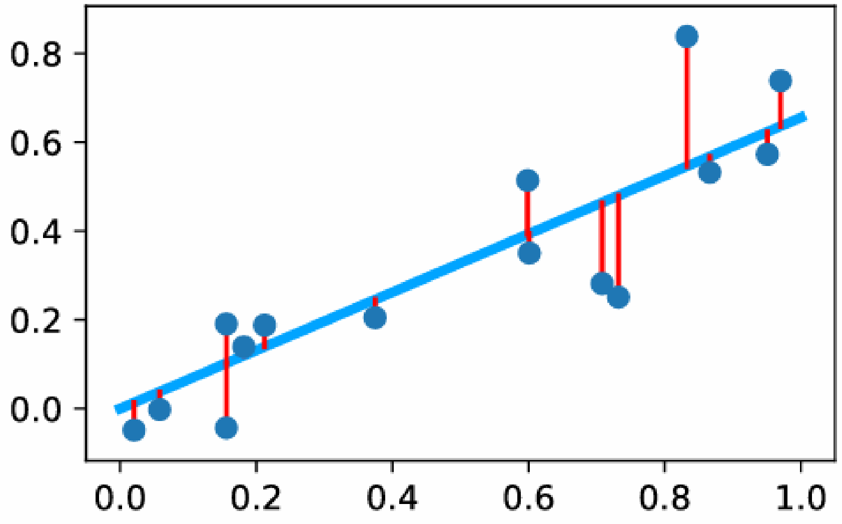
Логистическая регрессия

Линейные модели и решение задачи классификации

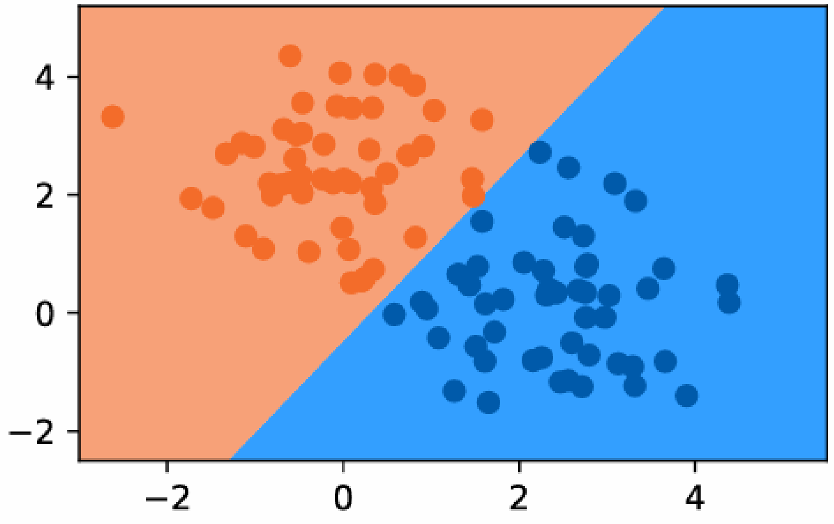
Regression:

$$\hat{f}(x) = \theta^{\mathrm{T}} x$$



Classification:

$$\hat{f}(x) = \mathbb{I}[\theta^{\mathrm{T}} x > 0]$$



Ограничения общих линейных моделей

- General linear models: $\hat{y} = b_0 + b_1 x$
- Важнейшее допущение: нормальность распределения ошибки (ε)
- Более общее допущение: нормальность распределения зависимой переменной (*y*)

- Что делать, если распределение зависимой переменной не нормальное?

Ограничения общих линейных моделей

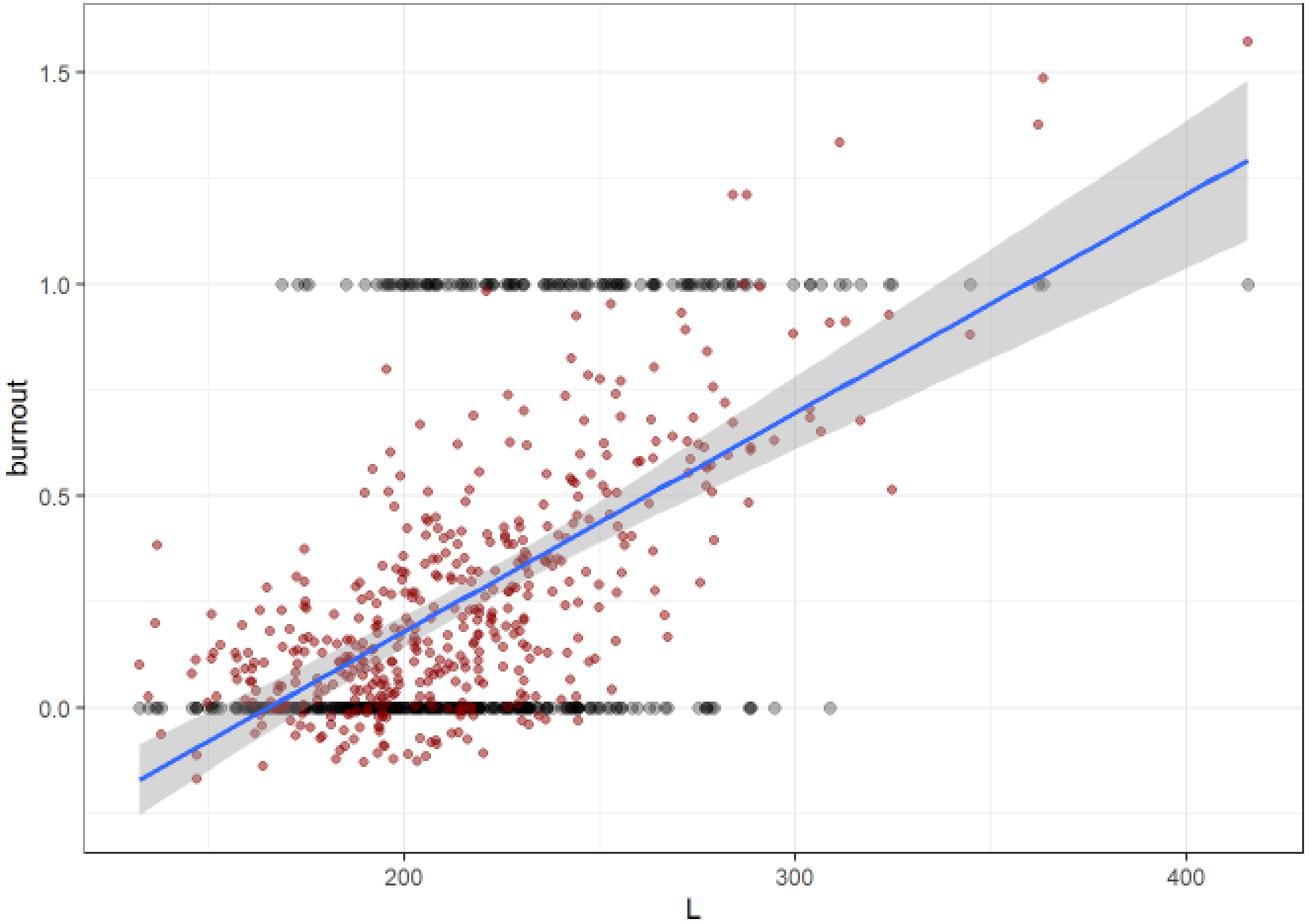
- Пусть зависимая переменная бинарная (принимает только одно из двух значений)
- Почему не можем их перекодировать (как?) и построить линейную регрессию?

