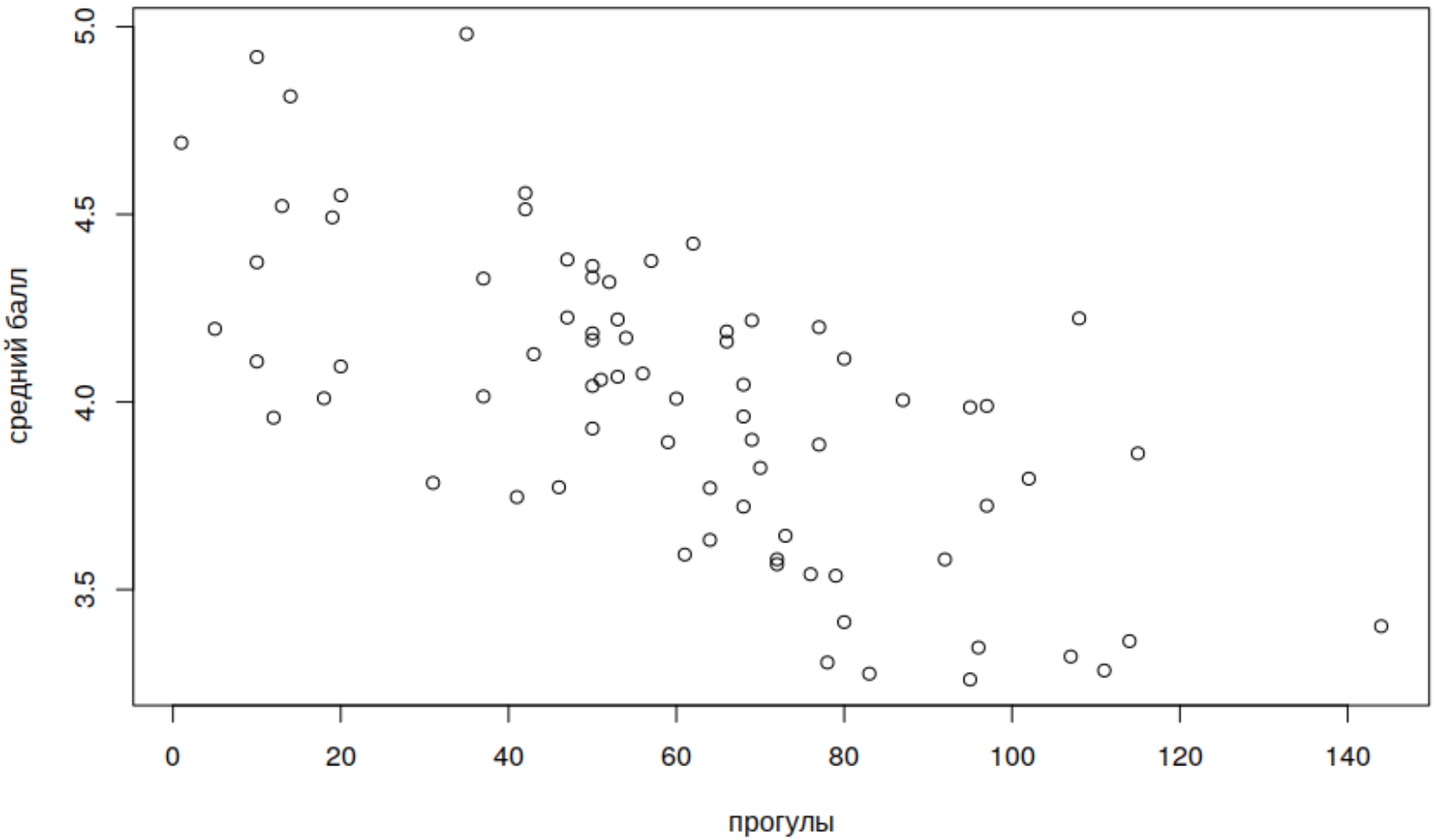
Прикладной статистический анализ

Линейная регрессия

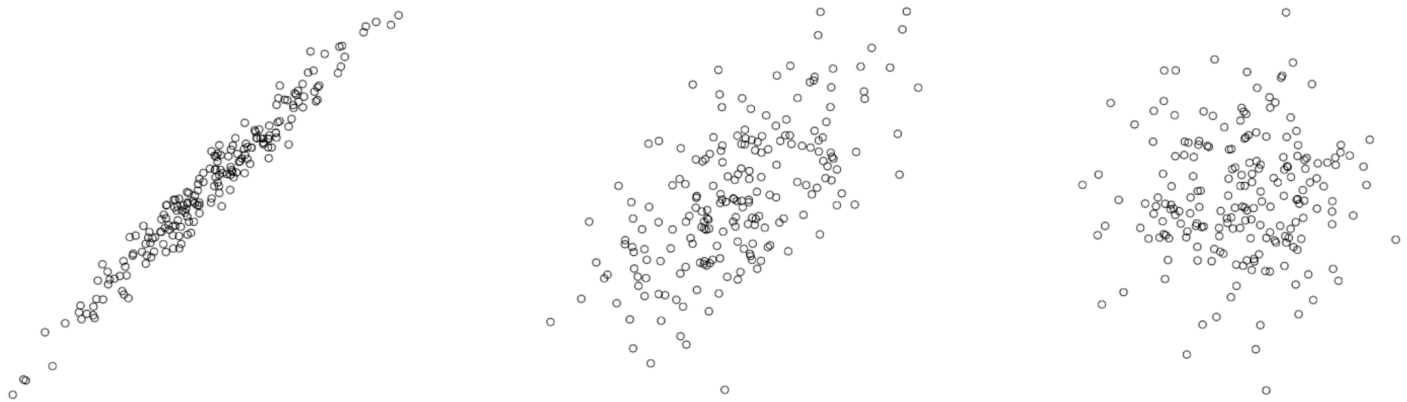
8 января 2021 г.

