

ВЗАИМОСВЯЗЬ НЕСКОЛЬКИХ ПРИЗНАКОВ

УСПЕВАЕМОСТЬ ШКОЛЬНИКОВ

- » Влияет ли уровень потребления алкоголя на успеваемость школьников?
- » Эксперимент:
 - ▶ Возьмём случайную выборку школьников
 - ▶ Назначим им случайную еженедельную дозу алкоголя
 - ▶ По окончании учебного года измерим корреляцию между дозой и успеваемостью

УСПЕВАЕМОСТЬ ШКОЛЬНИКОВ

» Эксперимент:

- ▶ Возьмём случайную выборку школьников
- ▶ Назначим им случайную еженедельную дозу алкоголя
- ▶ По окончании учебного года измерим корреляцию между дозой и успеваемостью

» Неэтично!

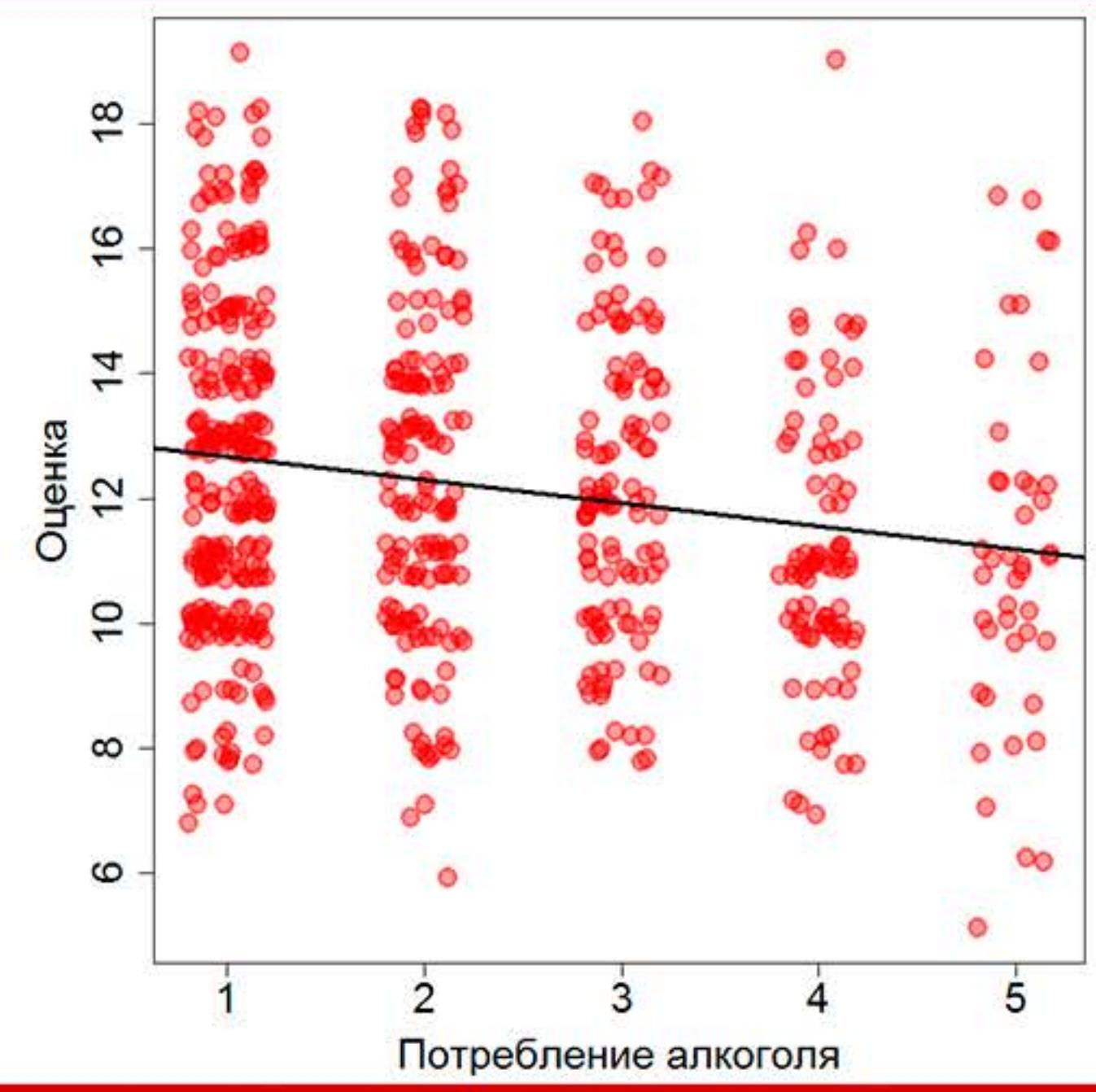
УСПЕВАЕМОСТЬ ШКОЛЬНИКОВ



- » Для 633 учеников старших классов двух португальских школ известны ряд демографических показателей и показателей успеваемости; известны уровень потребления алкоголя по выходным и финальная оценка по португальскому языку.

УСПЕВАЕМОСТЬ ШКОЛЬНИКОВ





УСПЕВАЕМОСТЬ ШКОЛЬНИКОВ



- › У нас есть ещё 29 признаков, потенциально влияющих на успеваемость.
- › Если учесть их влияние, остаётся ли у потребления алкоголя предсказательная сила?
- › Можно ли утверждать, что повышение потребления алкоголя вызывает снижение оценок?

ЛИНЕЙНАЯ РЕГРЕССИЯ

- › $1, \dots, n$ — объекты
- x_1, \dots, x_k — объясняющие переменные
- y — отклик
- › Ищем такой вектор β , что $y \approx \beta x$.

- › Модель линейной регрессии:

$$\mathbb{E}(y | x) = \beta_0 + \sum_{j=1}^k \beta_j x_j$$

- » Модель линейной регрессии:

$$\mathbb{E}(y|x) = \beta_0 + \sum_{j=1}^k \beta_j x_j$$

- » β_j показывает, насколько в среднем увеличивается y , если x_j увеличивается на единицу, а остальные факторы фиксированы.

- › Регрессию можно использовать для исследования остаточного влияния признака на отклик с учётом других признаков
- › Далее: горячие подробности

СВОЙСТВА РЕШЕНИЯ ЗАДАЧИ РЕГРЕССИИ

ЛИНЕЙНАЯ РЕГРЕССИЯ

» Модель линейной регрессии:

$$\mathbb{E}(y|x) = \beta_0 + \sum_{j=1}^k \beta_j x_j$$

$$X = \begin{pmatrix} x_{10} = 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

» Метод наименьших квадратов:

$$\|y - X\beta\|_2^2 \rightarrow \min_{\beta}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

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

- » $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ (Total Sum of Squares);
- » $ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ (Explained Sum of Squares);
- » $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ (Residual Sum of Squares);
- » $TSS = ESS + RSS$

- » Коэффициент детерминации:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ МНК

- » Линейность отклика:

$$y = X\beta + \varepsilon$$

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ МНК

- » Линейность отклика
- » Случайность выборки: наблюдения (x_i, y_i) независимы.

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ МНК

- › Линейность отклика
- › Случайность выборки
- › Полнота ранга X : ни один из признаков не является линейной комбинацией других признаков ($\text{rank } X = k + 1$).

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ МНК

- › Линейность отклика
- › Случайность выборки
- › Полнота ранга X

- › Случайность ошибок: $\mathbb{E}(\varepsilon | x) = 0$

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ МНК

- › Линейность отклика
- › Случайность выборки
- › Полнота ранга X
- › Случайность ошибок

⇒ МНК-оценки коэффициентов β несмешённые:

$$\mathbb{E}\hat{\beta}_j = \beta_j$$

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ МНК

⇒ МНК-оценки коэффициентов β несмешённые:

$$\mathbb{E}\hat{\beta}_j = \beta_j$$

и состоятельные:

$$\forall \gamma > 0 \lim_{n \rightarrow \infty} P\left(\left|\beta_j - \hat{\beta}_j\right| < \gamma\right) = 1$$

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ МНК

- › Линейность отклика
- › Случайность выборки
- › Полнота ранга X
- › Случайность ошибок

- › Гомоскедастичность ошибок: дисперсия ошибки не зависит от значений признаков: $\mathbb{D}(\varepsilon | x) = \sigma^2$
(предположения Гаусса-Маркова)

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ МНК

- › Линейность отклика
- › Случайность выборки
- › Полнота ранга X
- › Случайность ошибок
- › Гомоскедастичность ошибок

⇒ МНК-оценки имеют наименьшую дисперсию в классе оценок β , линейных по y .

ДИСПЕРСИЯ $\hat{\beta}_j$

(1)-(5) \Rightarrow

$$\mathbb{D}(\hat{\beta}_j) = \frac{\sigma^2}{TSS_j (1 - R_j^2)},$$

где $TSS_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$, R_j^2 — коэффициент

детерминации при регрессии x_j на все остальные
признаки.

ДИСПЕРСИЯ $\hat{\beta}_j$

- » Чем больше σ^2 , тем больше дисперсия $\hat{\beta}_j$.
- » Чем больше вариация значений x_j в выборке, тем меньше дисперсия $\hat{\beta}_j$.
- » Чем лучше признак x_j объясняется линейной комбинацией оставшихся признаков, тем больше дисперсия $\hat{\beta}_j$.

ДИСПЕРСИЯ $\hat{\beta}_j$

- » $R_j^2 < 1$ по предположению (3); тем не менее, может быть $R_j^2 \approx 1$.
- » В матричном виде:

$$\mathbb{D}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

- » Если столбцы X почти линейно зависимы, то матрица $X^T X$ плохо обусловлена, и дисперсия оценок $\hat{\beta}_j$ велика.