Что тут не так?

burnout=1 – выгорание

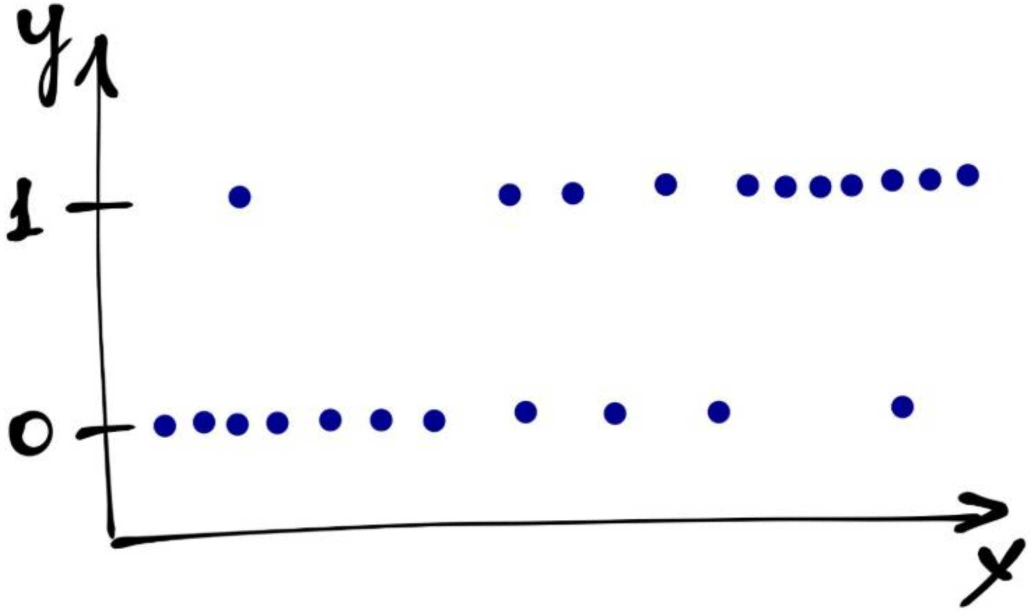
burnout=0 – нет выгорания

L – исследования +
преподавание + ...



