MML minor #12

Линейная регрессия

Thanks

Рябенко Евгений
riabenko.e@gmail.com

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

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

Неэтично!

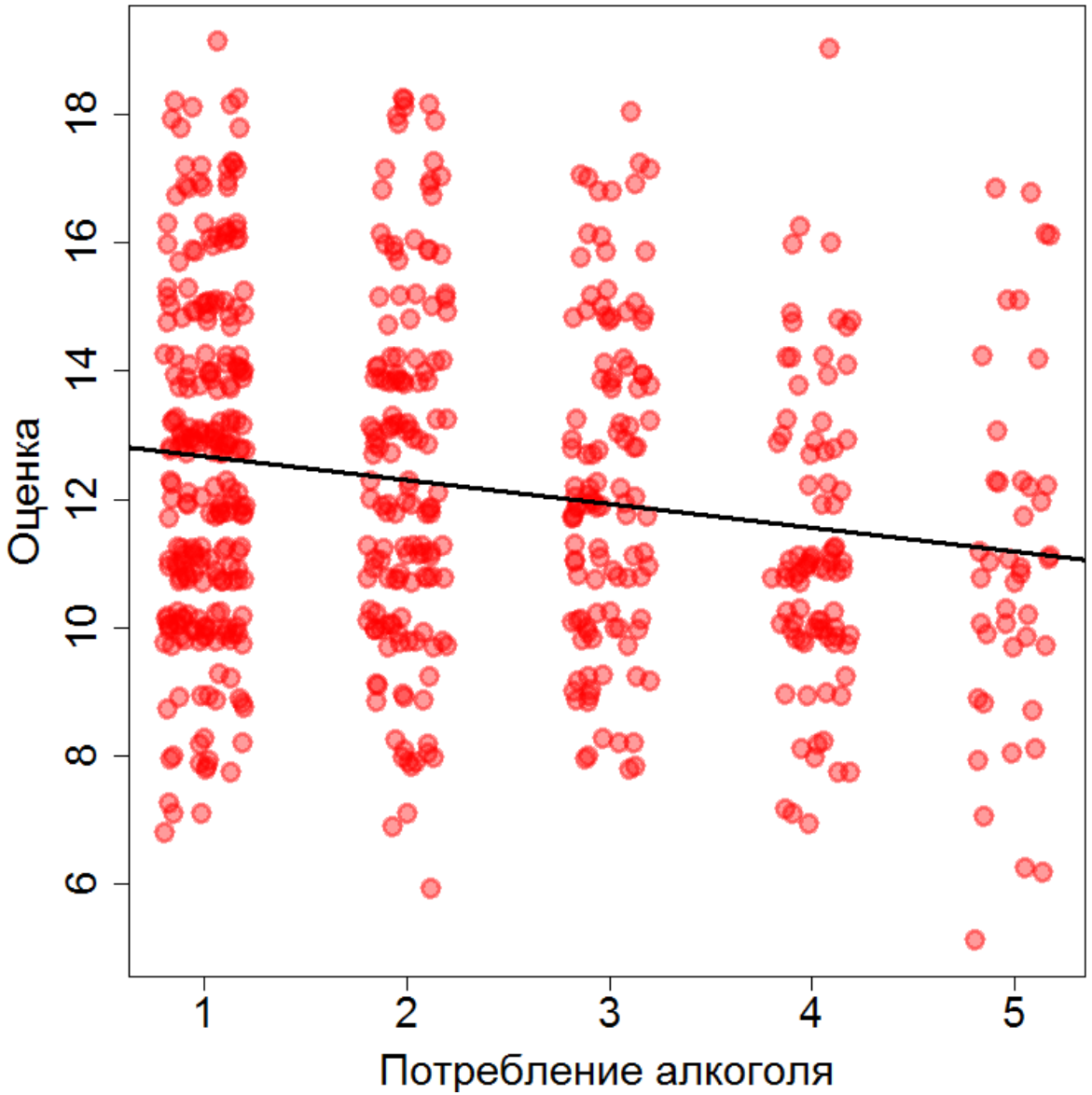
Успеваемость школьников

Эксперимент провести нельзя, но есть обзервационные данные.

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.

Cortez P., Silva A. (2008). *Using Data Mining to Predict Secondary School Student Performance*. Proceedings of 5th FUture BUsiness TEChnology Conference, pp. 5-12.

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

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

Модель линейной регрессии:

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

β_j показывает, насколько в среднем увеличивается y, если x_j увеличивается на единицу, а остальные факторы фиксированы.
$\Rightarrow$ регрессию можно использовать для исследования остаточного влияния признака на отклик с учётом других признаков.

Линейная регрессия

Модель линейной регрессии:

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

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

МНК

Метод наименьших квадратов:

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

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

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

Качество решения

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{(Total Sum of Squares);}$$

$$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{(Explained Sum of Squares);}$$

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{(Residual Sum of Squares);}$$

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}\,.$$

Коэффициент детерминации:

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

Приведённый коэффициент детерминации

Стандартный коэффициент детерминации всегда увеличивается при добавлении регрессоров в модель.

Для сравнения качества моделей, содержащих разное число признаков, можно использовать приведённый коэффициент детерминации:

$$R_a^2 = \frac{\mathrm{ESS}\,/(n-k-1)}{\mathrm{TSS}\,/(n-1)} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k-1}.$$

Базовые предположения МНК

1. Линейность отклика:

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

Базовые предположения МНК

1. Линейность отклика
2. Случайность выборки: наблюдения (x_i, y_i) независимы.