Корреляция

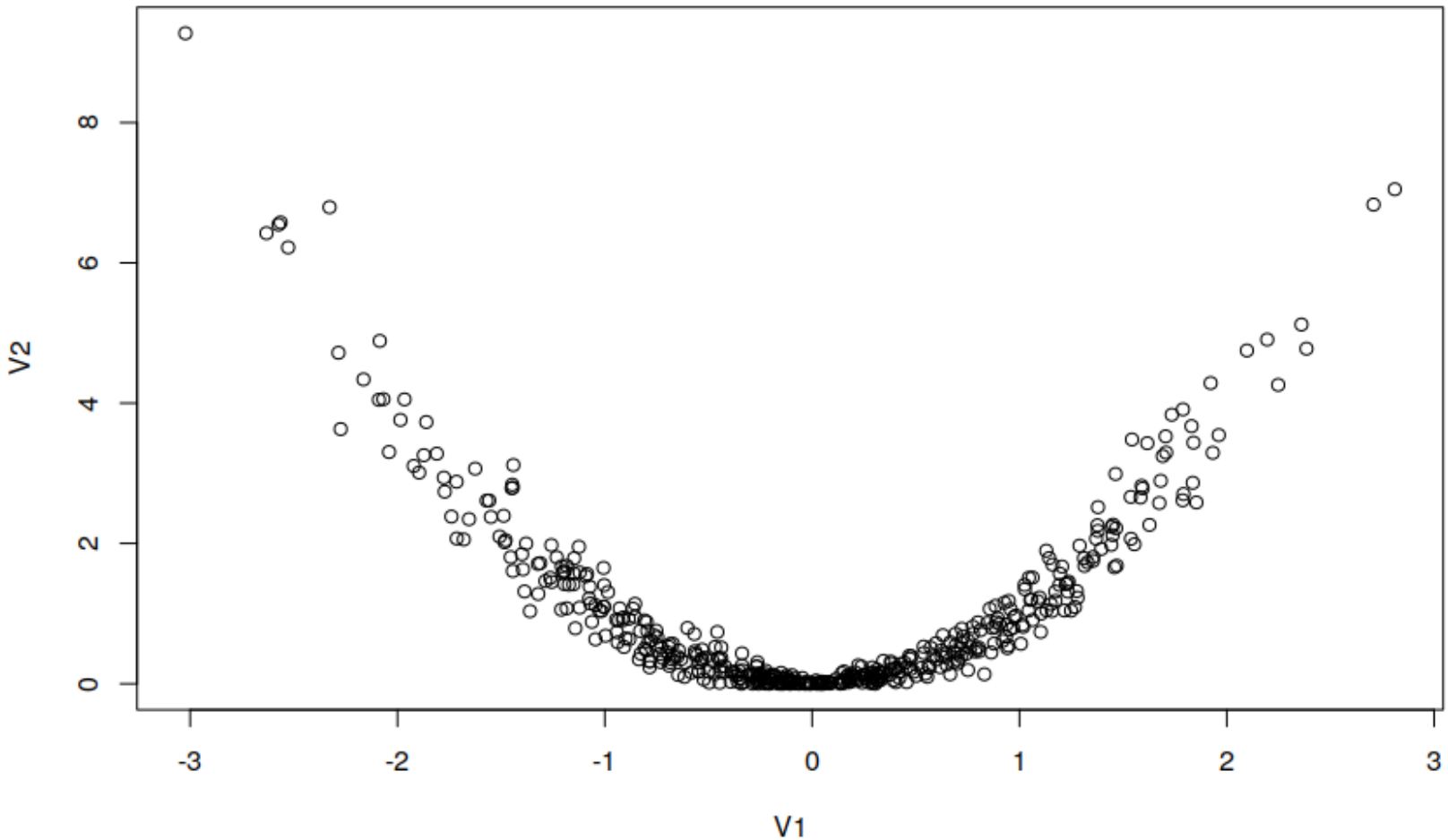


Корреляция

Для сравнения посмотрите на диаграмы данных с корреляциями (слева на право) 0.98, 0.64, 0.05:



Корреляция



Коэффициент корреляции: –0.06

Нахождение корреляции

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{s_x s_y},$$

где $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$, $s_x = \sqrt{\left(\overline{x^2} - (\overline{x})^2\right)}$.

Свойства корреляции

1. Абсолютное значение корреляции измеряет силу линейной связи между величинами;
2. Знак корреляции означает “направление” этой связи: положительная корреляция означает что рост одного параметра влечет рост другого, отрицательная - наоборот;
3. Коэффициент корреляции всегда лежит между –1 и 1, означающими идеальную линейную связь. 0 означает отсутствие *линейной* связи,
4. Коэффициент корреляции не зависит от смещения или единиц измерения числовых величин, т.е. он не меняется от линейного преобразования с положительным множителем;

Простая линейная регрессия

Попробуем описать наблюдаемую линейную зависимость $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^n$ подобрав прямую

$$y = \beta_0 + \beta_1 x,$$

минимизирующую ошибки $y_i - \beta_0 - \beta_1 x_i$. В качестве меры общей величины ошибок возьмем

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Нахождение параметров β_0, β_1 нашей прямой через минимизацию этого функционала называют *методом наименьших квадратов* (МНК).

Оценки параметров простой линейной регрессии

$$b_1 = \frac{s_y}{s_x} r_{x,y}, \quad b_0 = \bar{y} - b_1 \bar{x}.$$

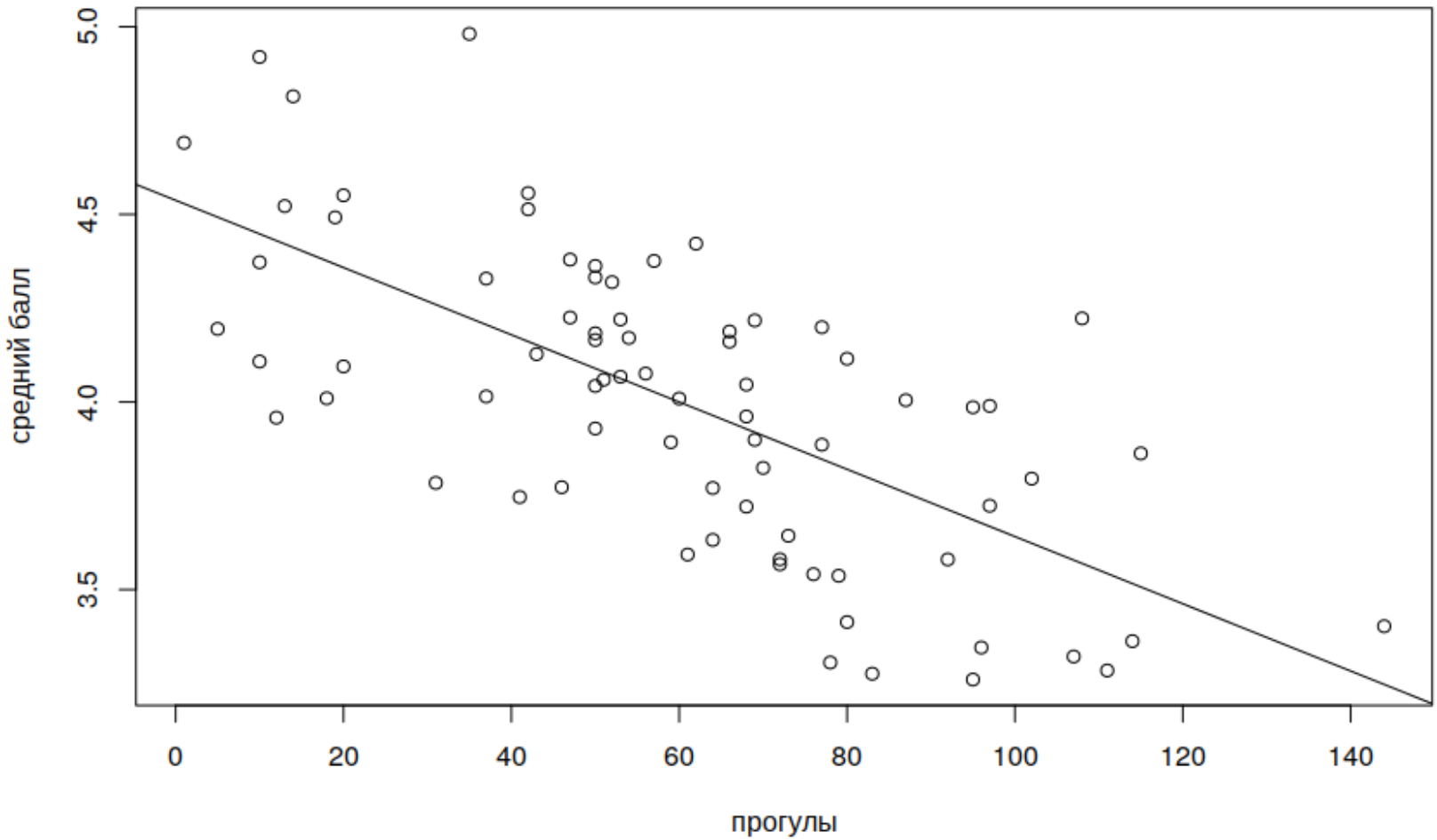
Линия регрессии всегда проходит через $(\bar{x}, \bar{y})$:

$$\hat{y} = b_0 + b_1 x$$

- x ← Фактор (explanatory)
- b_1 → Наклон (slope)
- b_0 → Константа (intercept)
- $\hat{y}$ → Предсказание модели (predicted response)