ДИСПЕРСИЯ $\hat{\beta}_j$

- » В матричном виде:

$$\mathbb{D}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

- » Если столбцы X почти линейно зависимы, то матрица $X^T X$ плохо обусловлена, и дисперсия оценок $\hat{\beta}_j$ велика.
- » Близкая к линейной зависимость между x_j — мультиколлинеарность.

НОРМАЛЬНОСТЬ

- › Линейность отклика
 - › Случайность выборки
 - › Полнота ранга X
 - › Случайность ошибок
 - › Гомоскедастичность ошибок
-
- › Нормальность ошибок:

$$\varepsilon | x \sim N(0, \sigma^2)$$

НОРМАЛЬНОСТЬ

- » Нормальность ошибок:

$$\varepsilon | x \sim N(0, \sigma^2)$$

- » Эквивалентная запись:

$$y | x \sim N(x\beta, \sigma^2)$$

НОРМАЛЬНОСТЬ

- › Линейность отклика
- › Случайность выборки
- › Полнота ранга X
- › Случайность ошибок
- › Гомоскедастичность ошибок
- › Нормальность ошибок

⇒ МНК-оценки совпадают с оценками максимального правдоподобия

НОРМАЛЬНОСТЬ

(1)-(6) \Rightarrow МНК-оценки совпадают с оценками максимального правдоподобия \Rightarrow

- › Имеют наименьшую дисперсию среди всех несмешённых оценок β
- › Имеют нормальное распределение

$$N \left(\beta, \sigma^2 (X^T X)^{-1} \right)$$

НОРМАЛЬНОСТЬ

- » Имеют нормальное распределение

$$N \left(\beta, \sigma^2 (X^T X)^{-1} \right)$$

- » $\hat{\sigma}^2 = \frac{1}{n - k - 1} \text{RSS}$ — несмешённая оценка σ^2 , и

$$\frac{\text{RSS}}{\sigma^2} \sim \chi_{n-k-1}^2$$

НОРМАЛЬНОСТЬ

» $\hat{\sigma}^2 = \frac{1}{n - k - 1} \text{RSS}$ — несмешённая оценка σ^2 , и

$$\frac{\text{RSS}}{\sigma^2} \sim \chi_{n-k-1}^2$$

» $\forall c \in \mathbb{R}^{k+1}, \frac{c^T (\beta - \hat{\beta})}{\hat{\sigma} \sqrt{c^T (X^T X)^{-1} c}} \sim St(n - k - 1)$

В предположениях (1)-(6) можно строить:

- ▶ доверительные для β_j
- ▶ доверительные интервалы для $\mathbb{E}(y | \mathbf{x})$
- ▶ предсказательные интервалы для $y | \mathbf{x}$

- › Предположения, при которых можно делать выводы по регрессионной модели
- › Никаких регуляризаторов :(
- › Далее: строим интервалы и проверяем гипотезы

ИНТЕРВАЛЫ И ГИПОТЕЗЫ В РЕГРЕССИИ

СВОЙСТВА МНК-ОЦЕНОК

(1)-(6) \Rightarrow

- $\frac{\text{RSS}}{\sigma^2} \sim \chi_{n-k-1}^2$
- $\hat{\beta} \sim N\left(\beta, \sigma^2 (X^T X)^{-1}\right)$
- $\forall c \in \mathbb{R}^{k+1} \quad \frac{c^T (\beta - \hat{\beta})}{\hat{\sigma} \sqrt{c^T (X^T X)^{-1} c}} \sim St(n - k - 1)$

ДОВЕРИТЕЛЬНЫЕ И ПРЕДСКАЗАТЕЛЬНЫЕ ИНТЕРВАЛЫ

- » 100(1 – α) доверительный интервал для σ^2 :

$$\frac{\text{RSS}}{\chi_{n-k-1, 1-\alpha/2}^2} \leq \sigma^2 \leq \frac{\text{RSS}}{\chi_{n-k-1, \alpha/2}^2}$$

ДОВЕРИТЕЛЬНЫЕ И ПРЕДСКАЗАТЕЛЬНЫЕ ИНТЕРВАЛЫ

› Возьмём $c = \begin{pmatrix} 0 \dots 0 & 1 & 0 \dots 0 \\ & j & \end{pmatrix}$;

100(1 - α) доверительный интервал для β_j :

$$\hat{\beta}_j \pm t_{n-k-1, 1-\alpha/2} \hat{\sigma} \sqrt{(X^T X)_{jj}^{-1}}$$

ДОВЕРИТЕЛЬНЫЕ И ПРЕДСКАЗАТЕЛЬНЫЕ ИНТЕРВАЛЫ

- » Для нового объекта x_0 возьмём $c = x_0$;
 $100(1 - \alpha)$ доверительный интервал для
 $\mathbb{E}(y | x = x_0)$:

$$x_0^T \hat{\beta} \pm t_{n-k-1, 1-\alpha/2} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}$$

ДОВЕРИТЕЛЬНЫЕ И ПРЕДСКАЗАТЕЛЬНЫЕ ИНТЕРВАЛЫ

- Чтобы построить предсказательный интервал для $y(x_0) = x_0^T \beta + \varepsilon(x_0)$, учтём ещё дисперсию ошибки:

$$x_0^T \hat{\beta} \pm t_{n-k-1, 1-\alpha/2} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

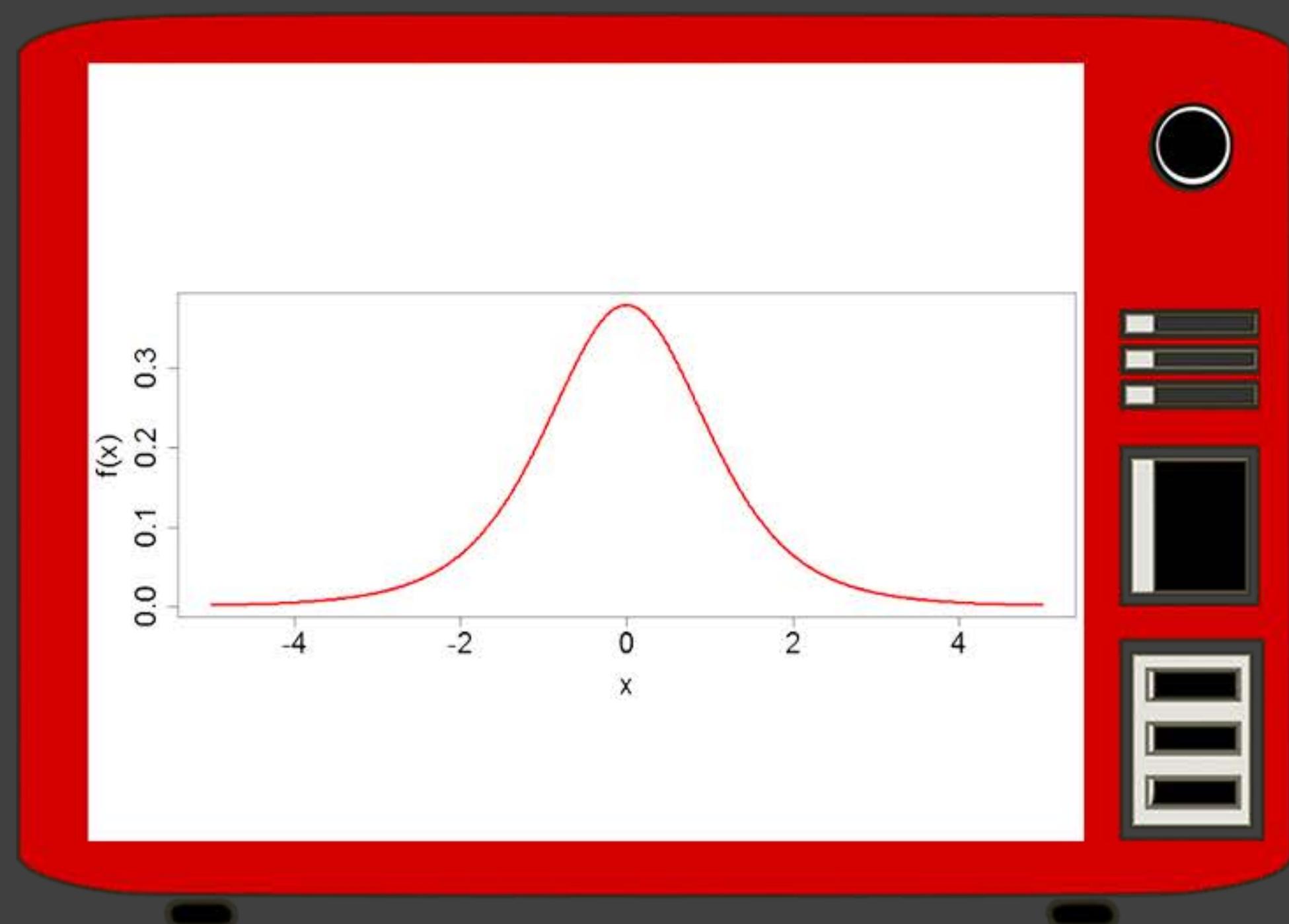
t-КРИТЕРИЙ СТЪЮДЕНТА

нулевая гипотеза: $H_0: \beta_j = 0$

альтернатива: $H_1: \beta_j <\neq> 0$

статистика: $T = \frac{\hat{\beta}_j}{\sqrt{\frac{\text{RSS}}{n-k-1} (X^T X)_{jj}^{-1}}}$

нулевое распределение: $T \sim St(n - k - 1)$



t-КРИТЕРИЙ СТЪЮДЕНТА

- » Пример: 12 испытуемых, x — результат прохождения испытуемым составного теста скорости реакции, y — результат его теста на симулятора транспортного средства.
Проведение составного теста значительно проще и требует меньших затрат, поэтому ставится задача предсказания y по x ; строится линейная регрессия

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

t-КРИТЕРИЙ СТЪЮДЕНТА

» Значима ли переменная x для предсказания y ?

