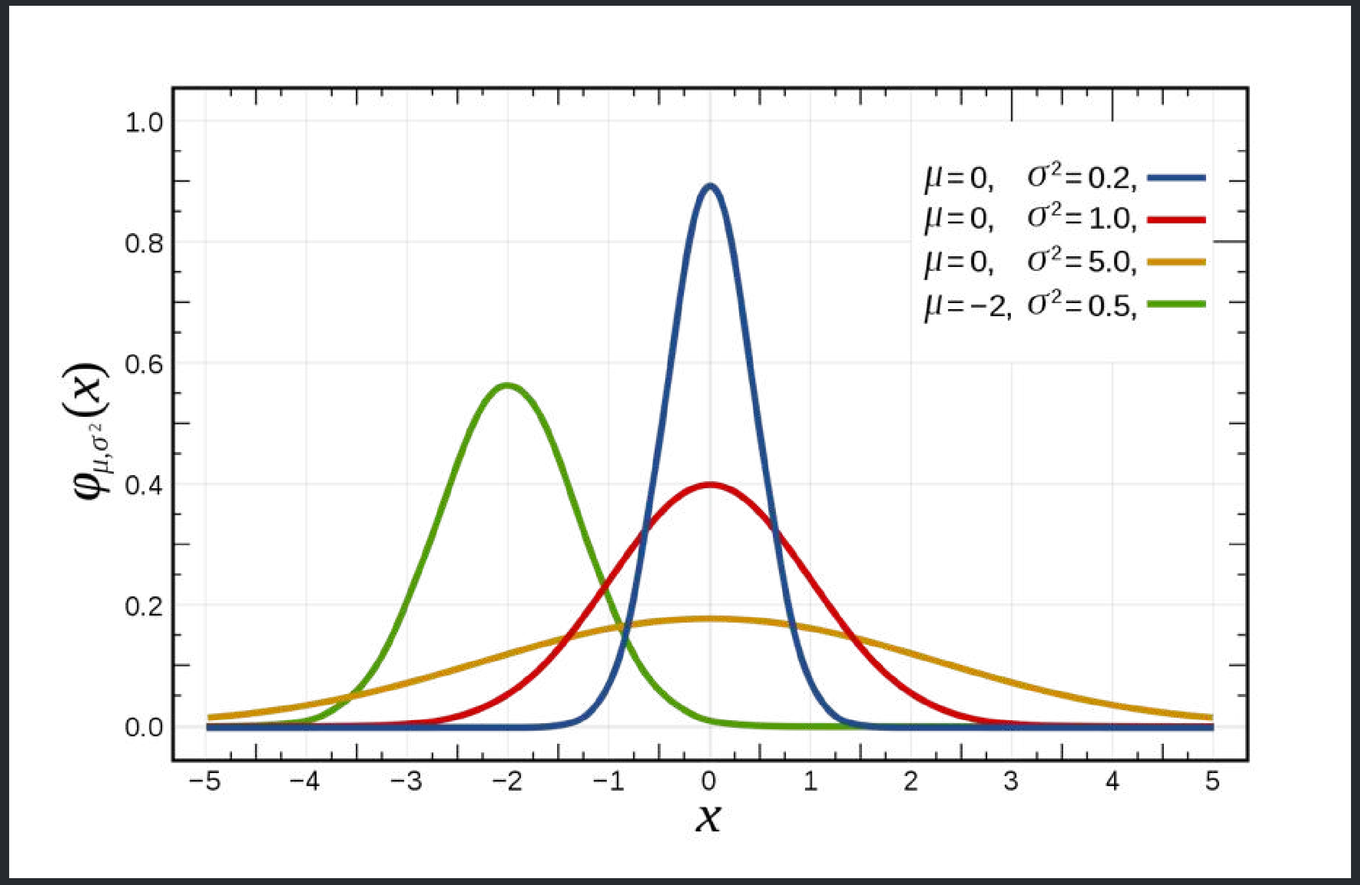
ВВЕДЕНИЕ В GLM

Что это такое и как всё становится хуже.

АЛЕКСАНДР МАНАЕНКОВ

ВСЁ СТАНОВИТСЯ ХУЖЕ

ЛИНЕЙНАЯ РЕГРЕССИЯ
LINEAR GAUSSIAN REGRESSION



ИМЯ РАСПРЕДЕЛЕНИЯ

Гауссовское / нормальное.

ЗНАЧЕНИЯ ЗП

(-∞; ∞).

ПАРАМЕТРЫ

μ – среднее.

σ – стандартное отклонение.

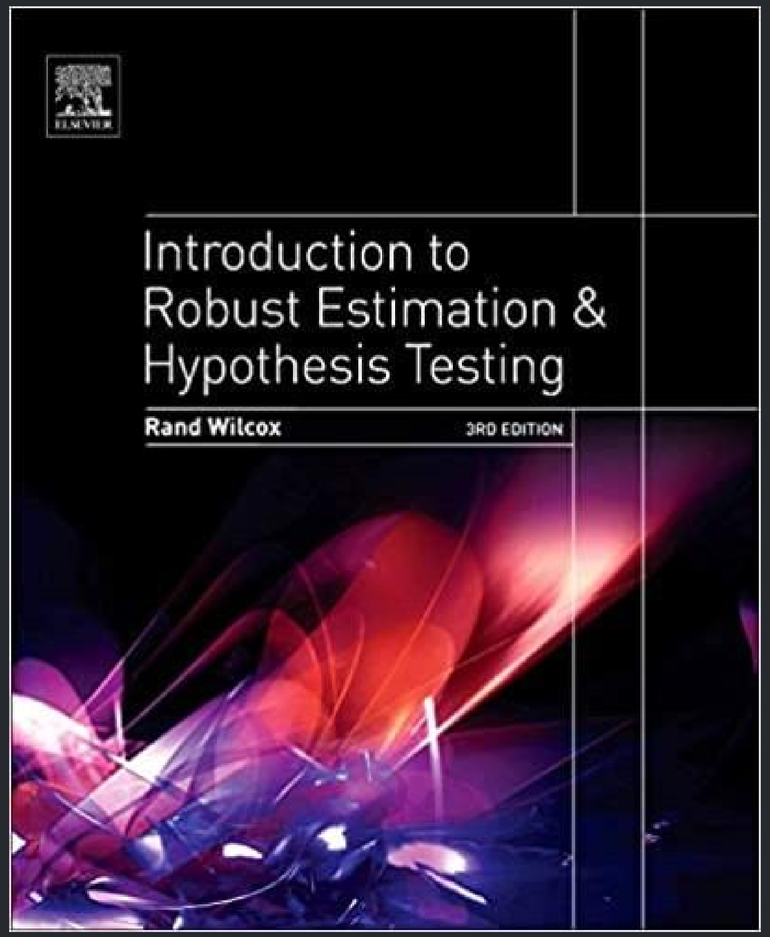
ТИПИЧНАЯ ФУНКЦИЯ СВЯЗИ

А это ещё что такое?

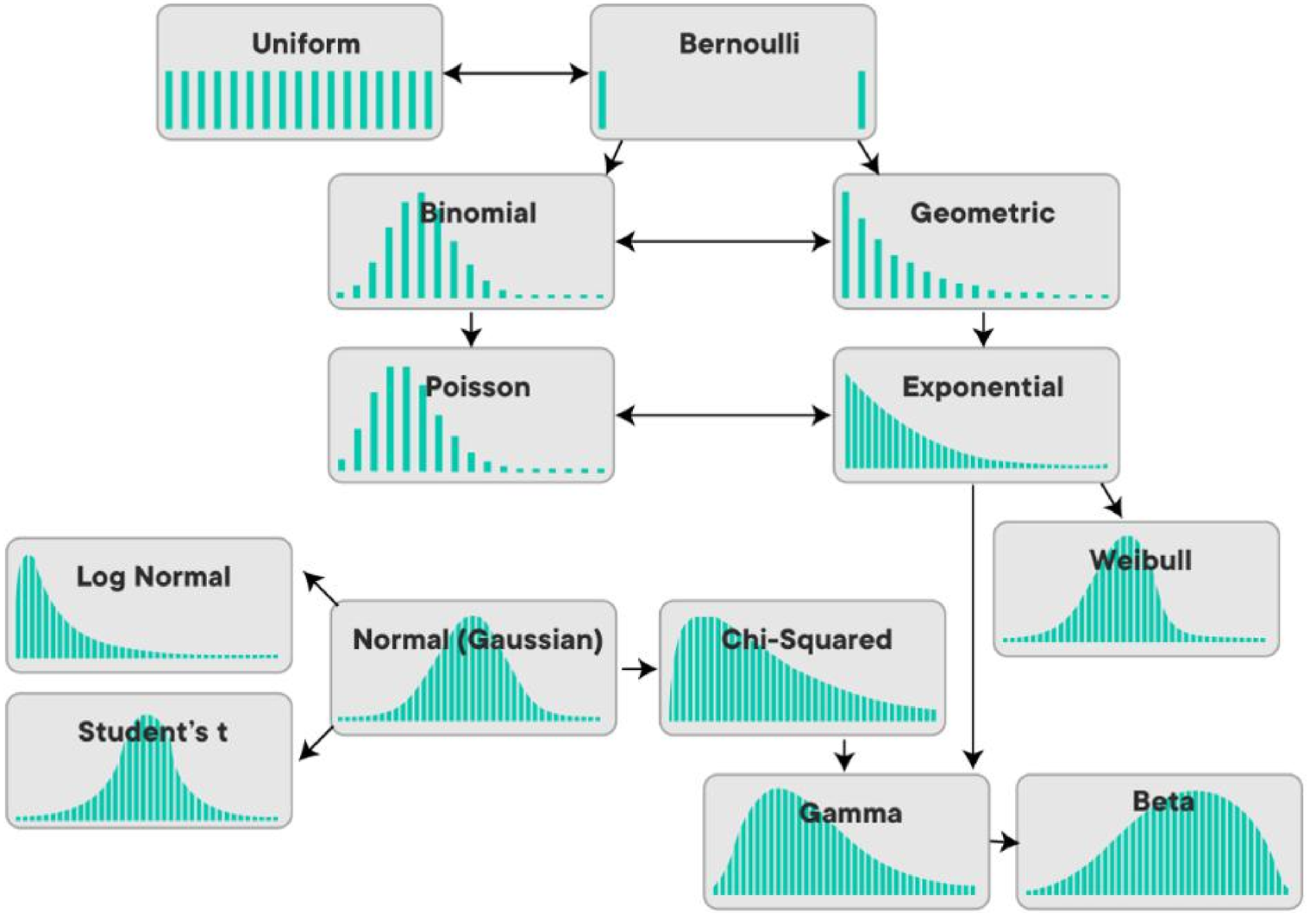
УСЛОВИЯ ПРИМЕНЕНИЯ

- Линейность взаимосвязи
- Нормальность распределения остатков модели
- Однородность дисперсии остатков на всех уровнях независимой переменной

УСТОЙЧИВЫЕ (РОБАСТНЫЕ) МЕТОДЫ



- Ранговые методы
- Квантильная регрессия
- Перестановочные тесты
- Sandwich estimators
- Iteratively Reweighted Least Squares
- Percentage-bend correlation
- Тесты Уэлча
- Тест Юэна
- Тонна других

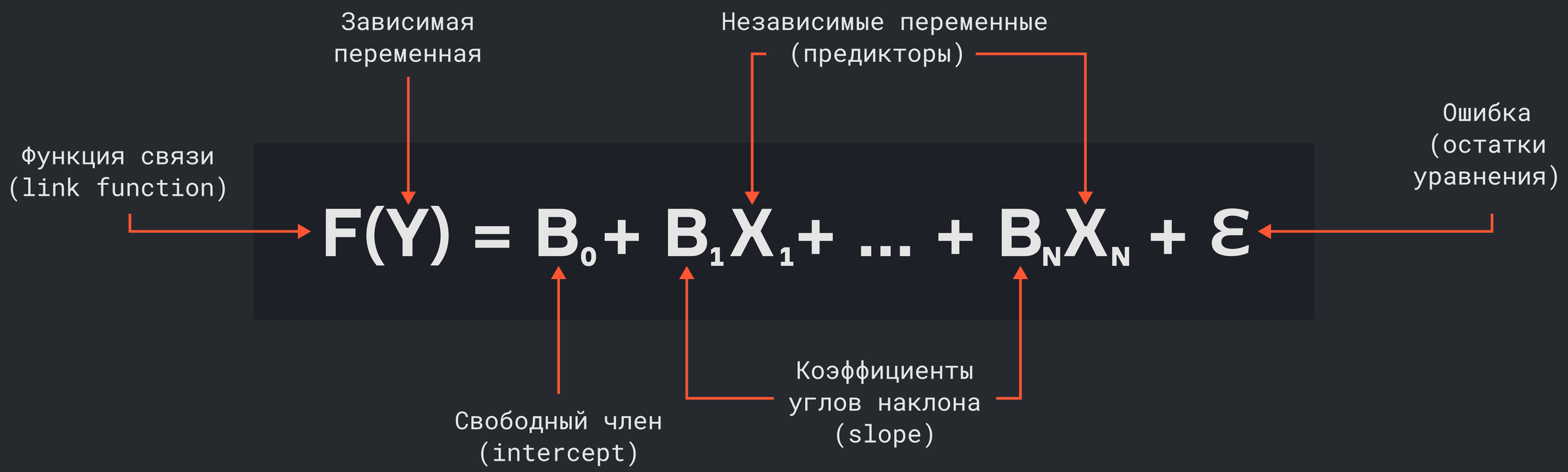

Uniform
Bernoulli
Binomial
Geometric
Poisson
Exponential
Weibull
Log Normal
Normal (Gaussian)
Chi-Squared
Student's t
Gamma
Beta

GLM

GENERALIZED LINEAR MODELS

ОБОБЩЁННЫЕ ЛИНЕЙНЫЕ МОДЕЛИ

ОБЩАЯ (УПРОЩЁННАЯ) ФОРМУЛА



ФУНКЦИЯ СВЯЗИ (LINK FUNCTION)

ПРЕОБРАЗУЕТ РАСПРЕДЕЛЕНИЕ ЗАВИСИМОЙ ПЕРЕМЕННОЙ ТАК, ЧТО:

- Оно принимает значения от -∞ до ∞
- Связь зависимой переменной с предикторами линейна

ПРИМЕРЫ ФУНКЦИЙ СВЯЗИ:

- Identity – без преобразования («функция связи» обычной линейной регрессии).
- Логарифмирование – данные не должны быть отрицательными или нулевыми, компенсирует правостороннюю асимметрию.
- Обратная функция (1/x) – данные могут быть отрицательными, но не могут быть нулевыми, меняет асимметрию на противоположную.
- Квадратный корень – только положительные данные, компенсирует правостороннюю асимметрию.
- Много, много других.