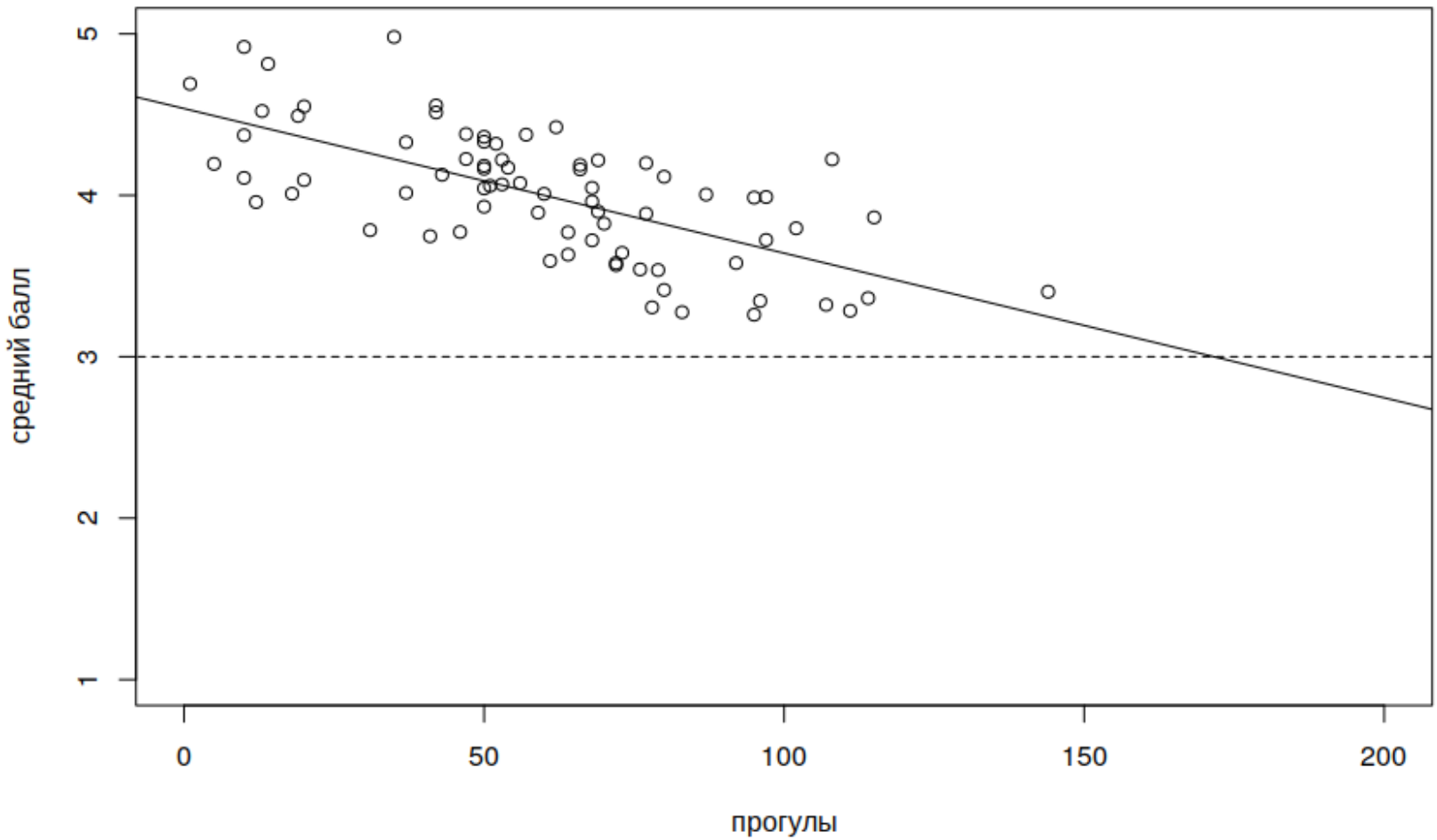
Линейная регрессия

$b_0 = 4.5372,\ b_1 = -0.008959$

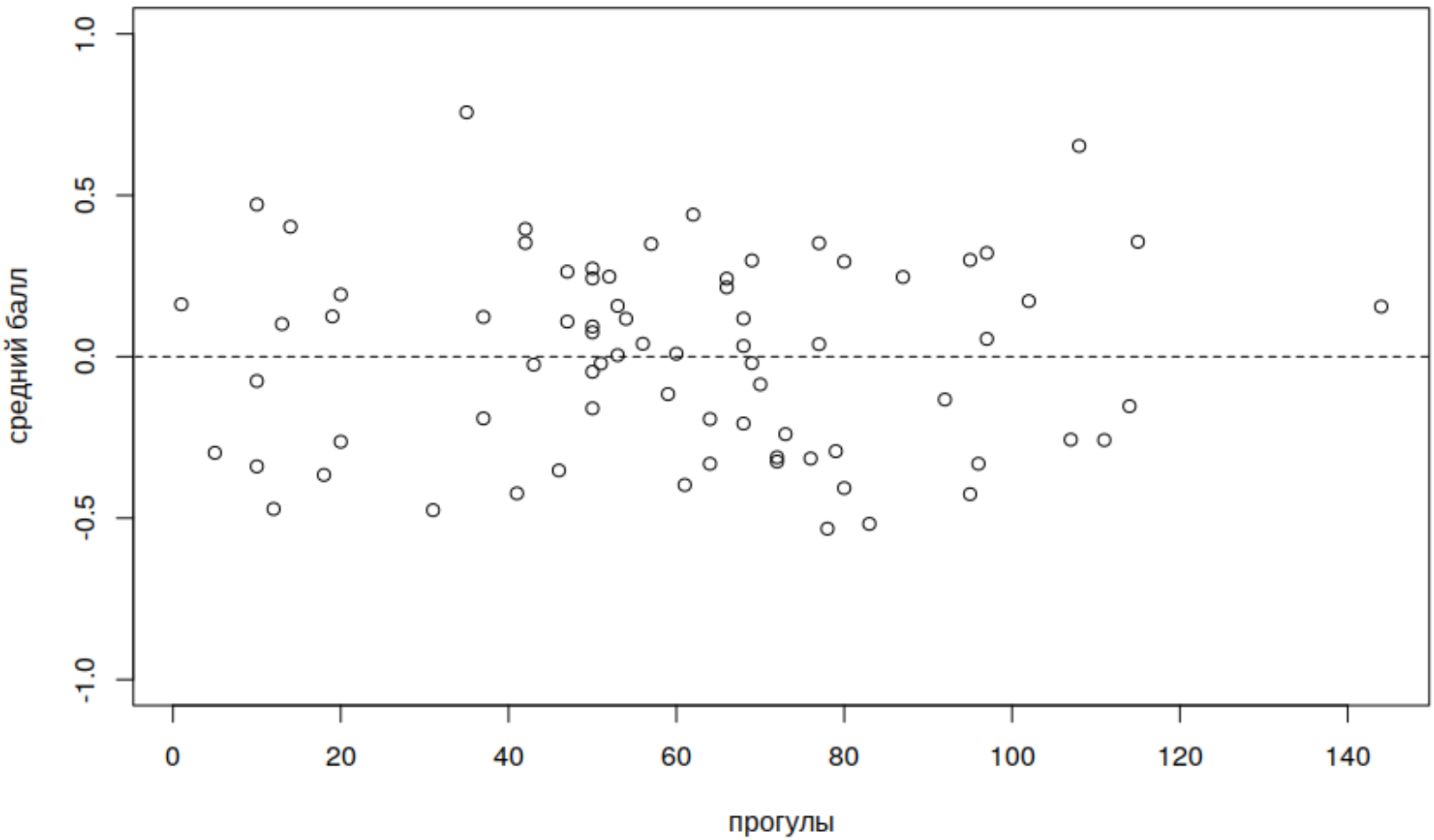


Линейная регрессия: экстраполяция

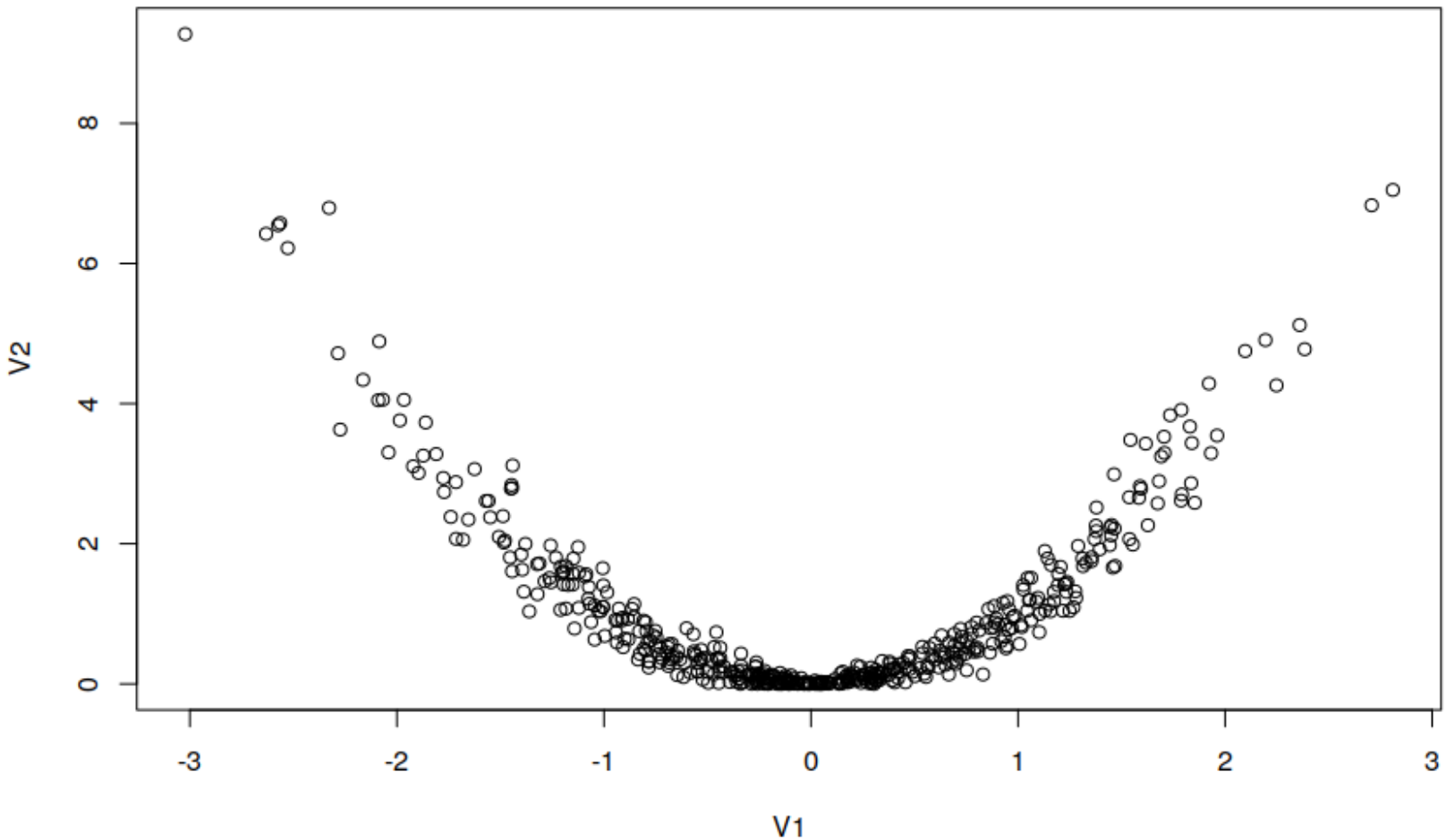