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0 \Rightarrow p = 2.2021 \times 10^{-5}.$$

КРИТЕРИЙ ФИШЕРА

$$X_{n \times (k+1)} = \begin{pmatrix} X_1_{n \times (k+1-k_1)}, & X_2_{n \times k_1} \end{pmatrix} \quad \beta^T_{(k+1) \times 1} = \begin{pmatrix} \beta_1^T_{(k+1-k_1) \times 1}, & \beta_2^T_{k_1 \times 1} \end{pmatrix}^T$$

нулевая гипотеза: $H_0: \beta_2 = 0$

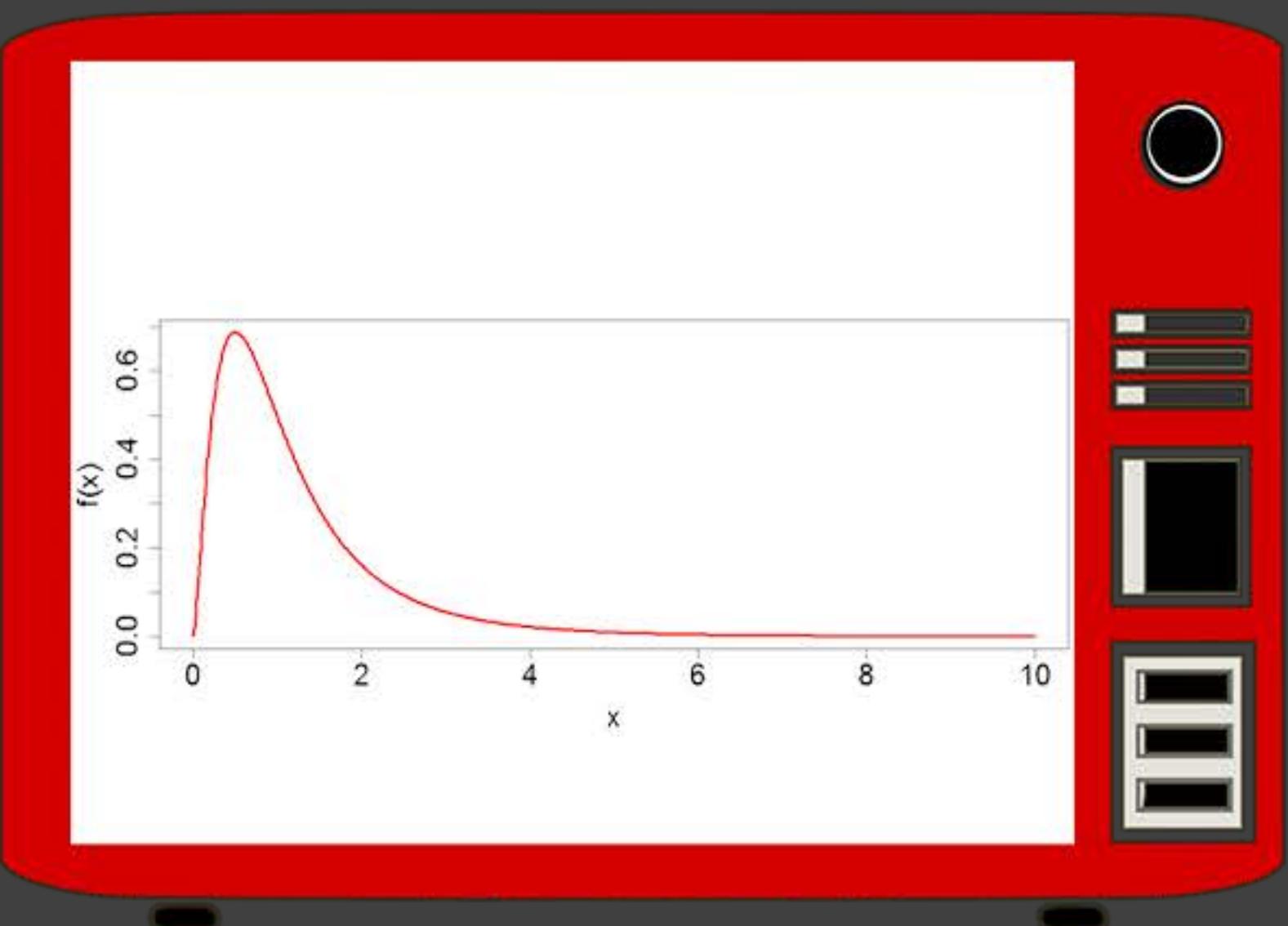
альтернатива: $H_1: H_0$ неверна

статистика: $\text{RSS}_r = \|y - X_1\beta_1\|_2^2$

$\text{RSS}_{ur} = \|y - X\beta\|_2^2$

$$F = \frac{(\text{RSS}_r - \text{RSS}_{ur}) / k_1}{\text{RSS}_{ur} / (n - k - 1)}$$

нулевое распределение: $F \sim F(k_1, n - k - 1)$



КРИТЕРИЙ ФИШЕРА

- » Пример: по данным о 1191 детей построена модель:

$$weight = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 inc + \beta_4 med + \beta_5 fed + \varepsilon$$

weight — вес ребёнка при рождении,

cigs — среднее число сигарет за один день беременности,

parity — номер ребёнка у матери,

inc — среднемесячный доход семьи,

КРИТЕРИЙ ФИШЕРА

$$weight = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 inc + \beta_4 med + \beta_5 fed + \varepsilon$$

weight — вес ребёнка при рождении,

cigs — среднее число сигарет за один день беременности,

parity — номер ребёнка у матери,

inc — среднемесячный доход семьи,

med — длительность получения образования матерью,

fed — отцом.

КРИТЕРИЙ ФИШЕРА

- Пример: по данным о 1191 детей построена модель:

$$weight = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 inc + \beta_4 med + \beta_5 fed + \varepsilon$$

- Зависит ли вес ребёнка при рождении от уровня образования родителей?

КРИТЕРИЙ ФИШЕРА

- » Пример: по данным о 1191 детей построена модель:

$$weight = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 inc + \beta_4 med + \beta_5 fed + \varepsilon$$

- » Зависит ли вес ребёнка при рождении от уровня образования родителей?

$$p = 0.2421$$

$$H_0: \beta_4 = \beta_5 = 0$$

$$H_1: H_0 \text{ неверна.}$$

КРИТЕРИЙ ФИШЕРА

- » Зависит ли вес ребёнка при рождении от уровня образования родителей?

$$H_0: \beta_4 = \beta_5 = 0$$

$H_1: H_0$ неверна.

- » Критерий Фишера: $p = 0.2421$

КРИТЕРИЙ ФИШЕРА И СТЪЮДЕНТА

- › При $k_1 = 1$ критерий Фишера эквивалентен критерию Стьюдента для двусторонней альтернативы

КРИТЕРИЙ ФИШЕРА И СТЬЮДЕНТА

- › Иногда критерий Фишера отвергает гипотезу незначимости признаков X_2 , а критерий Стьюдента не признаёт значимым ни один из них.

Возможные объяснения:

- ▶ отдельные признаки из X_2 недостаточно хорошо объясняют y , но совокупный эффект значим
- ▶ признаки в X_2 мультиколлинеарны

КРИТЕРИЙ ФИШЕРА И СТЪЮДЕНТА

- › Иногда критерий Фишера отвергает гипотезу незначимости признаков X_2 , а критерий Стьюдента не признаёт значимым некоторые из них.

Возможные объяснения:

- ▶ незначимые признаки в X_2 маскируют влияние значимых

КРИТЕРИЙ ФИШЕРА И СТЪЮДЕНТА

Возможные объяснения:

- ▶ незначимые признаки в X_2 маскируют влияние значимых
- ▶ значимость отдельных признаков в X_2 — результат множественной проверки гипотез