НОМИНАТИВНЫЕ ДАННЫЕ (2 КАТЕГОРИИ)

ПРИМЕРЫ

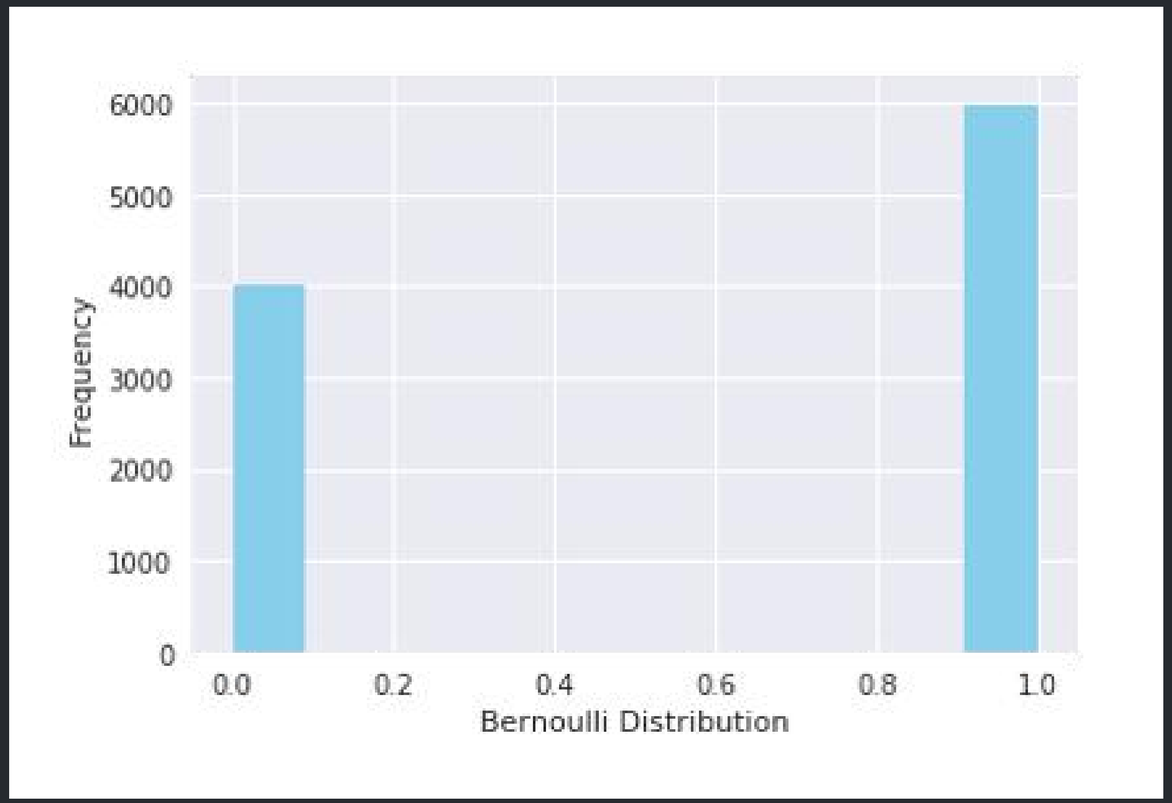
- Подписался ли клиент на сервис или отказался?
- Заболел ли человек или остался здоров?
- Красная таблетка или синяя?

ОСОБЕННОСТИ

- Нет градаций, данные дискретны
- За их пределами нет вообще ничего
- Работают не с самими значениями данных, а с их вероятностью

БИНОМИАЛЬНАЯ РЕГРЕССИЯ (BINOMIAL)

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ LOGISTIC



ИМЯ РАСПРЕДЕЛЕНИЯ

Бернулли (биномиальное с n = 1).

ЗНАЧЕНИЯ ЗП

(0; 1).