$b_0 = 4.5372,\ b_1 = -0.008959$



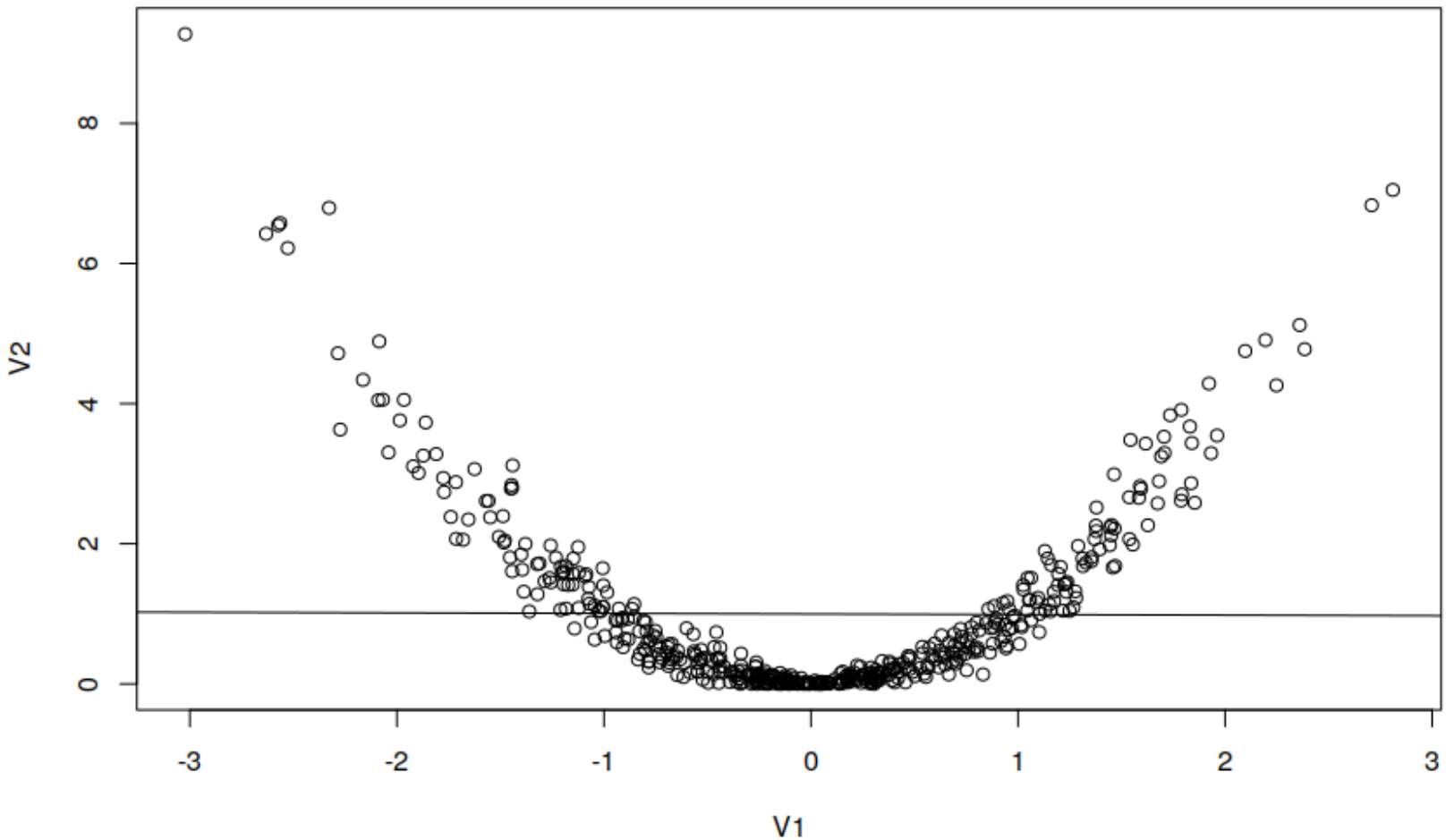
Линейная регрессия: остатки



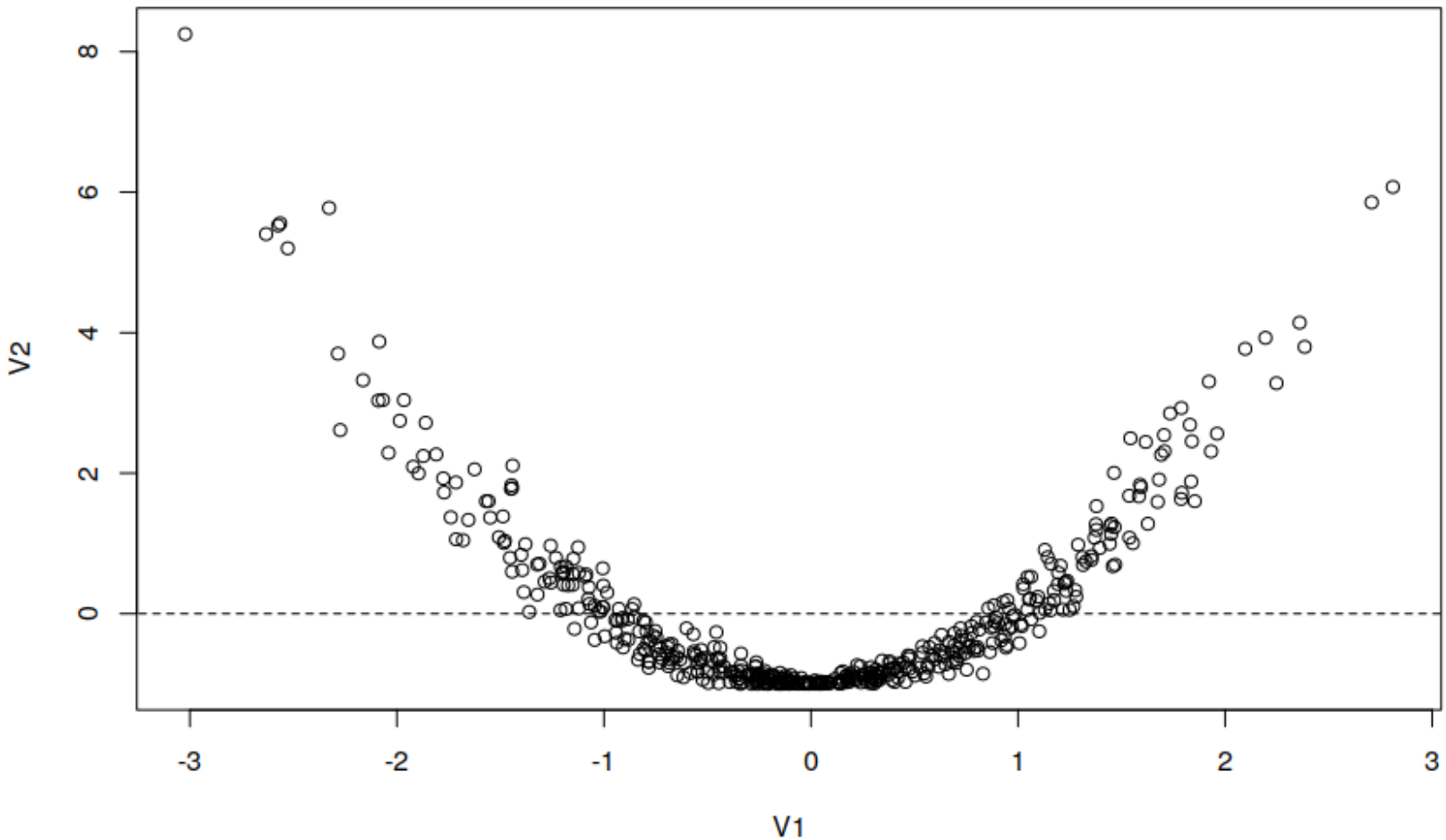
Линейная регрессия: остатки



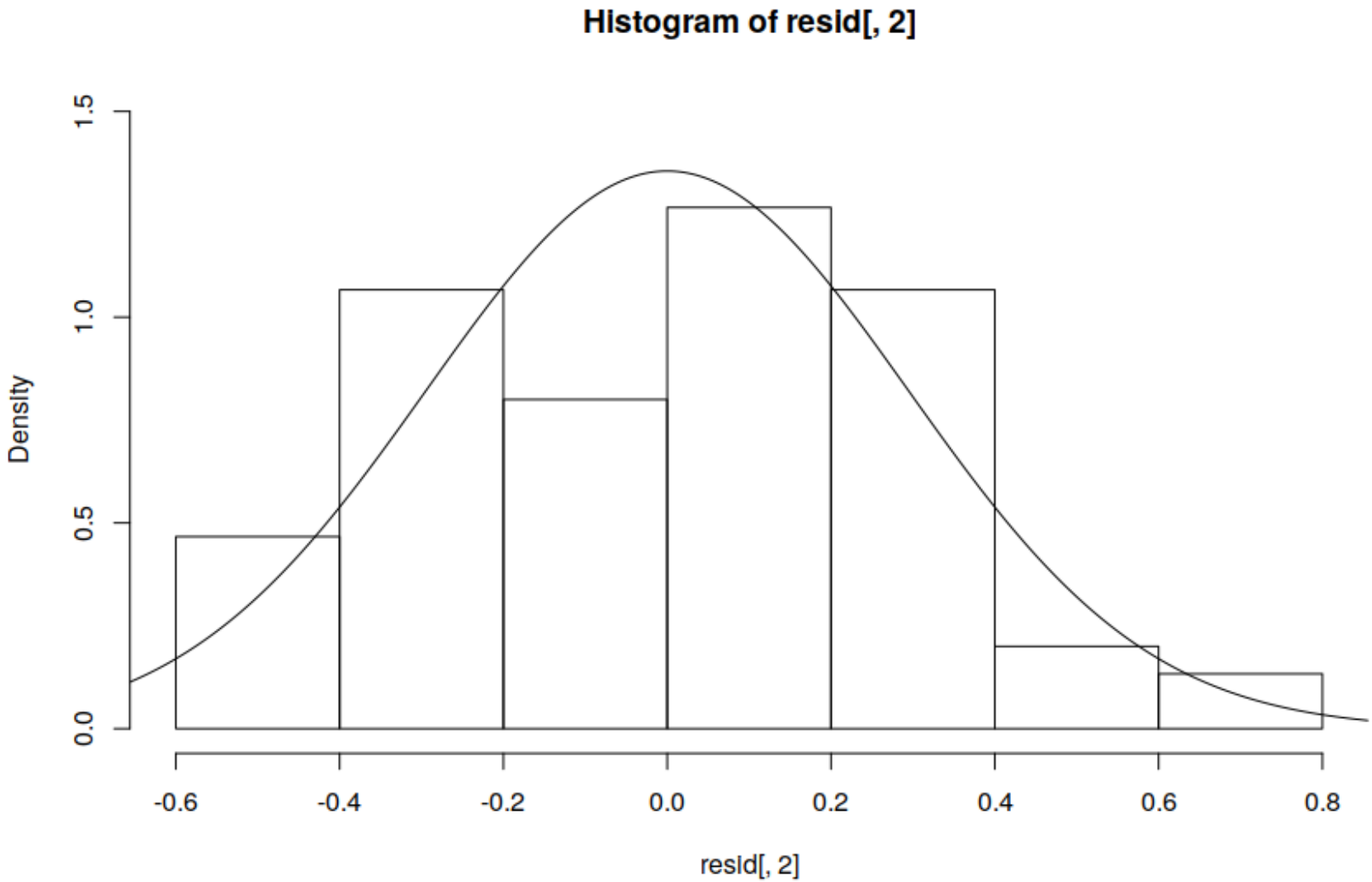