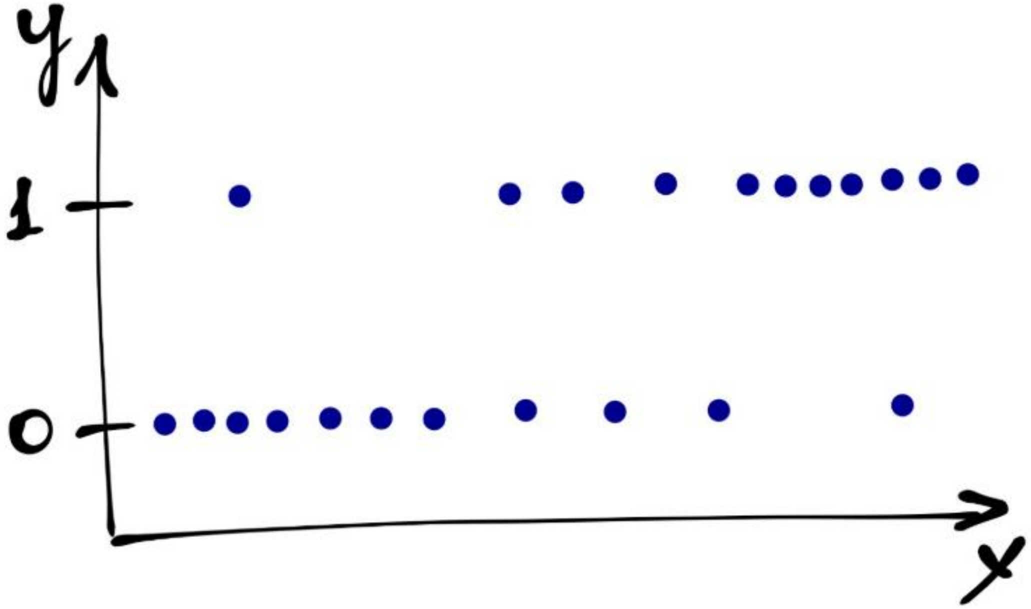
Дискретные → непрерывные величины

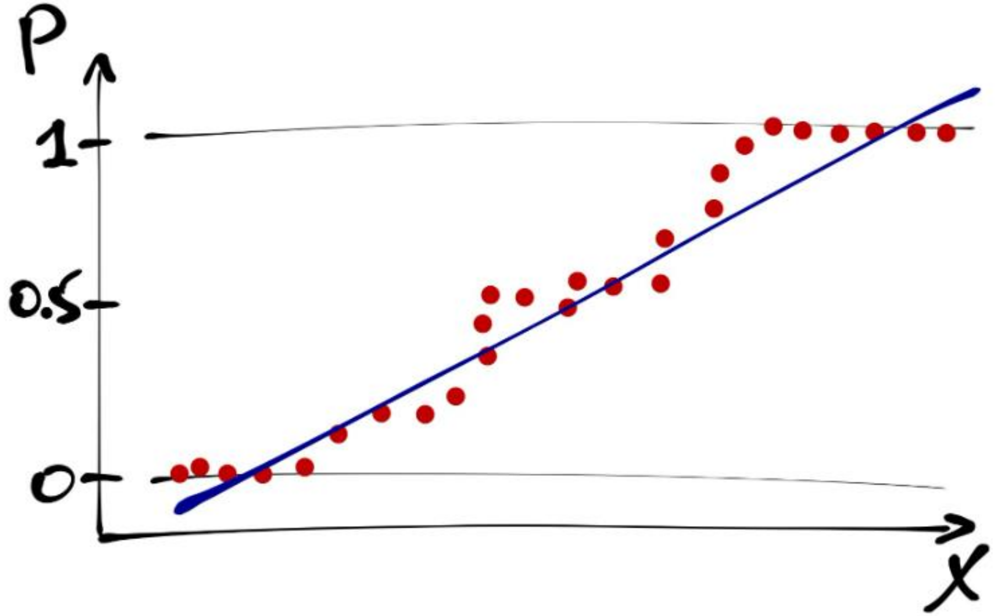
- Попытаемся превратить бинарную шкалу в непрерывную → моделируем *вероятность* получения единиц



Дискретные → непрерывные величины

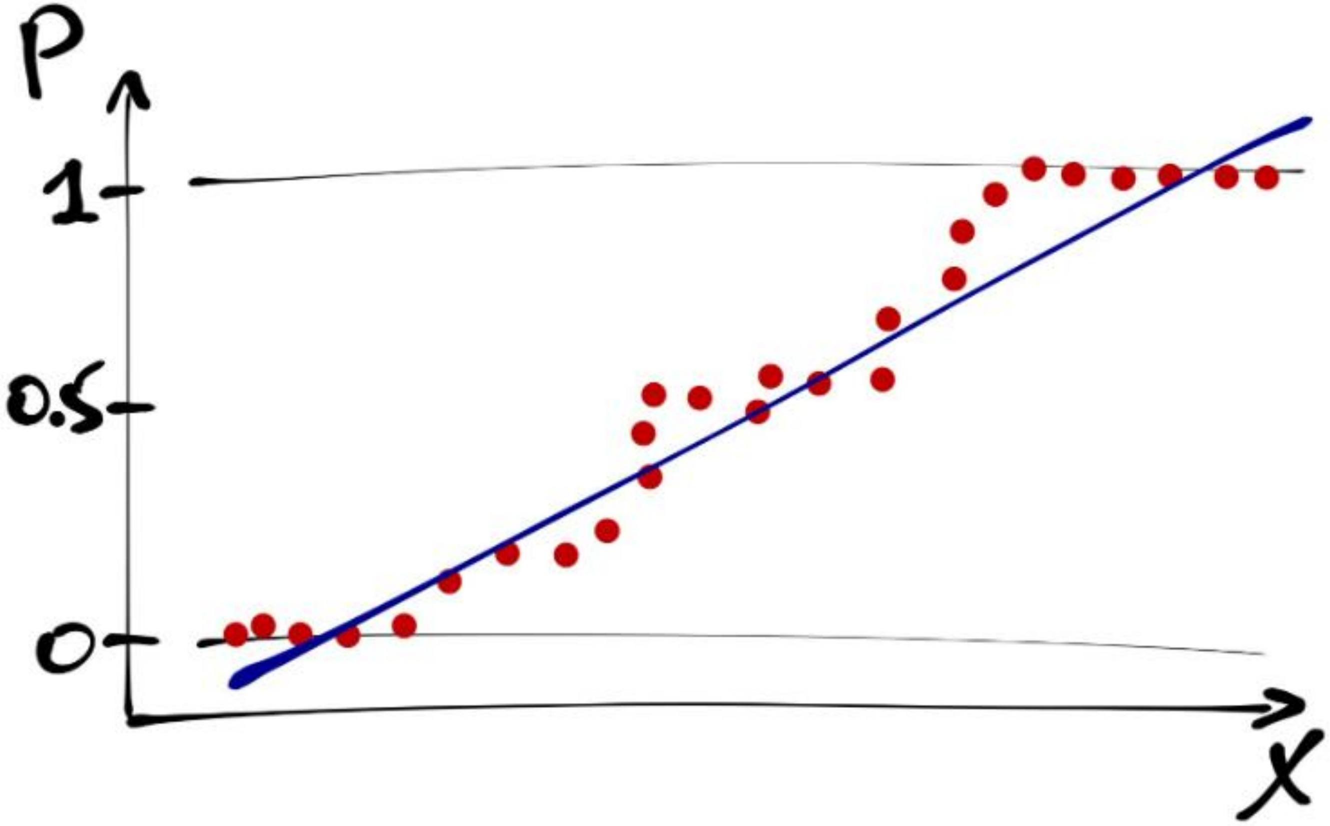
- Попытаемся превратить бинарную шкалу в непрерывную → моделируем *вероятность* получения единиц:
- p_i – вероятность события y=1
- $1 - p_i$ – вероятность события y=0