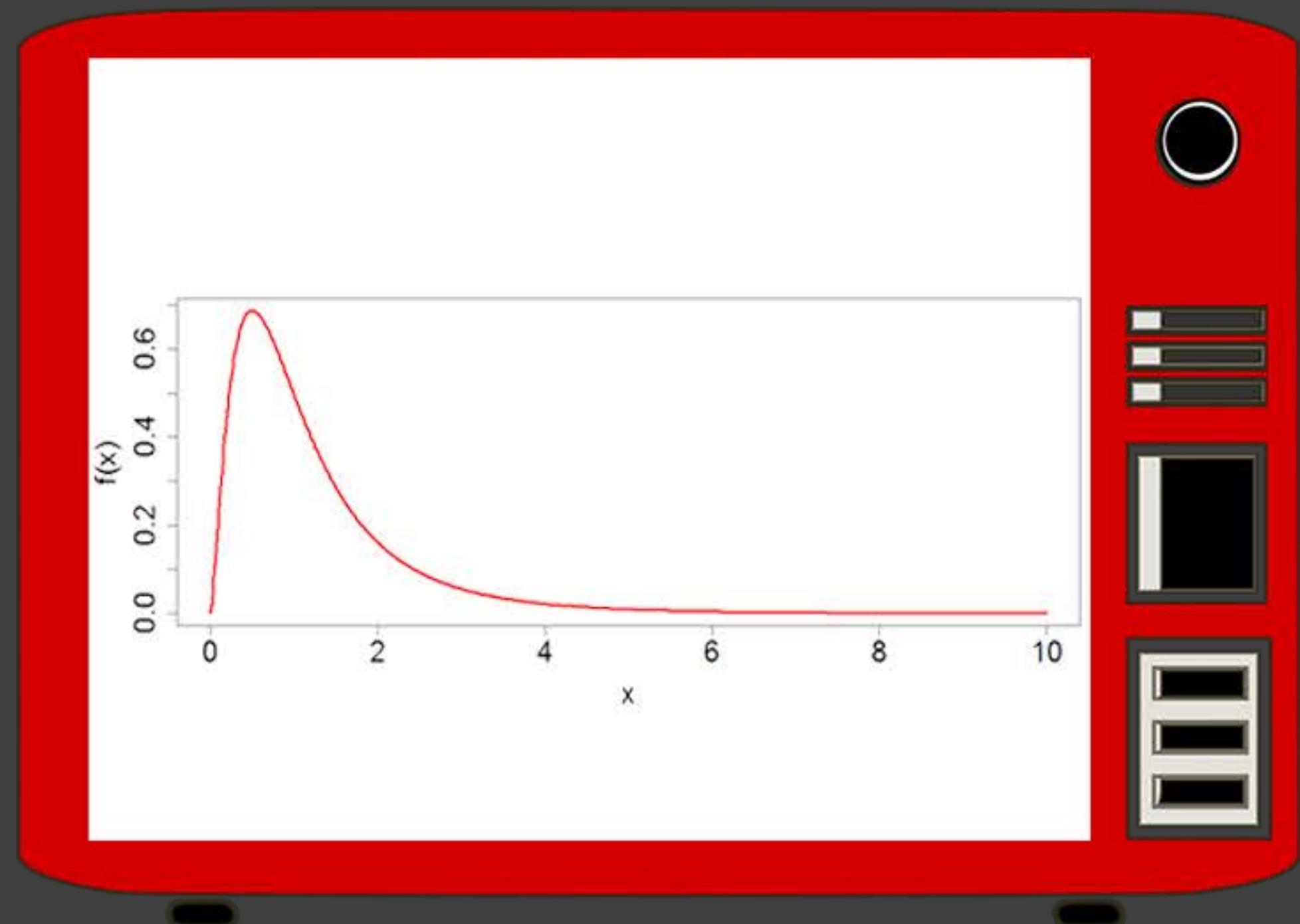
t-КРИТЕРИЙ СТЪЮДЕНТА

нулевая гипотеза: $H_0: \beta_1 = \dots = \beta_k = 0$

альтернатива: $H_1: H_0$ неверна

статистика: $F = \frac{R^2/k}{(1 - R^2) / (n - k - 1)}$

нулевое распределение: $F \sim F(k, n - k - 1)$



КРИТЕРИЙ ФИШЕРА

- › Пример: имеет ли вообще смысл модель веса ребёнка при рождении, рассмотренная выше?

$$H_0: \beta_1 = \dots = \beta_5 = 0$$

H_1 : H_0 неверна.

- › Критерий Фишера: $p = 6 \times 10^{-9}$

- › Доверительные и предсказательные интервалы для отклика
- › Доверительные интервалы и гипотеза о коэффициентах модели
- › Далее: выполнение предположений (1)-(6)

ПРОВЕРКА ПРЕДПОЛОЖЕНИЙ РЕГРЕССИИ

ПРОВЕРКА ПРЕДПОЛОЖЕНИЙ РЕГРЕССИИ

- › Линейность отклика
- › Случайность выборки
- › Полнота ранга X
- › Случайность ошибок
- › Гомоскедастичность ошибок
- › Нормальность ошибок

ЛИНЕЙНОСТЬ ОТКЛИКА

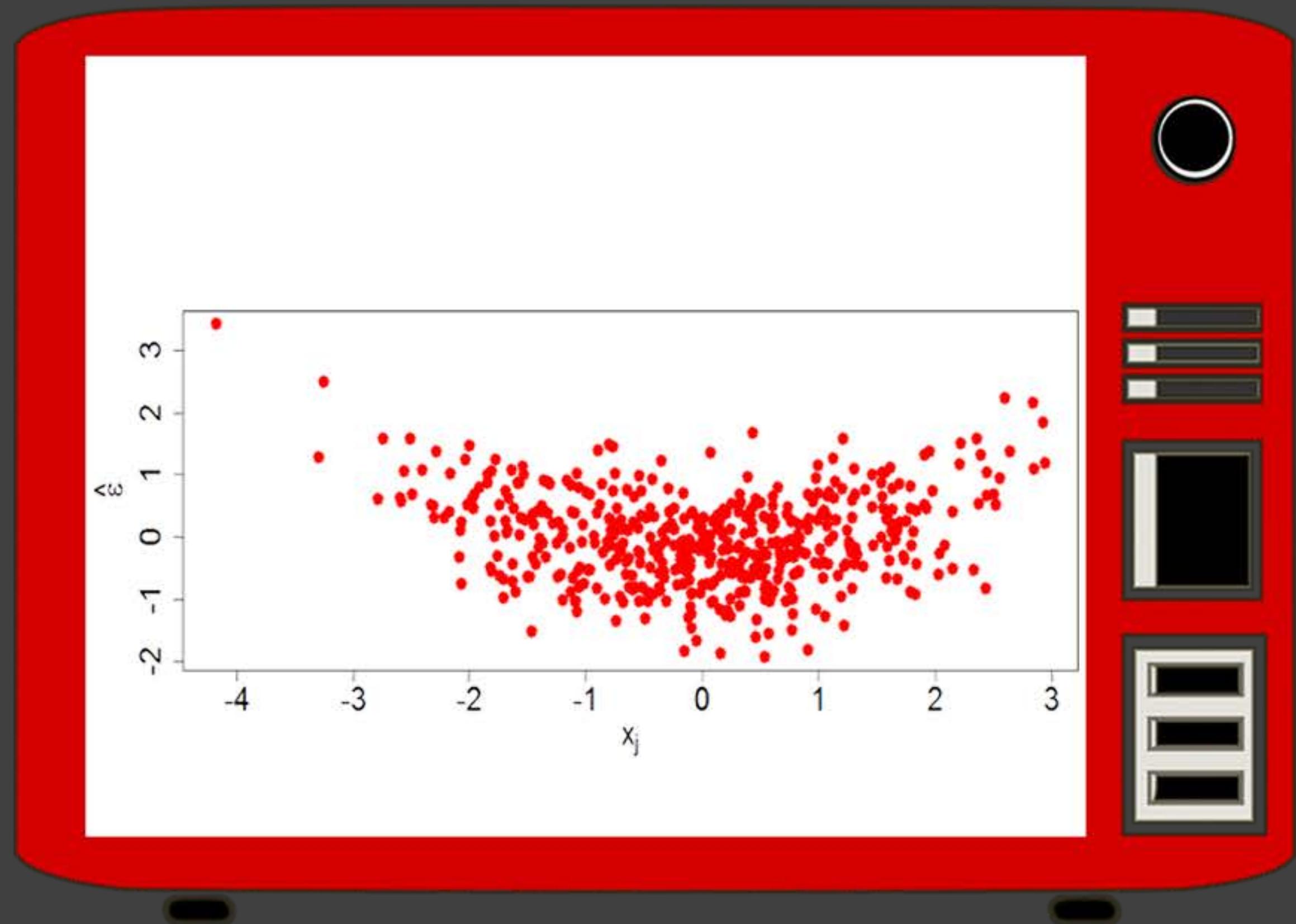
$$y = X\beta + \varepsilon$$

- › В точности не выполняется никогда — все модели неверны.
- › Чтобы убедиться в отсутствии больших отклонений от линейности, нужно анализировать остатки:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n.$$

ЛИНЕЙНОСТЬ ОТКЛИКА

- » Стоит добавить квадрат признака x_j



СЛУЧАЙНОСТЬ ВЫБОРКИ

- » Наблюдения (x_i, y_i) независимы.
- ▶ Если наблюдения зависимы, дисперсия недооценивается, а критерии не работают
- ▶ Фильтровать выборку по признаку z можно только если $\mathbb{E}(y|x, z) = \mathbb{E}(y|x)$

$$\text{rank } X = k + 1$$

- › Если есть линейно зависимые признаки, то дисперсия оценки коэффициентов при них будет бесконечной
- › Никакого one-hot encoding!

ФИКТИВНЫЕ ПЕРЕМЕННЫЕ

- » Если признак x_j принимает m различных значений, то его нужно кодировать $m - 1$ фиктивной переменной.
- » Пусть y — уровень заработной платы, x — должность.

ФИКТИВНЫЕ ПЕРЕМЕННЫЕ

- Пусть y — уровень заработной платы, x — должность.

Dummy-кодирование		
Тип должности	x_1	x_2
рабочий	0	0
инженер	1	0
управляющий	0	1

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

ФИКТИВНЫЕ ПЕРЕМЕННЫЕ

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

β_1, β_2 оценивают среднюю разницу в уровнях зарплат инженера и управляющего с рабочим.