Базовые предположения МНК

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

Базовые предположения МНК

1. Линейность отклика
2. Случайность выборки
3. Полнота ранга X
4. Случайность ошибок:

$$\mathbb{E}\left(\varepsilon \,|x\right) = 0$$

Базовые предположения МНК

1. Линейность отклика
2. Случайность выборки
3. Полнота ранга X
4. Случайность ошибок

$\Rightarrow$ МНК-оценки коэффициентов β несмещённые:

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

и состоятельные:

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

Базовые предположения МНК

1. Линейность отклика
2. Случайность выборки
3. Полнота ранга X
4. Случайность ошибок
5. Гомоскедастичность ошибок: дисперсия ошибки не зависит от значений признаков: $\mathbb{D}(\varepsilon \,|x\,) = \sigma^2$

(предположения Гаусса-Маркова)

Базовые предположения МНК

1. Линейность отклика
2. Случайность выборки
3. Полнота ранга X
4. Случайность ошибок
5. Гомоскедастичность ошибок

$\Rightarrow$ МНК-оценки имеют наименьшую дисперсию в классе оценок β, линейных по y.

Дисперсия $\hat{\beta}_j$

(1)-(5)$\Rightarrow$

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

где $TSS_j = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$, R_j^2 — коэффициент детерминации при регрессии x_j на все остальные признаки.

- 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:

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

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

1. Линейность отклика
2. Случайность выборки
3. Полнота ранга X
4. Случайность ошибок
5. Гомоскедастичность ошибок
6. Нормальность ошибок:

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

Эквивалентная запись: $y \,|x \sim N\left(x\beta, \sigma^2\right)$

Нормальность

1. Линейность отклика
2. Случайность выборки
3. Полнота ранга X
4. Случайность ошибок
5. Гомоскедастичность ошибок
6. Нормальность ошибок

$\Rightarrow$ МНК-оценки совпадают с оценками максимального правдоподобия

Нормальность

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

- имеют наименьшую дисперсию среди всех несмещённых оценок β
- имеют нормальное распределение $N\left(\beta, \sigma^2 \left(X^T X\right)^{-1}\right)$
- $\hat{\sigma}^2 = \frac{1}{n-k-1} \mathrm{RSS}$ — несмещённая оценка σ^2, и

$$\frac{\mathrm{RSS}}{\sigma^2} \sim \chi^2_{n-k-1}$$

- $\forall c \in \mathbb{R}^{k+1}$

$$\frac{c^T \left(\beta - \hat{\beta}\right)}{\hat{\sigma}\sqrt{c^T \left(X^T X\right)^{-1} c}} \sim St(n-k-1)$$

Последствия

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

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

Слайды далее на лекции не смотрели.
Основной посыл этой презентации такой:
При определенных предположениях можно
строить доверительные интервалы для коэффициентов и предсказаний.

Листайте до слайда 38

Доверительные и предсказательные интервалы

- $100(1-\alpha)\%$ доверительный интервал для σ^2:

$$\frac{\mathrm{RSS}}{\chi^2_{n-k-1,1-\alpha/2}} \leqslant \sigma^2 \leqslant \frac{\mathrm{RSS}}{\chi^2_{n-k-1,\alpha/2}}$$

- Возьмём $c = \left(0 \dots 0 \underset{j}{1} 0 \dots 0\right)$; $100(1-\alpha)\%$ доверительный интервал для β_j:

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

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

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

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

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

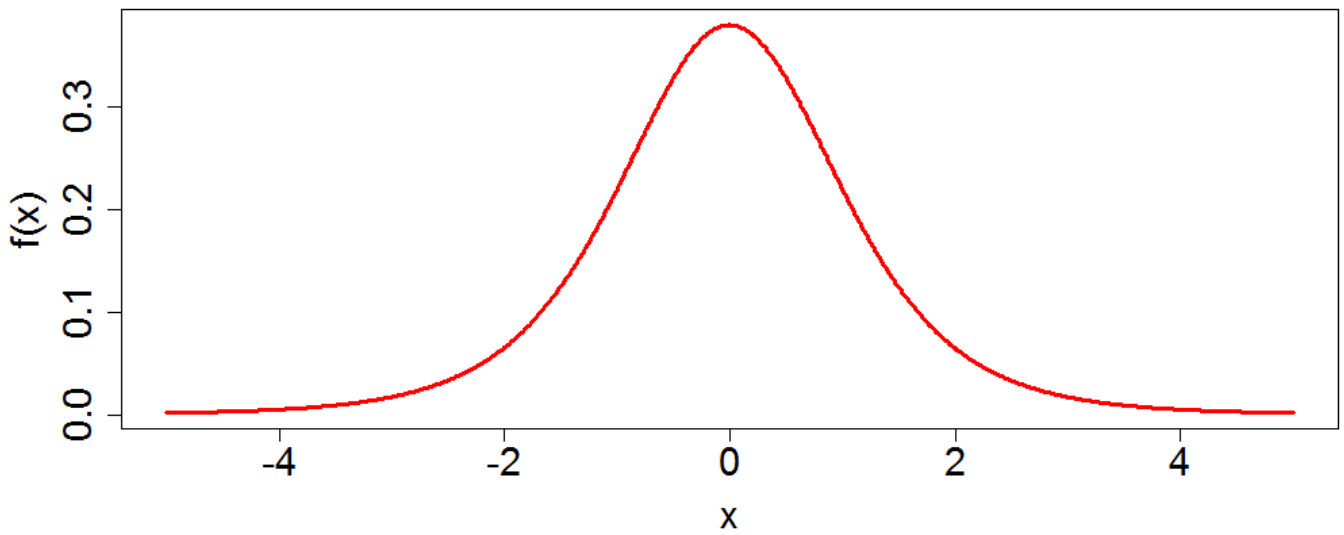
t-критерий Стьюдента

нулевая гипотеза: $H_0 : \beta_j = 0$

альтернатива: $H_1 : \beta_j <\neq> 0$

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

нулевое распределение: $St(n-k-1)$



t-критерий Стьюдента

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

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

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

$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2}\right) \qquad \underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T}\right)^T$$

