ШУМ И GAM: ОБОБЩЁННЫЕ АДДИТИВНЫЕ МОДЕЛИ

МАНАЕНКОВ АЛЕКСАНДР

ИСТОРИЯ О ЧАСТНЫХ СЛУЧАЯХ

В ПРЕДЫДУЩЕЙ СЕРИИ

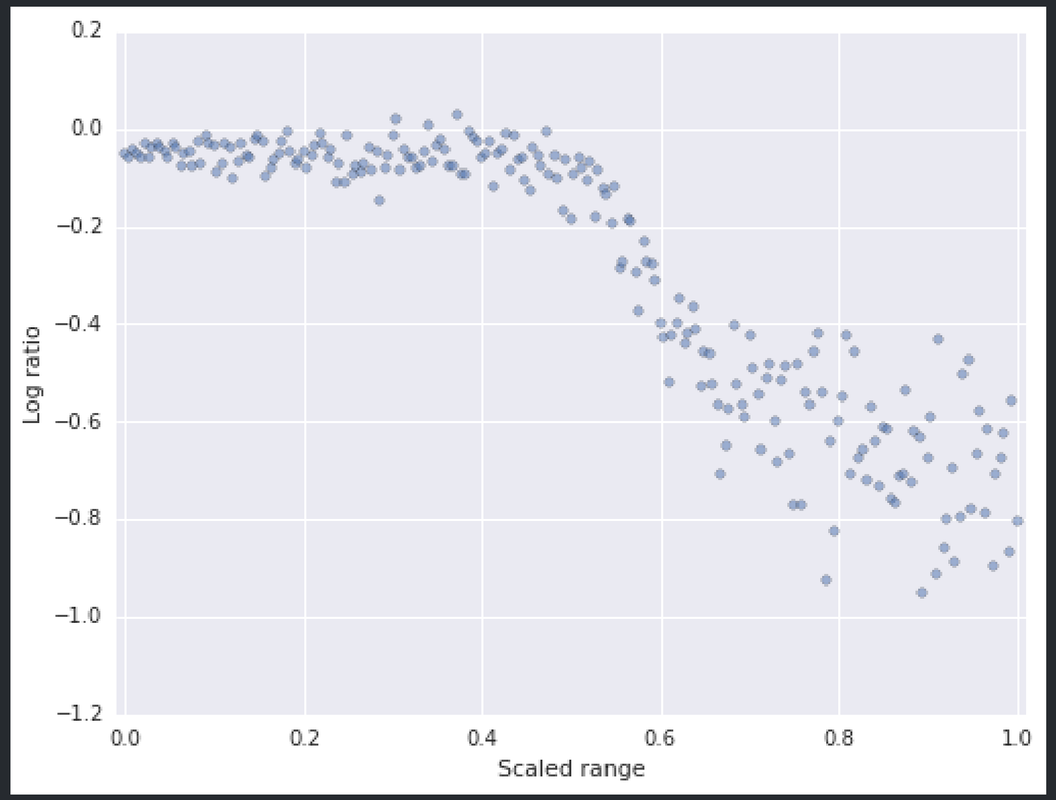
- Мы познакомились с GLM
- Это как обычные регрессионные модели, но с другим распределением
- Существует своё распределение чуть ли не под каждый рабочий случай
- Отчасти мы обходим допущение нелинейности – главное, чтобы преобразованные данные имели линейную взаимосвязь

В ПРЕДЫДУЩЕЙ СЕРИИ

- Мы познакомились с GLM
- Это как обычные регрессионные модели, но с другим распределением
- Существует своё распределение чуть ли не под каждый рабочий случай
- Отчасти мы обходим допущение нелинейности – главное, чтобы преобразованные данные имели линейную взаимосвязь

ОТЧАСТИ

НЕЛИНЕЙНЫЕ ВЗАИМОСВЯЗИ



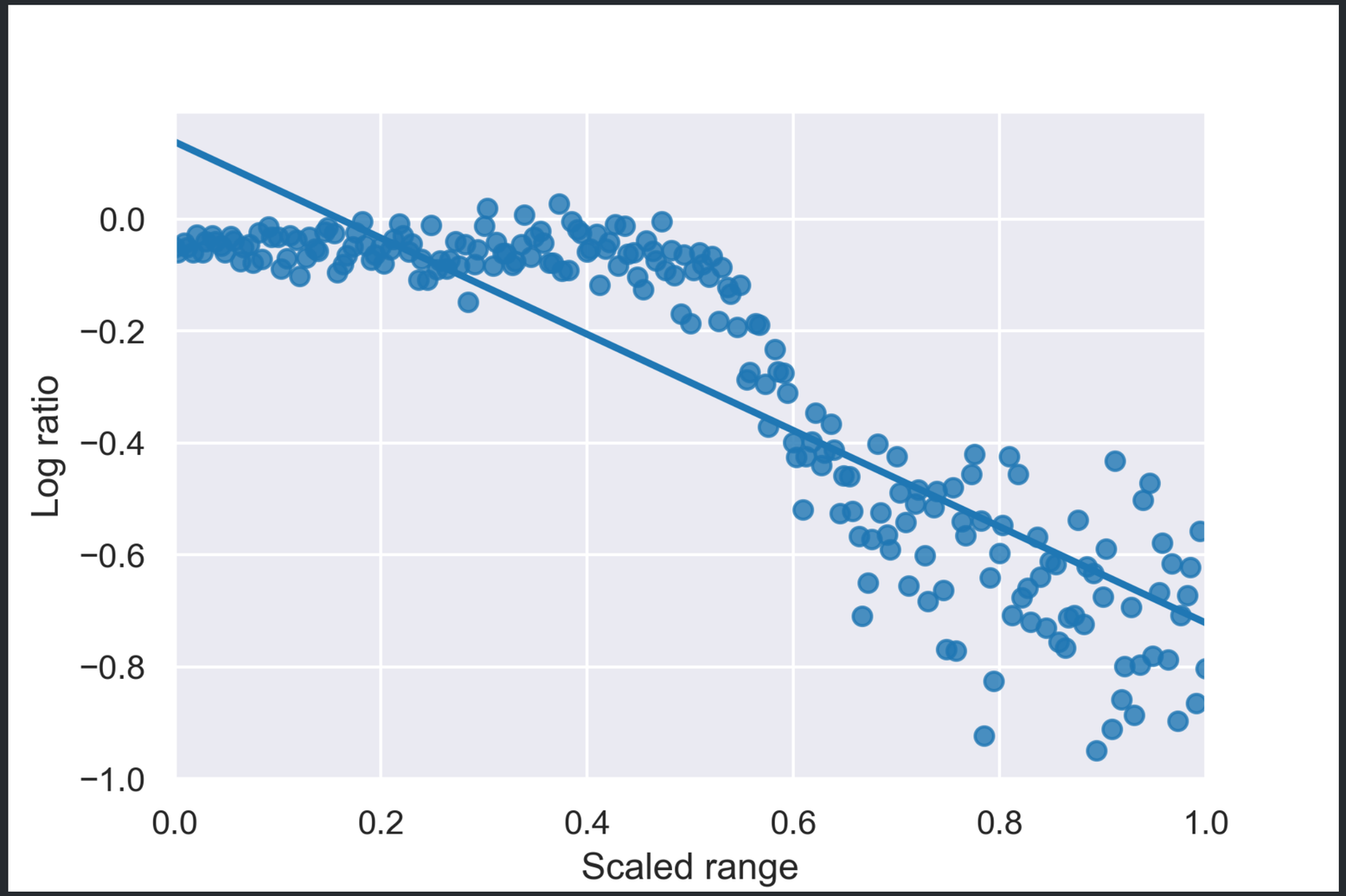
Независимы от характера распределения.

Наши данные могут быть хоть нормальные, хоть биномиальные, хоть пуассоновские - для всех них связи могут быть как (примерно) линейные, так и крайне нелинейные.

Таким образом, ни классические линейные модели, ни GLM с этой ситуацией не справляются.

