]. — — Режим доступа: (дата обращения: 21.11.2024) `https://elibrary.ru/item.asp?id=49547912`.
7. Линейная регрессия [Электронный ресурс]. — — Режим доступа: (дата обращения: 21.11.2024) `https://wiki.loginom.ru/articles/linear-regression.html`.
8. *Л.П. Г.* Особенности линейной регрессии и ее применение в экономике. — 2020.
9. Логистическая регрессия [Электронный ресурс]. — — Режим доступа: (дата обращения: 21.11.2024) `https://elib.sfu-kras.ru/bitstream/handle/2311/7199/s021-083.pdf?sequence=1`.
10. Логистическая регрессия [Электронный ресурс]. — — Режим доступа: (дата обращения: 21.11.2024) `https://developers.google.com/machine-learning/crash-course/logistic-regression`.
11. Применение логистической регрессии для оценки кредитного риска [Электронный ресурс]. — — Режим доступа: (дата обращения: 02.12.2024) `https://elibrary.ru/item.asp?id=27478789`.
12. О возможностях применения модели логистической регрессии [Электронный ресурс]. — — Режим доступа: (дата обращения: 02.12.2024) `https://www.elibrary.ru/item.asp?id=25090019`.

13. *Симонов П.*, *Лазуков С.* Оценка кредитного риска: актуальные практические вопросы // Вестник Пермского университета. Сер. Экономика. — 2009. — № 1. — С. 61—67.

14. *Мурадов Д.* Logit-прогностические модели прогнозирования банкротства предприятий // Труды Российского государственного университета нефти и газа им. И.М. Губкина. — 2011. — № 3. — С. 160—172.

15. *Середний С.* Оценивание вероятности дефолта по кредитным операциям с использованием логистической регрессии и кластерного анализа // Достижения информационных технологий. — 2011. — № 1. — С. 126—132.

16. *Сапонов Д.* Опыт конкурентной борьбы как фактор академической успеваемости // Мониторинг общественного мнения: экономические и социальные перемены. — 2013. — 6 (118). — С. 113—126.

17. *Симонова С.* Интеллектуальный анализ данных для задач CEM // International Journal of Open Information Technologies. — 2015. — Т. 3, № 2. — С. 17—22.

18. *Осиков М.*, *Ахматов К.* Использование логистической регрессии в оценке изменений психологического статуса у больных хронической почечной недостаточностью, находящихся на гемодиализе // Вестник Южно-Уральского государственного университета. Сер. Образование, здравоохранение, физическая культура. — 2010. — 19 (195). — С. 34—37.

19. *Богданов Л.* Оценка эффективности бинарных классификаторов на основе логистической регрессии методом КОС-анализа // Вестник СГТУ. — 2010. — Т. 4, 2с. — С. 92—97.

20. *Фридман Д. Х.* Многомерные Адаптивные Регрессионные Сплайны // Ежегодный журнал статистики. — 1991. — Т. 19, № 1.

21. К вопросу применимости аппарата MAPC к задаче прогнозирования банкротства: зарубежный опыт [Электронный ресурс]. — — Режим доступа: (дата обращения: 02.12.2024) `https://www.imi-samara.ru/wp-content/uploads/2018/04/11_Romanova_91-98.pdf`.

22. Адаптивная регрессия [Электронный ресурс]. — — Режим доступа: (дата обращения: 26.11.2024) `https://bradleyboehmke.github.io/HOML/mars.html`.

23. *Элит Д.*, *Летвик Д.* Прогнозирование распространения видов на основе музейных и гербарных записей с использованием многомерных адаптивных регрессионных сплайнов // Разнообразие и распределение. — 2007. — Т. 13, № 3. — С. 265—275.

24. *Деконинк Э.*, *Коомонс Д.*, *Хейден Й.* Исследование методов линейного моделирования и их комбинаций с многомерными адаптивными регрессионными сплайнами для прогнозирования желудочно-кишечной абсорбции лекарств // Журнал фармацевтического и биомедицинского анализа. — 2007. — Т. 43, № 1. — С. 119—130.

25. *Хаас Х.*, *Кубин Г.* Многополосная нелинейная осцилляторная модель для речи // Сборник материалов 32-й конференции Asilomar по сигналам, системам и компьютерам. Т. 1. — 1998. — С. 338—342.