Дискретные → непрерывные величины

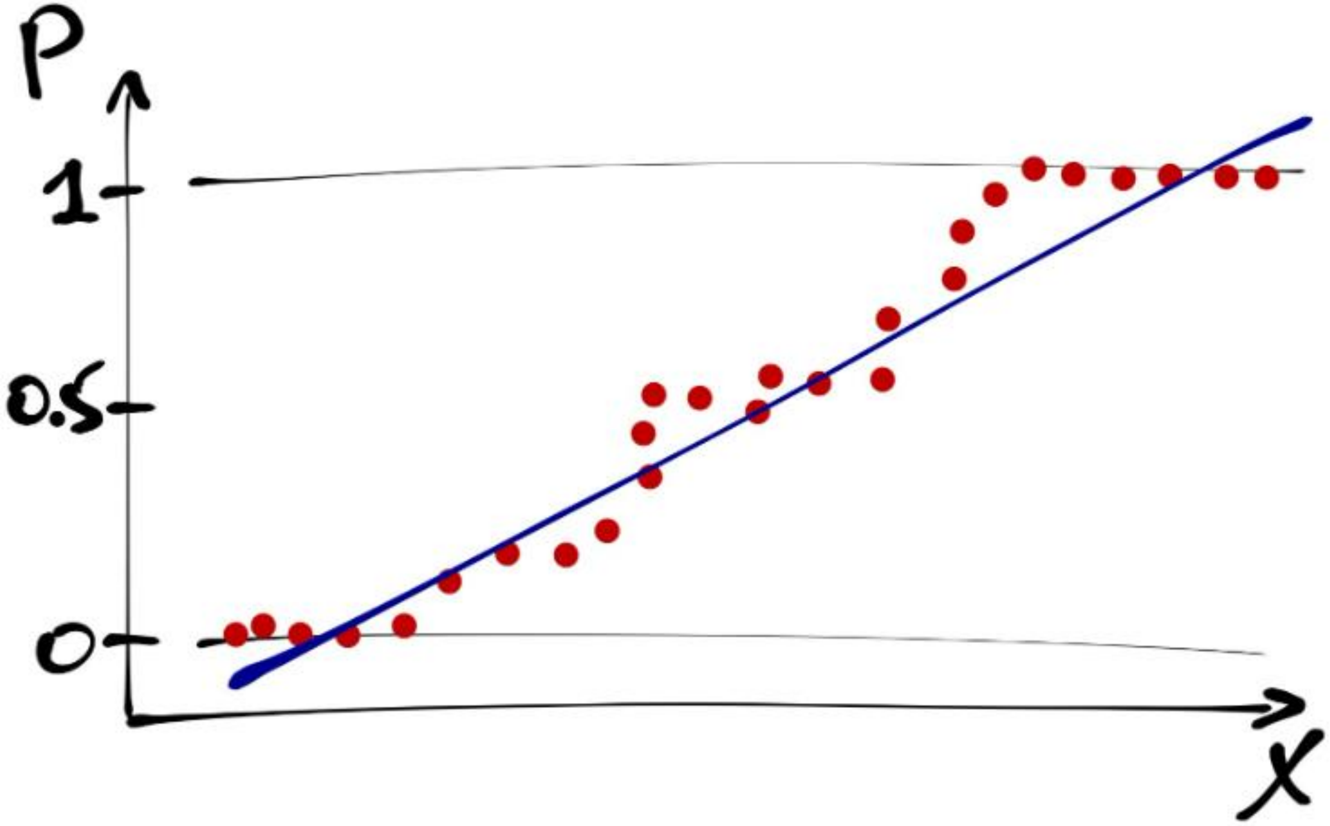
- p_i – вероятность события y=1
- $1 - p_i$ – вероятность события y=0
- Построим регрессионную модель
- В чем проблема?



Дискретные → непрерывные величины

- p_i – вероятность события y=1
- $1 - p_i$ – вероятность события y=0
- Построим регрессионную модель
- В чем проблема?