(данные и рисунок - отсюда)

БОЛЬ



😂 МНОГОЧЛЕНЫ 😂

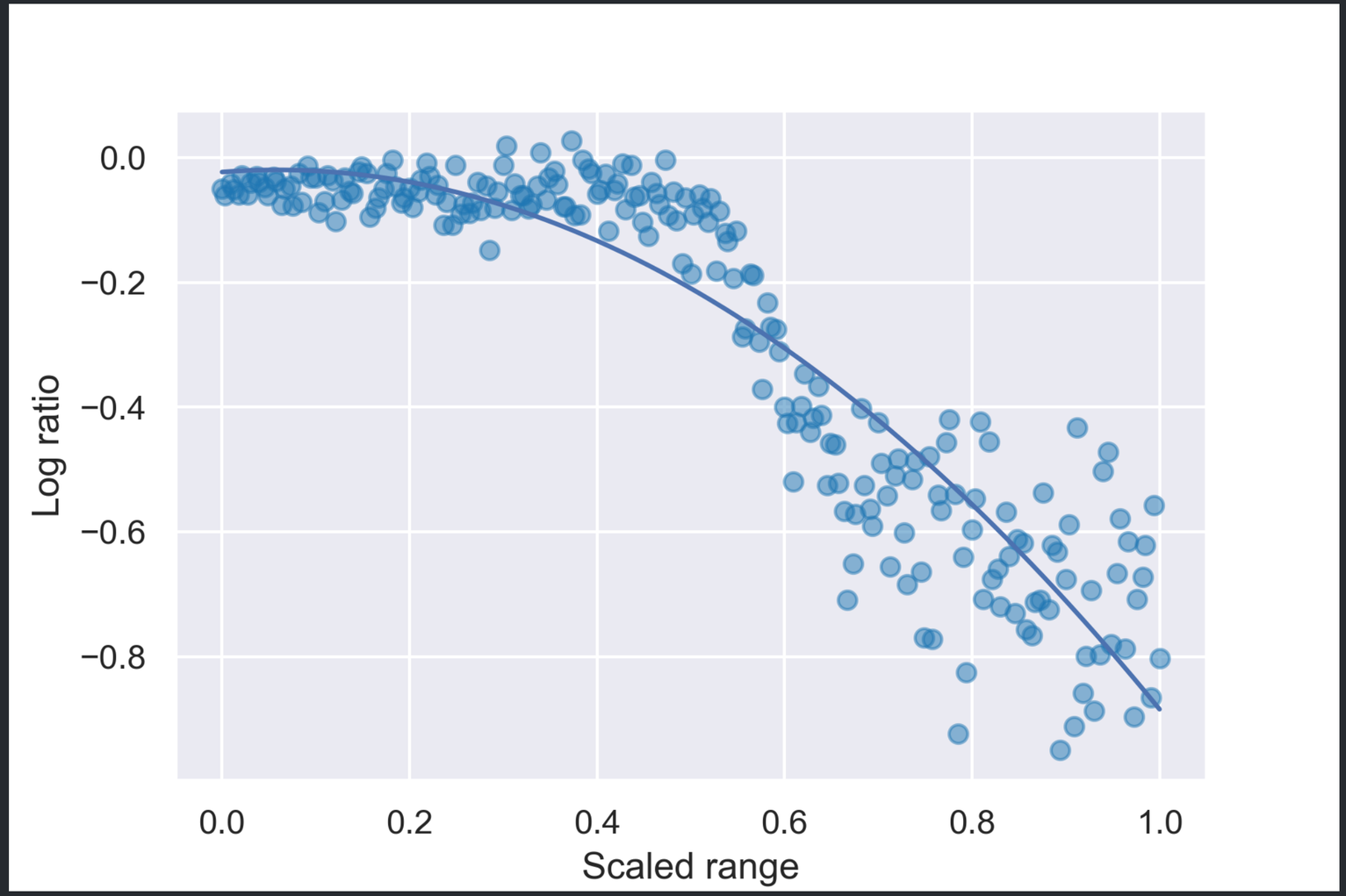
$ax + b$ – многочлен первого порядка (линия)

$ax^2 + bx + c$ – многочлен второго порядка (квадратный)

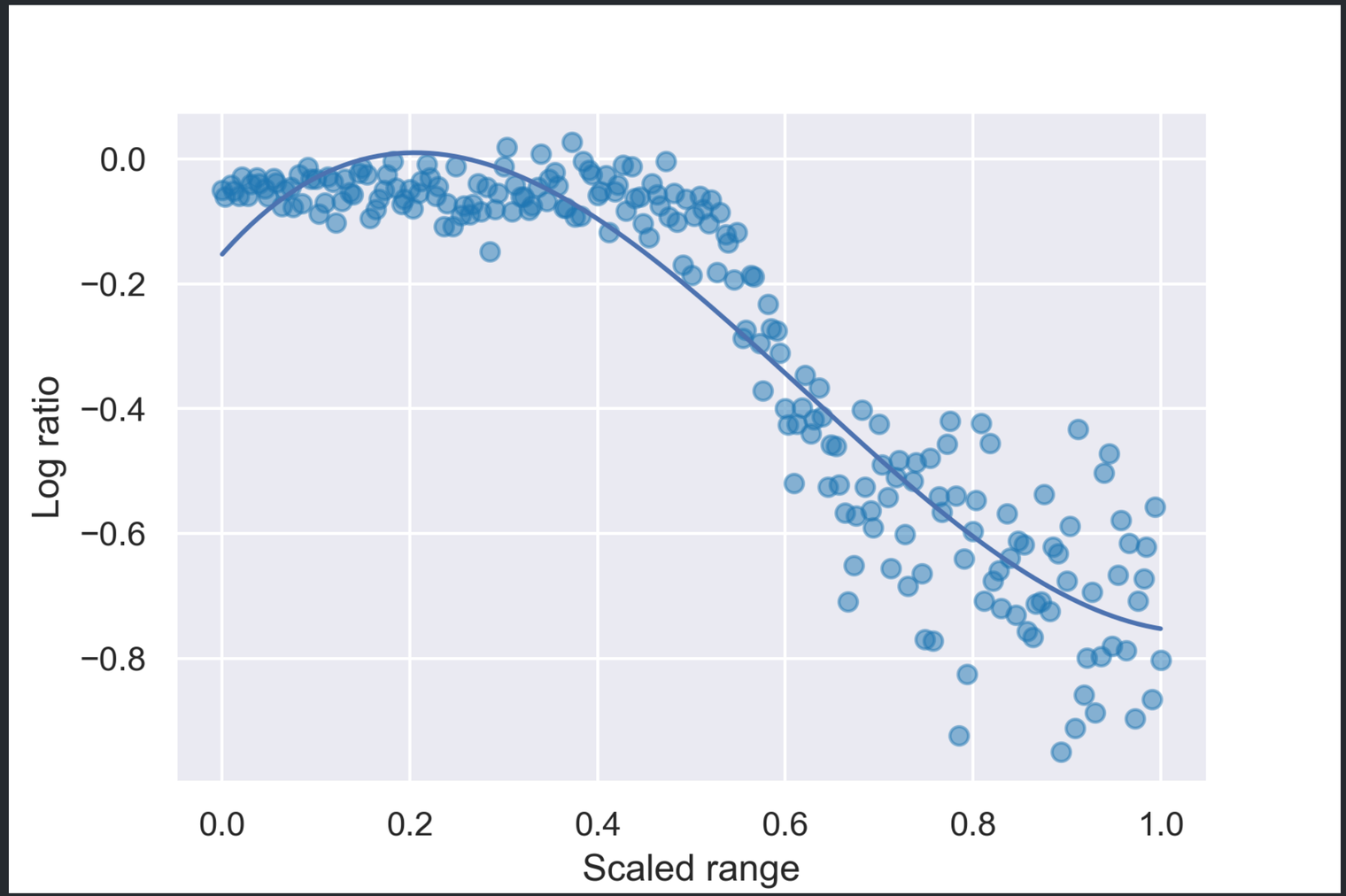
$ax^3 + bx^2 + cx + d$ – многочлен третьего порядка (кубический)