СЛУЧАЙНОСТЬ ОШИБОК

$$\mathbb{E}(\varepsilon | \mathbf{x}) = 0$$

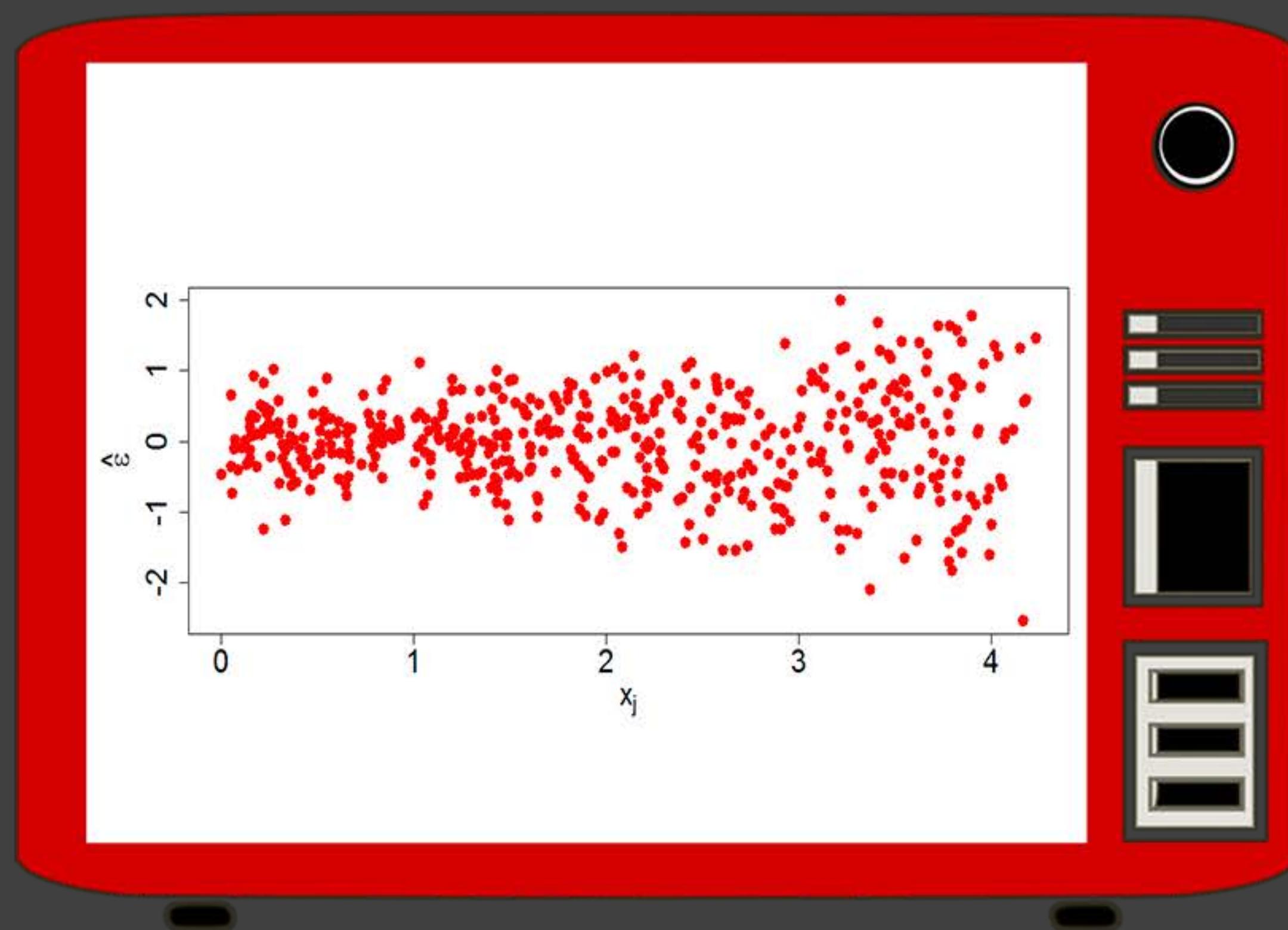
- › Гипотезу $H_0: \mathbb{E}(\varepsilon | \mathbf{x}) = 0$ можно проверить по остаткам критерием Стьюдента.

ГОМОСКЕДАСТИЧНОСТЬ ОШИБОК

$$\mathbb{D}(\varepsilon | x) = \sigma^2$$

› Проверка:

- ▶ Визуальный анализ
- ▶ Критерий Брайша-Пагана



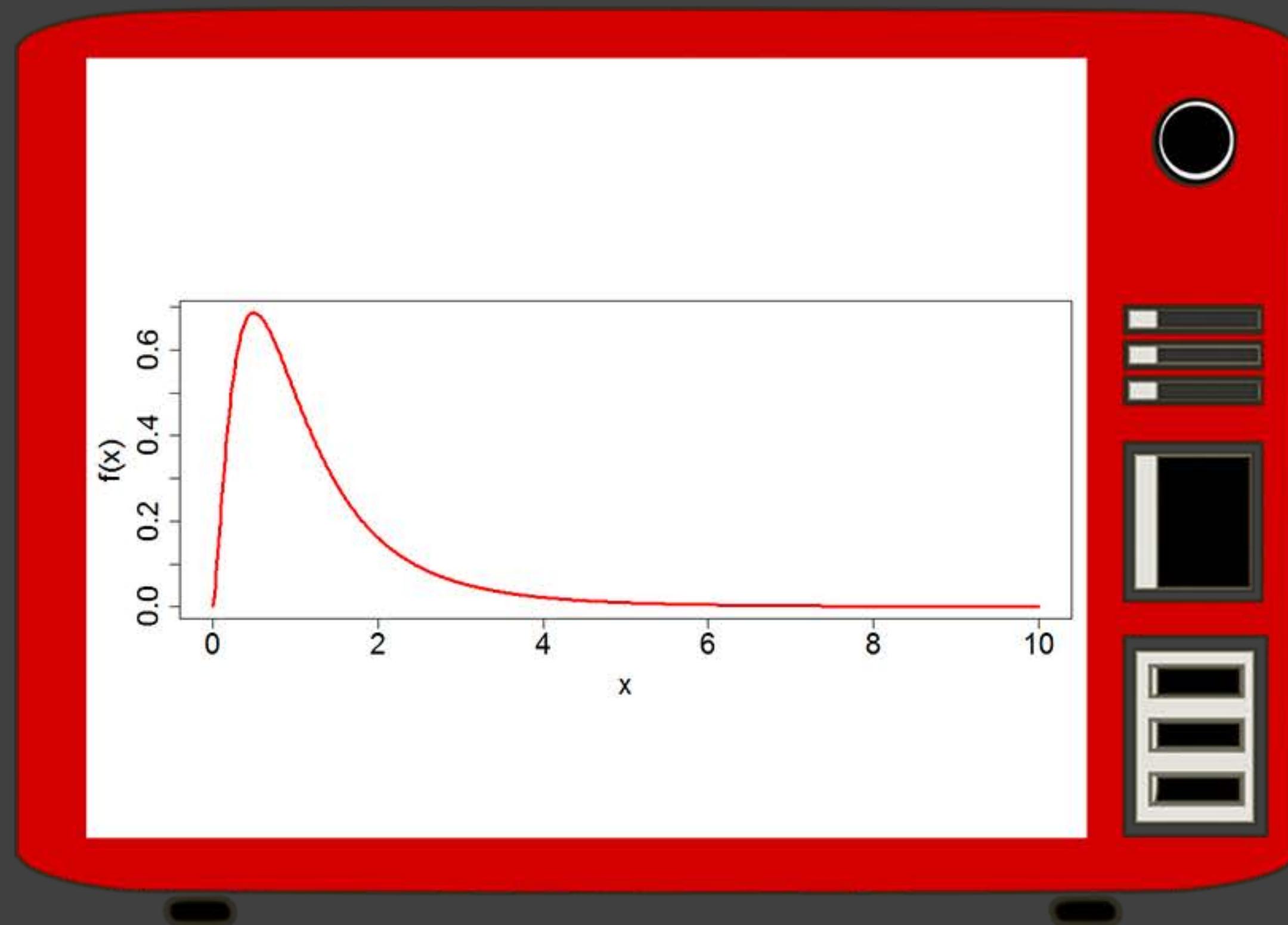
КРИТЕРИЙ БРОЙША-ПАГАНА

нулевая гипотеза: $H_0: \mathbb{D}\varepsilon = \sigma^2$

альтернатива: $H_1: H_0$ неверна

статистика: $LM = nR_{\hat{\varepsilon}^2}^2$, $R_{\hat{\varepsilon}^2}^2$ — коэффициент
детерминации при регрессии $\hat{\varepsilon}^2$ на x

нулевое распределение: $LM \sim \chi_k^2$



НОРМАЛЬНОСТЬ ОШИБОК

$$\varepsilon | x \sim N(0, \sigma^2)$$

› Проверка:

- ▶ Ку-ку график
- ▶ Критерий Шапиро-Уилка

РЕЗЮМЕ

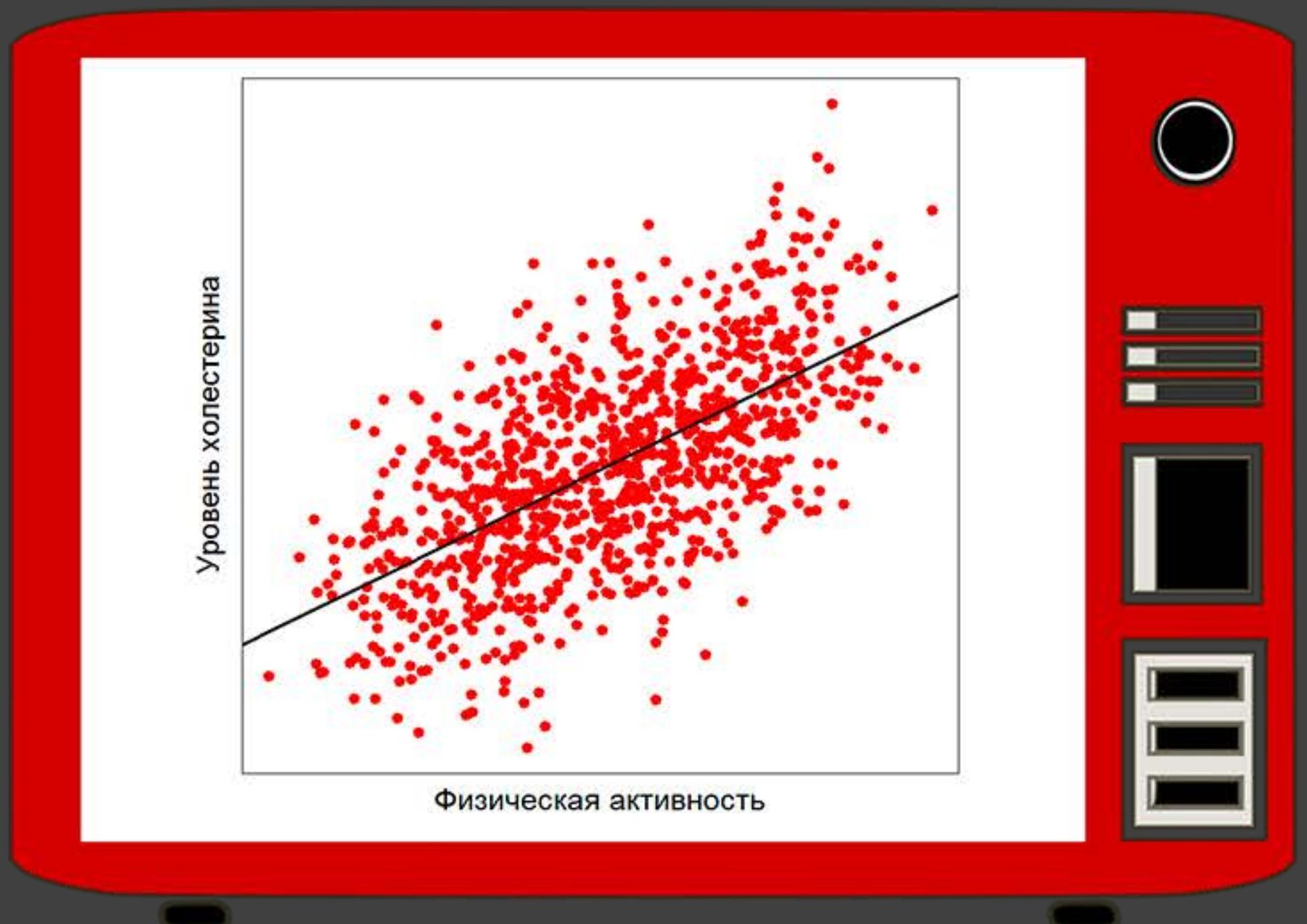
- » Инструменты для проверки предположений (1)-(6)
- » Далее: причинно-следственная интерпретация регрессии

РЕГРЕССИЯ И ПРИЧИННО-СЛЕДСТВЕННЫЕ СВЯЗИ

$$chol = \beta_0 + \beta_1 ex$$

- › $H_0: \beta_1 = 0$
- › $H_0: \beta_1 > 0$

- › Критерий Стьюдента: $p = 2 \times 10^{-16}$



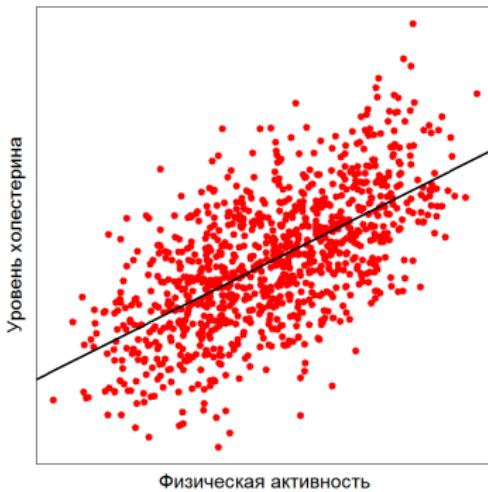
УПРАЖНЕНИЯ И ХОЛЕСТЕРИН

$$chol = \beta_0 + \beta_1 ex + \beta_2 age$$

- › $H_0: \beta_1 = 0$
- › $H_0: \beta_1 < 0$
- › Критерий Стьюдента: $p = 2 \times 10^{-16}$

Регрессия и причинно-следственные связи

Упражнения и холестерин



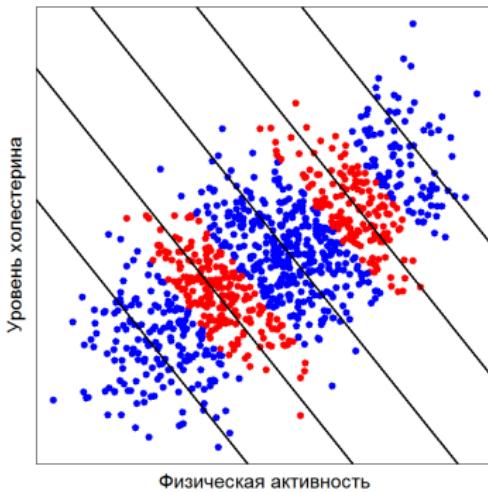
$$chol = \beta_0 + \beta_1 ex$$

$$H_0: \beta_1 = 0$$

$$H_0: \beta_1 > 0$$

Критерий Стьюдента: $p = 2 \times 10^{-16}$.

Упражнения и холестерин



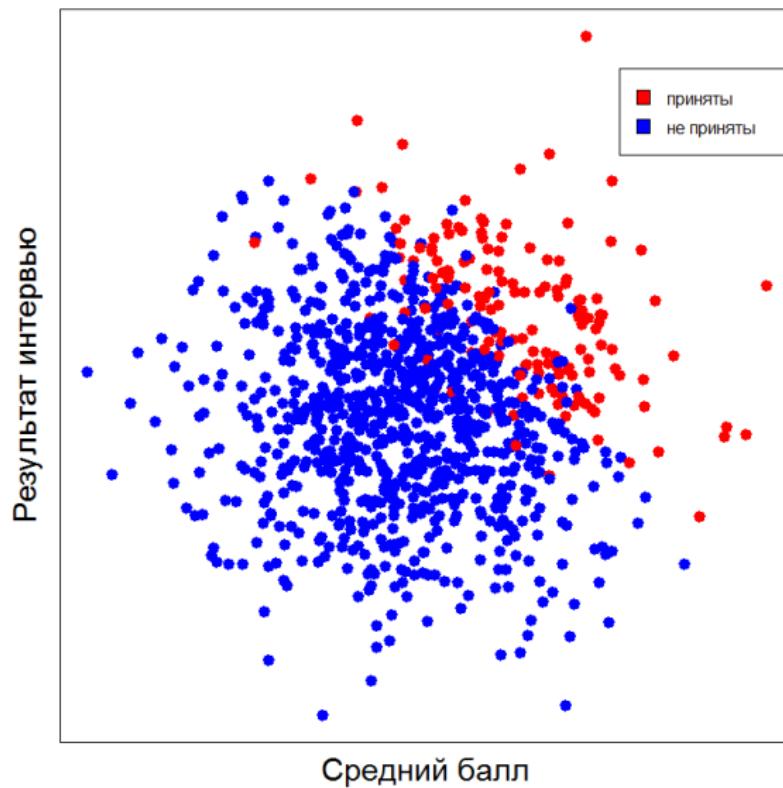
$$chol = \beta_0 + \beta_1 ex + \beta_2 age$$

$$H_0: \beta_1 = 0$$

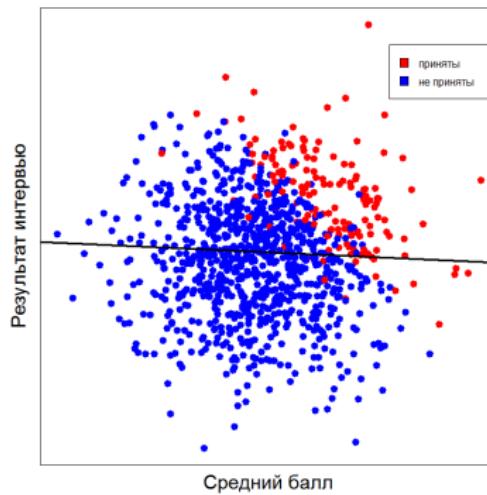
$$H_0: \beta_1 < 0$$

Критерий Стьюдента: $p = 2 \times 10^{-16}$.

Средний балл и мотивация



Средний балл и мотивация



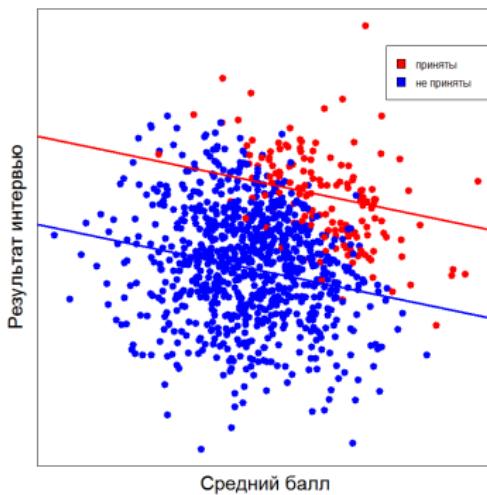
$$mot = \beta_0 + \beta_1 SAT$$

$$H_0: \beta_1 = 0$$

$$H_0: \beta_1 \neq 0$$

Критерий Стьюдента: $p = 0.1452$.

Средний балл и мотивация



$$mot = \beta_0 + \beta_1 SAT + \beta_2 acc$$

$$H_0: \beta_1 = 0$$

$$H_0: \beta_1 \neq 0$$

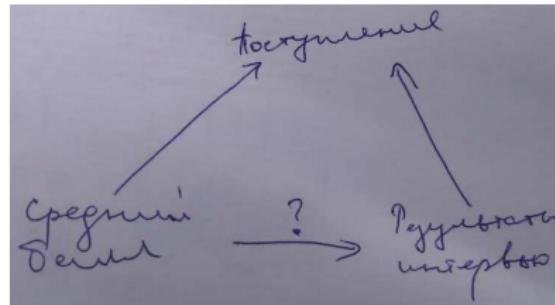
Критерий Стьюдента: $p = 2 \times 10^{-16}$.

В чём разница?