26. *Крино С.*, *Браун Д.* Глобальная оптимизация с использованием многомерных адаптивных регрессионных сплайнов // Труды IEEE: Системы, человек и кибернетика, часть B: Кибернетика. — 2007. — Т. 37, № 2. — С. 333—340.

27. *С. Ф. А.* Модели потребления электроэнергии в многоквартирных жилых домах на основе многомерных адаптивных регрессионных сплайнов [Электронный ресурс]. — 2011. — — Режим доступа: (дата обращения: 02.12.2024) `https://elibrary.ru/download/elibrary_26342121_69582650.pdf`.

# ПРИЛОЖЕНИЕ А

Презентация к научно–исследовательской работе состоит из 7 слайдов.

Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)

# Решения задачи регрессии

**Студент**: Везирова Йована Недялкова
**Группа**: ИУ7И-71Б
**Руководитель НИР**: Строганов Юрий Владимирович

# Цель и задачи работы

**Цель** – провести анализ методов и подходов к решению задачи регрессии.

**Задачи:**

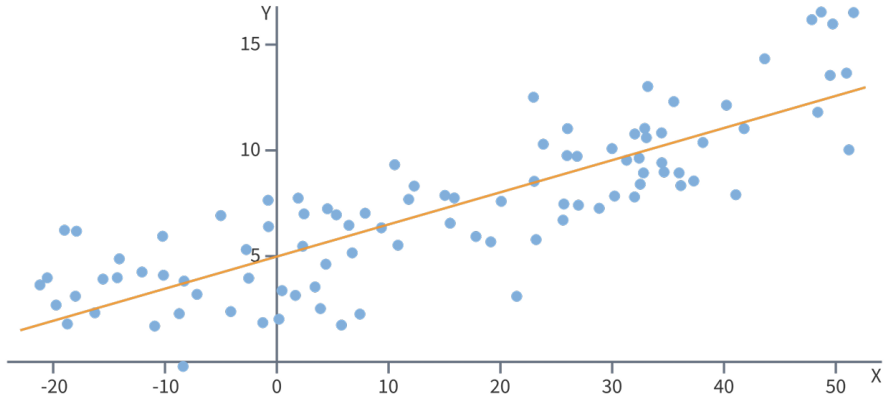
- провести анализ предметной области;
- провести обзор существующих решений задачи регрессии;
- сформулировать критерии сравнения решений задачи регрессии;
- классифицировать существующие решения задачи регрессии.

# Анализ предметной области

1. Регрессия – это метод прогнозирования и анализа зависимости целевой переменной от одной или нескольких независимых переменных. Этот подход широко применяется в задачах предсказания, моделирования и объяснения зависимости переменных, позволяя строить аналитические модели, описывающие взаимодействия в сложных системах.
2. Модель погоды, прогнозирующая количество осадков в миллиметрах, является регрессионной моделью.
3. В презентации будут рассмотрены три основных типа регрессии: линейная, логистическая и адаптивная регрессия.

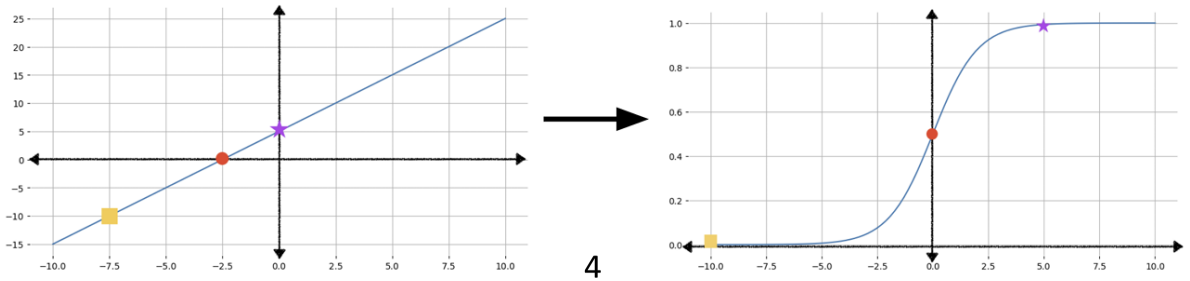
# Линейная регрессия

1. Линейная регрессия – это статистический метод, используемый для поиска взаимосвязи между переменными. В контексте машинного обучения линейная регрессия находит связь между функциями и меткой.
2. Уравнение регрессии: $y = ax + b$, где a – коэффициент наклона, b – свободный член.
3. Линейная регрессия может использоваться для прогнозирования будущих значений на основе существующих данных.
4. Этот метод применяется для оценки силы связи между зависимой и независимыми переменными.