и так далее

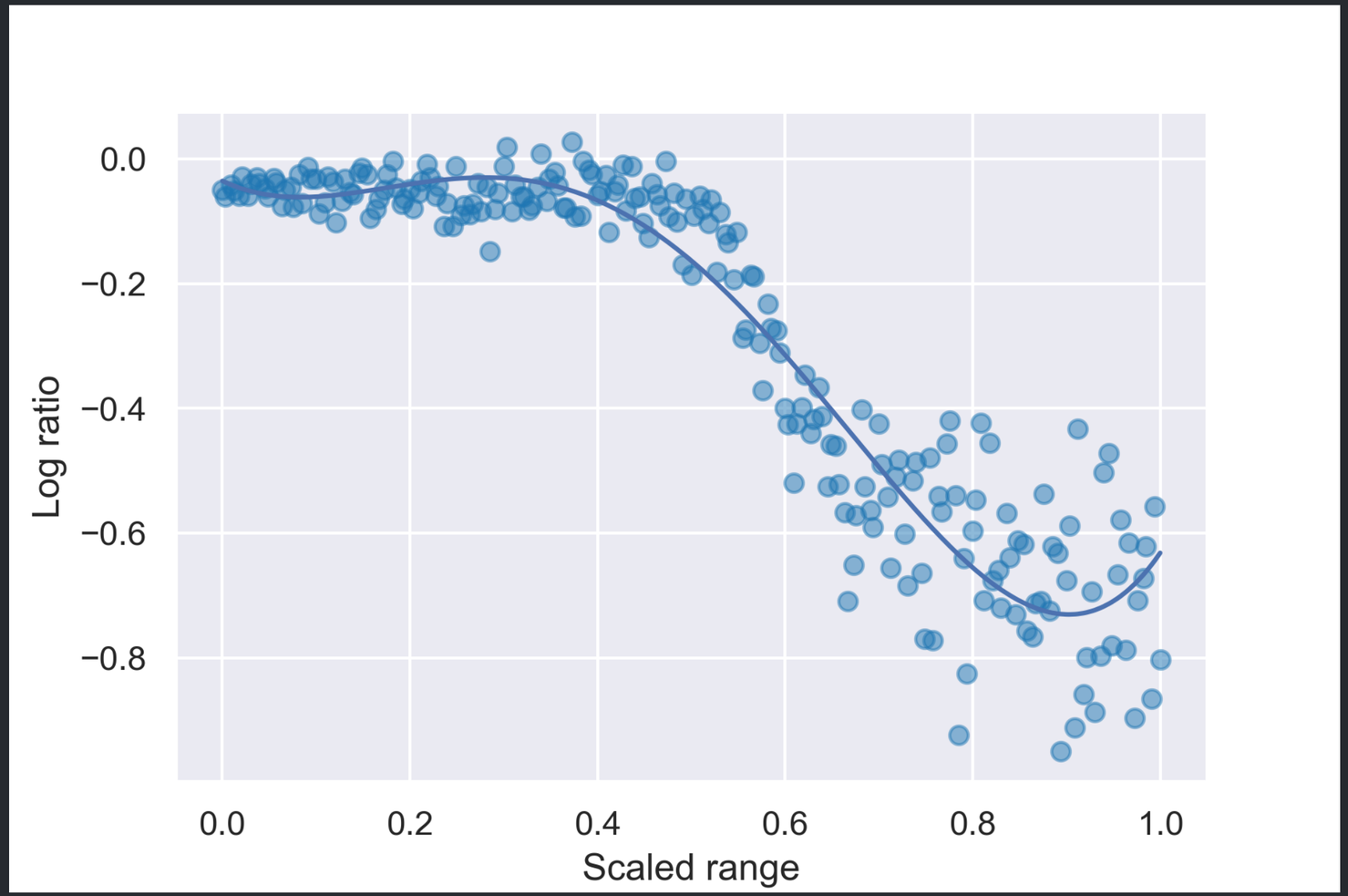
ВТОРАЯ СТЕПЕНЬ



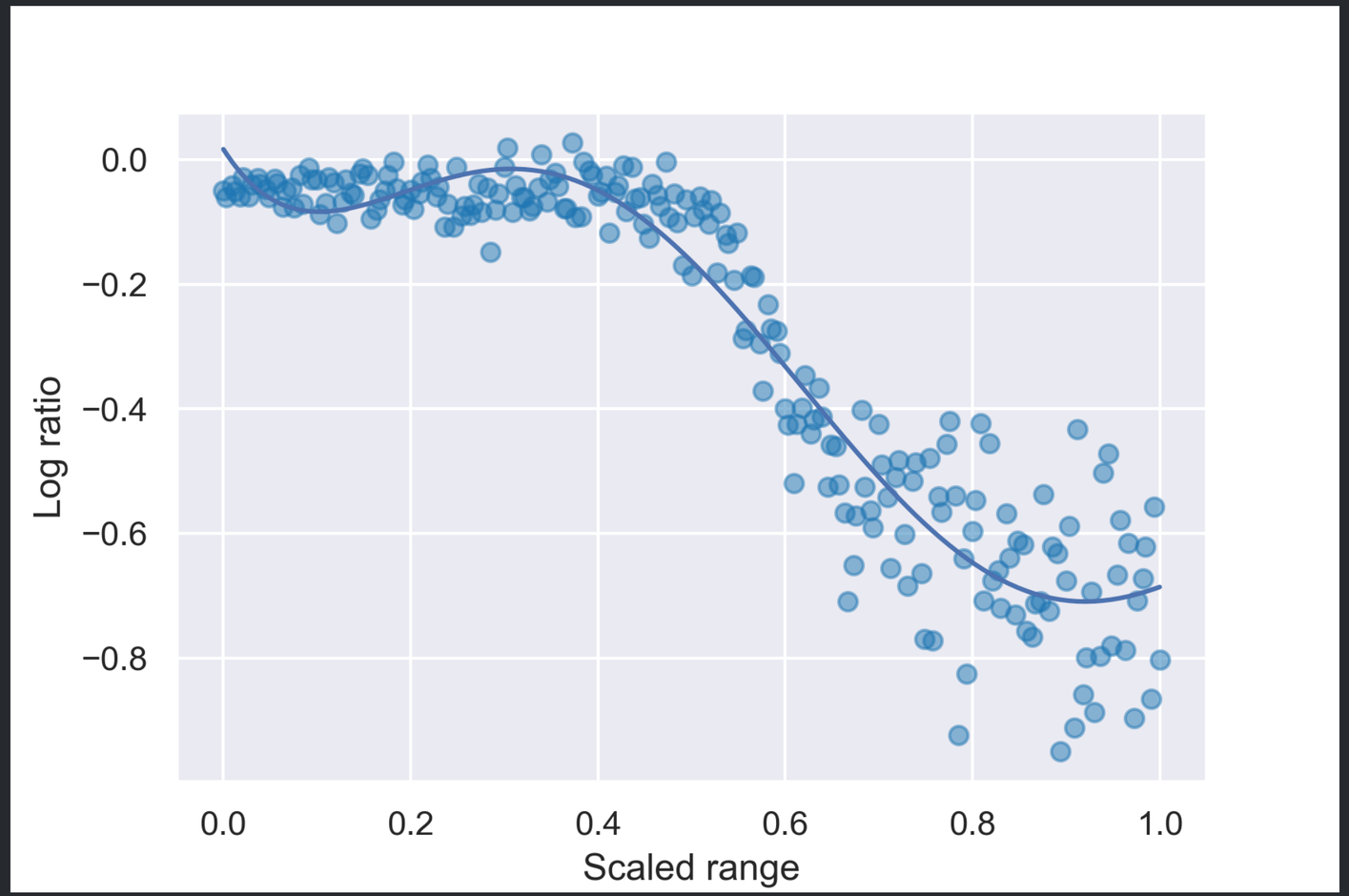
ТРЕТЬЯ СТЕПЕНЬ



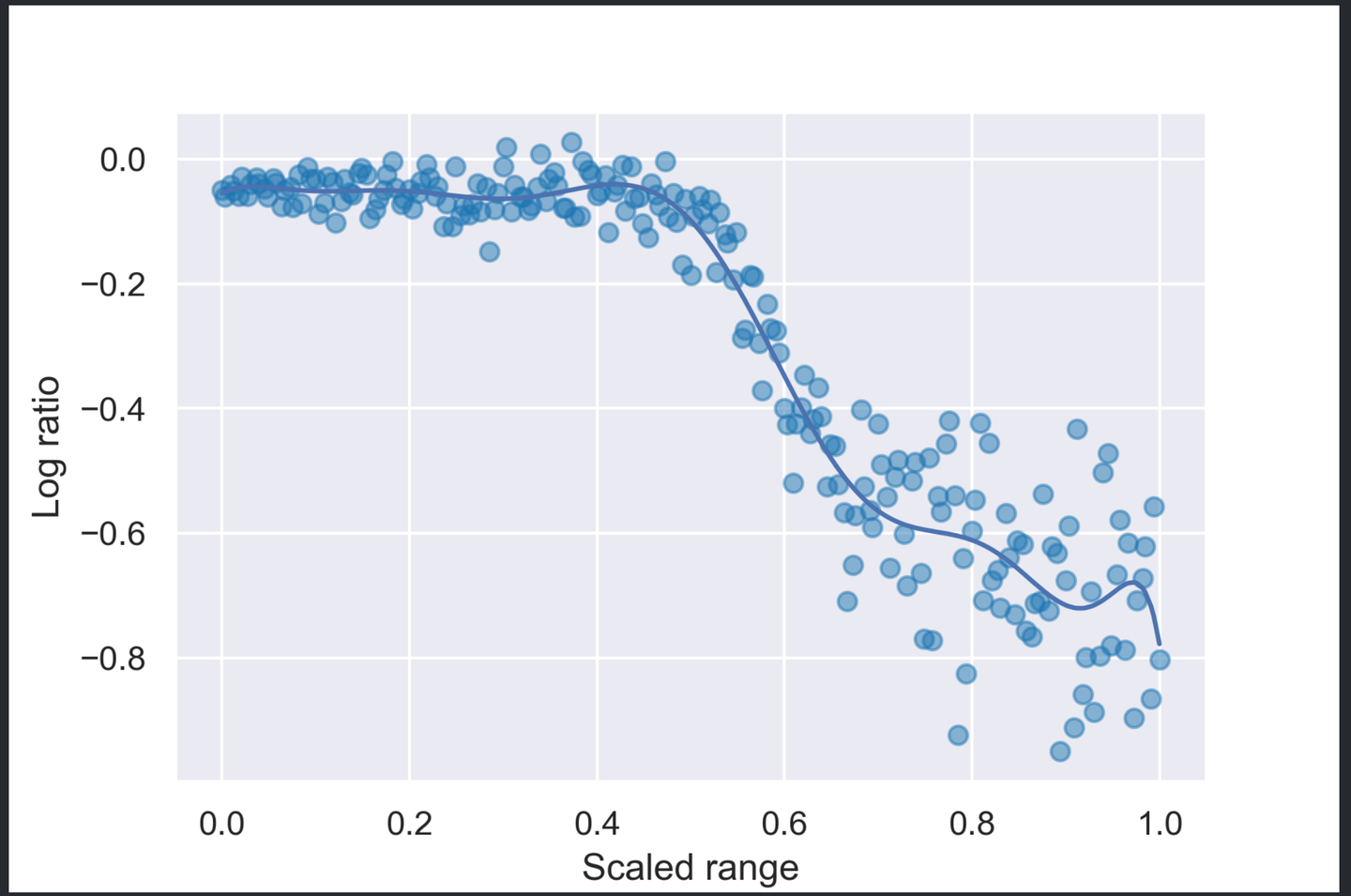
ЧЕТВЁРТАЯ СТЕПЕНЬ