ПАРАМЕТРЫ

p – вероятность успеха

n – количество попыток

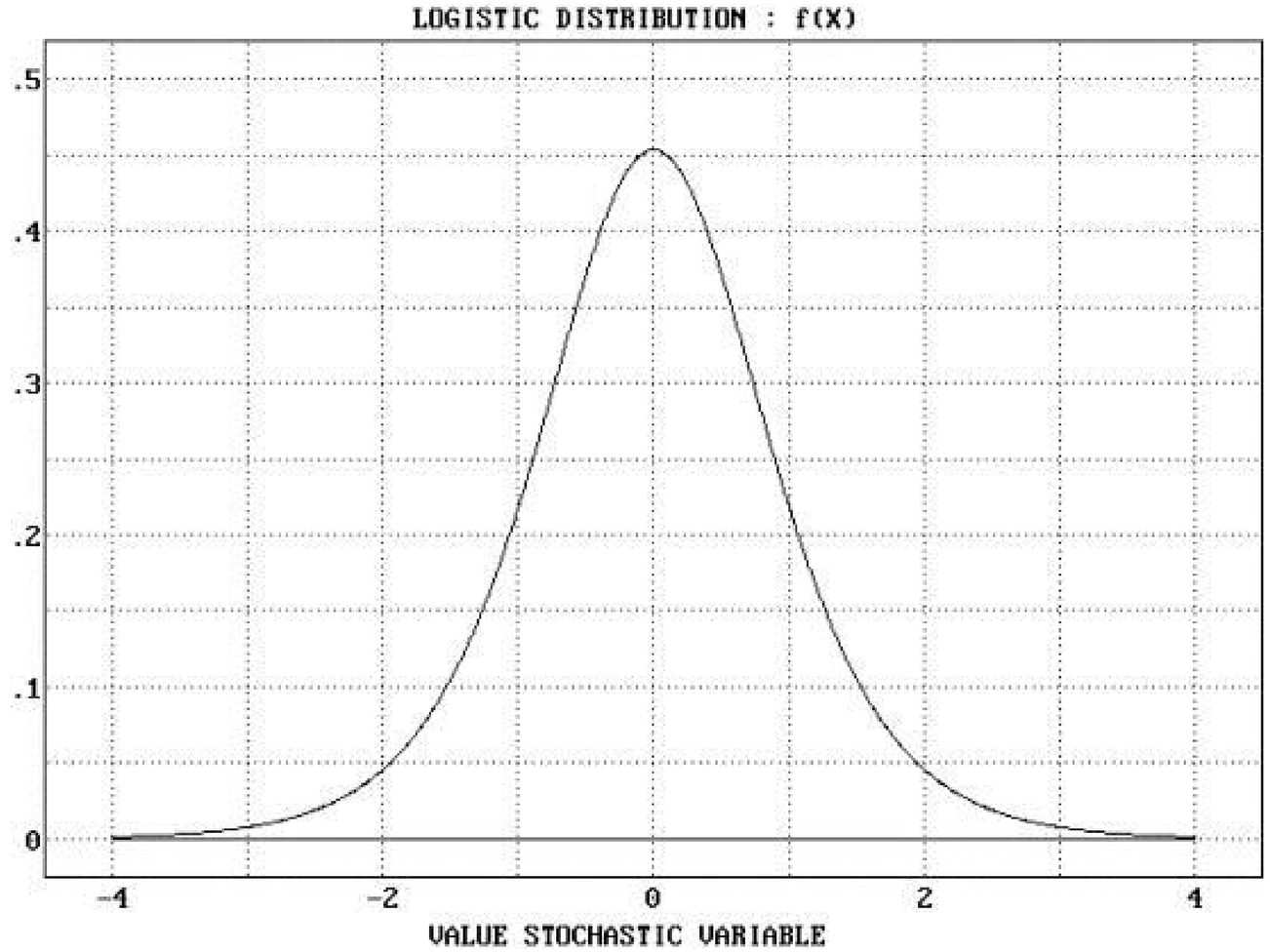
ТИПИЧНАЯ ФУНКЦИЯ СВЯЗИ

Логит

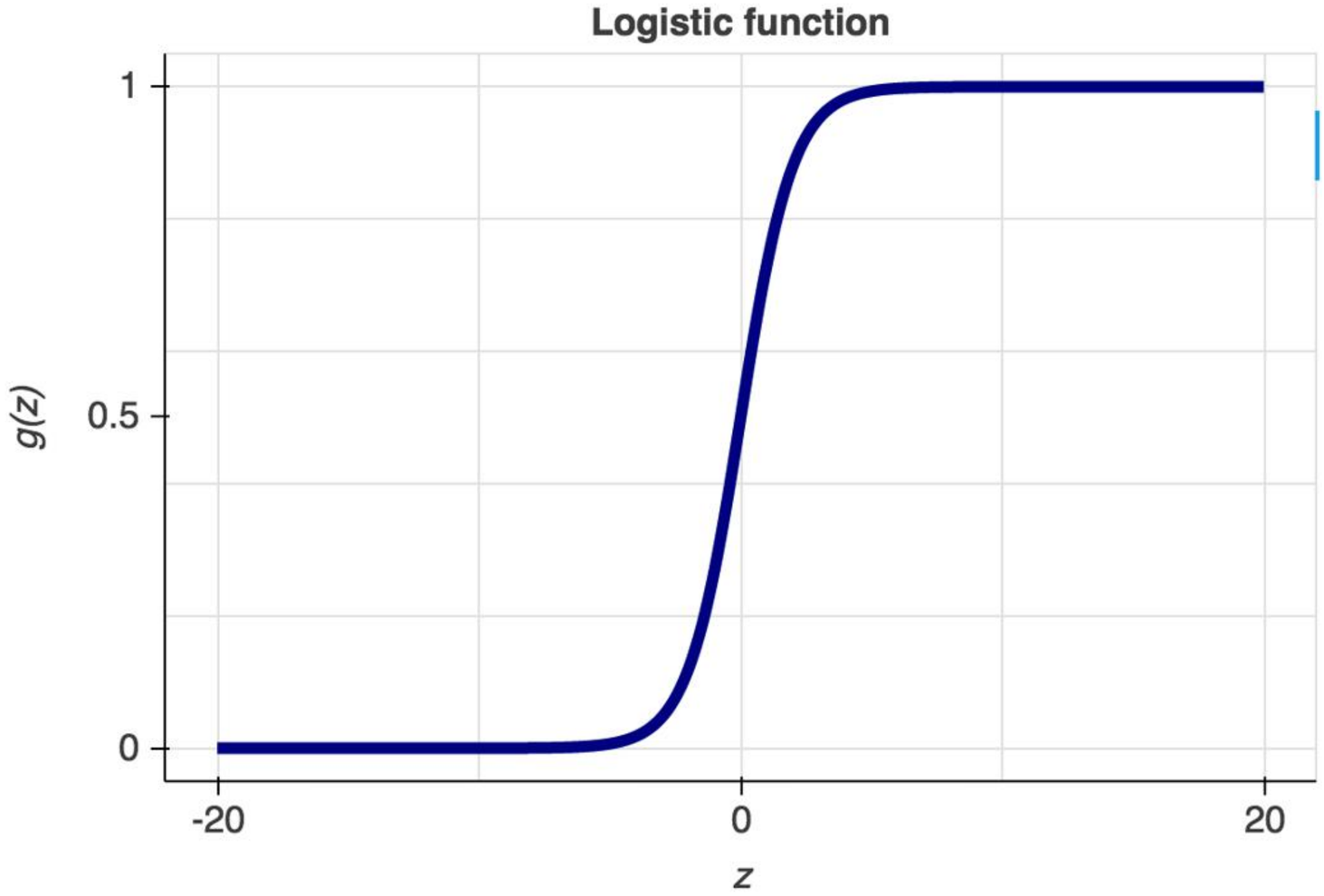
КАК СЧИТАТЬ ЛОГИТ

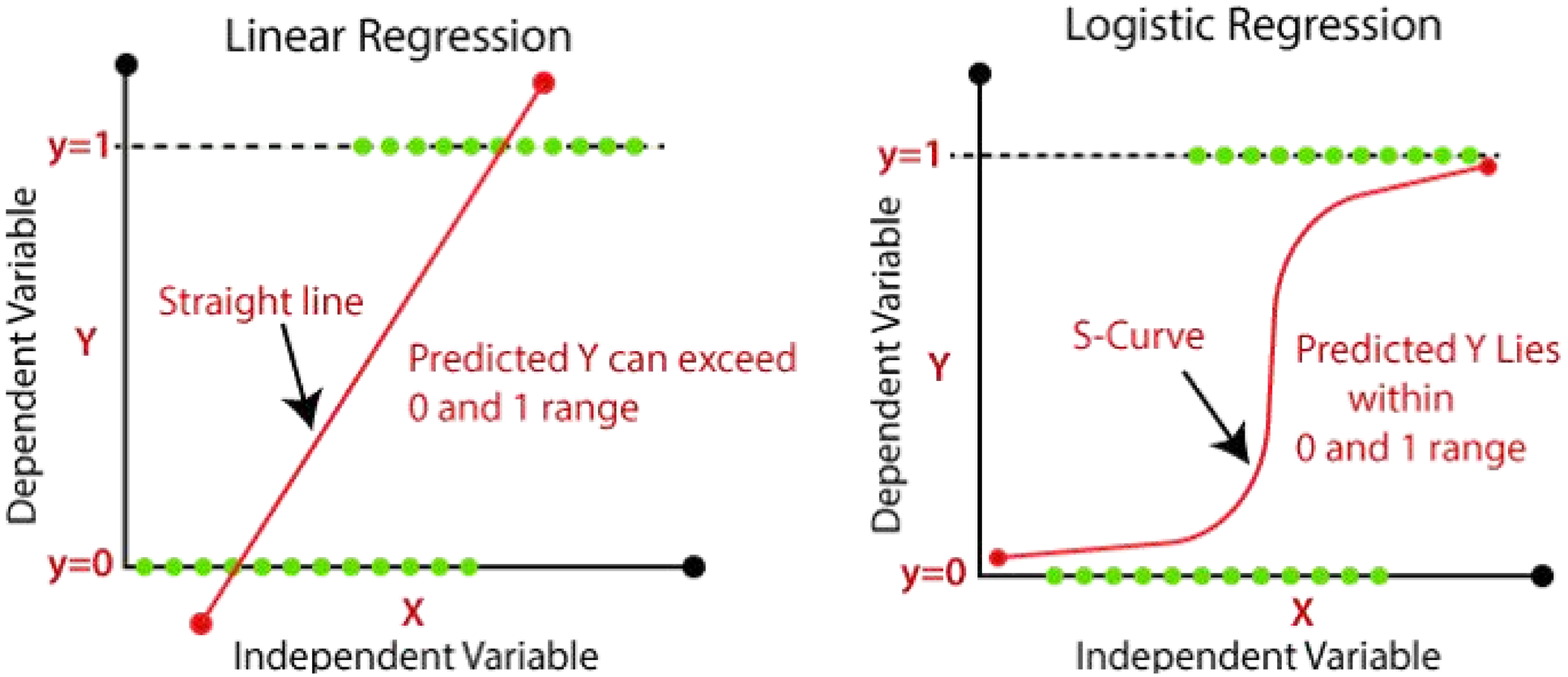
ЭЛЕМЕНТ ФОРМУЛЫ	ИНТЕРПРЕТАЦИЯ И ГРАНИЦЫ	ПРИМЕР
q	Вероятность успеха (0; 1)	0.75
$\frac{q}{1 - q}$	Шансы успеха (0; ∞)	$\frac{0.75}{1-0.75} = \frac{0.75}{0.25} = \frac{3}{1}$
$\ln\left(\frac{q}{1 - q}\right)$	Логит (-∞; ∞)	ln(3) = 1.097

ЛОГИСТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ



ЛОГИСТИЧЕСКАЯ ФУНКЦИЯ

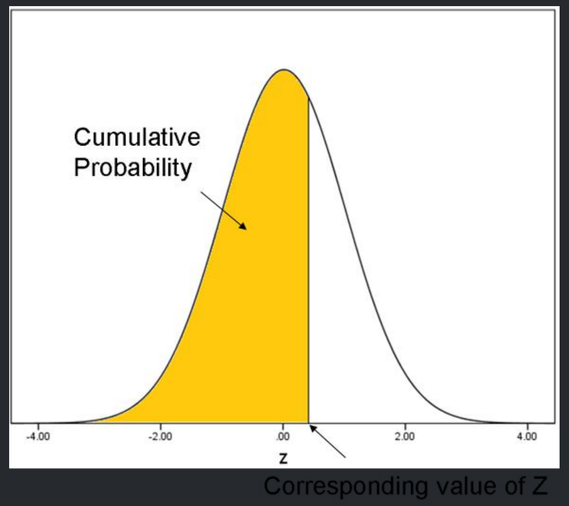



Linear Regression
y=1
y=0
Dependent Variable
Y
X
Independent Variable
Straight line
Predicted Y can exceed
0 and 1 range
Logistic Regression
y=1
y=0
Dependent Variable
Y
X
Independent Variable
S-Curve
Predicted Y Lies
within
0 and 1 range

БИНОМИАЛЬНАЯ РЕГРЕССИЯ = ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ?

ПРОБИТ-РЕГРЕССИЯ (PROBIT)

- Способ перевести вероятности в нормальное распределение (среднее 0, стандартное отклонение 1).
- В данном случае используется обратная функция кумулятивного стандартного нормального распределения.

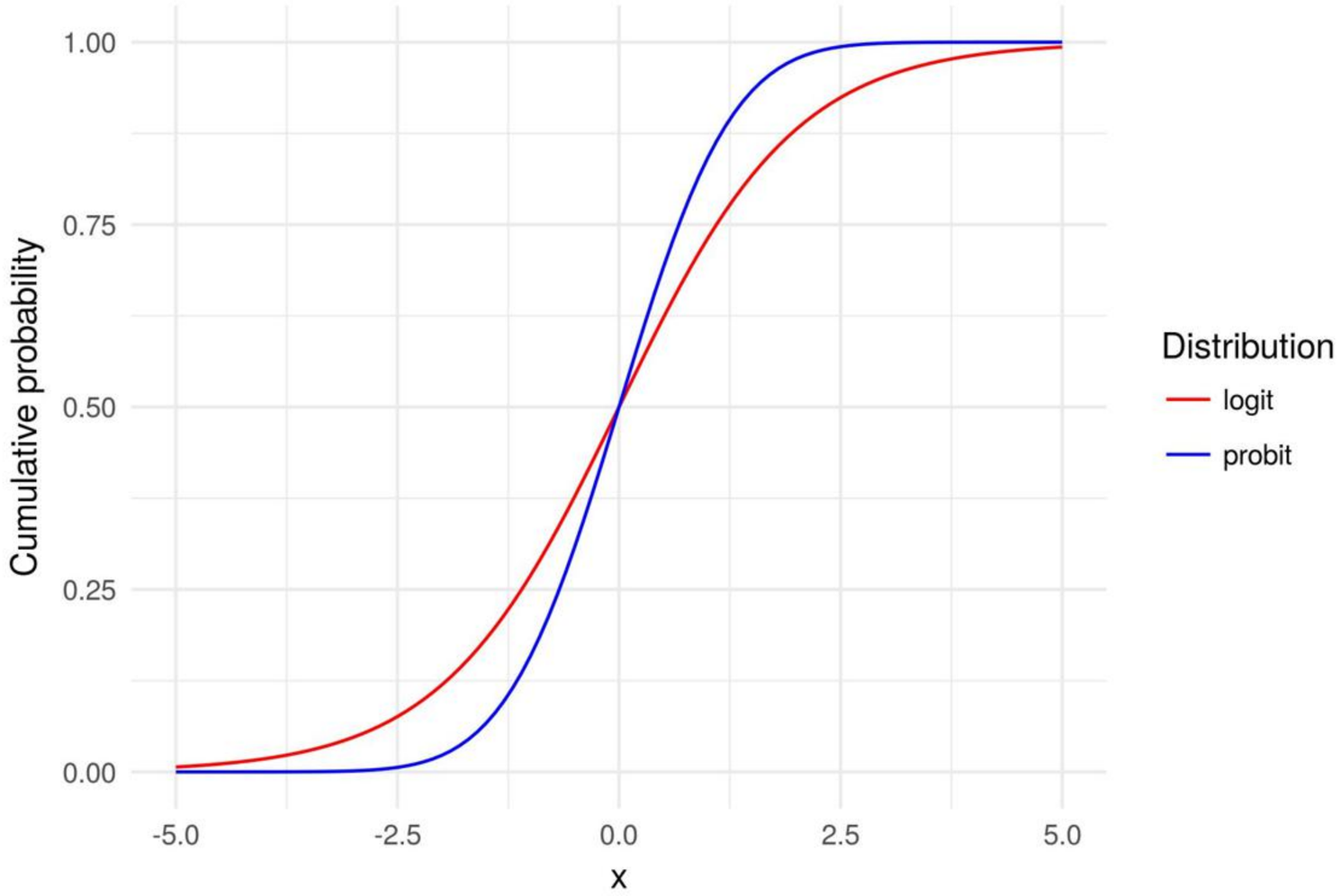


$\Phi^{-1}(p)$

Согласно моей модели, шансы позитивного исхода в этом случае ощутимо вырастут.

Согласно моей модели, в этом случае z-оценка данной величины будет большей, что отражает сдвиг с единицах стандартного отклонения нормального распределения... Ну куда вы уходите 😞

ЛОГИСТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ ПО СРАВНЕНИЮ С НОРМАЛЬНЫМ



НОМИНАТИВНЫЕ ДАННЫЕ (3 КАТЕГОРИИ И БОЛЬШЕ)

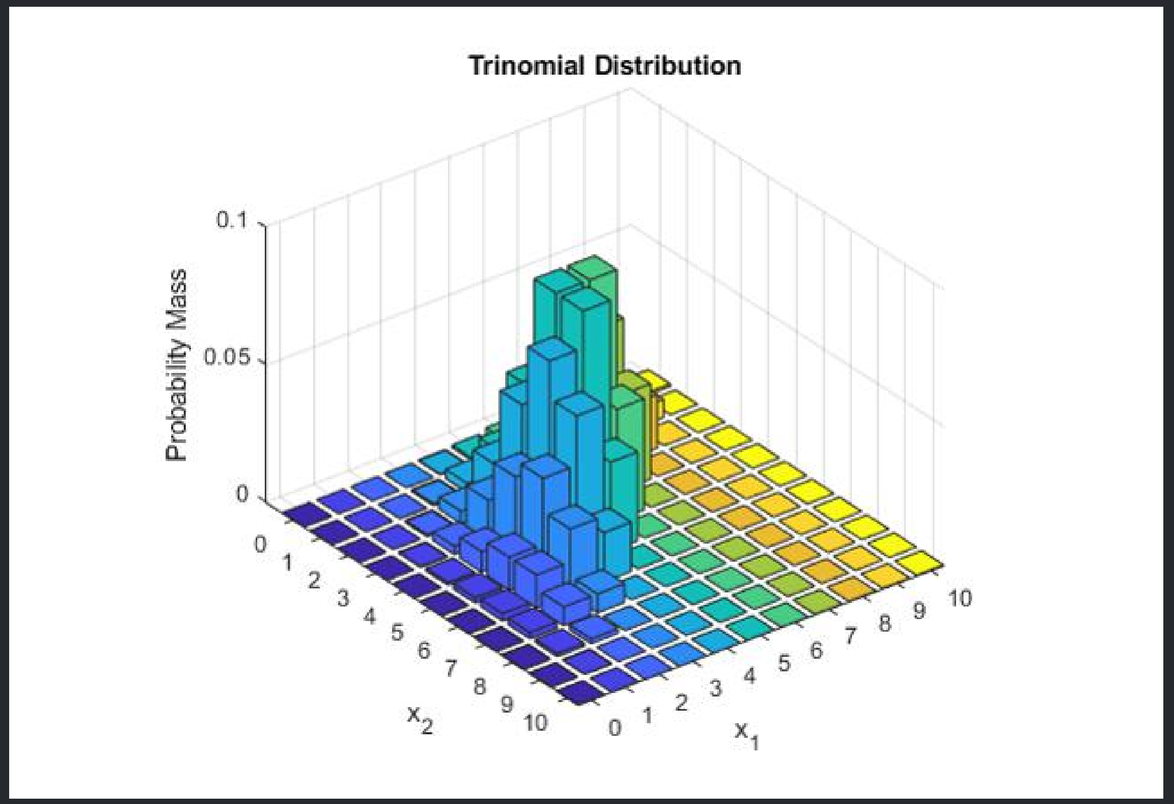
ПРИМЕРЫ

- Какой тариф у мобильного оператора использует клиент?
- В какую сеть магазинов чаще ходит покупатель?
- В каком районе предпочитают жить люди?

ОСОБЕННОСТИ

- Такие же, как и у бинарных номинативных данных.
- Вероятность нужно рассчитывать для каждой категории отдельно.

МУЛЬТИНОМИАЛЬНАЯ РЕГРЕССИЯ MULTINOMIAL



ИМЯ РАСПРЕДЕЛЕНИЯ

Мультиномиальное.

ЗНАЧЕНИЯ ЗП

(0; 1).

ПАРАМЕТРЫ

p_1 ... p_k – вероятность каждого события.

n – количество попыток.

ТИПИЧНАЯ ФУНКЦИЯ СВЯЗИ

Мультиномиальный логит.