нулевая гипотеза: $H_0 \colon \beta_2 = 0$

альтернатива: $H_1 \colon H_0$ неверна

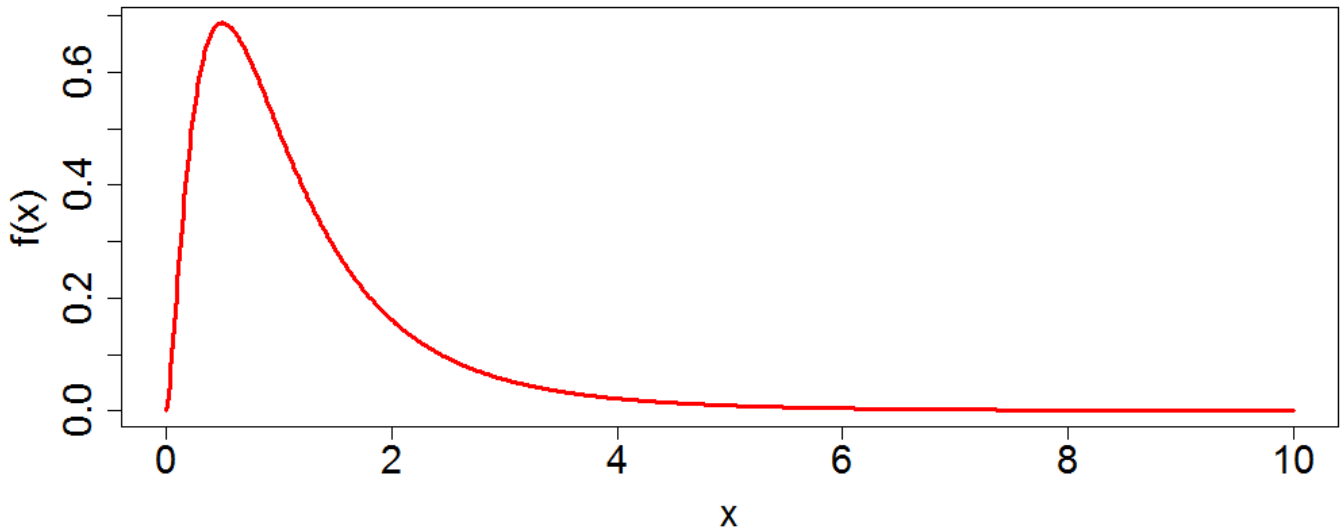
статистика:

$$\mathrm{RSS}_r = \|y - X_1\beta_1\|_2^2$$

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

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

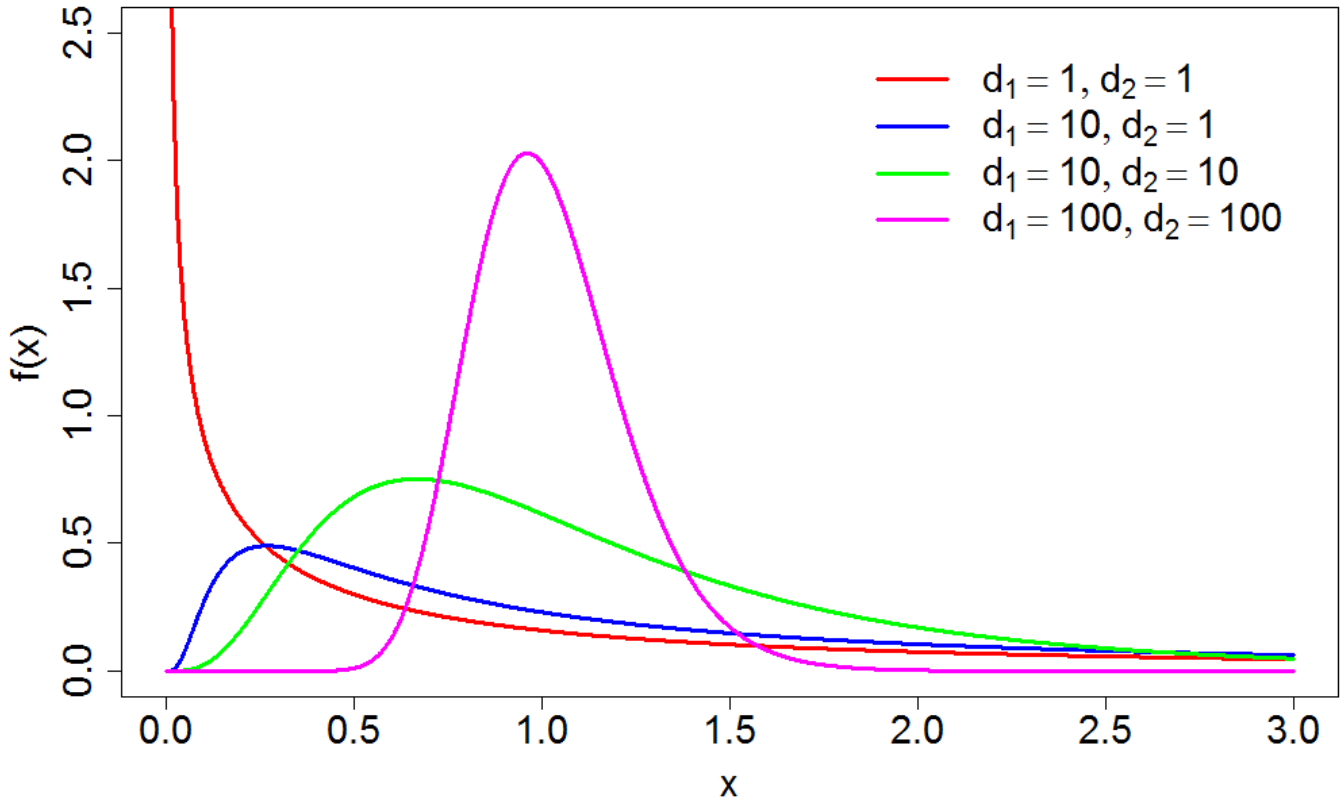
нулевое распределение: $F(k_1, n-k-1)$



Критерий Фишера

$X_1 \sim \chi^2_{d_1}, X_2 \sim \chi^2_{d_2}$ независимы,

$X = \frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2)$ — распределение Фишера с d_1, d_2 степенями свободы.