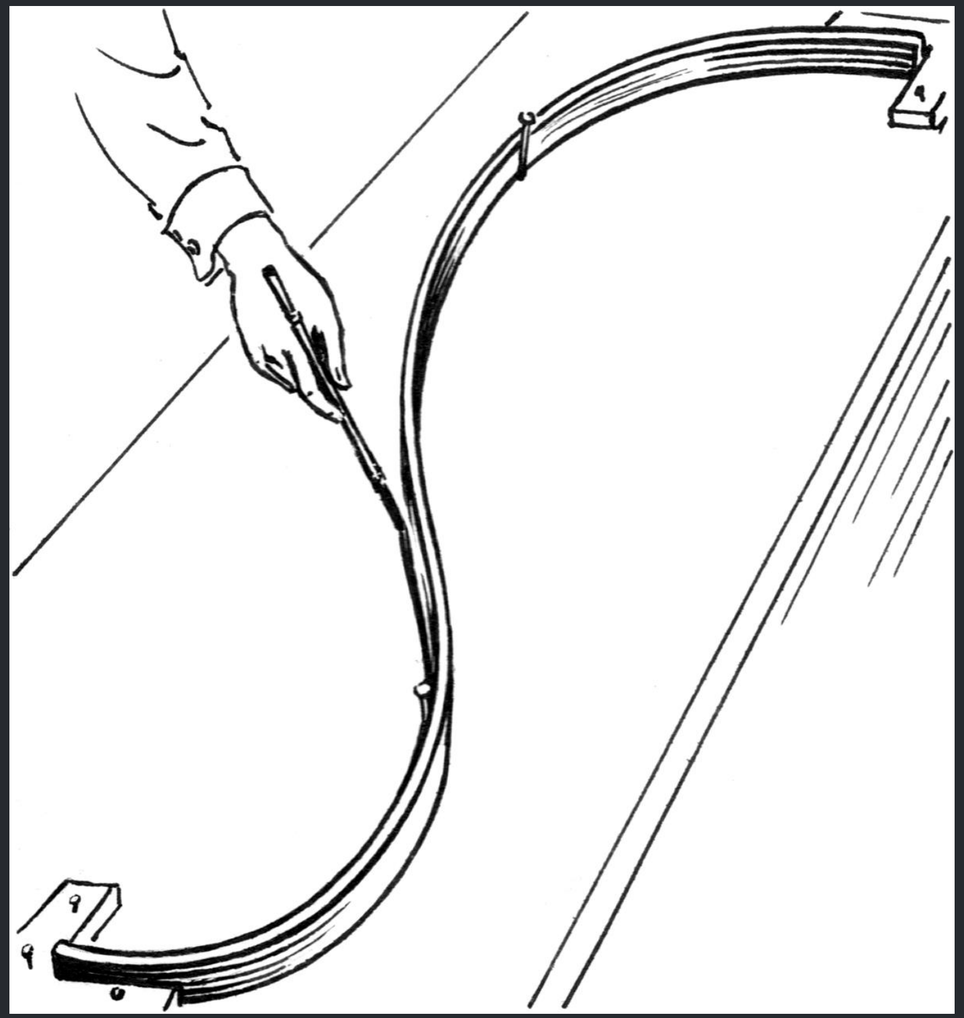
ПЯТАЯ СТЕПЕНЬ



ОДИННАДЦАТАЯ СТЕПЕНЬ



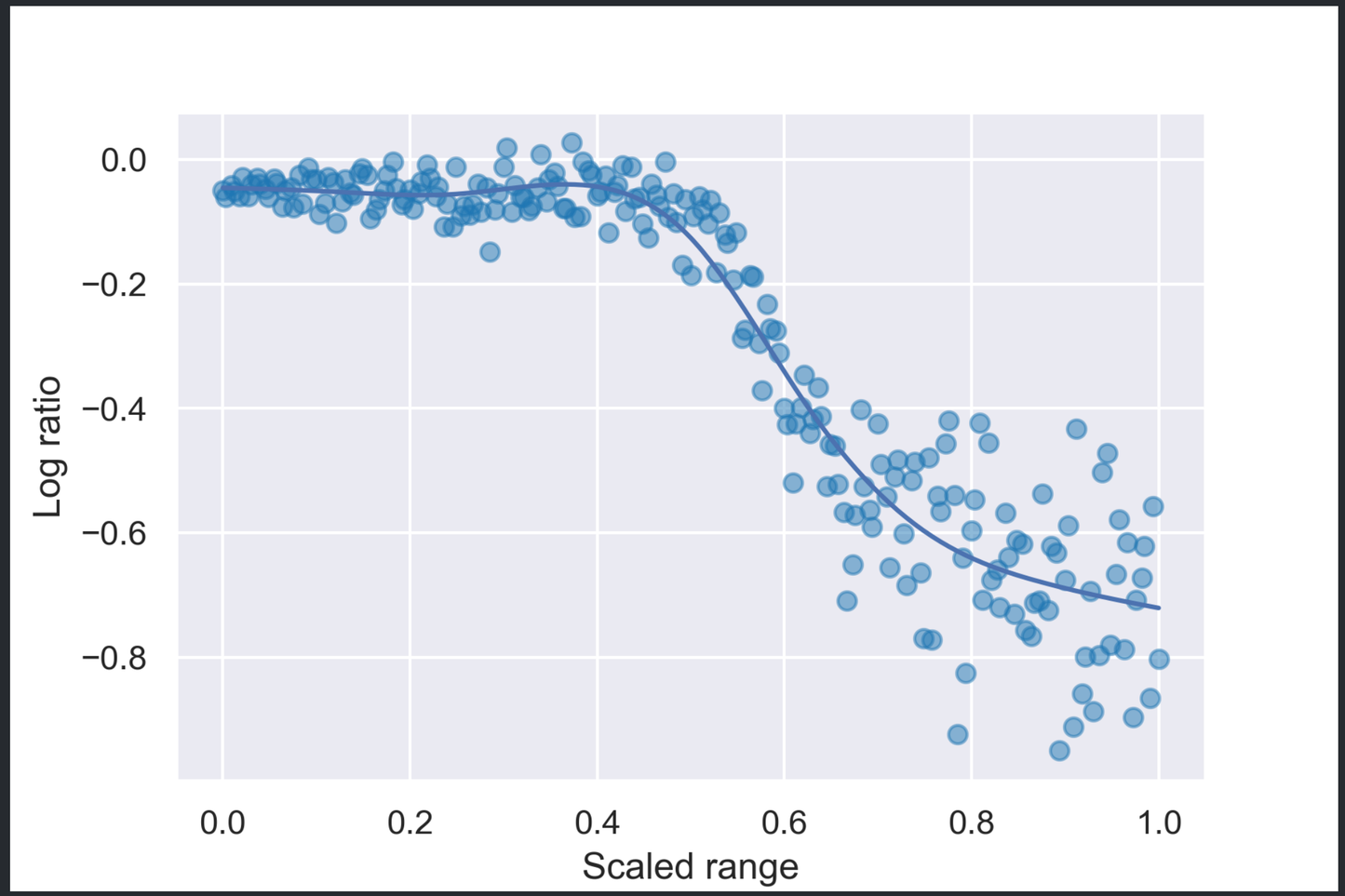
ГИБКОЕ ЛЕКАЛО (СПЛАЙН)



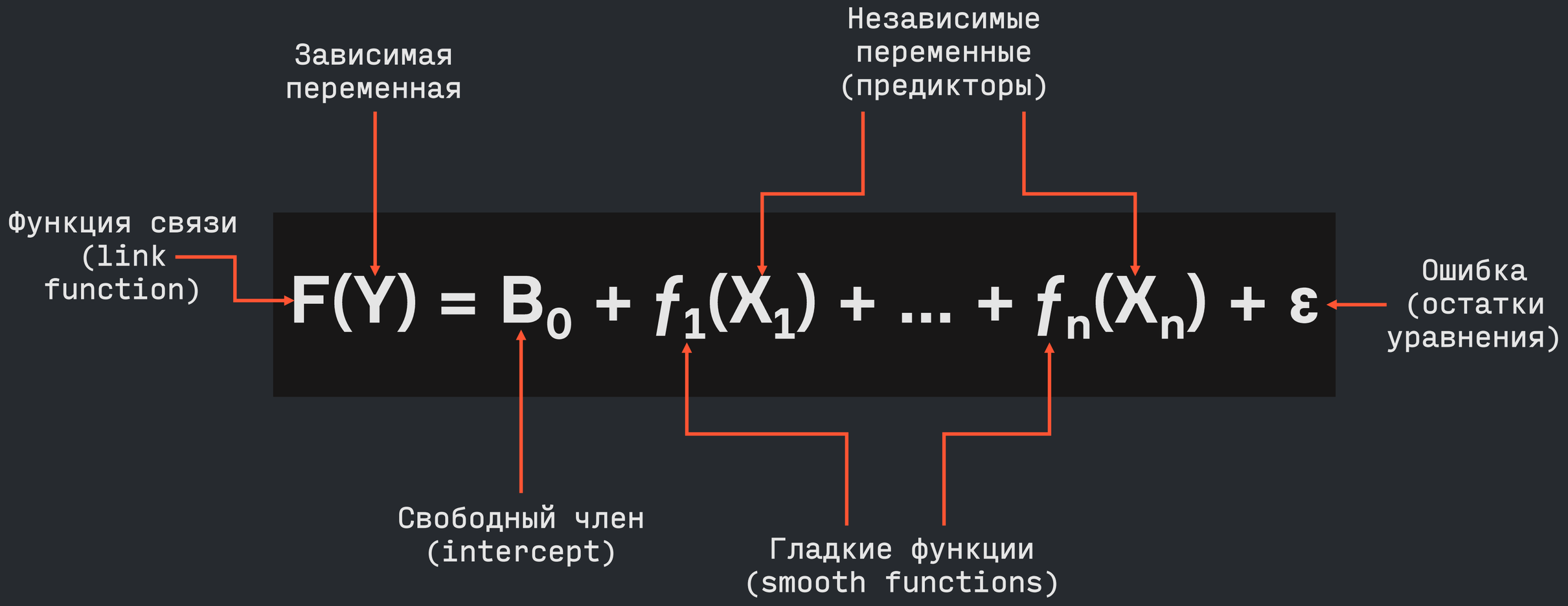
ВИДЫ СПЛАЙНОВ

- B-splines
- P-splines (penalized B-splines)
- Cyclic splines
- Thin plate splines
- Duchon splines
- Soap film splines
- Сферические сплайны
- Много другого безобразия

P-SPLINE (≈7 СТЕПЕНЕЙ СВОБОДЫ)