Линейная регрессия: остатки



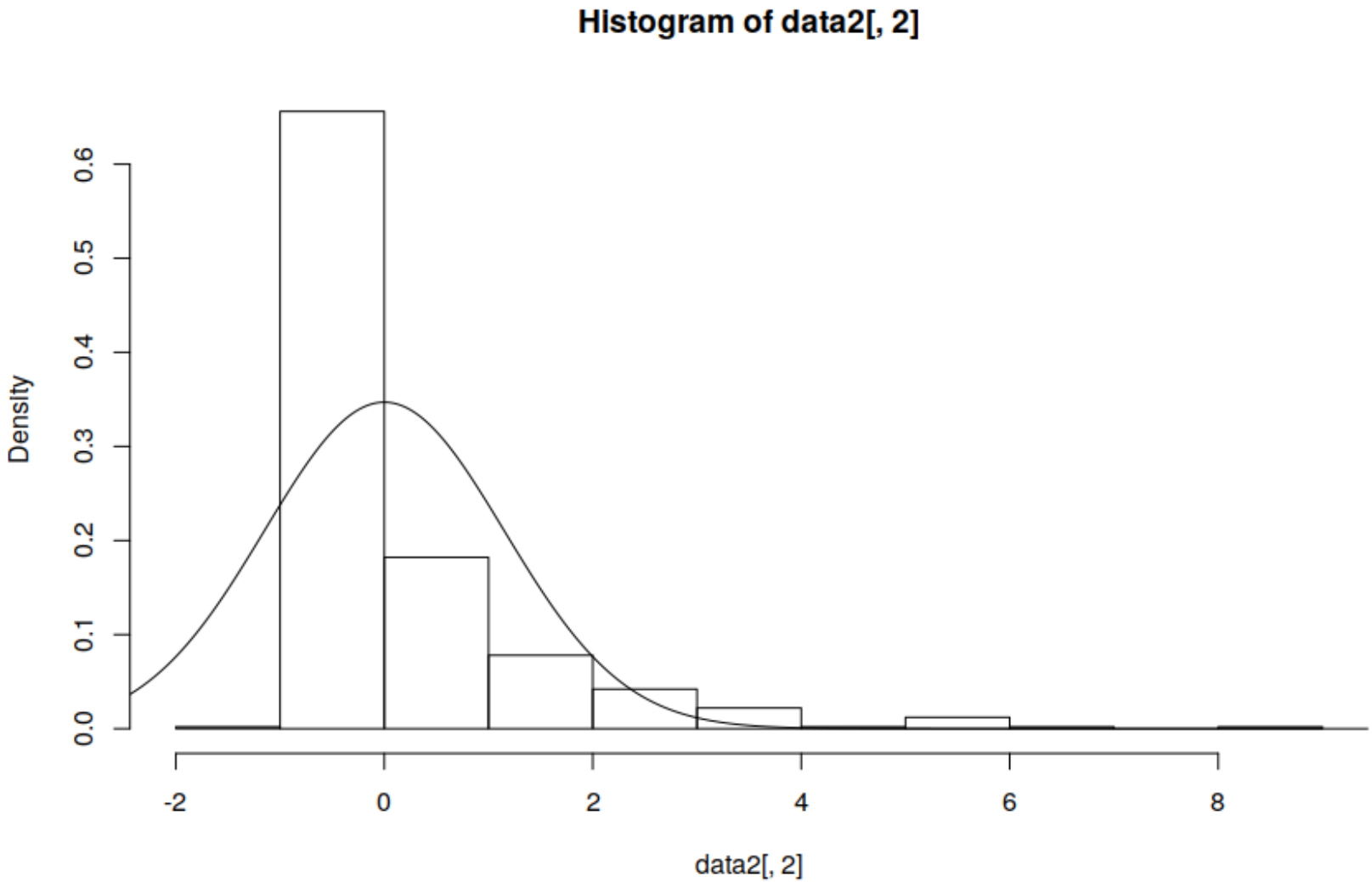
Линейная регрессия: остатки



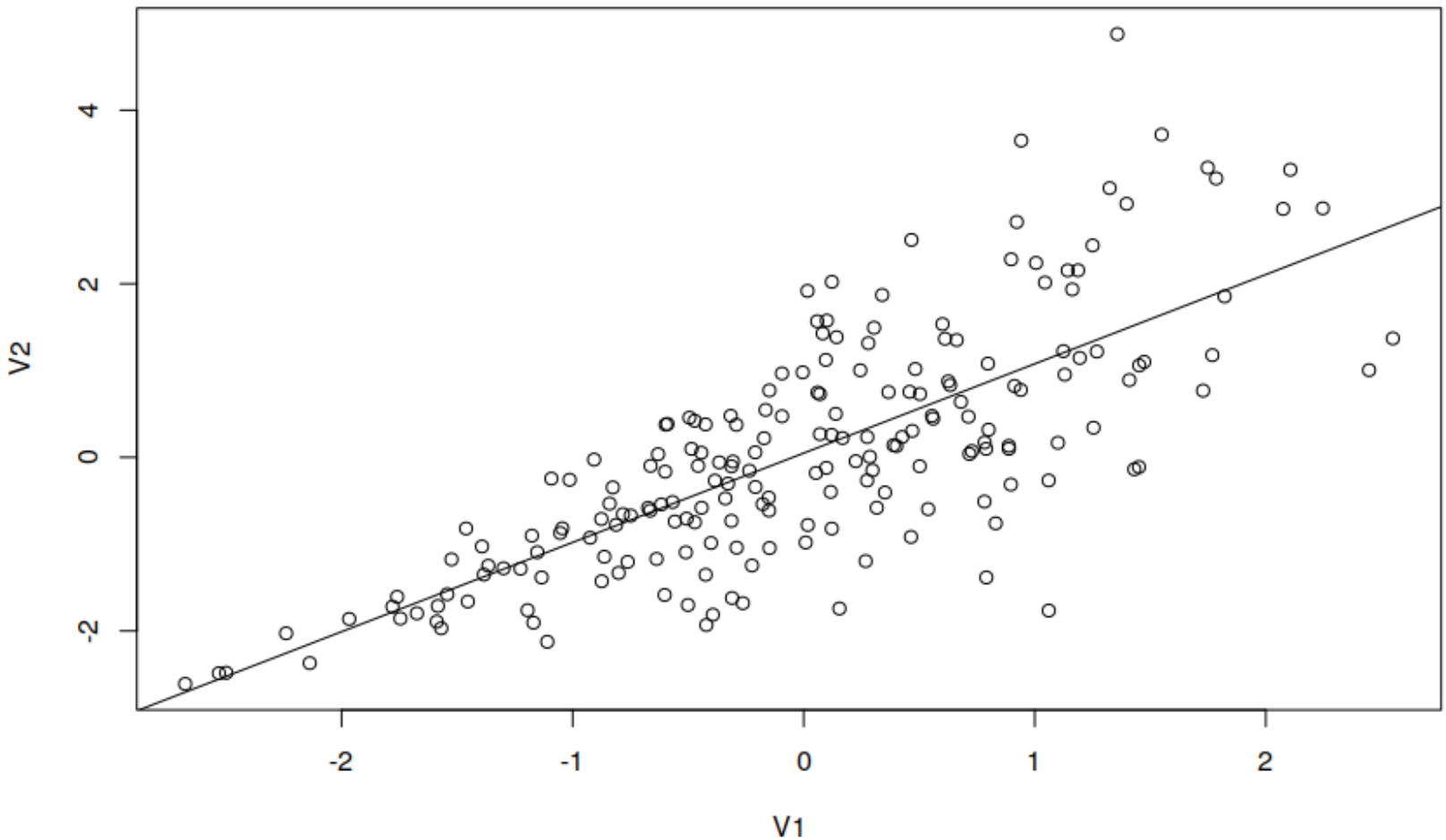
Линейная регрессия: гистограмма остатков



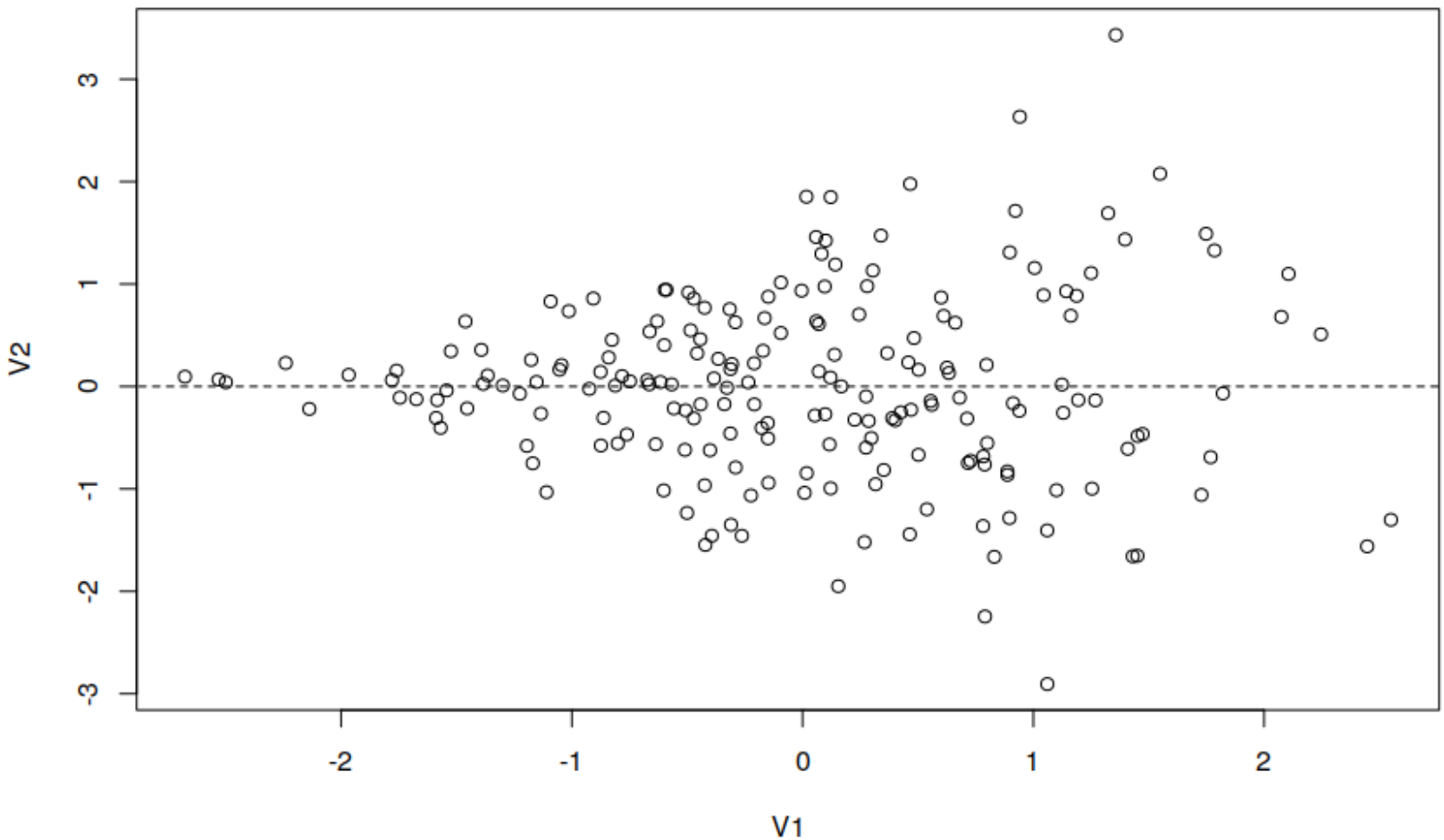
Линейная регрессия: гистограмма остатков



Линейная регрессия: гомоскедастичность



Линейная регрессия: гомоскедастичность



Линейная регрессия: насколько хорошо наша модель описывает данные?