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,
med — длительность получения образования матерью, fed — отцом.

Зависит ли вес ребёнка при рождении от уровня образования родителей?

$H_0 : \beta_4 = \beta_5 = 0.$
$H_1 : H_0$ неверна.

Критерий Фишера: $p = 0.2421.$

Критерии Фишера и Стьюдента

- При $k_1 = 1$ критерий Фишера эквивалентен критерию Стьюдента для двусторонней альтернативы
- 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
- Иногда критерия Фишера не отвергает гипотезу незначимости признаков X_2, а критерий Стьюдента признаёт значимыми некоторые из них.
  Возможные объяснения:
  - незначимые признаки в X_2 маскируют влияние значимых
  - значимость отдельных признаков в X_2 — результат множественной проверки гипотез

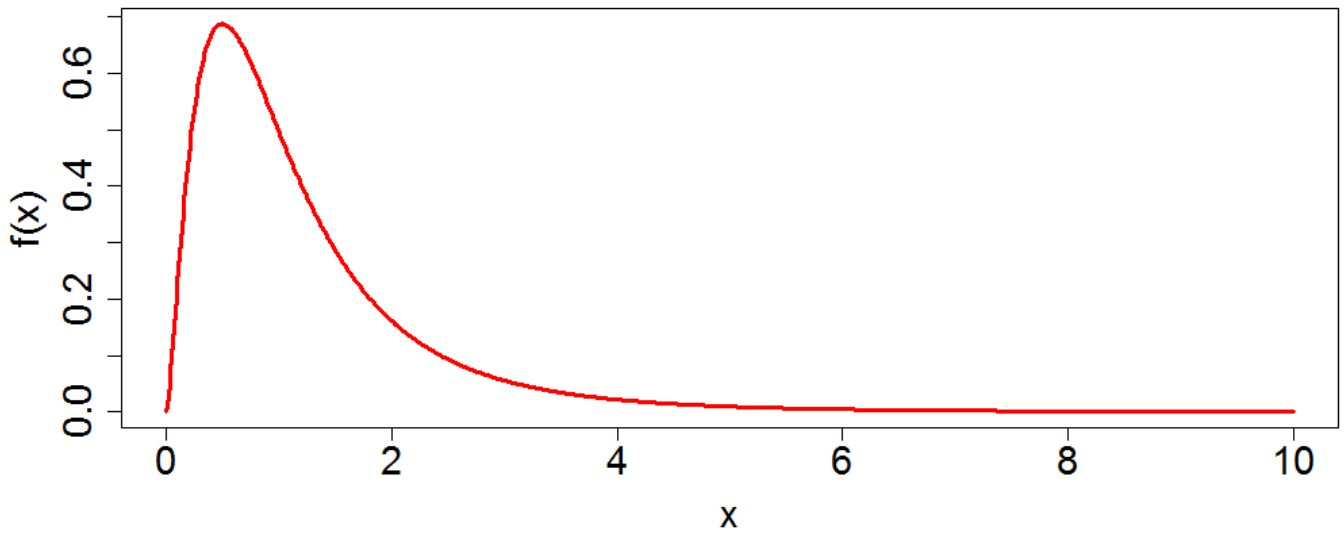
Критерий Фишера

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

альтернатива: $H_1 :$ H_0 неверна

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

нулевое распределение: $F(k, n-k-1)$



Критерий Фишера

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

H_0: $\beta_1 = \cdots = \beta_5 = 0$.
H_1: H_0 неверна.

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

Отбор признаков

Незначимые признаки можно исключать из модели — доказательств тому, что они влияют на y, нет!

Недоопределение: если зависимая переменная определяется моделью

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{j-1} x_{j-1} + \beta_j x_j + \beta_{j+1} x_{j+1} + \cdots + \beta_k x_k + \varepsilon,$$

а вместо этого используется модель

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + \cdots + \beta_k x_k + \varepsilon,$$

то МНК-оценки $\hat{\beta}_0, \ldots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_k$ являются смещёнными и несостоятельными оценками $\beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k$.

Переопределение: если признак x_j не влияет на y, т. е. $\beta_j = 0$, то МНК-оценка $\hat{\beta}$ остаётся несмещённой состоятельной оценкой β, но дисперсия её возрастает.

$\Rightarrow$ исключая незначимые признаки, мы рискуем получить смещённые оценки оставшихся коэффициентов, но уменьшаем их дисперсию.

Best subset

Для каждого k полным перебором по R^2 можно выбрать лучшую модель с k признаками, затем по R_a^2 среди них можно выбрать одну лучшую модель.

Полный перебор требует больших вычислительных затрат.

Пошаговая регрессия

- **Шаг 0**. Настраивается модель с одной только константой, а также все модели с одной переменной. Рассчитывается F-статистика каждой модели и достигаемый уровень значимости. Выбирается модель с наименьшим достигаемым уровнем значимости. Соответствующая переменная X_{e1} включается в модель, если этот достигаемый уровень значимости меньше порогового значения $p_E = 0.05$.
- **Шаг 1**. Рассчитывается F-статистика и достигаемый уровень значимости для всех моделей, содержащих две переменные, одна из которых X_{e1}. Аналогично принимается решение о включении X_{e2}.
- **Шаг 2**. Если была добавлена переменная X_{e2}, возможно, X_{e1} уже не нужна. В общем случае просчитываются все возможные варианты исключения одной переменной, рассматривается вариант с наибольшим достигаемым уровнем значимости, соответствующая переменная исключается, если он превосходит пороговое значение $p_R = 0.1$.
- ...