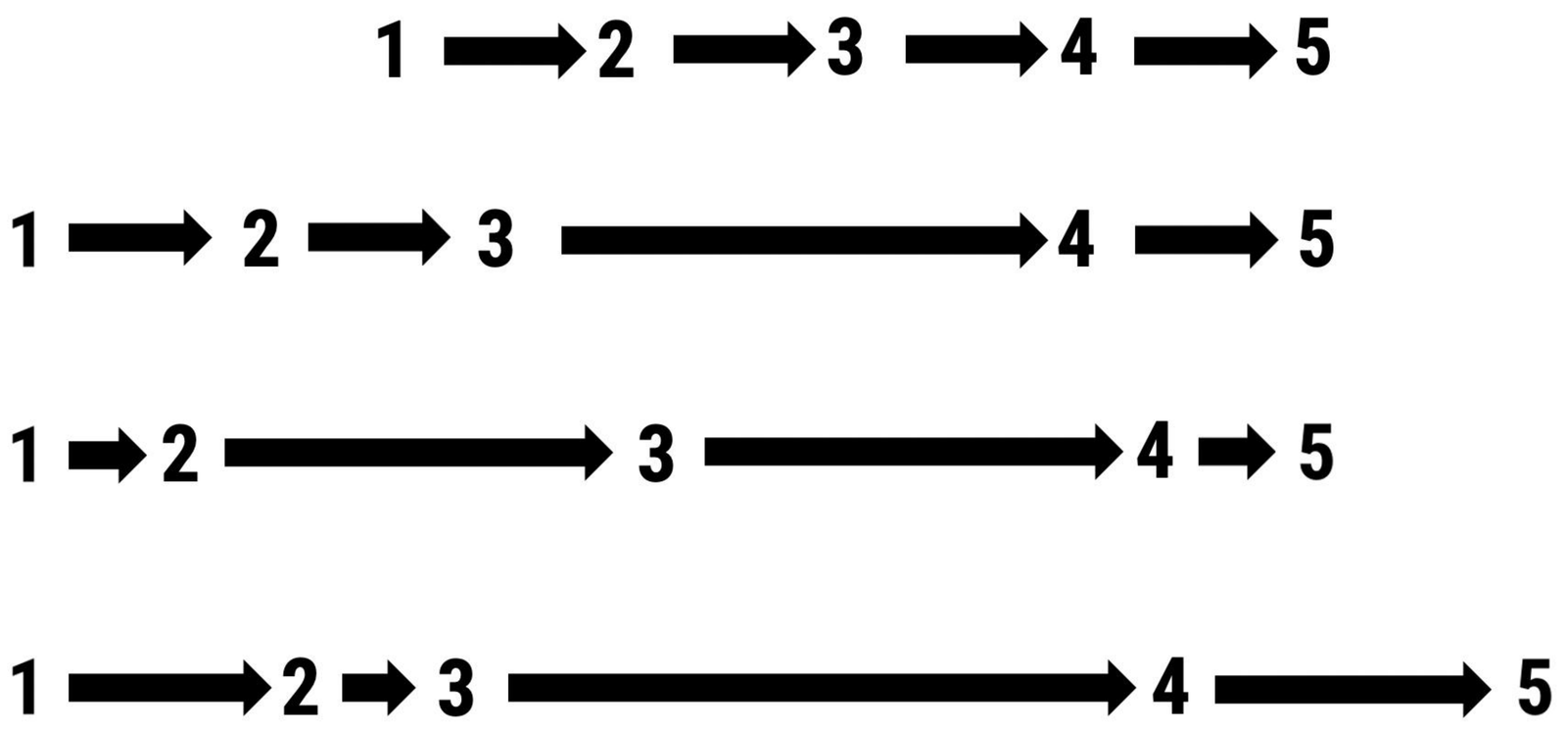
ПОРЯДКОВЫЕ ДАННЫЕ

ПРИМЕРЫ

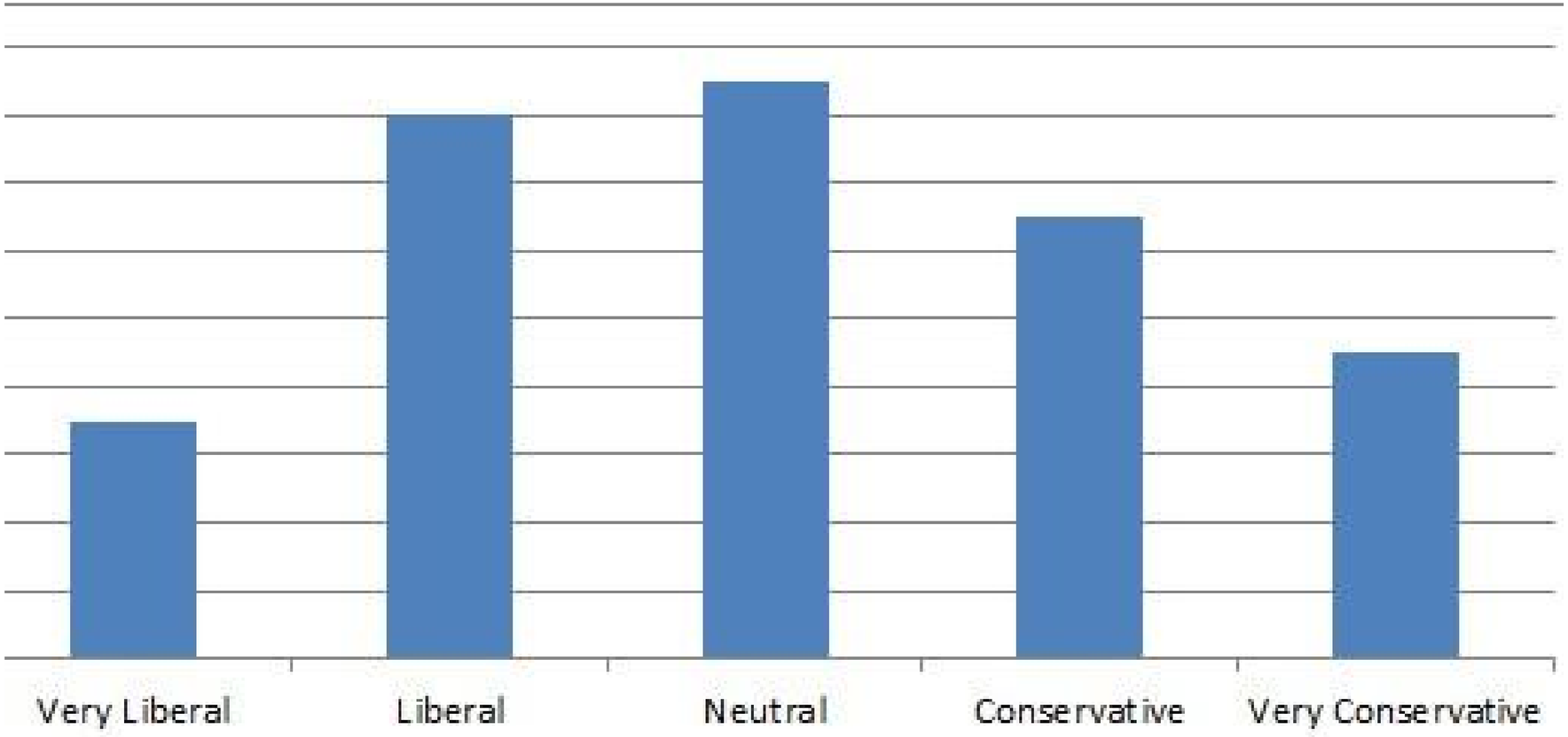
- Шкала опросника (совершенно не согласен – совершенно согласен).
- Рейтинги – на сколько баллов оценили сервис?
- В какой возрастной категории испытуемый?

ОСОБЕННОСТИ

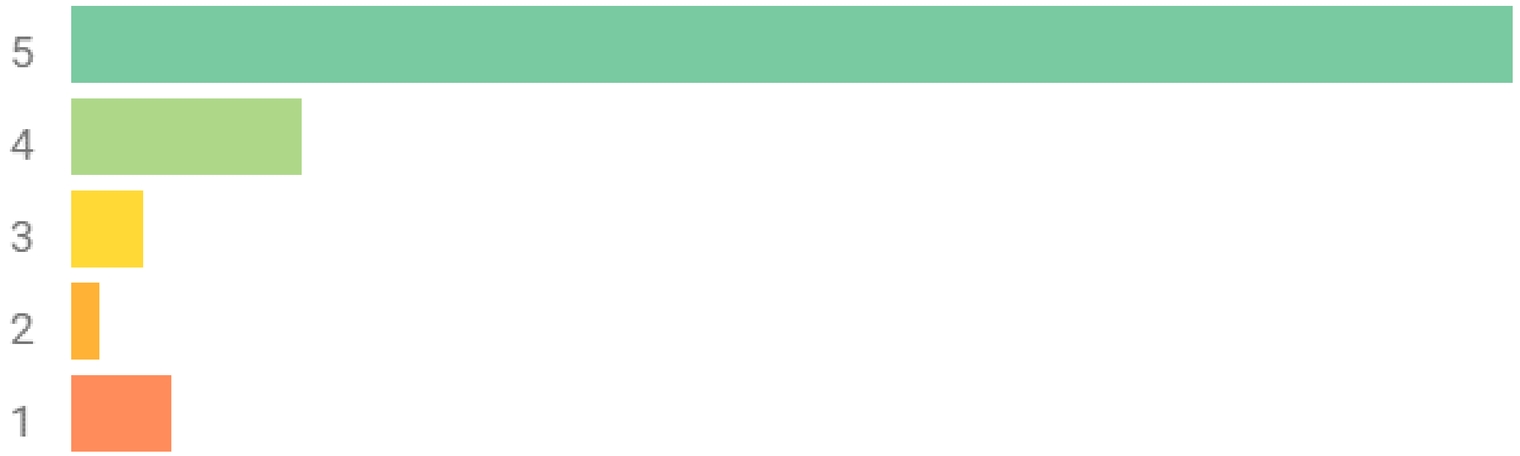
- Данные дискретны.
- Есть порядок в категориях – какие-то категории больше или меньше других.
- Не можем понять, на сколько больше или меньше работают не с самими значениями данных, а с их вероятностью.


1 2 3 4 5
1 2 3 4 5
1 2 3 4 5
1 2 3 4 5

ШКАЛА ЛАЙКЕРТА

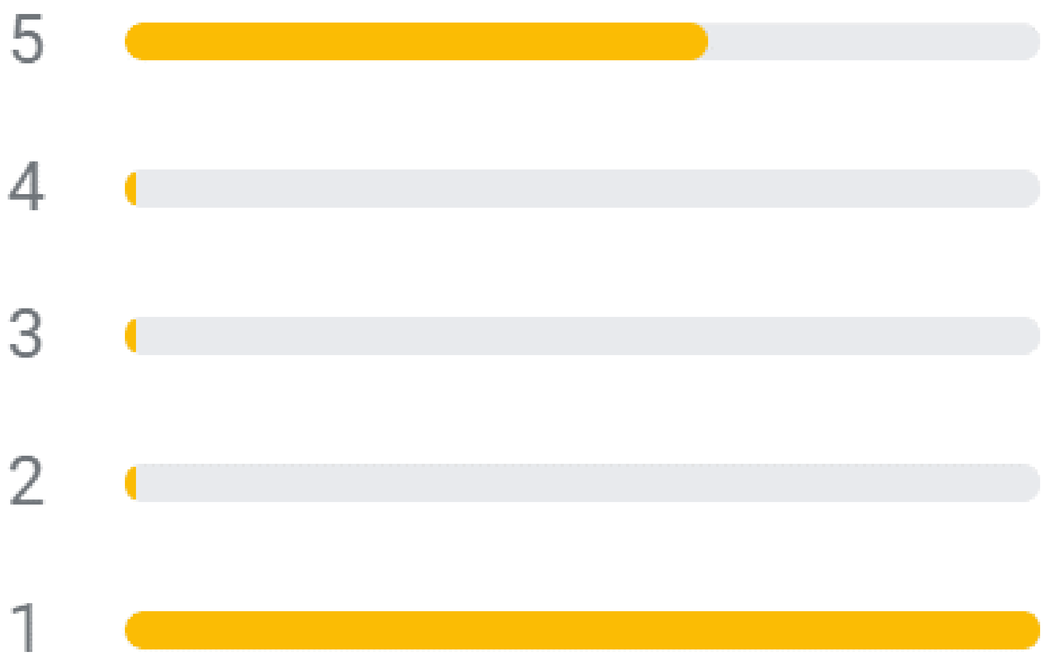


РЕЙТИНГ ПРИЛОЖЕНИЯ «SPOTIFY»



РЕЙТИНГ ПРИЛОЖЕНИЯ «СОЦИАЛЬНЫЙ МОНИТОРИНГ»