В качестве оценки описательной силы линейной модели для наших данных применяется

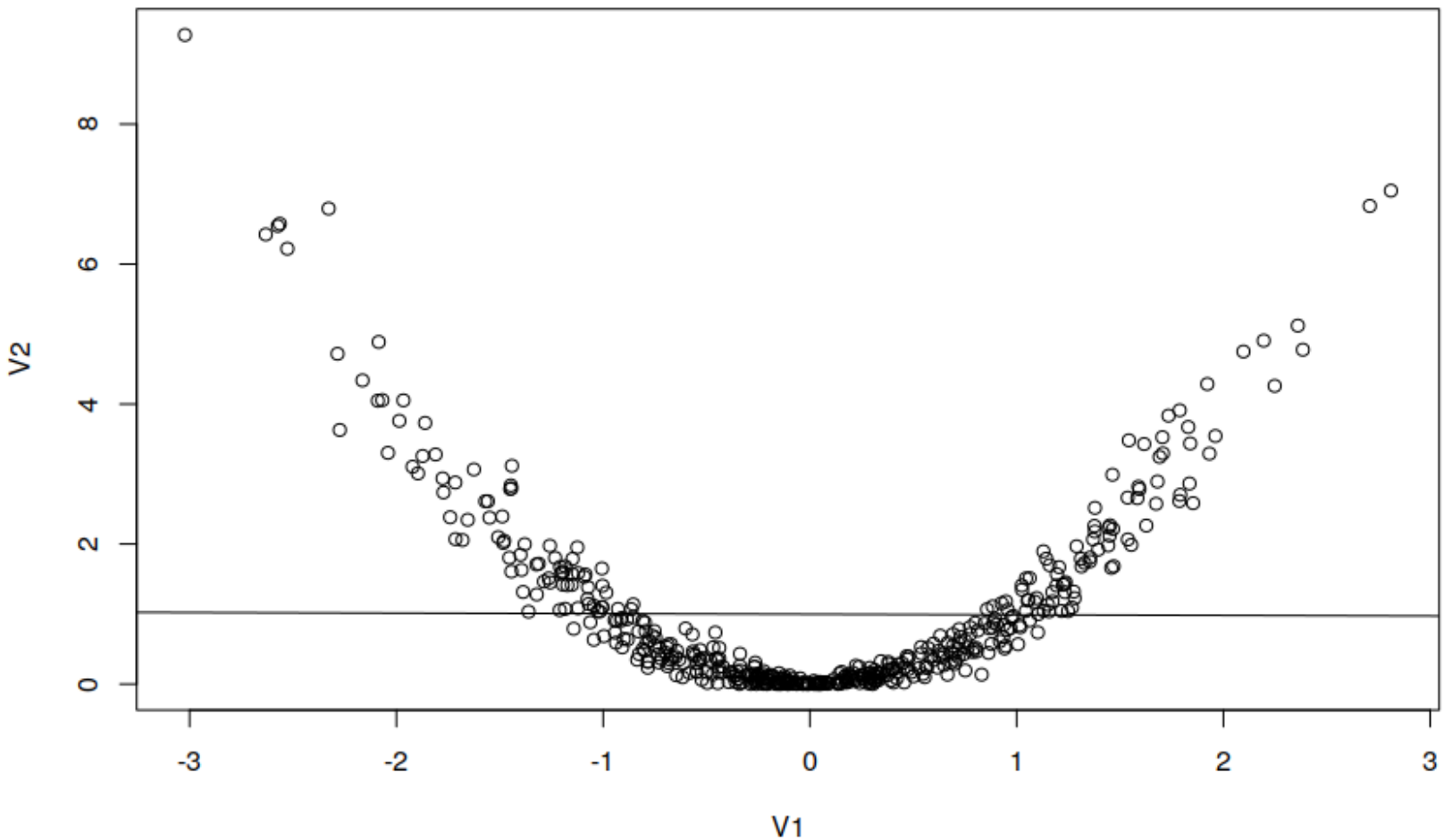
$$R^2 = r^2_{x,y}.$$

Эта величина описывает долю вариации предсказываемой величины y, которую описывает наша модель.

$$0 \leqslant R^2 \leqslant 1.$$

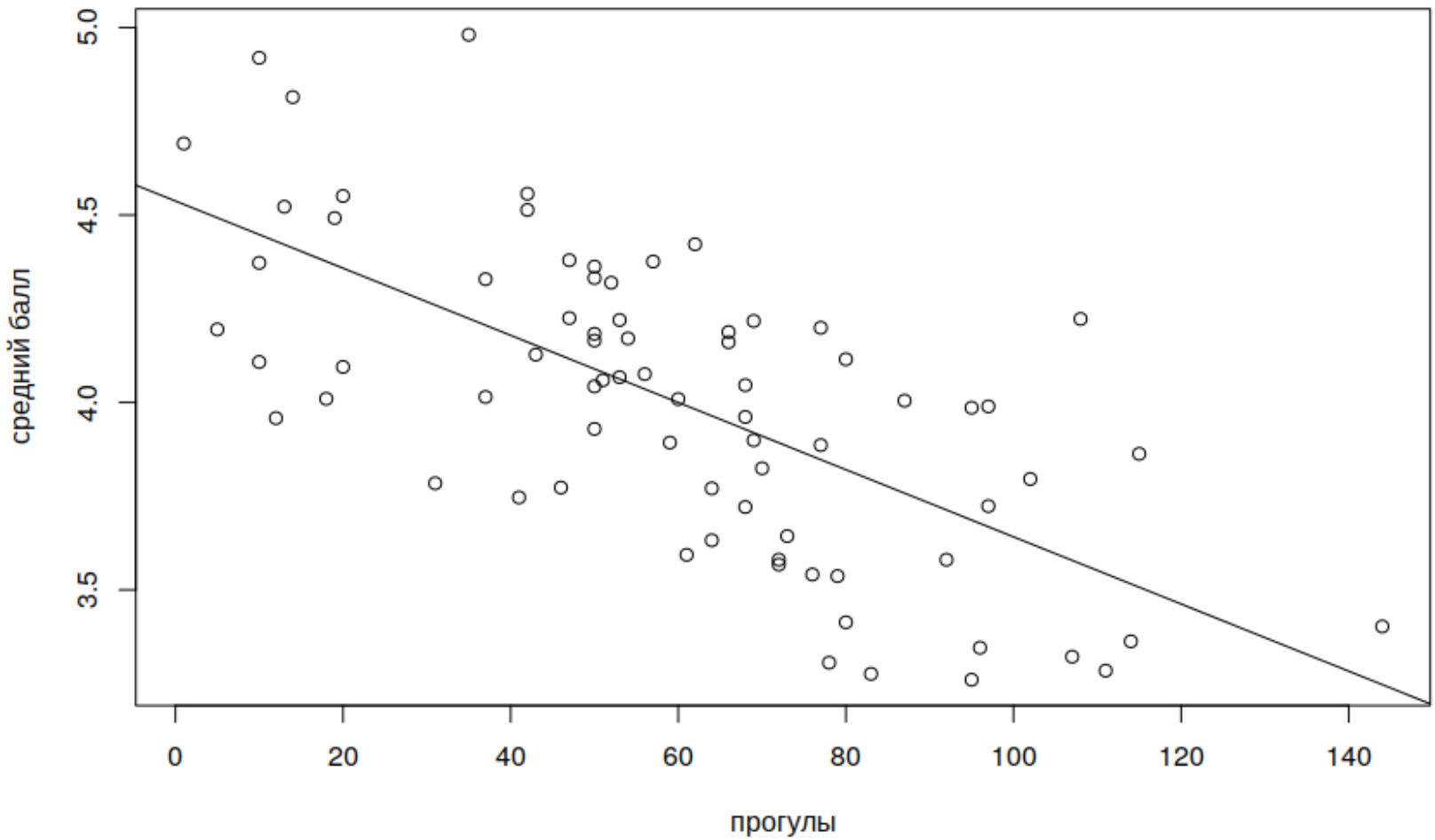
Линейная регрессия: насколько хорошо наша модель описывает данные?