Эксперимент Фридмана

(Freedman, 1983): best subset и пошаговая регрессия несовместимы с проверкой гипотез о значимости коэффициентов: критерии Фишера и Стьюдента антиконсервативны , если вычисляются на той же самой выборке, на которой настраивалась модель.

Если мы хотим считать значимость признаков, признаки должны отбираться не слишком интенсивно (или на другой выборке).

Проверка предположений регрессии

На лекции не успели, самостоятельно изучите

1. Линейность отклика
2. Случайность выборки
3. Полнота ранга X
4. Случайность ошибок
5. Гомоскедастичность ошибок
6. Нормальность ошибок

1. Линейность отклика

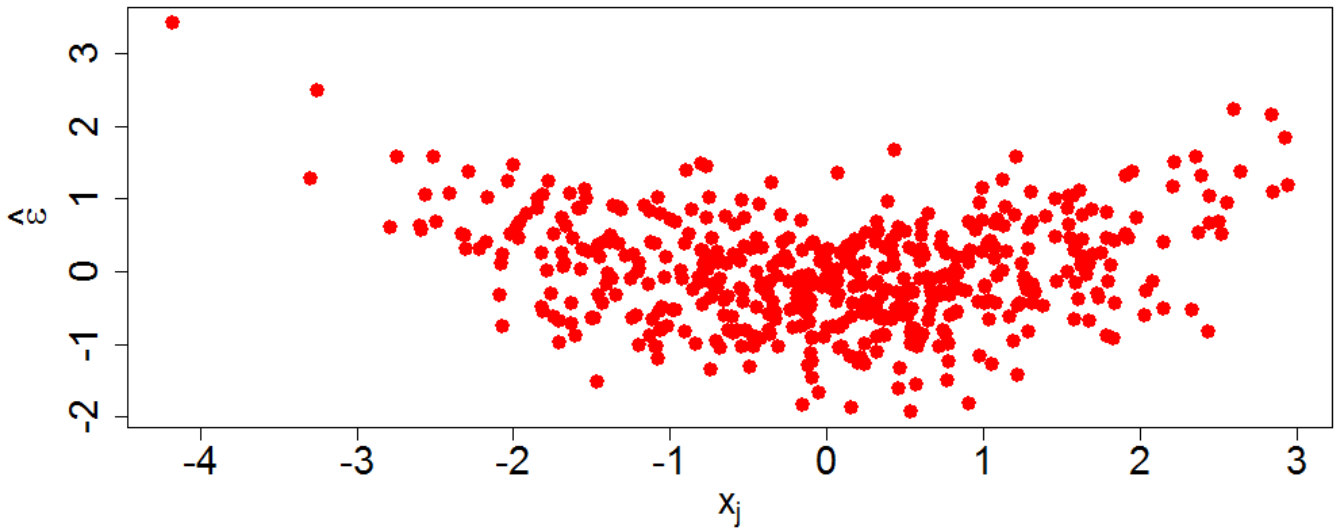
$$y = X\beta + \varepsilon$$

В точности не выполняется никогда — все модели неверны.

Чтобы убедиться в отсутствии больших отклонений от линейнойсти, нужно анализировать остатки:

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

1. Линейность отклика



Стоит добавить квадрат признака x_j

2. 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 $\mathbb{E}(y\,|x, z) = \mathbb{E}(y\,|x)$

3. Полнота ранга

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

- 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.

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

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

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

4. Случайность ошибок

$$\mathbb{E}\left(\varepsilon \,|x\right) = 0$$

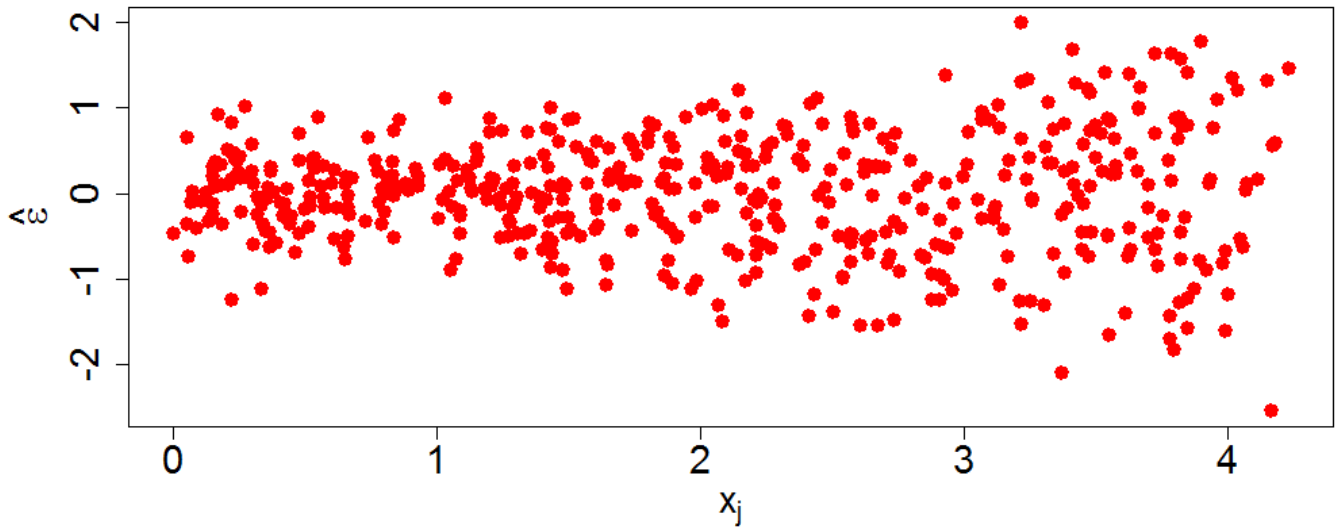
Гипотезу $H_0 : \mathbb{E}\left(\varepsilon \,|x\right) = 0$ можно проверить по остаткам критерием Стьюдента.