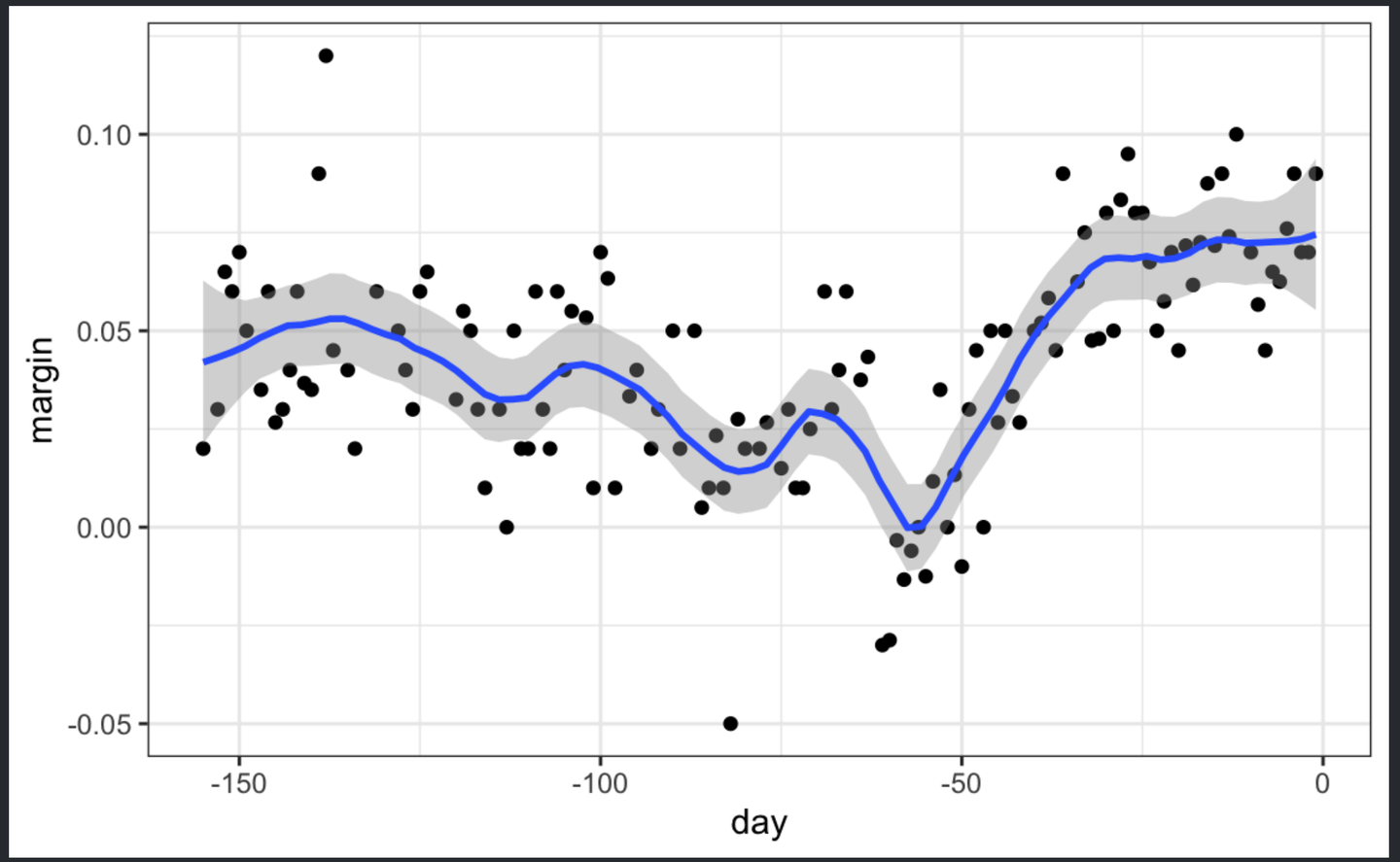


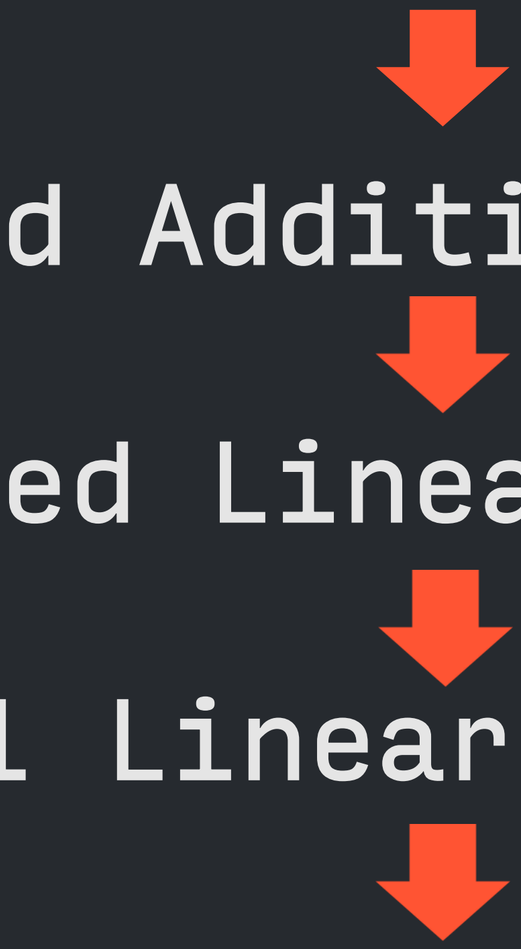
ОБЩАЯ (УПРОЩЁННАЯ) ФОРМУЛА GAM



KERNEL METHODS

LOESS/LOWESS




Generalized Additive Models (GAM)
Generalized Linear Models (GLM)
General Linear Models (LM)
ANOVA

ПАКЕТЫ ДЛЯ GAM В PYTHON

Как обычно бывает, это statsmodels, а также pyGAM и gammy.

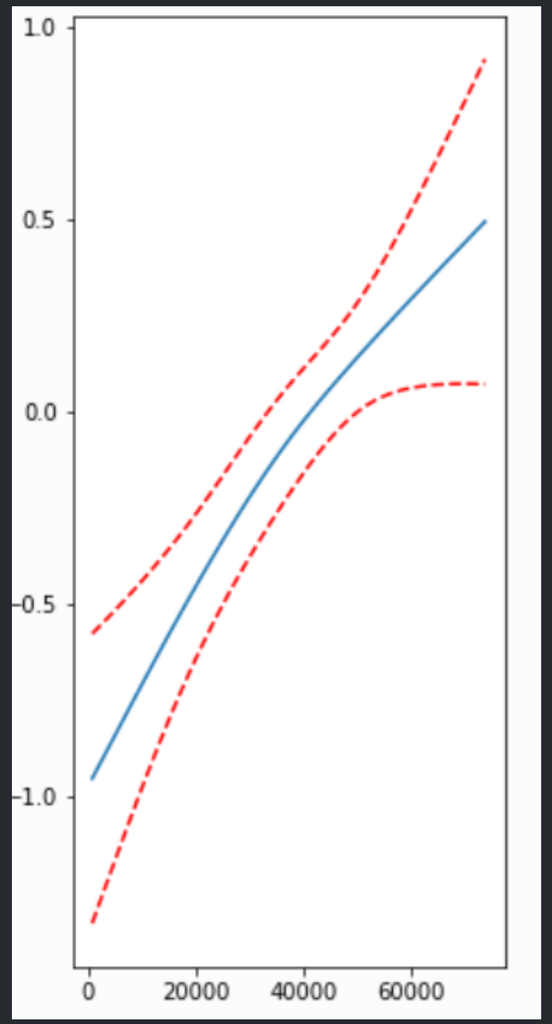
Из всего этого мы будем использовать pyGAM - на мой взгляд, им пользоваться приятнее всего.

Одновременно похож на scikit-learn и на mgcv из R.

Документация: https://pygam.readthedocs.io/en/latest/

ОСНОВНЫЕ ФОРМЫ ПРЕДИКТОРОВ

s()

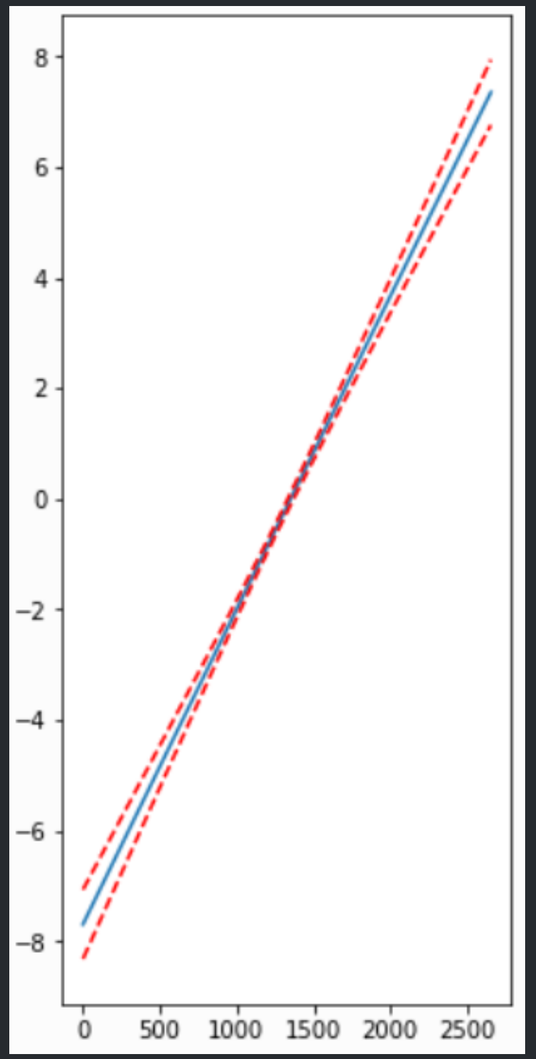


Основная рабочая лошадка пакета - делает сплайны.

Есть ряд задаваемых параметров:

- `n_splines` - число сплайнов, должен быть больше порядка многочлена (по умолчанию 20)
- `spline_order` - порядок базового многочлена (по умолчанию 3)
- `lam` - константа регуляризации (по умолчанию 0.6)
- `basis` - тип сплайна (по умолчанию p-spline)
- другое...