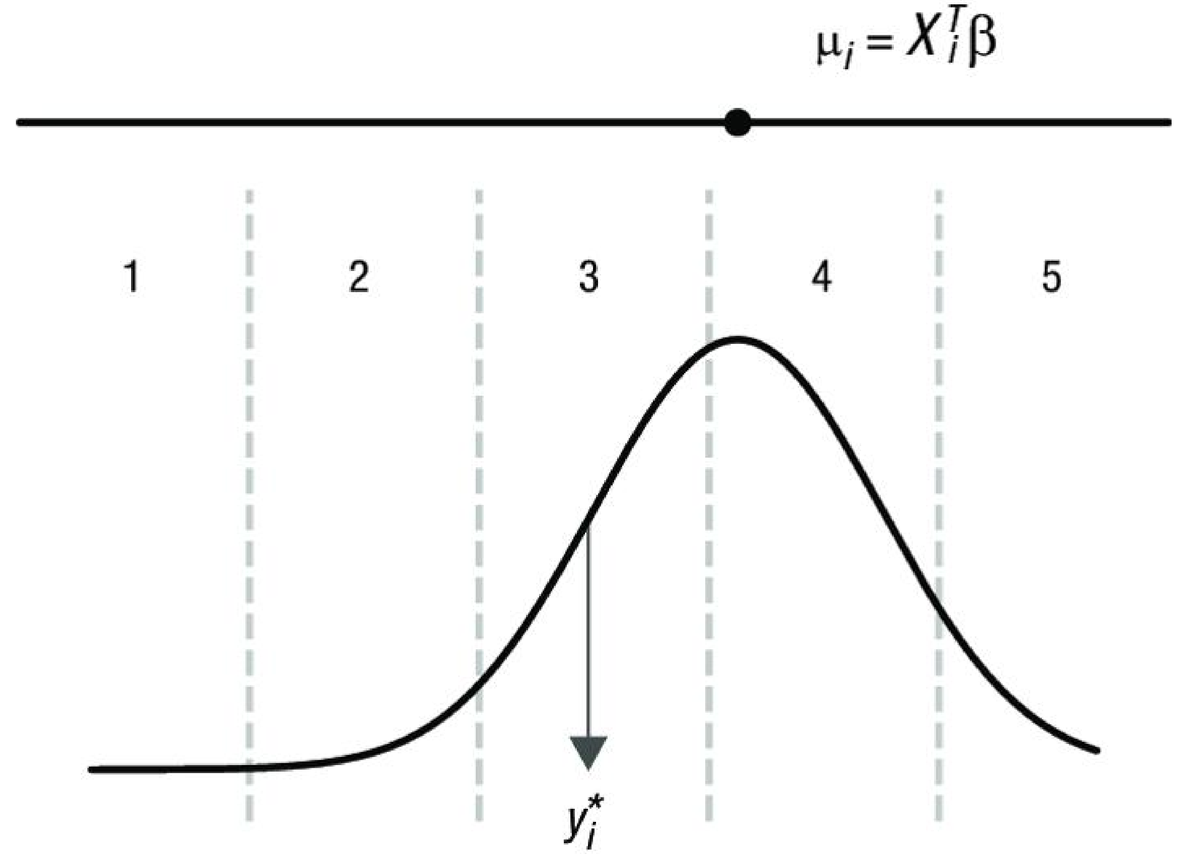
РЕЙТИНГ ВОЕНКОМАТА РАЙОНА КОПТЕВО



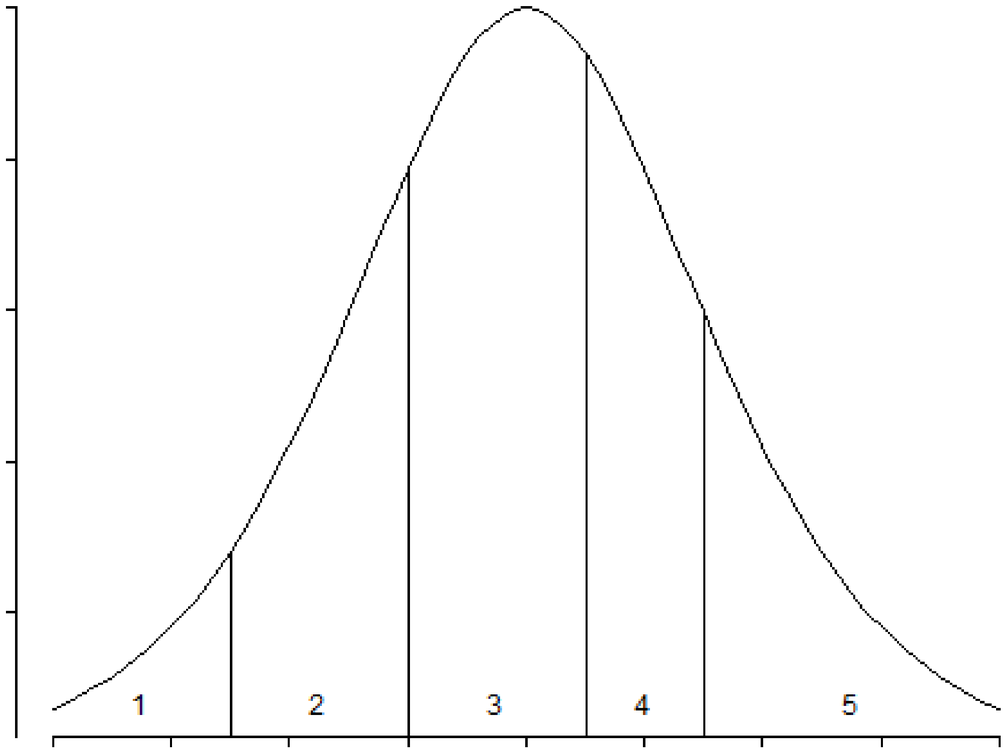
ЧТО ДЕЛАТЬ?

- Если есть возможность – не анализируйте такие данные.
- Если приходится, но гипотезы простые – пользуйтесь ранговыми непараметрическими методами.
- Если иначе, то остаётся только GLM.

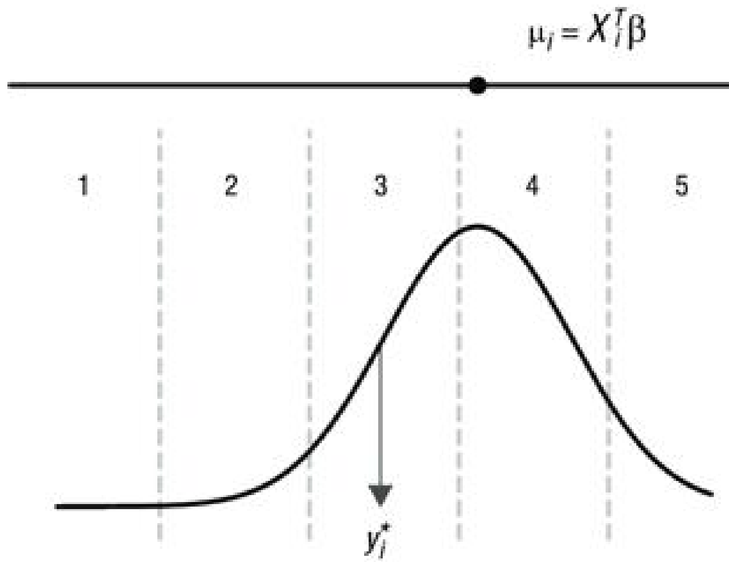
ПОРОГИ В РАСПРЕДЕЛЕНИИ (THRESHOLDS)



ПОРОГИ В РАСПРЕДЕЛЕНИИ (THRESHOLDS)



ПОРЯДКОВАЯ РЕГРЕССИЯ ORDINAL



ИМЯ РАСПРЕДЕЛЕНИЯ

Кумулятивное пороговое.

ЗНАЧЕНИЯ ЗП

(0; 1).

ПАРАМЕТРЫ

p_1 ... p_k – вероятность каждой категории.

ТИПИЧНАЯ ФУНКЦИЯ СВЯЗИ

Порядковый логит.

КОЛИЧЕСТВА (COUNT DATA)

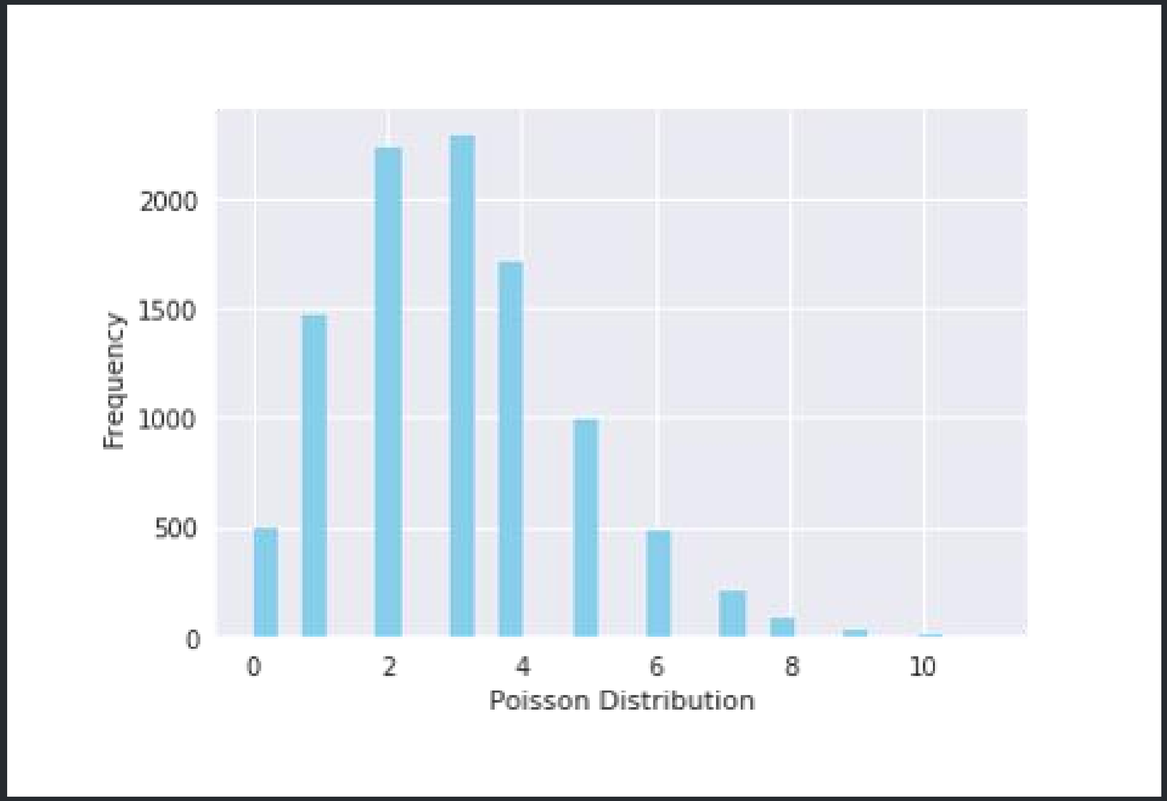
ПРИМЕРЫ

- Куда чаще всего едут на такси?
- Какое количество молока покупают ежемесячно?
- Сколько истребителей было на тихоокеанском театре военных действий?

ОСОБЕННОСТИ

- Данные дискретны.
- Данные неотрицательные.
- Чаще всего наблюдается правосторонняя асимметрия.

РЕГРЕССИЯ ПУАССОНА POISSON



ИМЯ РАСПРЕДЕЛЕНИЯ

Пуассоновское.

ЗНАЧЕНИЯ ЗП

(0; ∞).

ПАРАМЕТРЫ

λ – темп (rate).

ТИПИЧНАЯ ФУНКЦИЯ СВЯЗИ

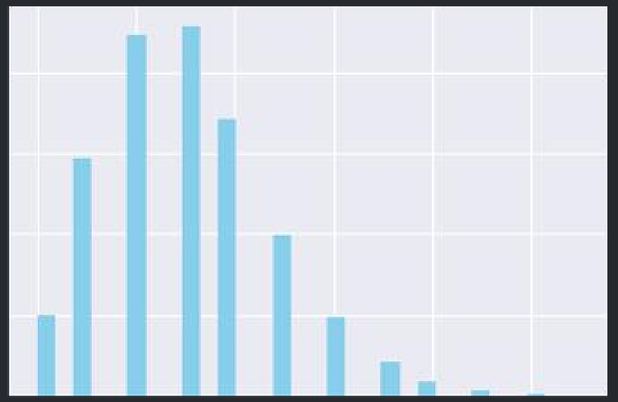
Логарифм.

ПРОБЛЕМЫ С ДИСПЕРСИЕЙ

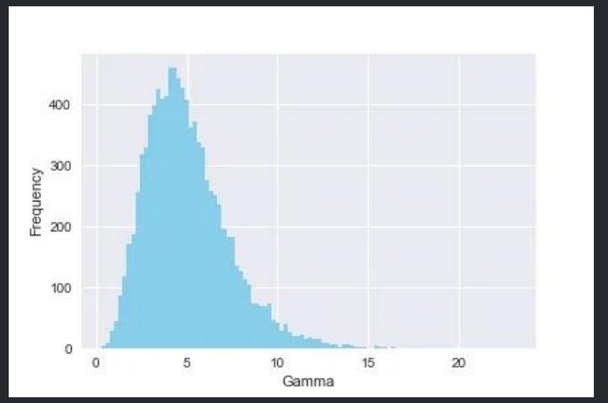
На среднее и дисперсию у распределения Пуассона один параметр (λ).

- Если дисперсия в реальности меньше среднего — недодисперсия (underdispersion, встречается редко).
- Если дисперсия в реальности больше среднего — сверхдисперсия (overdispersion, очень частый случай).

В обоих случаях наша модель неадекватна данным!