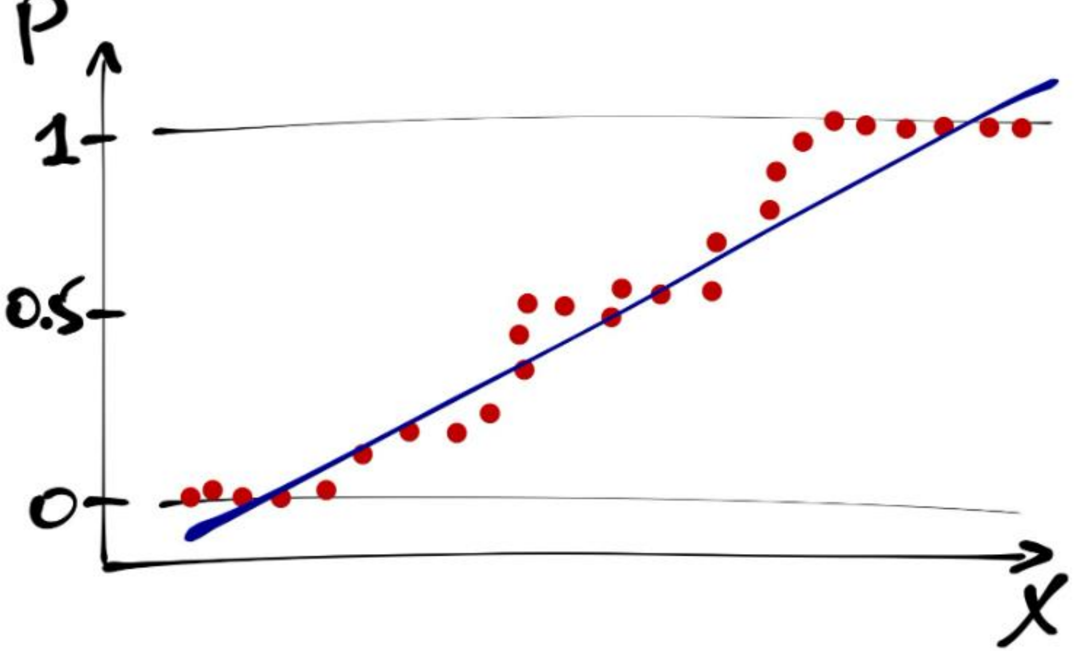
- Вероятность меняется в пределах от 0 до 1

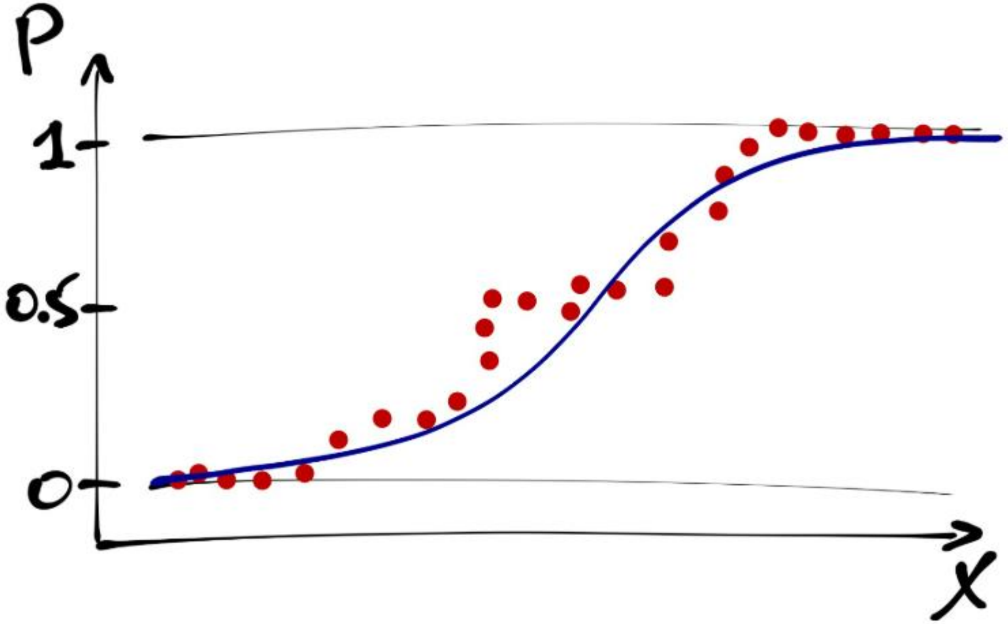


Дискретные → непрерывные величины

- Логистическая кривая (**sigmoid**)

$$p_i = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$





Дискретные → непрерывные величины

Хотим обойти ограниченность логистической кривой и перейти от [0; 1] к (-∞; +∞), заменим вероятности шансами.

Шансы (odds ratio) – отношение вероятности успеха к вероятности неудачи. Варьируется в интервале [0; +∞).

$$Odds\ ratio = \frac{p_i}{1-p_i}$$

Дискретные → непрерывные величины

Хотим обойти ограниченность логистической кривой и перейти от [0; 1] к (-∞; +∞), заменим вероятности шансами.

Шансы (odds ratio) – отношение вероятности успеха к вероятности неудачи. Варьируется в интервале [0; +∞).

Чтобы подвинуть нижнюю границу, используем логарифм!