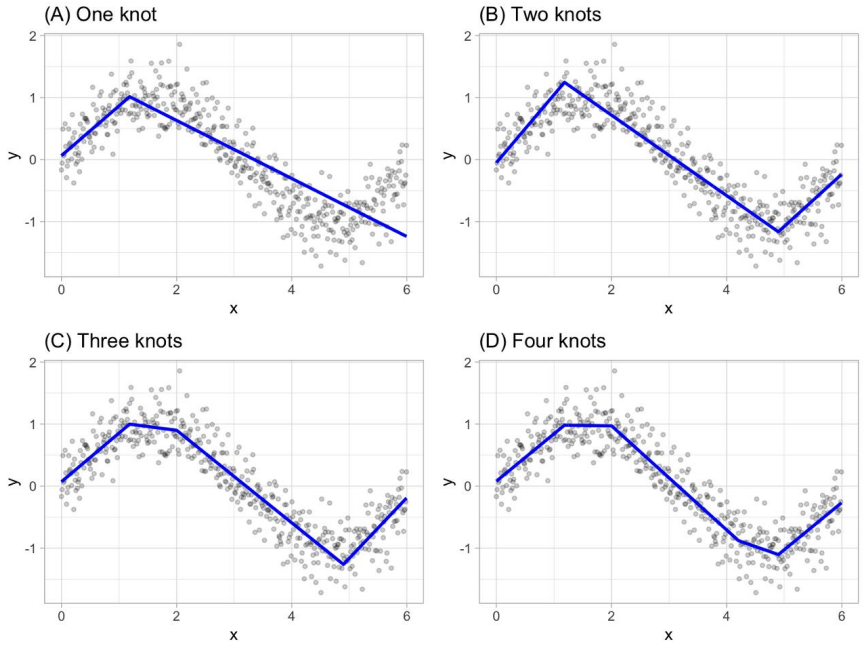
# Логистическая регрессия

1. Логистическая регрессия – это статистическая модель, используемая для предсказания вероятности возникновения некоторого события путем подгонки данных к логистической кривой.
2. Линейный предиктор (z): основной элемент логистической регрессии, вычисляется как взвешенная сумма независимых переменных и коэффициентов: $z = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$,
   где n – число входных переменных, z – результат линейного уравнения (логарифм шансов), $b_i$ – коэффициент регрессии для i–го признака, $x_i$ – значения признаков.
3. Сигмоидальная функция: преобразует значение *z* в вероятность в диапазоне от 0 до 1: $y' = \frac{1}{(1 + e^{-z})}$.
4. Логистическая регрессия используется для прогнозирования вероятности принадлежности объекта к определенному классу на основе существующих данных.
5. Этот метод применяется для анализа связи между зависимой переменной и независимыми, а также для решения задач классификации, в частности, для бинарной классификации (например, "да"/"нет", "истина"/"ложь").



# Адаптивная регрессия

1. Адаптивная регрессия — это метод, который моделирует нелинейные зависимости между переменными, автоматически подстраиваясь под данные для повышения точности предсказаний.
2. Одним из популярных подходов является метод многомерных адаптивных регрессионных сплайнов (MARC), который определяет вид и параметры функции, описывающей зависимости в данных.
3. Уравнения MAP-сплайнов: $y = b_0 + \sum_{m=1}^{M} b_m \, h_m(x)$, где $b_0$ — свободный член, $b_m$ — коэффициенты регрессии, определяемые методом наименьших квадратов, $h_m(x)$ — базисная функция, M — число базисных функций.
4. MARC используются для прогнозирования будущих значений на основе существующих данных, моделируя нелинейные зависимости между переменными.
5. Этот метод позволяет выявлять сложные взаимодействия и зависимости между зависимой и независимыми переменными, улучшая точность предсказаний.



# Заключение

Цель работы достигнута.

В ходе выполнения работы были решены следующие задачи:

- был проведен анализ предметной области;
- был проведен обзор существующих решений задачи регрессии;
- были сформулированы критерии сравнения решений задачи регрессии;
- были классифицированы существующие решения задачи регрессии.