I()

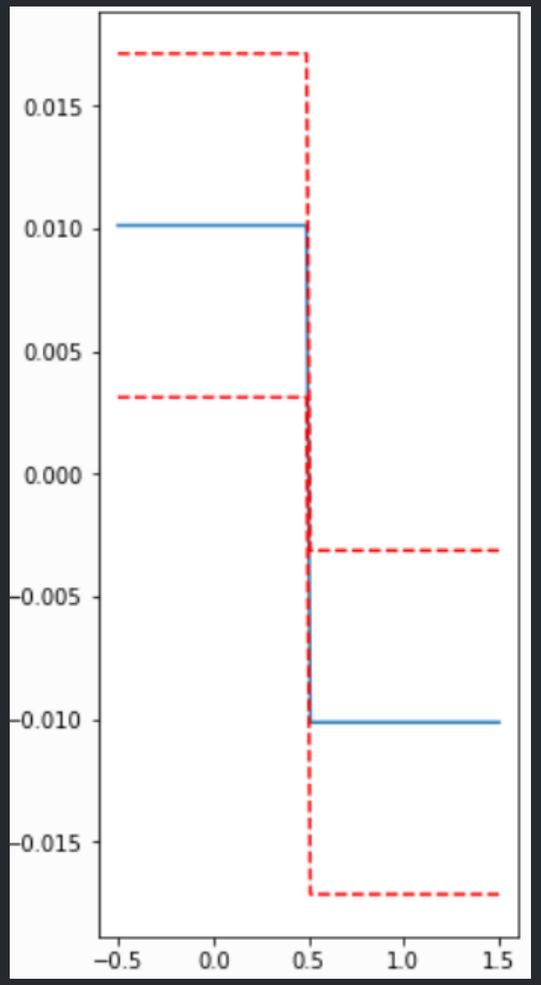


Нужен для тех случаев, когда мы хотим указать именно линейную взаимосвязь.

Также подвержен регуляризации (есть параметр lam).

Доверительные интервалы обычно шире, чем у сплайнов.

f()

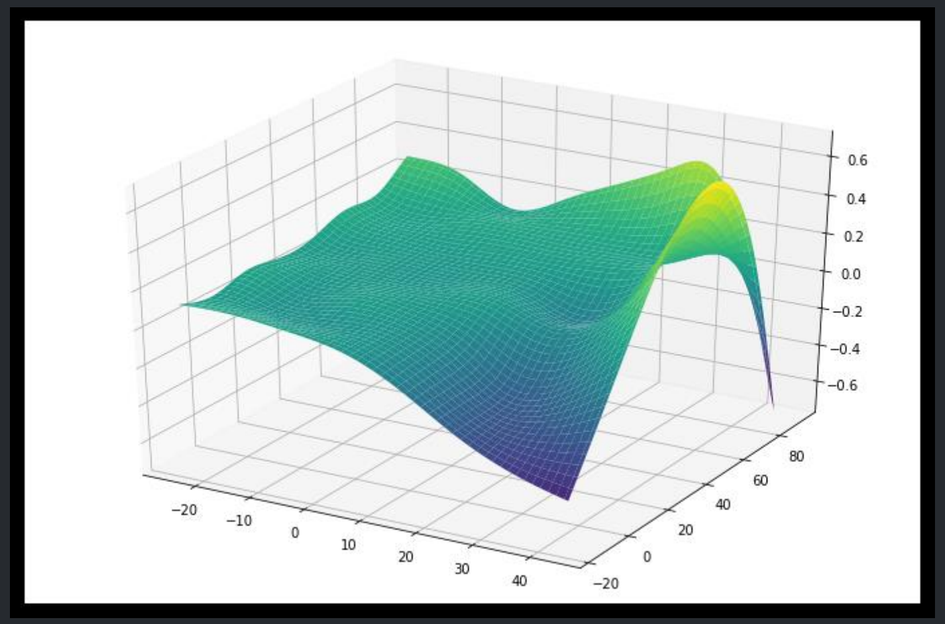


Для факторных (категориальных) переменных.

«Под капотом» использует one-hot encoding.

Также подвержен регуляризации, как и все остальные предикторы.

te()

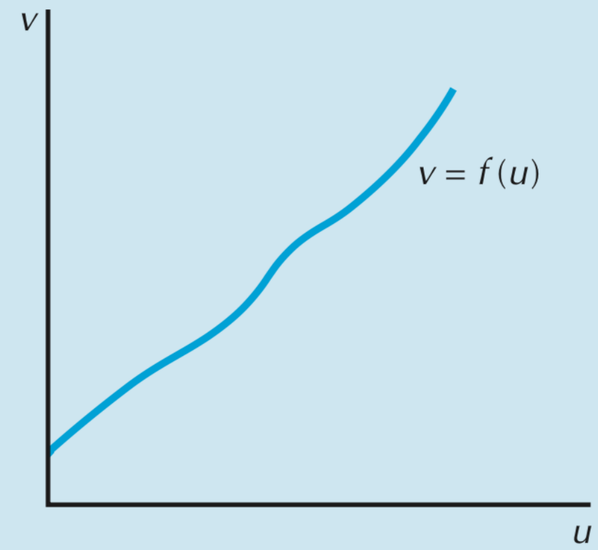


От слова «тензор» – генерализация матрицы на много измерений.

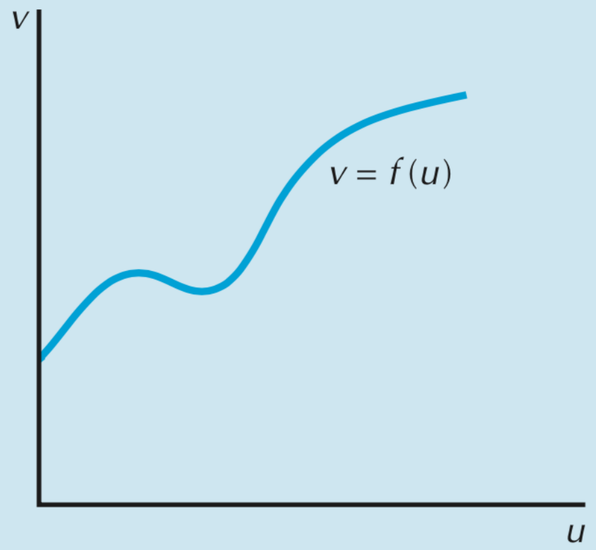
Полезно для кодирования взаимодействия нескольких предикторов (число произвольно).

Сложно рисовать.