Легким движением превращаем шансы в **логиты**:

$$logit(p) = \ln\left(\frac{p_i}{1 - p_i}\right)$$

TL; DR

1. От дискретных событий (0 и 1) переходим к вероятностям
2. Связь вероятностей с предиктором описываем логистической кривой
3. Совершаем переход: вероятности → шансы → логиты
4. Оцениваем логиты с помощью линейной регрессии

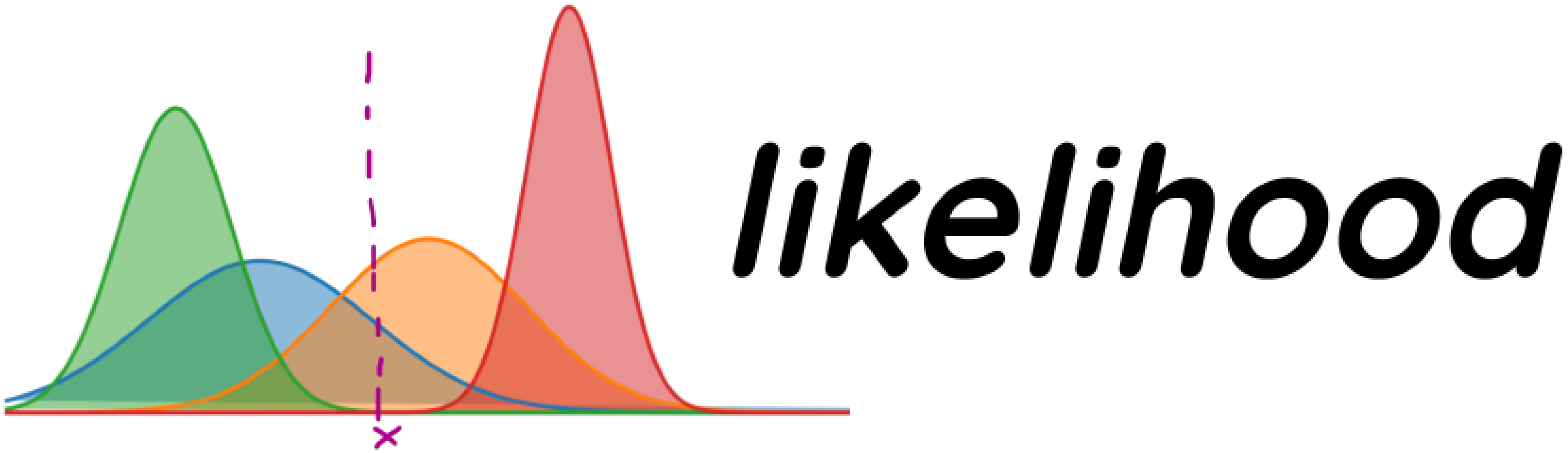
TL; DR

1. От дискретных событий (0 и 1) переходим к вероятностям
2. Связь вероятностей с предиктором описываем логистической кривой
3. Совершаем переход: вероятности → шансы → логиты
4. Оцениваем логиты с помощью линейной регрессии

NB! Функция, которая позволяет линеаризовать связь между предикторами и зависимой переменной называется связывающей функцией (**linked function**). Например, logit

Как подобрать коэффициенты?

- Не можем использовать МНК, MSE и т.д.
- Используем **метод максимального правдоподобия (maximum likelihood method)**
- Интуитивно: правдоподобие – попытка оценить: насколько вероятно получить данные, которые мы насобирали, используя нашу модель с полученными параметрами.



Статистическая значимость модели и предикторов

Псевдо R^2 – интерпретируется аналогично коэффициенту детерминации

Статистическая значимость модели и предикторов

Интерпретация коэффициентов при предикторах: нужно перейти от логарифмов обратно к отношению шансов (чтобы интерпретация была понятной). Для этого коэффициенты нужно возвести в экспоненту

Отношение шансов: во сколько раз изменится отношение шансов при увеличении значения переменной на единицу

Статистическая значимость модели и предикторов

Интерпретация коэффициентов при предикторах: нужно перейти от логарифмов обратно к отношению шансов (чтобы интерпретация была понятной). Для этого коэффициенты нужно возвести в экспоненту

Интерсепт показывает *логарифм отношения шансов для случая, когда все остальные предикторы равны нулю*

Коэффициент при предикторе показывает, *на сколько единиц изменяется логарифм отношения шансов* при изменении значения предиктора на единицу.

Интерпретация коэффициентов модели

Coefficient	exp(Coefficient) == odds ratio	Пример
>0	>1	Odds ratio = 1.6 → Шансы больше на 60%
=0	= 1	Нет разницы в шансах
< 0	(0; 1)	Odds ratio = 0.6 → шансы меньше на 40%

Анализ девиансы

Девианса – мера различия двух моделей

- Нулевая модель (среднее): $y \sim 1$
- Насыщенная модель (идеальная, overfit) : $y \sim b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_1 * x_2 * \ldots x_m$

Интуитивно: можем определить значимость нашей модели по тому, отличается ли она от нулевой модели

Проверка допущений модели

1. Мультиколлинеарность
2. Сверхдисперсия (~ гетероскедастичность)
3. Нормальность распределения остатков
4. Отсутствие влиятельных наблюдений

Метрики качества логистической регрессии

Вспомним про таблицы смежности:

		True Values	
		Positive	Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Метрики качества логистической регрессии. **Accuracy**

Вспомним про таблицы смежности:

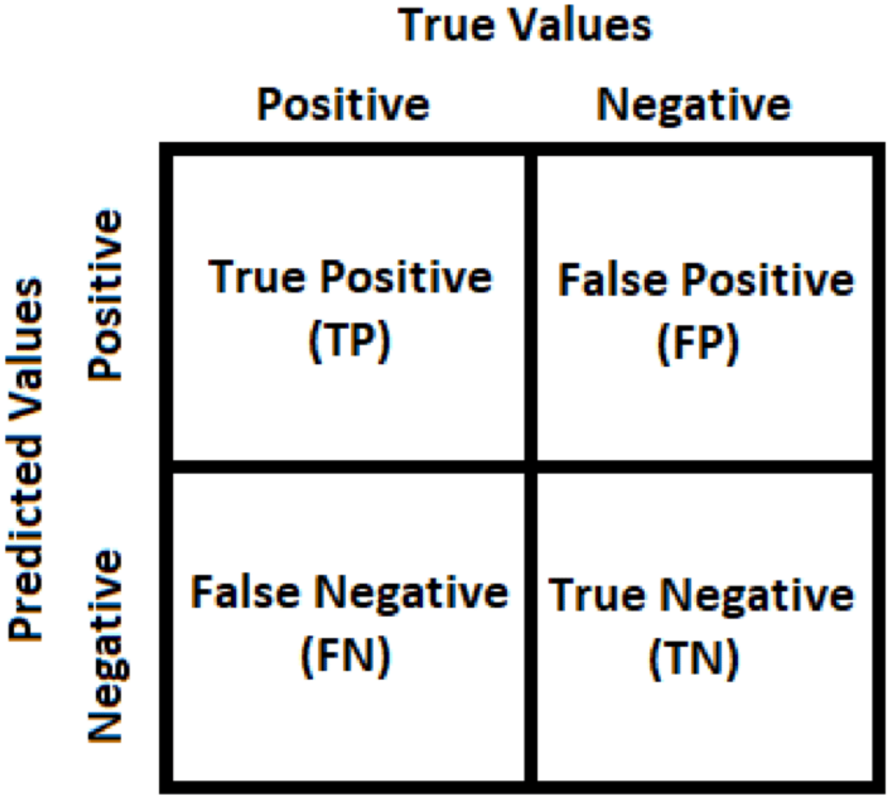
		True Values: Positive	True Values: Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
Predicted Values	Negative	False Negative (FN)	True Negative (TN)

Из них и будем исходить.

Accuracy – доля верно классифицированных объектов

$$\text{Acuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

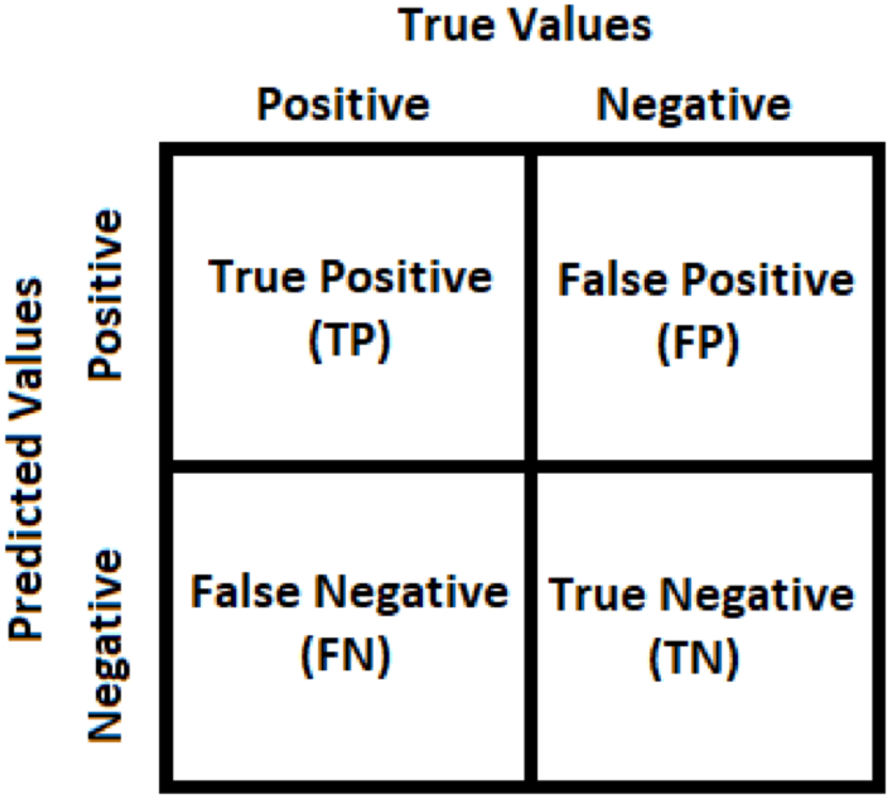
Метрики качества логистической регрессии. **Precision**



Precision: сколько положительных объектов действительно положительные?

$$\text{Precision} = \frac{TP}{TP + FP}$$

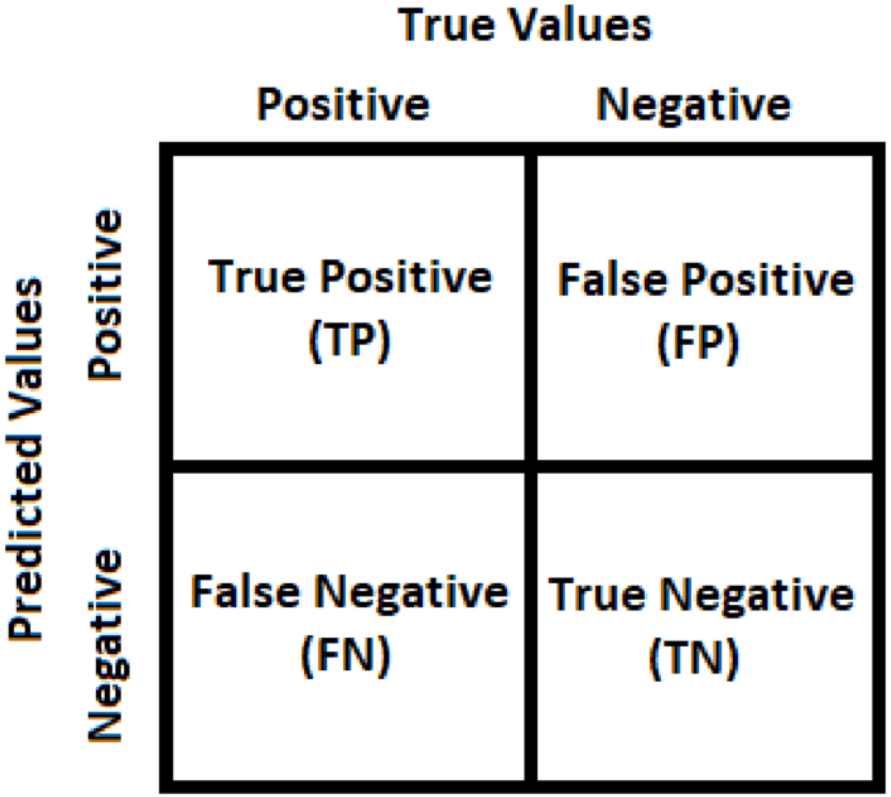
Метрики качества логистической регрессии. **Recall**