5. Гомоскедастичность ошибок

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

Проверка:

- визуальный анализ:



- критерий Бройша-Пагана

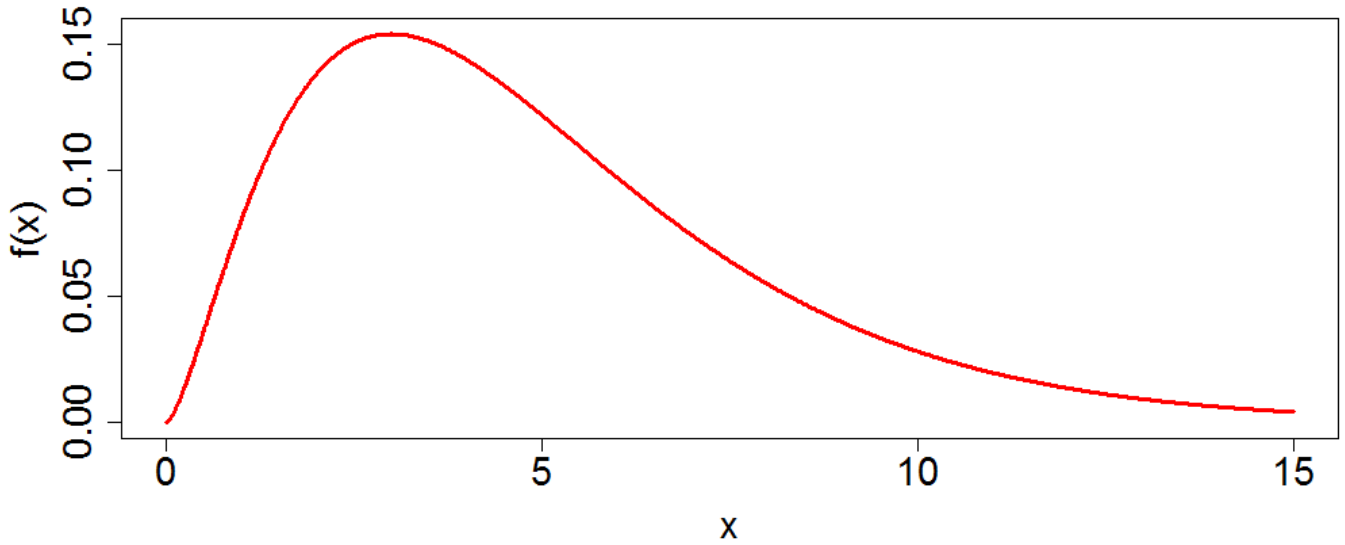
Критерий Бройша-Пагана

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

альтернатива: $H_1 : H_0$ неверна

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

нулевое распределение: χ^2_k



Гетероскедастичность

Гетероскедастичность может быть следствием недоопределения модели.

Последствия гетероскедастичности:

- МНК-оценки β и R^2 остаются несмещёнными и состоятельными
- нарушаются предположения критериев Стьюдента и Фишера и методов построения доверительных интервалов для σ и β (независимо от объёма выборки)

Варианты:

- переопределить модель, добавить признаки, преобразовать отклик
- использовать модифицированные оценки дисперсии коэффициентов

Устойчивая оценка дисперсии Уайта

Далее читать не нужно

Если не удаётся избавиться от гетероскедастичности, при анализе моделей (дальше) можно использовать устойчивые оценки дисперсии.
White's heteroscedasticity-consistent estimator (HCE):

$$\mathbb{D}\left(\hat{\beta}\middle| X\right) = \left(X^T X\right)^{-1} \left(X^T \operatorname{diag}\left(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\right) X\right) \left(X^T X\right)^{-1}.$$

Асимптотика устойчивой оценки:

$$\sqrt{n}\left(\beta - \hat{\beta}\right) \xrightarrow{d} N\left(0, \Omega\right),$$

$$\hat{\Omega} = n\left(X^T X\right)^{-1} \left(X^T \operatorname{diag}\left(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\right) X\right) \left(X^T X\right)^{-1}.$$

Другие устойчивые оценки дисперсии

Элементы диагональной матрицы могут задаваться разными способами:

const	$\hat{\sigma}^2$
HC0	$\hat{\varepsilon}_i^2$
HC1	$\frac{n}{n-k}\hat{\varepsilon}_i^2$
HC2	$\frac{\hat{\varepsilon}_i^2}{1-h_i}$
HC3	$\frac{\hat{\varepsilon}_i^2}{(1-h_i)^2}$
HC4	$\frac{\hat{\varepsilon}_i^2}{(1-h_i)^{\min\left(4, \frac{nh_i}{k}\right)}}$

const — случай гомоскедастичной ошибки,
HC0 — оценка Уайта,
HC1–HC3 — модификации МакКиннона-Уайта,
HC4 — модификация Крибари-Нето.

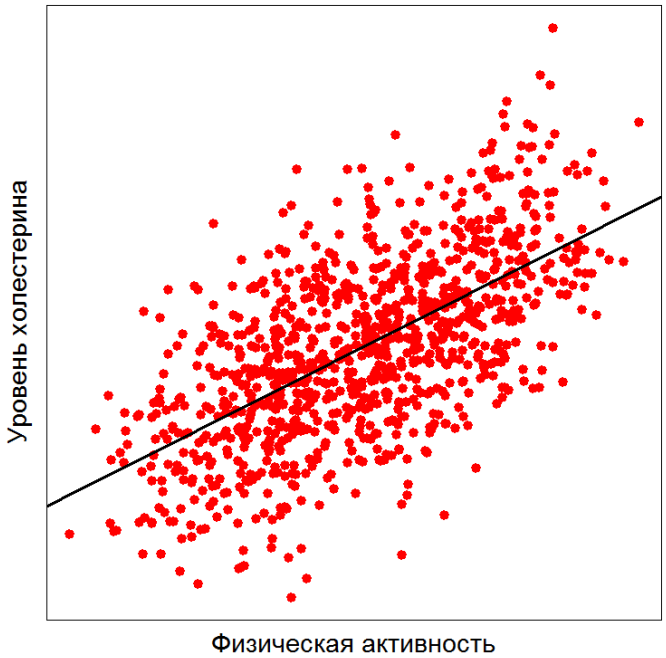
6. Нормальность ошибок

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

Проверка:

- ку-ку график
- критерий Шапиро-Уилка

Упражнения и холестерин

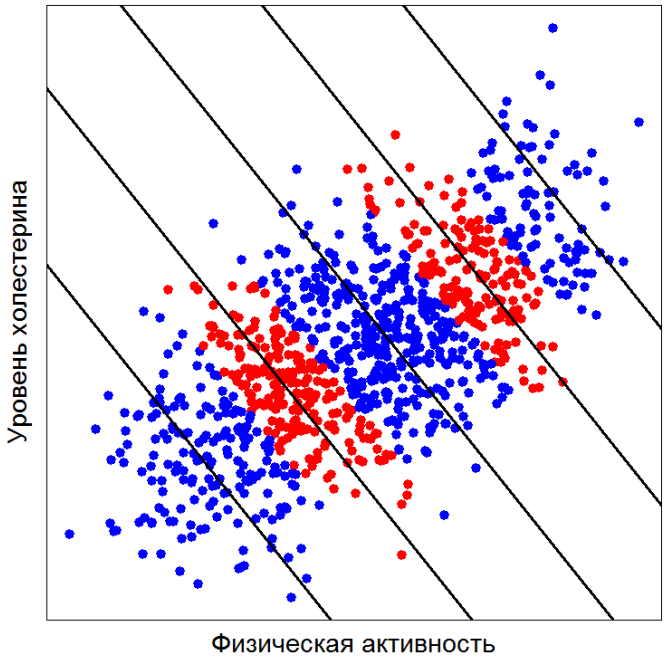


$$chol = \beta_0 + \beta_1 ex$$

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

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

Упражнения и холестерин

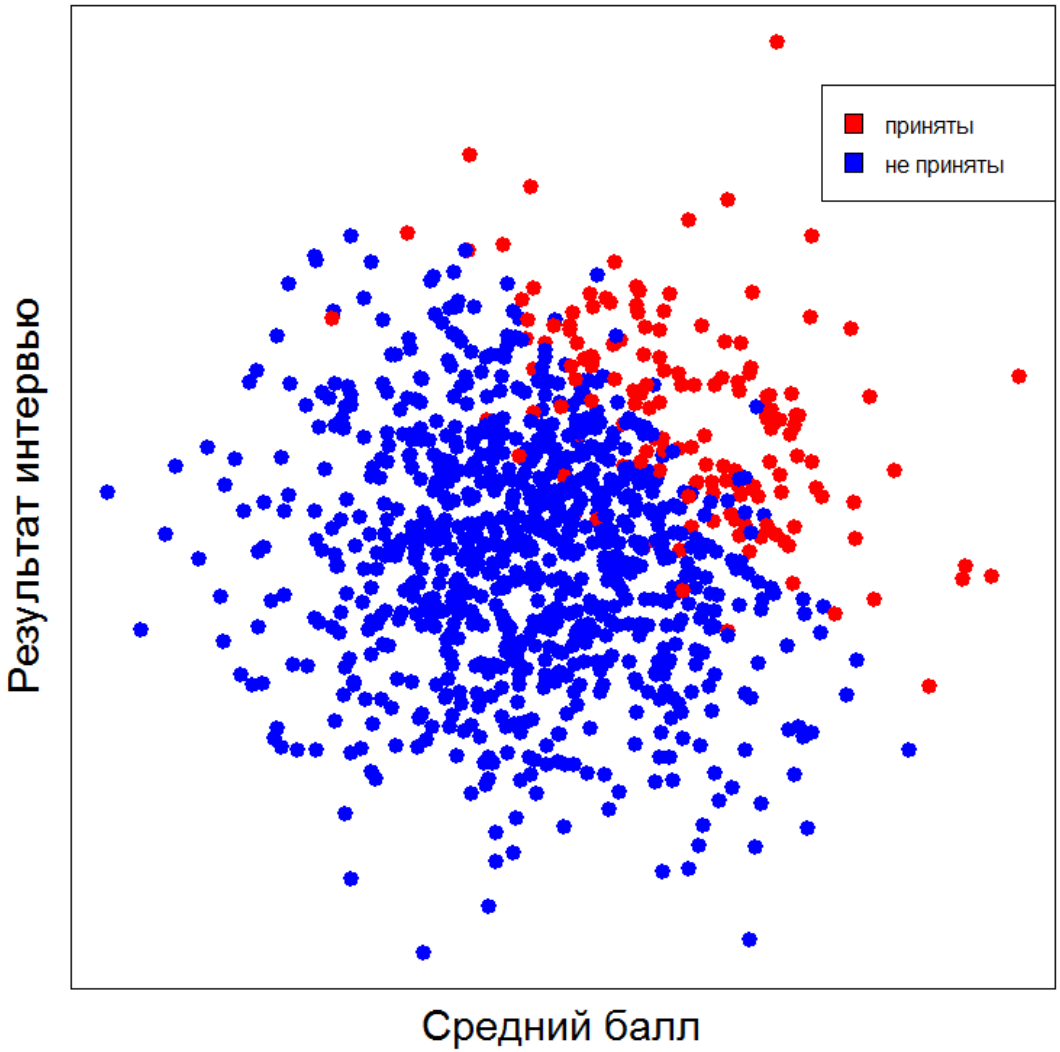


$$chol = \beta_0 + \beta_1 ex + \beta_2 age$$

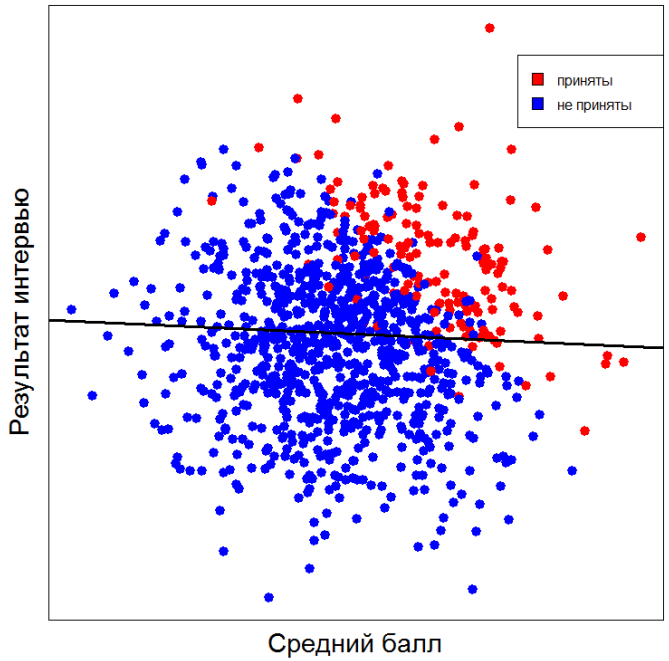
$H_0 : \beta_1 = 0$
$H_0 : \beta_1 < 0$

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

Средний балл и мотивация



Средний балл и мотивация

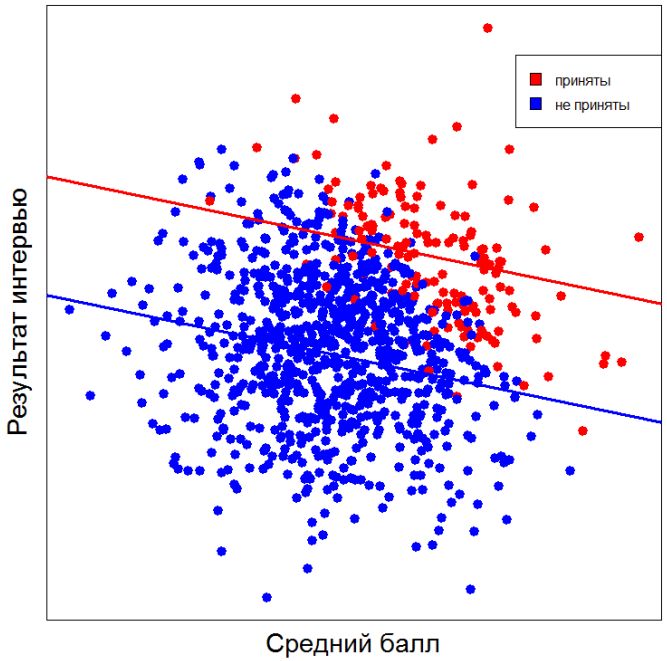


$$mot = \beta_0 + \beta_1 SAT$$

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

Критерий Стьюдента: $p = 0.1452$.

Средний балл и мотивация



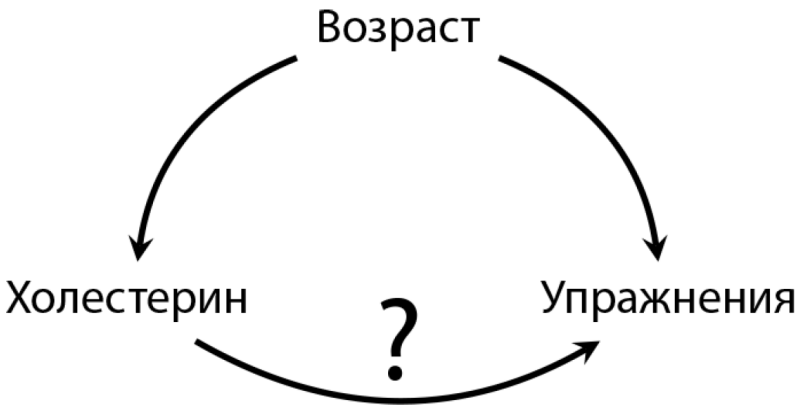
$$mot = \beta_0 + \beta_1 SAT + \beta_2 acc$$

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