ОГРАНИЧЕНИЯ ПО МОНОТОННОСТИ



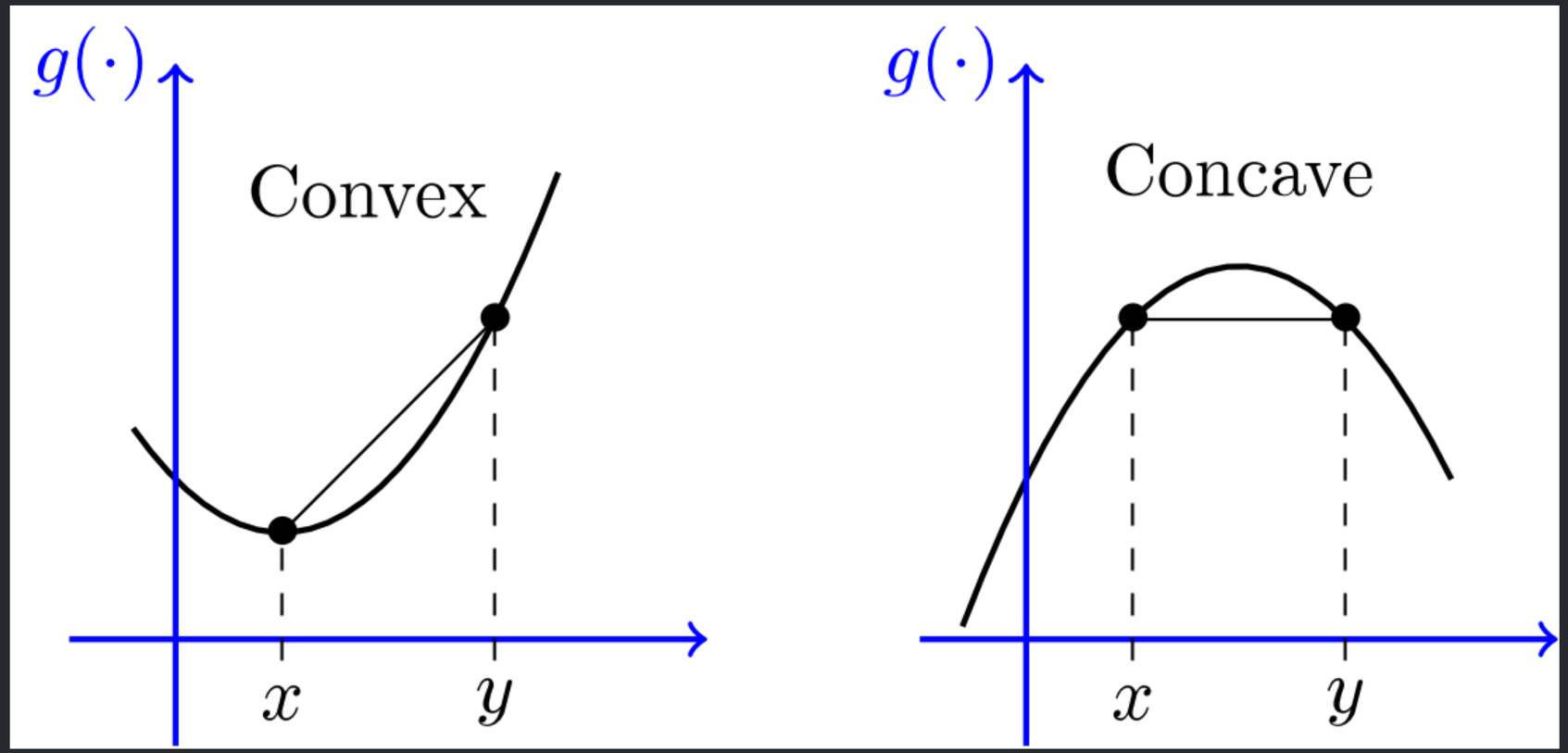
A



B

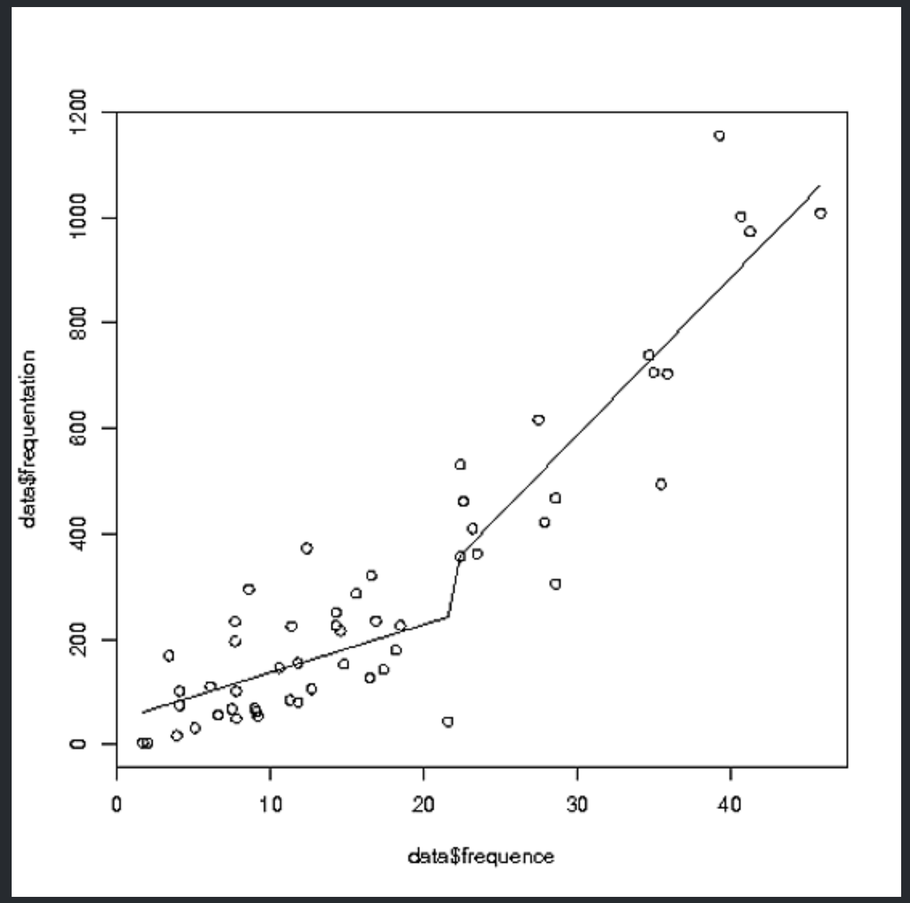
A positive monotonic transformation. Panel A illustrates a monotonic function—one that is always increasing. Panel B illustrates a function that is *not* monotonic, since it sometimes increases and sometimes decreases.

ВЫПУКЛЫЕ И ВОГНУТЫЕ ОГРАНИЧЕНИЯ



ПОХОЖИЕ МЕТОДЫ

СЕГМЕНТИРОВАННАЯ РЕГРЕССИЯ



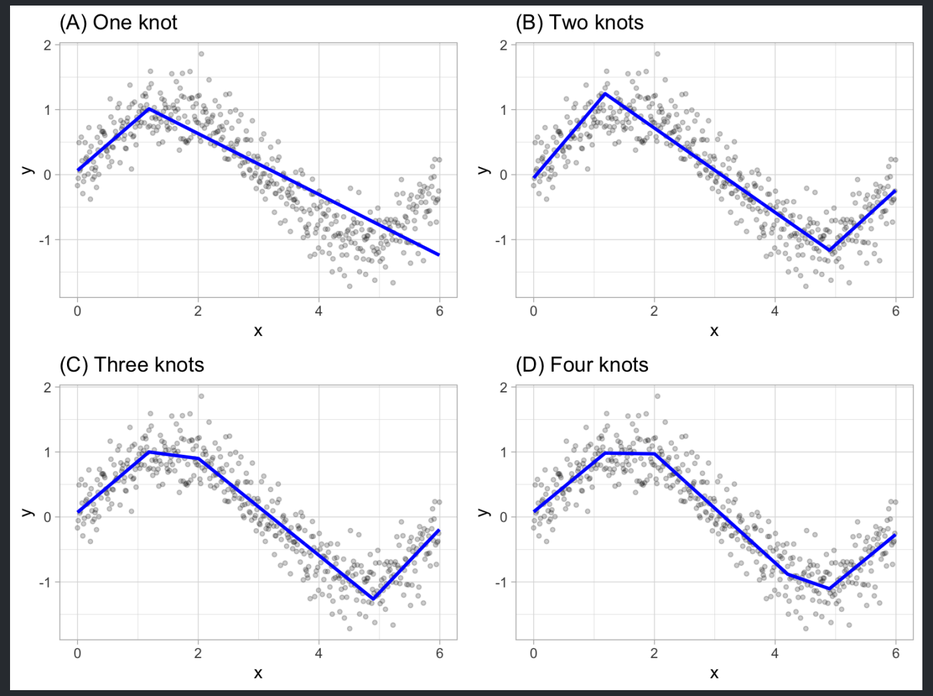
Также известна как piecewise или broken-stick.

Использует линейный базис, иногда отдельные элементы не соединены.

Места «перелома» представляют собой основной интерес.

Комбинируется с GLM.

MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)



Считается алгоритмом машинного обучения.

Также использует линейный базис.

В основе лежат т.н. «шарнирные функции» (hinge functions).

Автоматически отбирает переменные.

Непараметрический метод.

Документация, установка.

ВСЁ ПРОДОЛЖАЕТ СТАНОВИТЬСЯ СЛОЖНЕЕ

ВСЁ ПРОДОЛЖАЕТ СТАНОВИТЬСЯ СЛОЖНЕЕ

Standard Deviation Not Enough For Perverted Statistician