Recall: из всех объектов положительного класса сколько было выявлено?

$$\text{Precision} = \frac{TP}{TP + FN}$$

Метрики качества логистической регрессии. **F1-score**

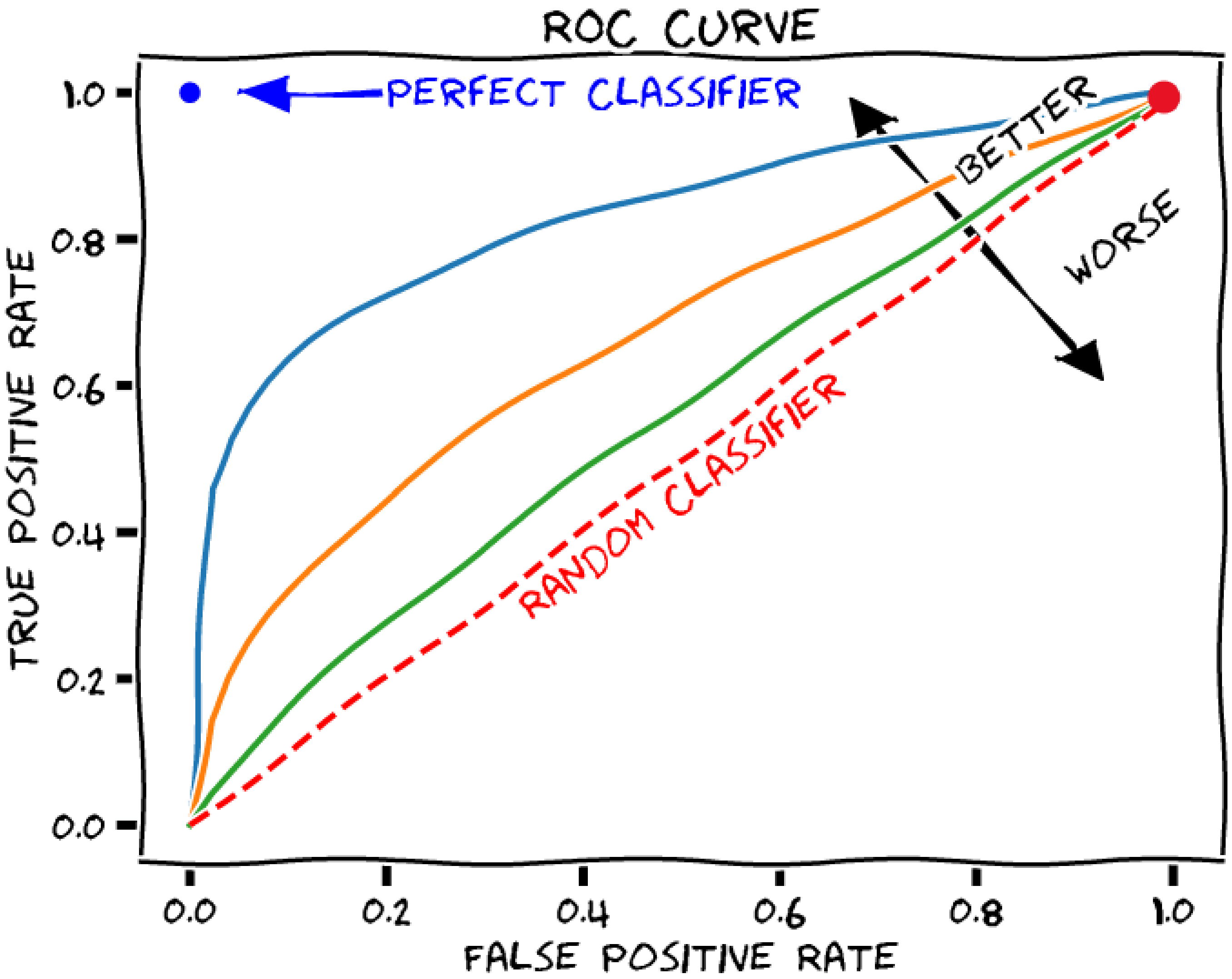


$$\mathrm{F1} = \frac{2 * precision \ * recall}{precision + recall}$$

ROC-AUC score

Все прошлые метрики зависят от порога, который мы выбираем для определения класса.

ROC-AUC работает не с классами, а с **вероятностями (!)** и позволяет оценить, насколько модель выучила закономерность



- ROC-AUC
- Complete separation
- переобучение

Синтаксис glm

```
glm(y ~ x, family = binomial(link = 'logit'))
```