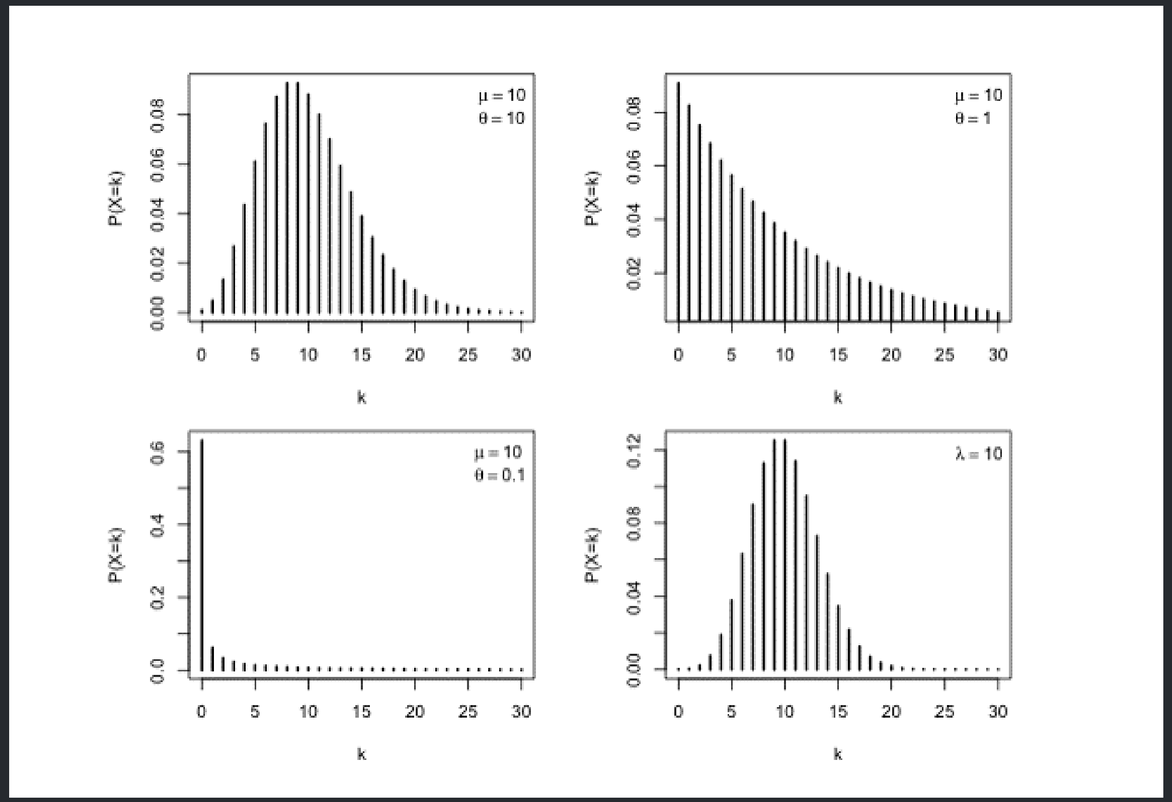
ПУАССОНА



ГАММА

ОТРИЦАТЕЛЬНОЕ БИНОМИАЛЬНОЕ

ОТРИЦАТЕЛЬНО-БИНОМИАЛЬНАЯ РЕГРЕССИЯ (NEGATIVE BINOMIAL)



ИМЯ РАСПРЕДЕЛЕНИЯ

Отрицательное биномиальное.

ЗНАЧЕНИЯ ЗП

(0; ∞).

ПАРАМЕТРЫ

μ – среднее.

θ/α – форма/дисперсия.

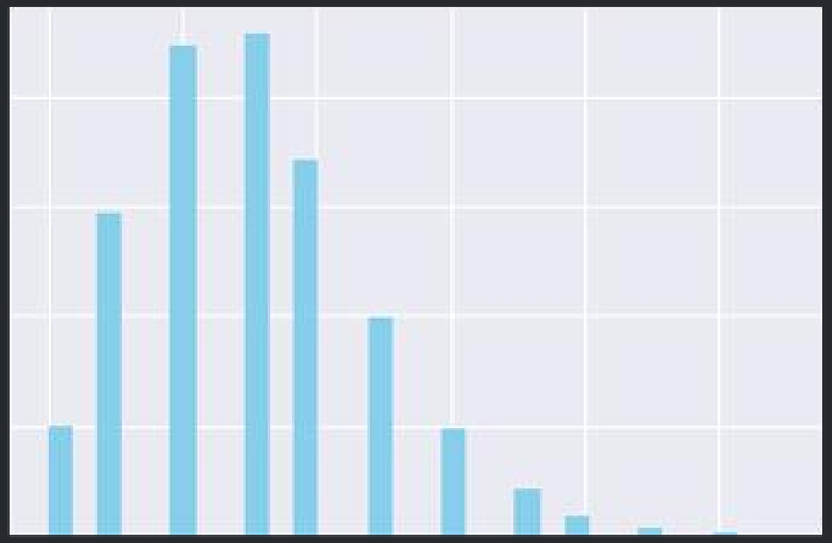
ТИПИЧНАЯ ФУНКЦИЯ СВЯЗИ

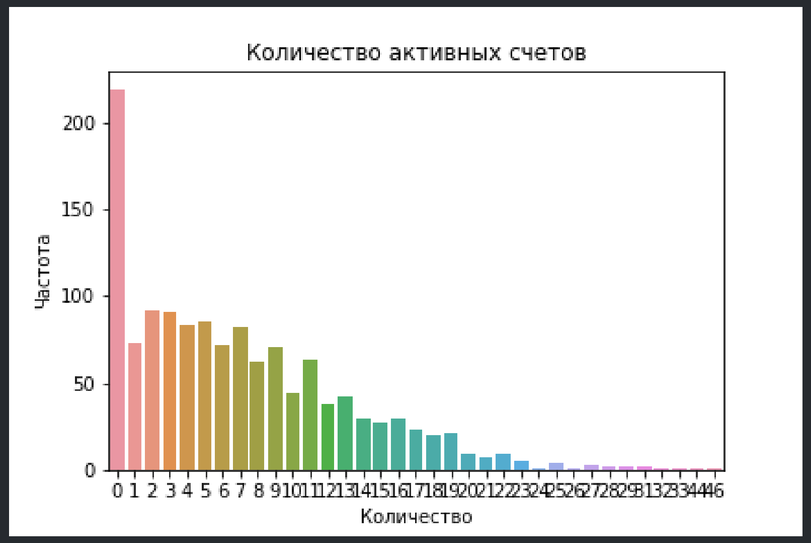
Логарифм.

ПРОБЛЕМЫ С НУЛЯМИ

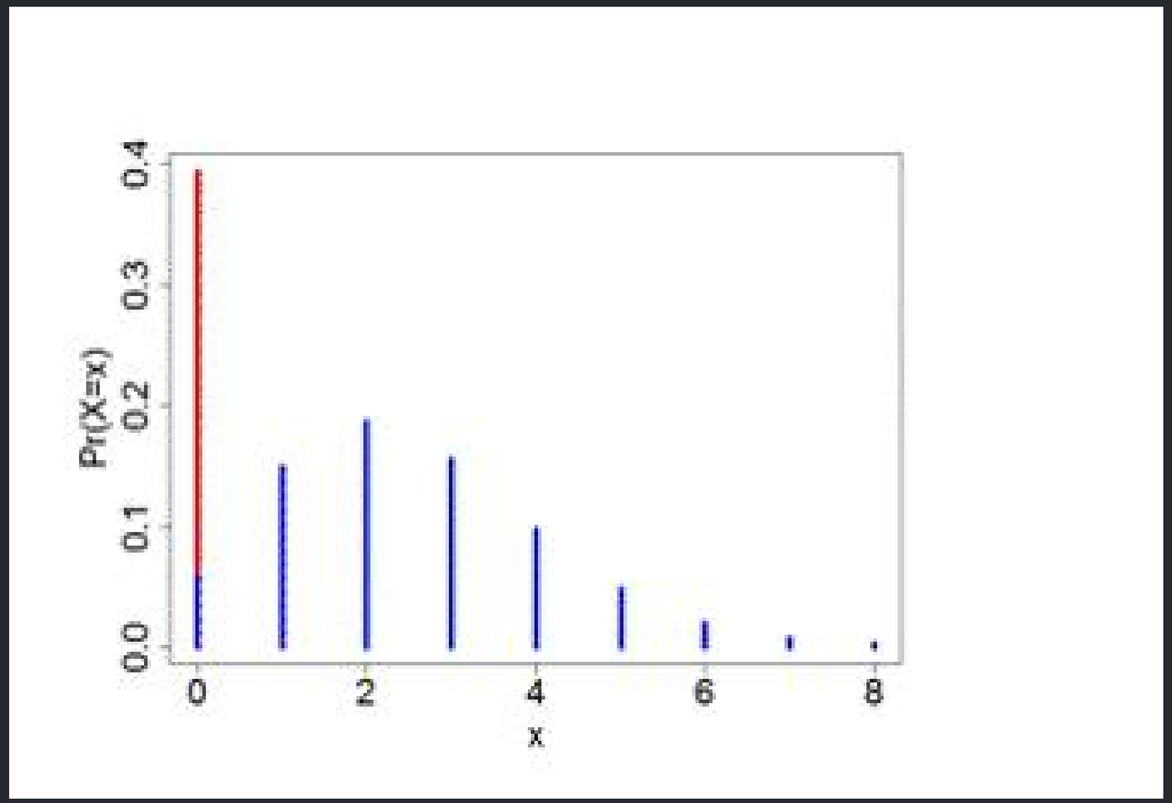
В таком случае предполагают, что действуют два процесса:

- Основной (Пуассоновский или отрицательно-биномиальный).
- Дополнительный (генерирующий лишние нули).





РЕГРЕССИЯ С ПОВЫШЕННЫМ КОЛИЧЕСТВОМ НУЛЕЙ (ZERO-INFLATED)



ИМЯ РАСПРЕДЕЛЕНИЯ

Пуассоновское/отрицательное биномиальное с повышенным количеством нулей.

ЗНАЧЕНИЯ ЗП

(0; ∞).

ПАРАМЕТРЫ

Такие же, как для прошлых двух моделей + π (вероятность принадлежности процессу).

ТИПИЧНАЯ ФУНКЦИЯ СВЯЗИ

Логарифм.

ОТТОК КЛИЕНТОВ

Клиенты часто покидают компании – либо полностью перестают пользоваться их услугами, либо отписываются от конкретного сервиса.

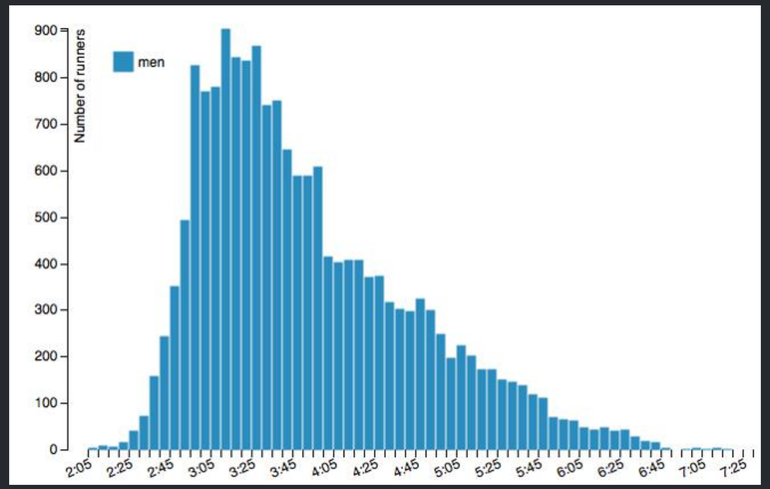
Подходит ли логистическая регрессия для такой задачи? Технически да.

НО...

У НАС ЕСТЬ НЕ ТОЛЬКО СОБЫТИЯ, НО И ВРЕМЯ ДО НИХ

Но (2) мы не можем определить время до события у тех, у кого оно не произошло. Но (3) исключить их из модели мы тоже не можем.

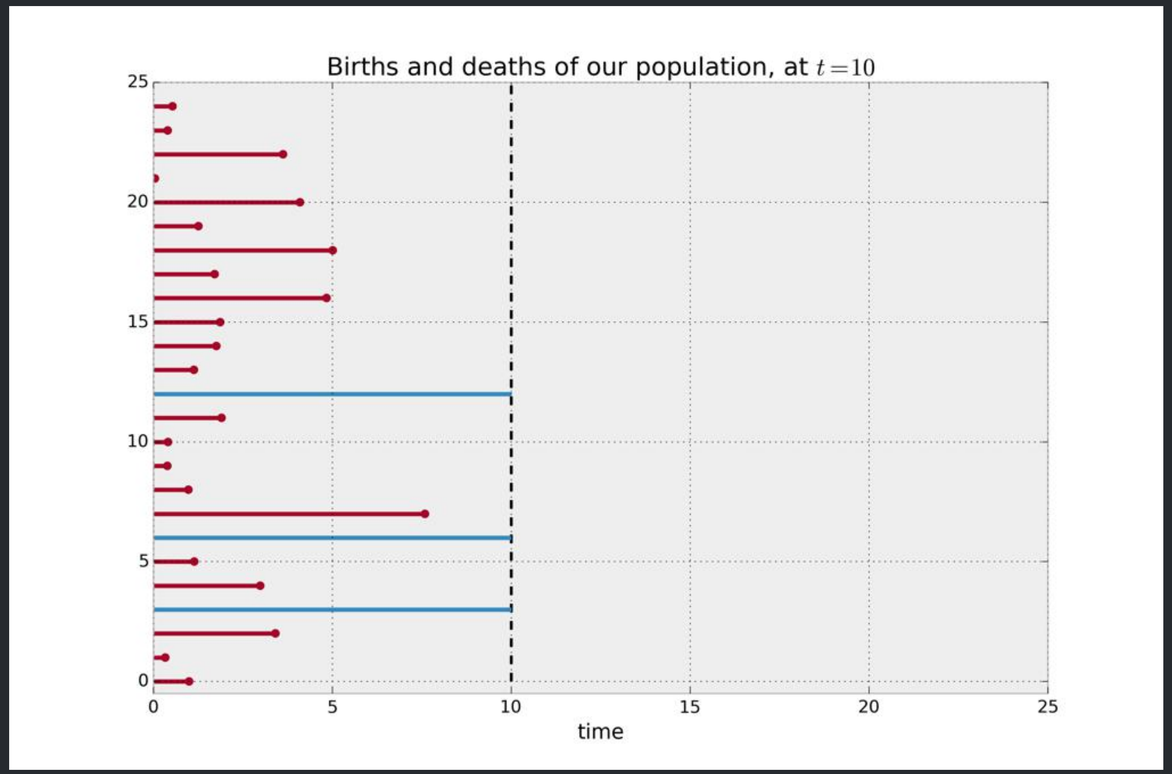
Что делать?