Вилка:



Коллайдер:



Причинно-следственная связь

$\hat{\beta}_1$ — оценка среднего эффекта от увеличения x_1 на единицу, если среди x_2, \dots, x_k :

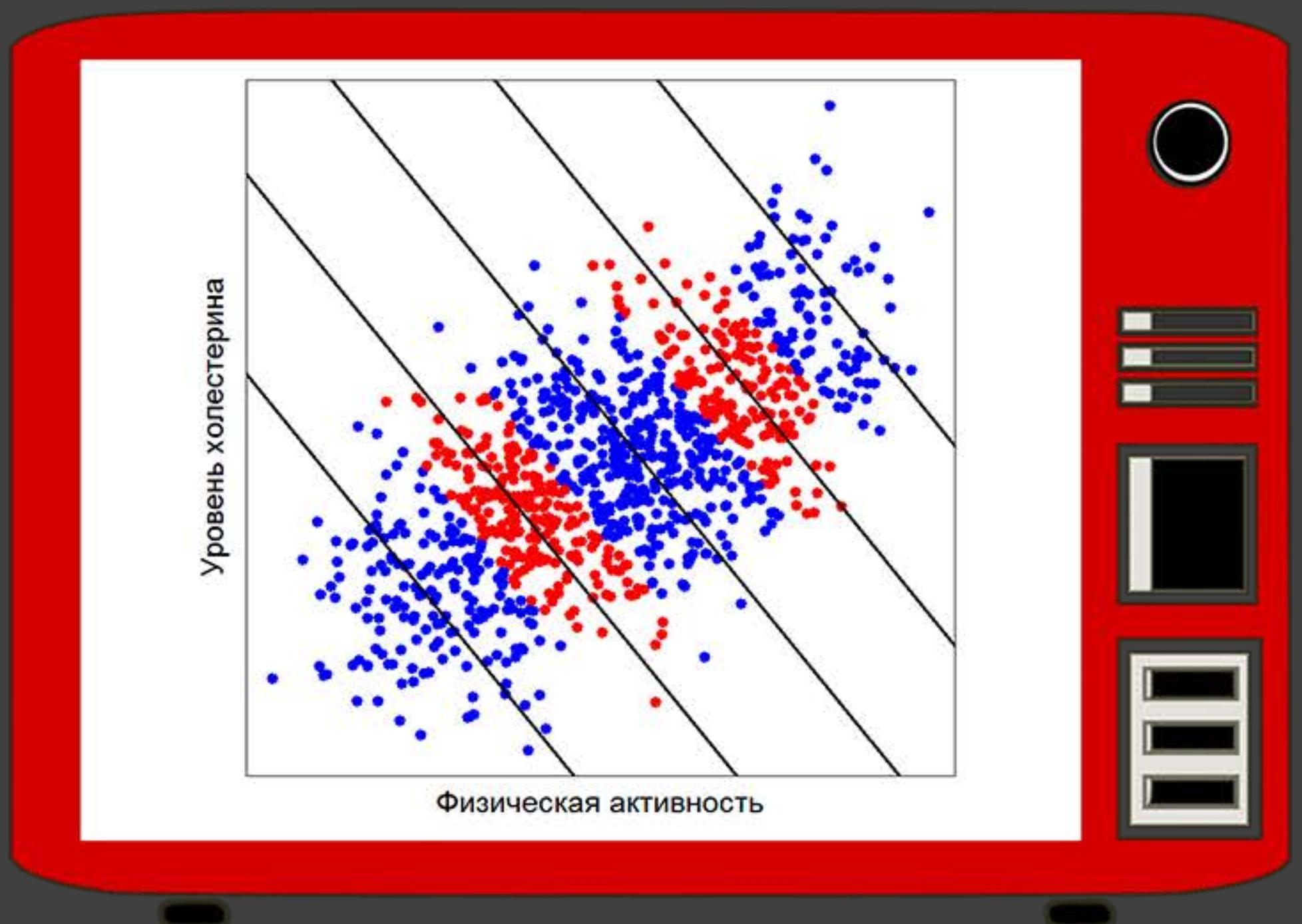
- ▶ содержатся все признаки, являющиеся причинами x_1
- ▶ не содержится признаков, являющихся следствиями одновременно x_1 и y

Резюме

- ▶ иногда регрессия позволяет обнаруживать причинно-следственные связи!
- ▶ плохо подобранные признаки могут привести к противоположным выводам