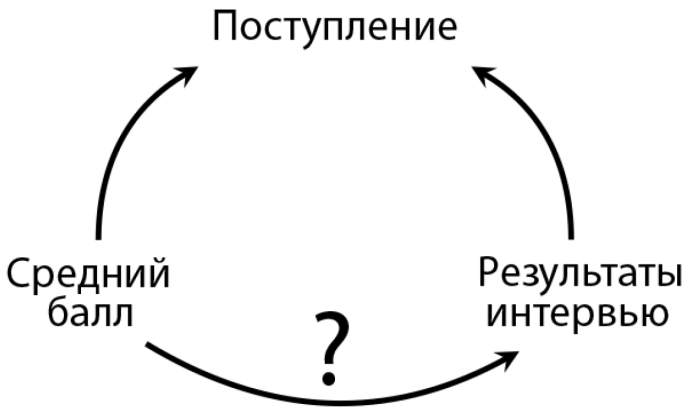
Критерий Стьюдента: $p = 2 \times 10^{-16}$.

В чём разница?

Вилка:



Коллайдер:



Причинно-следственная связь

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

- 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

Литература

- линейная регрессия в целом — Wooldridge (много примеров, без матричной алгебры);
- преобразование Бокса-Кокса (Box-Cox transformation) — Дрейпер, гл. 14;
- устойчивые оценки дисперсии — White;
- расстояние Кука (Cook's distance) — Cook.

Дрейпер Н.Р., Смит Г. *Прикладной регрессионный анализ*, 2007.

Cook D.R., Weisberg S. *Residuals and influence in regression*, 1982.

White H. (1980). *A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity*. Econometrica: Journal of the Econometric Society, 48(4), 817–838.

Wooldridge J. *Introductory Econometrics: A Modern Approach*, 2016.