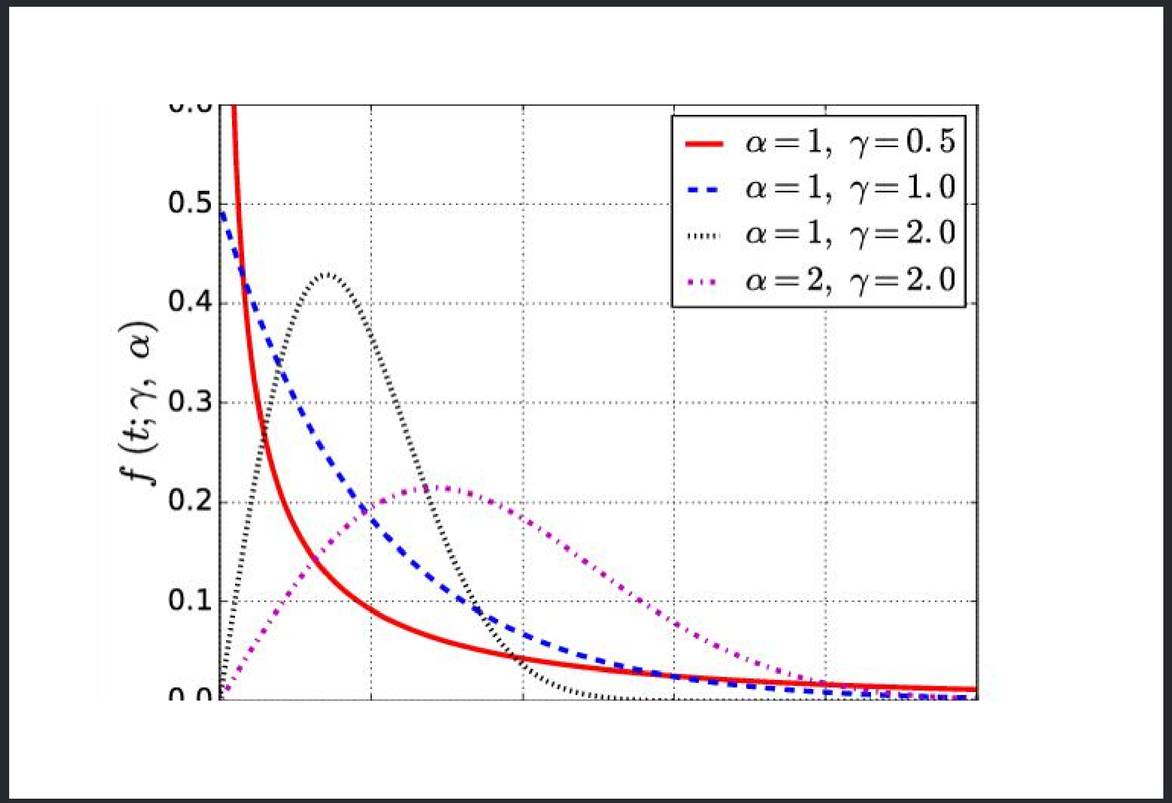
РЕГРЕССИЯ ВРЕМЕНИ ДО СОБЫТИЯ (TIME-TO-EVENT)

РЕГРЕССИЯ ВЫЖИВАЕМОСТИ (SURVIVAL)

Ту часть данных, для которой событие не произошло, называют цензурированной. Она также влияет на показатели итоговой модели.



МОДЕЛЬ УСКОРЕНИЯ ВРЕМЕНИ НЕУДАЧИ (ACCELERATED FAILURE TIME MODEL, AFT)



ИМЯ РАСПРЕДЕЛЕНИЯ

Вейбулловское (как вариант).

ЗНАЧЕНИЯ ЗП

(0; ∞).

ПАРАМЕТРЫ

α – дисперсия.

γ – форма.

ТИПИЧНАЯ ФУНКЦИЯ СВЯЗИ

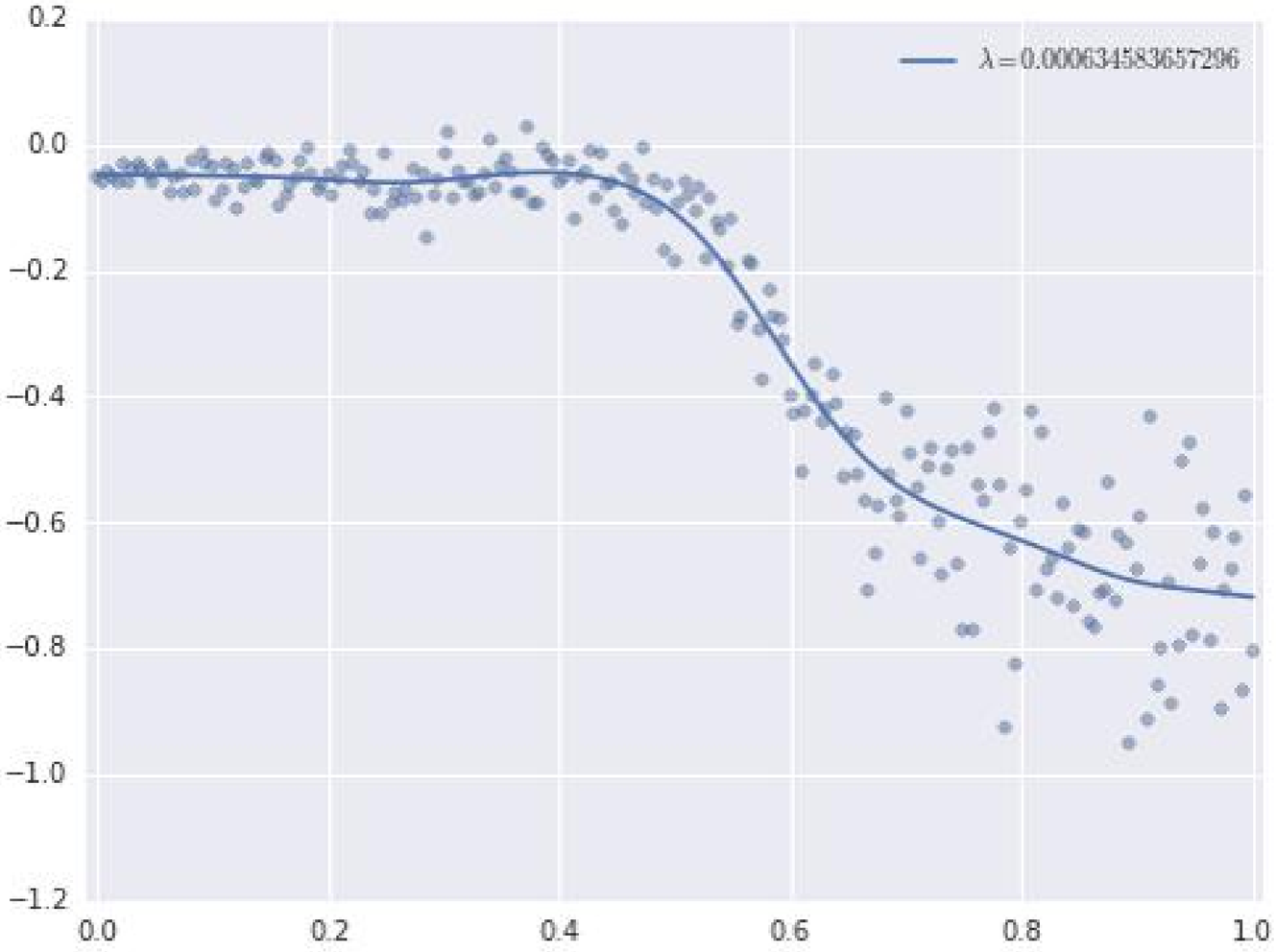
Логарифм.

СУММИРУЯ

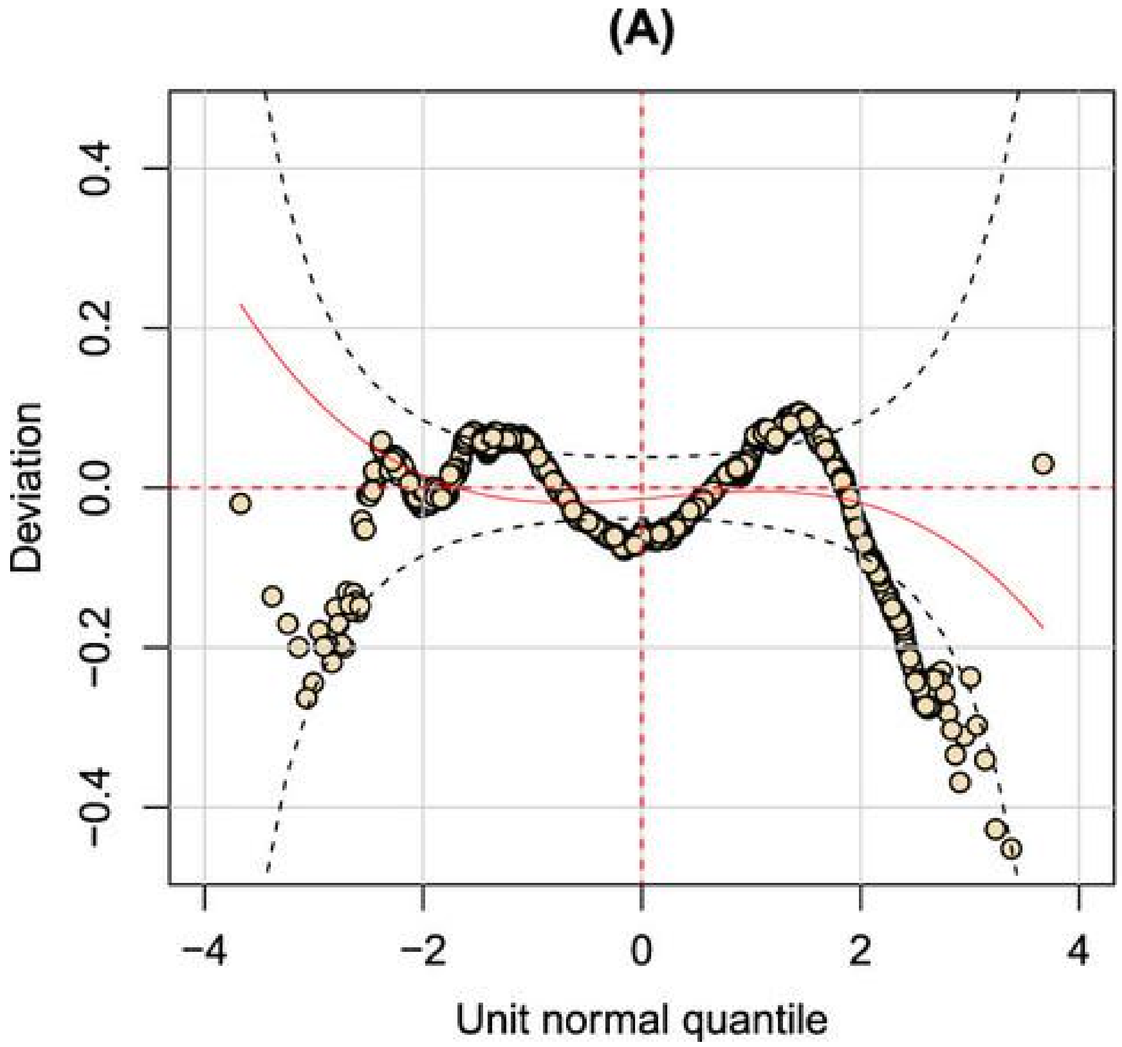
- Переменная из двух категорий – биномиальная регрессия.
- Категорий больше – мультиномиальная.
- В категориях есть явное убывание или нарастание – порядковая
- Считаем количество чего-то – Пуассоновская. Дисперсия больше среднего – отрицательно-биномиальная. Слишком много нулей – zero-inflated-модель.
- У нас есть какое-то событие и время до него – анализ выживаемости.