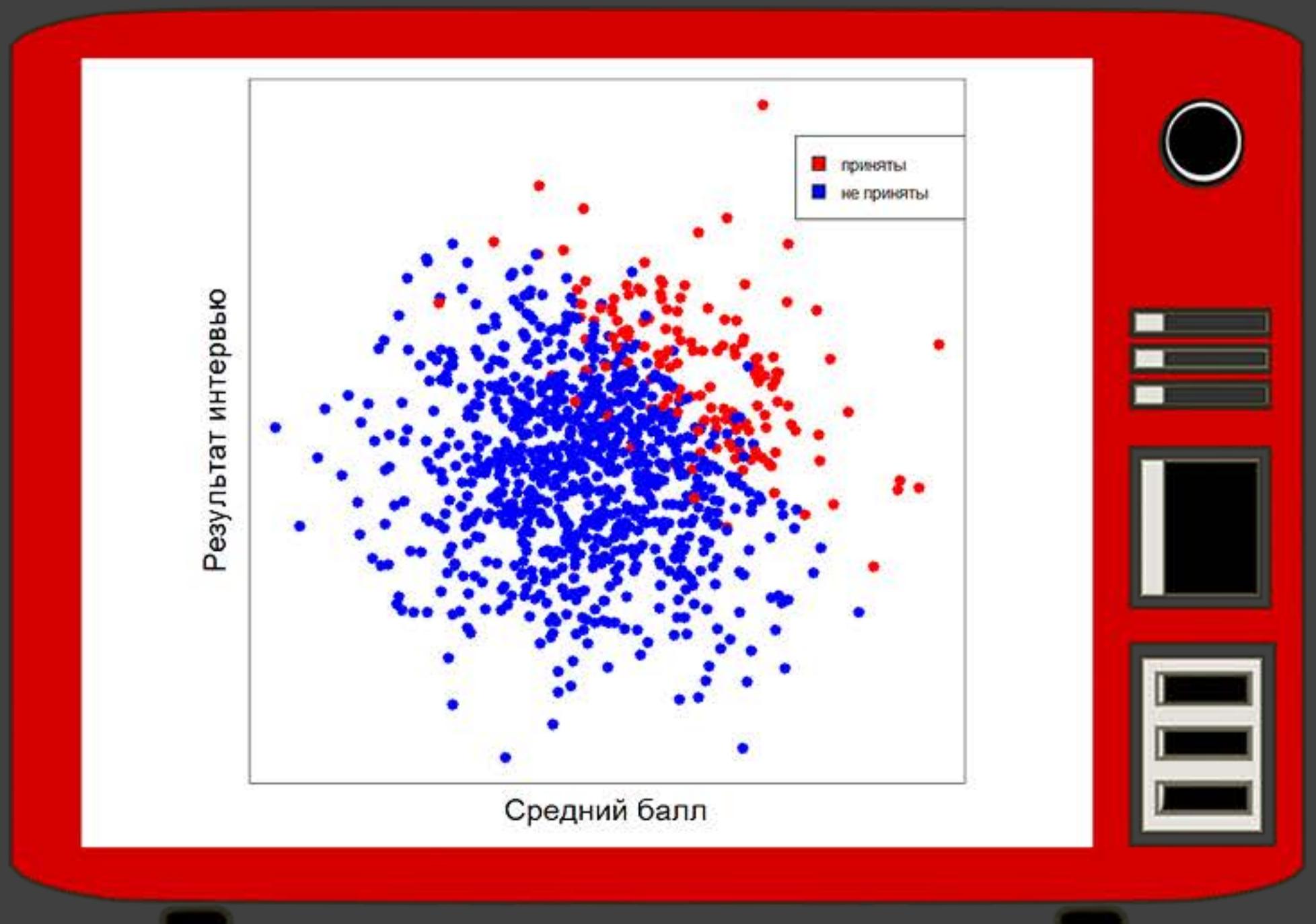
Далее в программе

- ▶ пример



СРЕДНИЙ БАЛЛ И МОТИВАЦИЯ



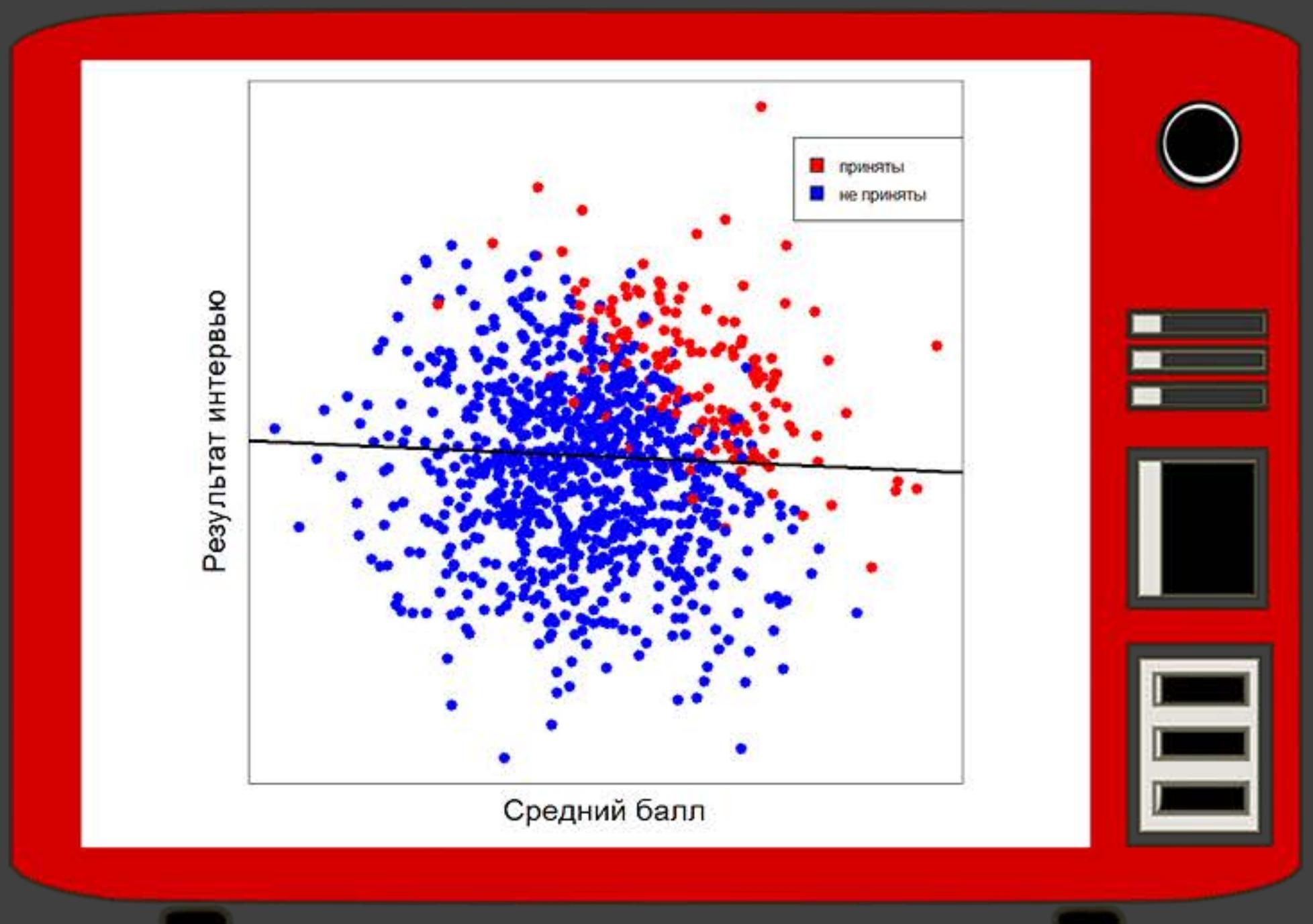


СРЕДНИЙ БАЛЛ И МОТИВАЦИЯ

$$mot = \beta_0 + \beta_1 SAT$$

- › $H_0: \beta_1 = 0$
- › $H_0: \beta_1 \neq 0$

- › Критерий Стьюдента: $p = 0.1452$

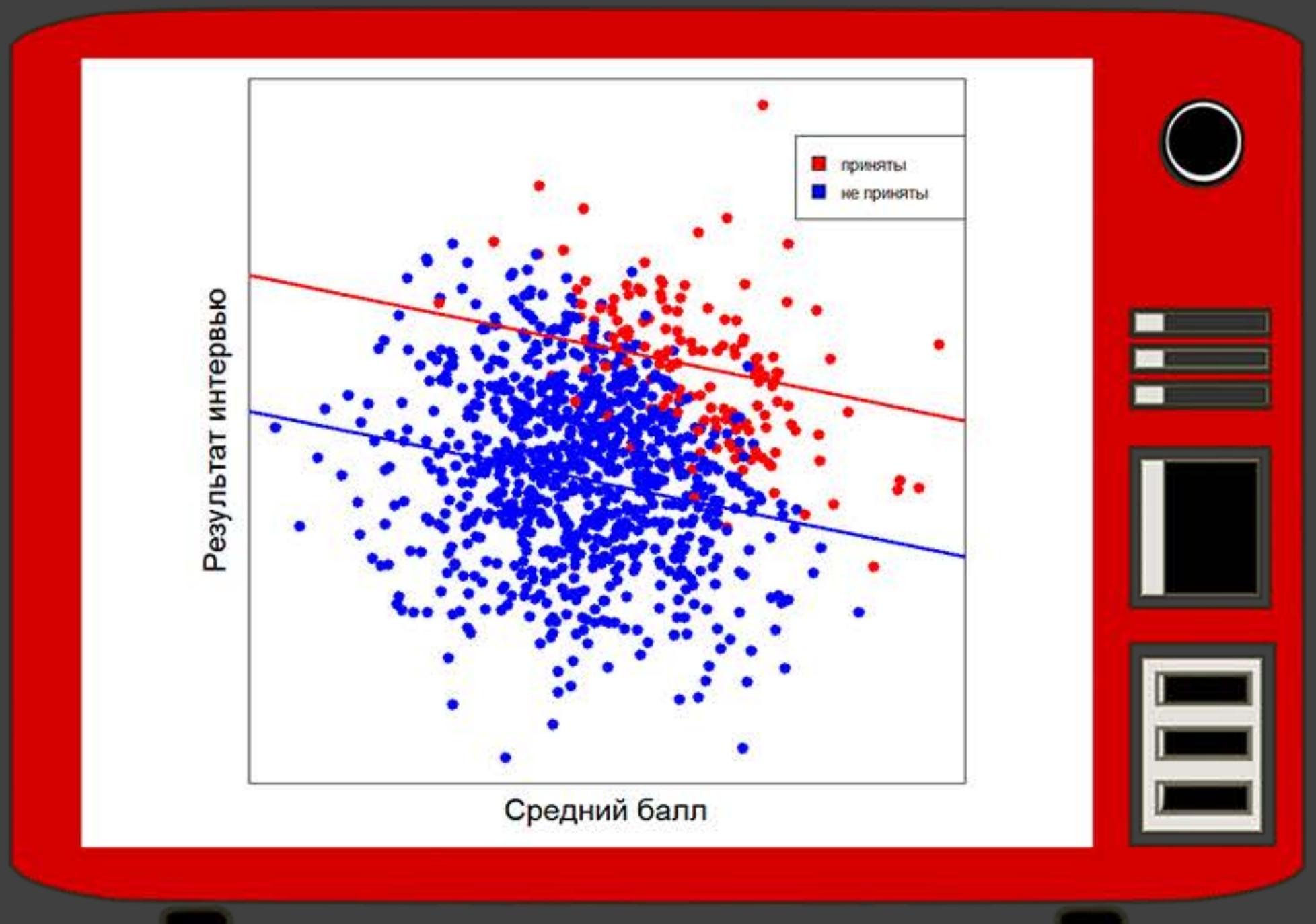


СРЕДНИЙ БАЛЛ И МОТИВАЦИЯ

$$mot = \beta_0 + \beta_1 SAT + \beta_2 acc$$

- $H_0: \beta_1 = 0$
- $H_0: \beta_1 \neq 0$

- Критерий Стьюдента: $p = 2 \times 10^{-16}$

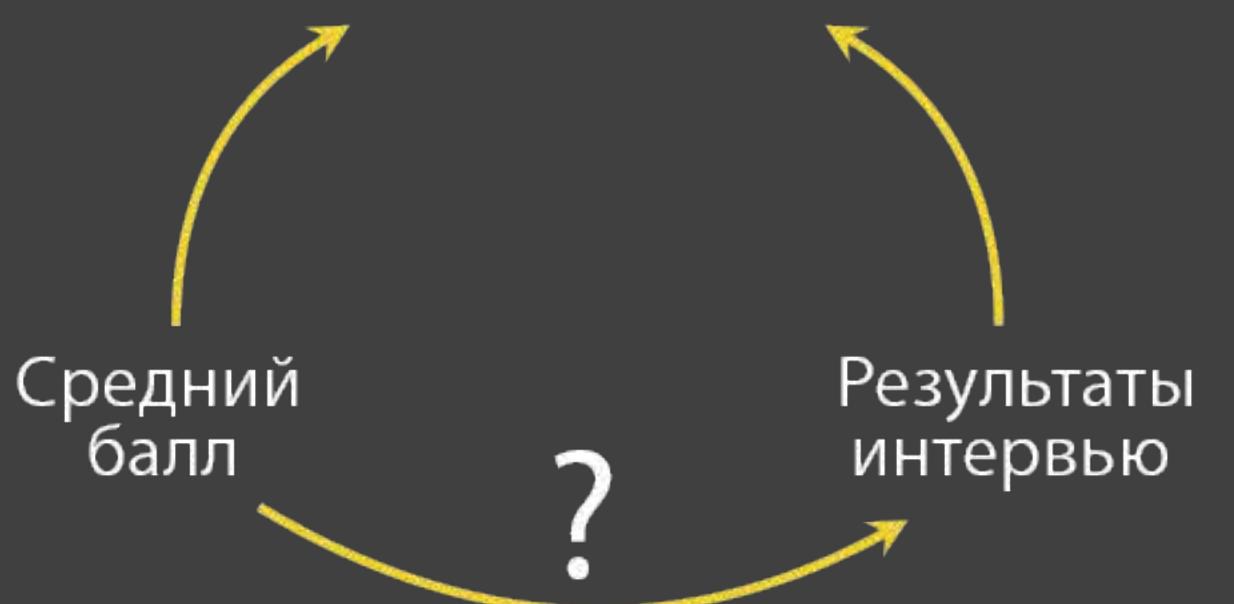


В ЧЁМ РАЗНИЦА?

» Вилка:



» Коллайдер:



ПРИЧИННО-СЛЕДСТВЕННАЯ СВЯЗЬ

- $\hat{\beta}_1$ — оценка среднего эффекта от увеличения x_1 на единицу, если среди x_2, \dots, x_k
- ▶ Содержатся все признаки, являющиеся причинами x_1
- ▶ Не содержится признаков, являющихся следствиями одновременно x_1 и y

- › Иногда регрессия позволяет обнаруживать причинно-следственные связи!
- › Плохо подобранные признаки могут привести к противоположным выводам
- › Далее: пример