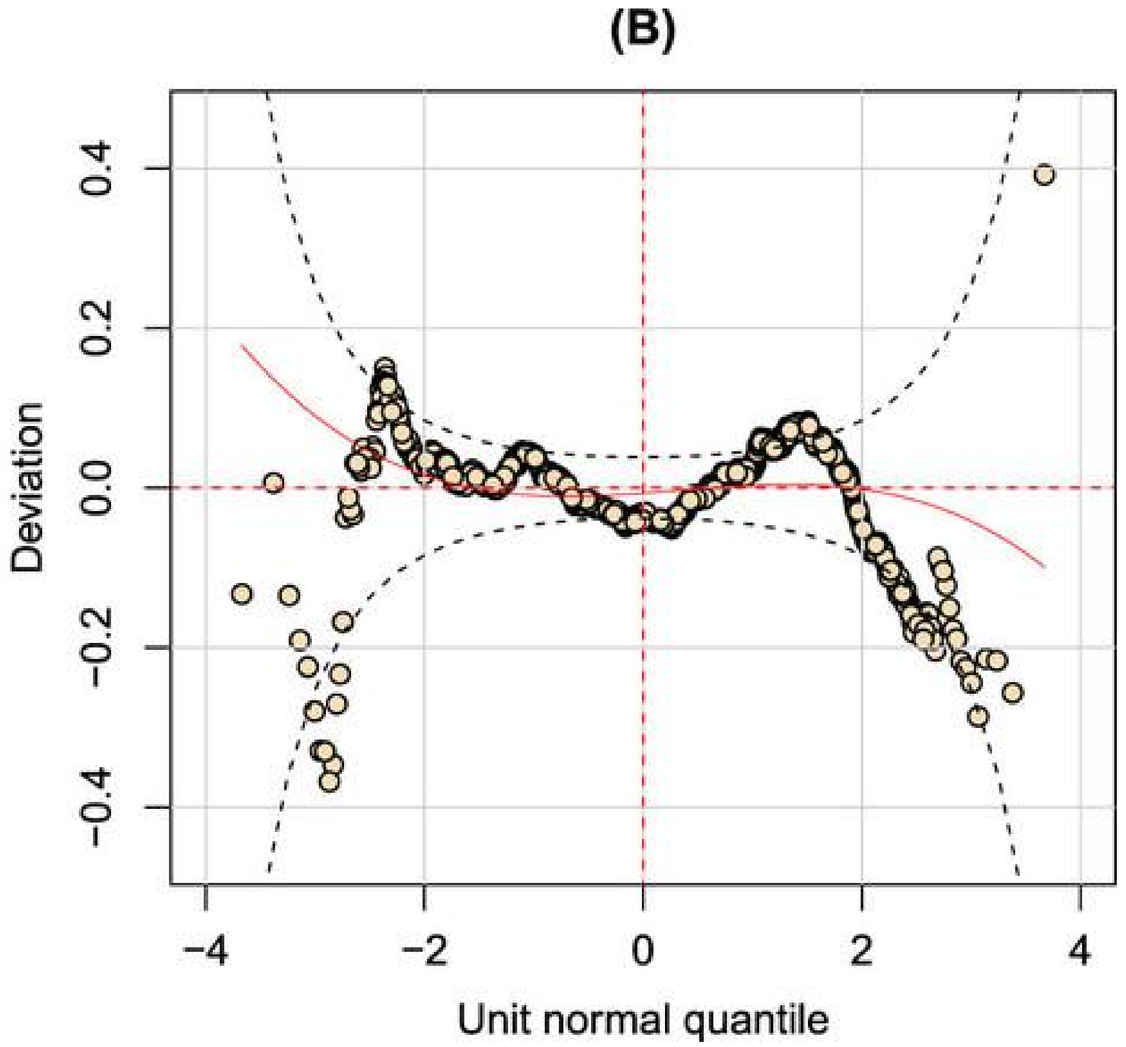
МОРАЛЬ ИСТОРИИ

GAM



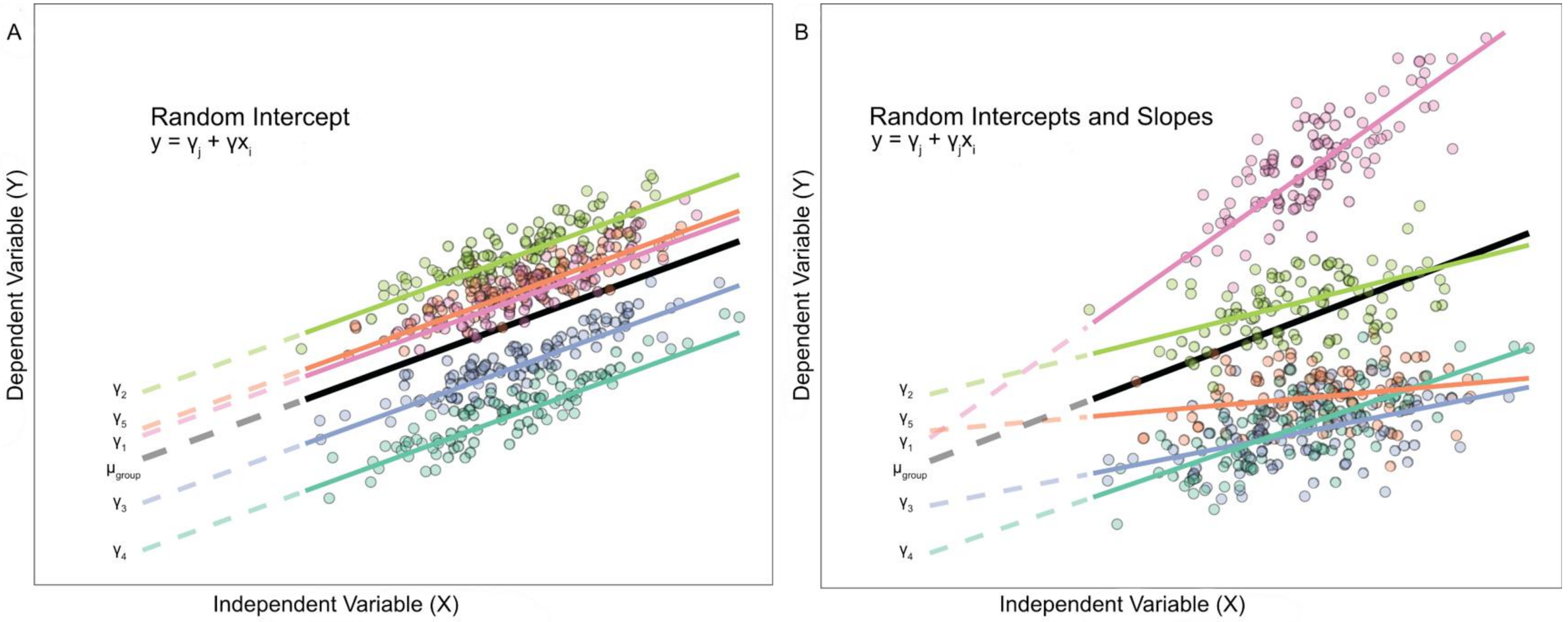
GAMLSS


(A)
Deviation
0.4
0.2
0.0
−0.2
−0.4
−4
−2
0
2
4
Unit normal quantile

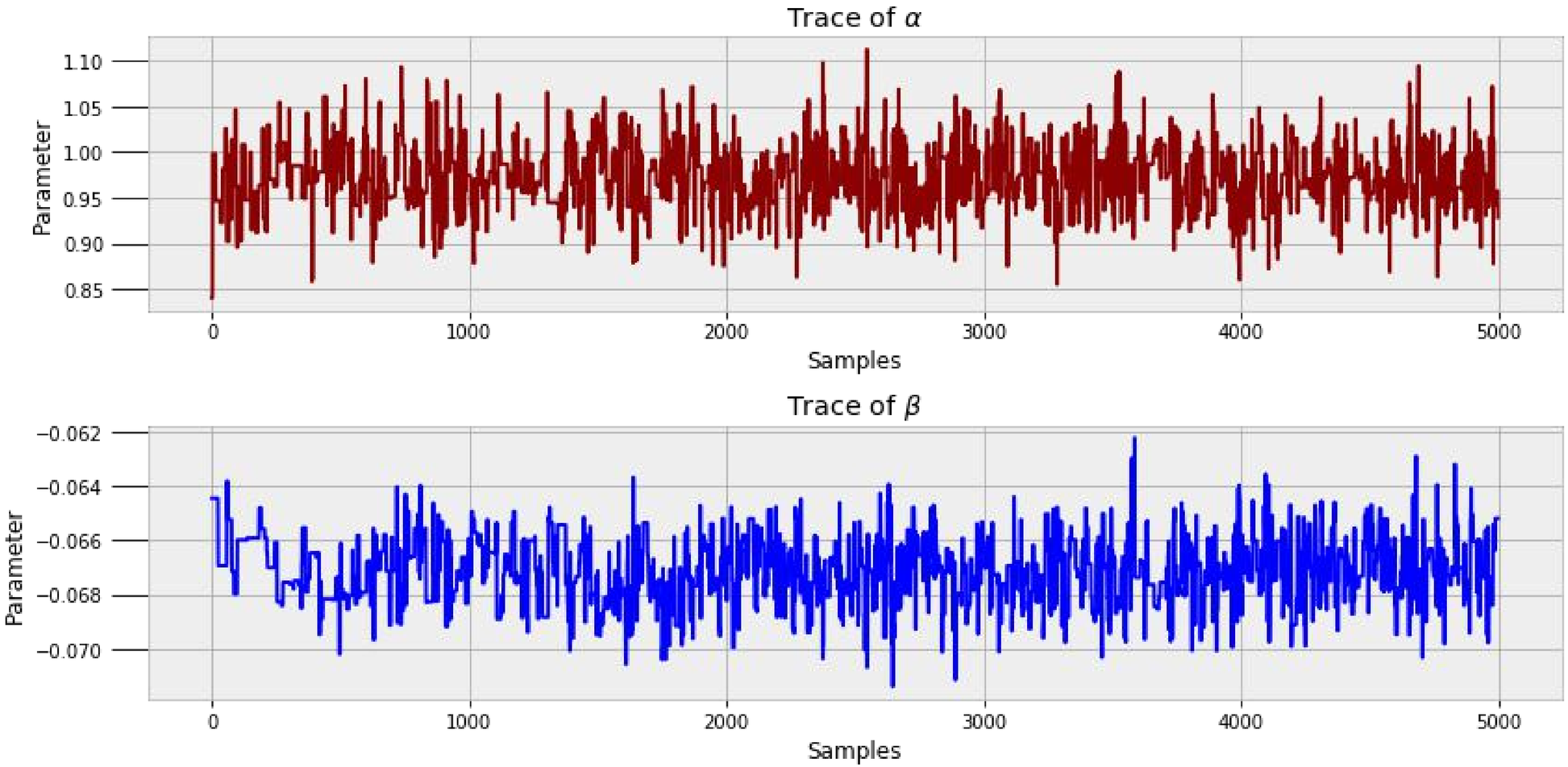

(B)
Deviation
0.4
0.2
0.0
−0.2
−0.4
−4
−2
0
2
4
Unit normal quantile

ARIMA/(G)ARCH/PROPHET

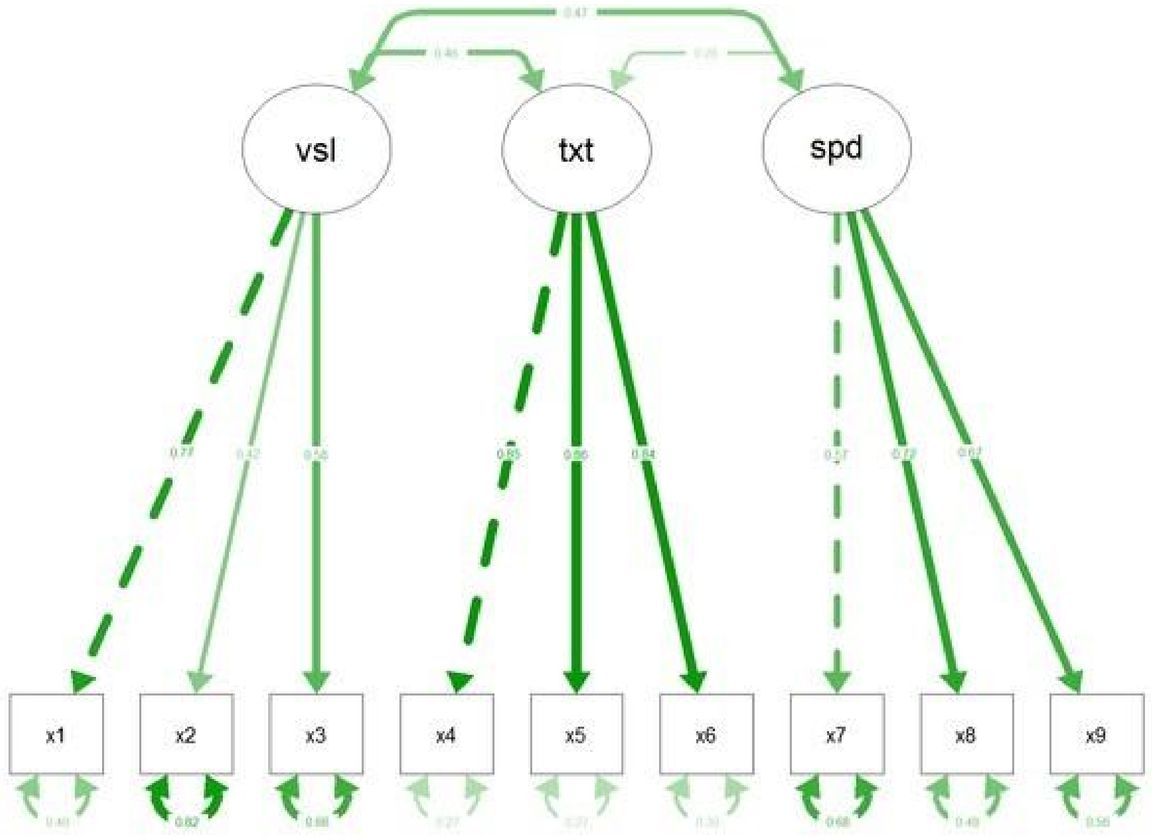
MIXED-EFFECTS



BAYESIAN (MCMC)



SEM



ВСЁ СТАНОВИТСЯ ХУЖЕ

ВСЁ СТАНОВИТСЯ ~~ЖУЖЕ~~

ВСЁ СТАНОВИТСЯ СЛОЖНЕЕ