$R^2 = 0.003787605$



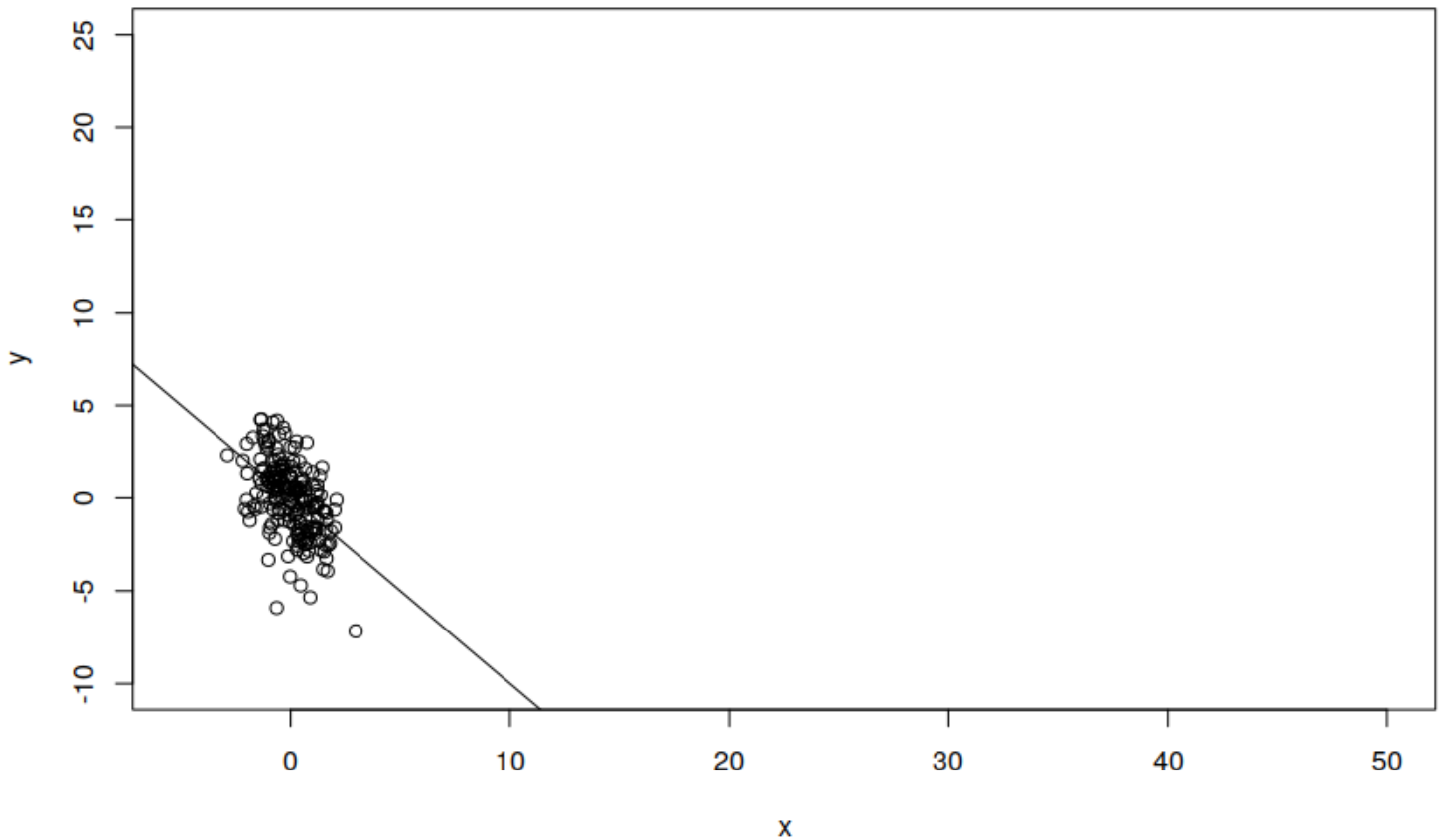
Линейная регрессия: насколько хорошо наша модель описывает данные?

$R^2 = 0.4509775$



Выбросы

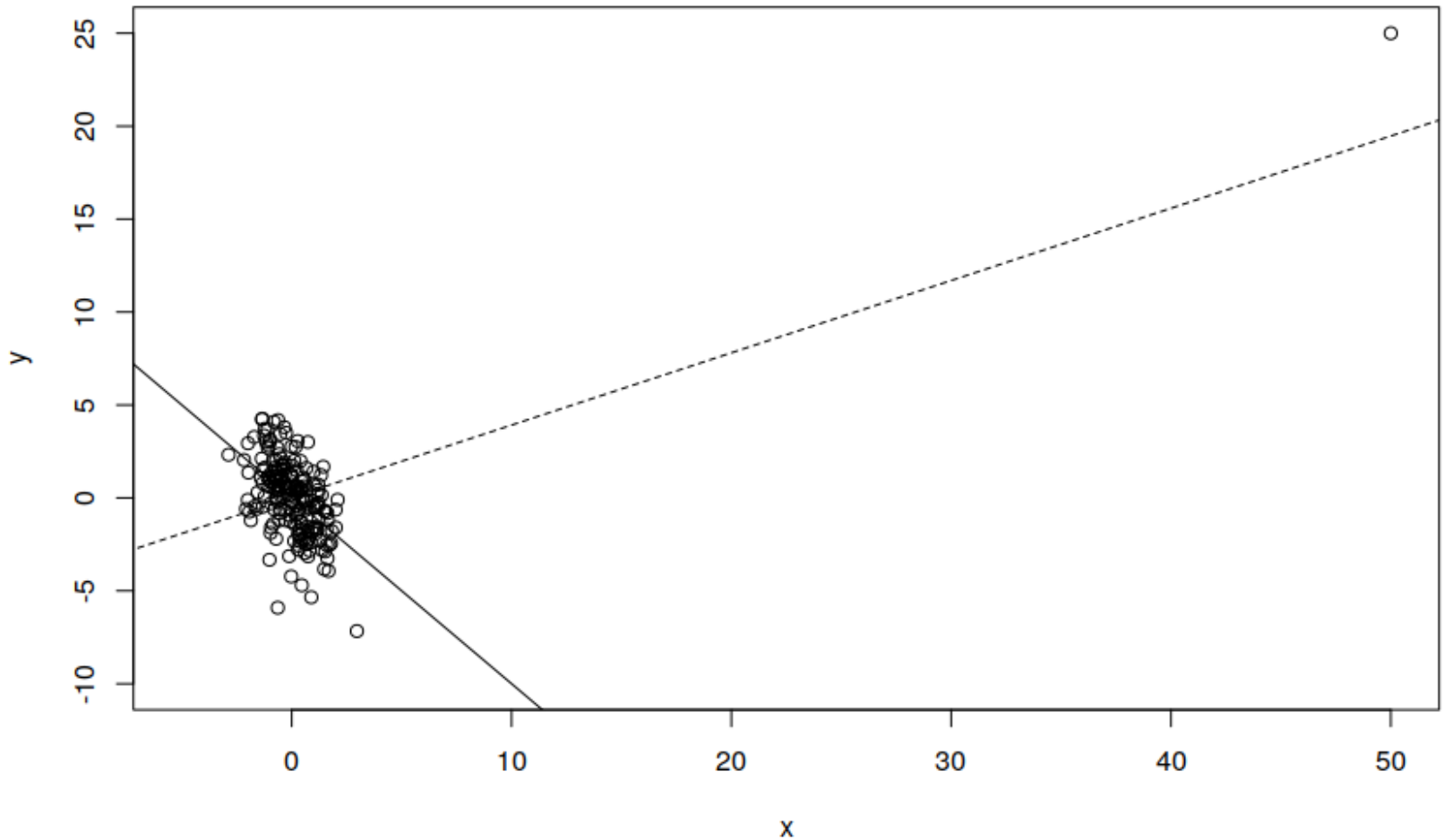
Выбросы могут значительно повлиять на результат регрессии. Рассмотрим сначала отрицательно коррелированный результат 200 наблюдений:



$R^2 = 0.25$.

Выбросы

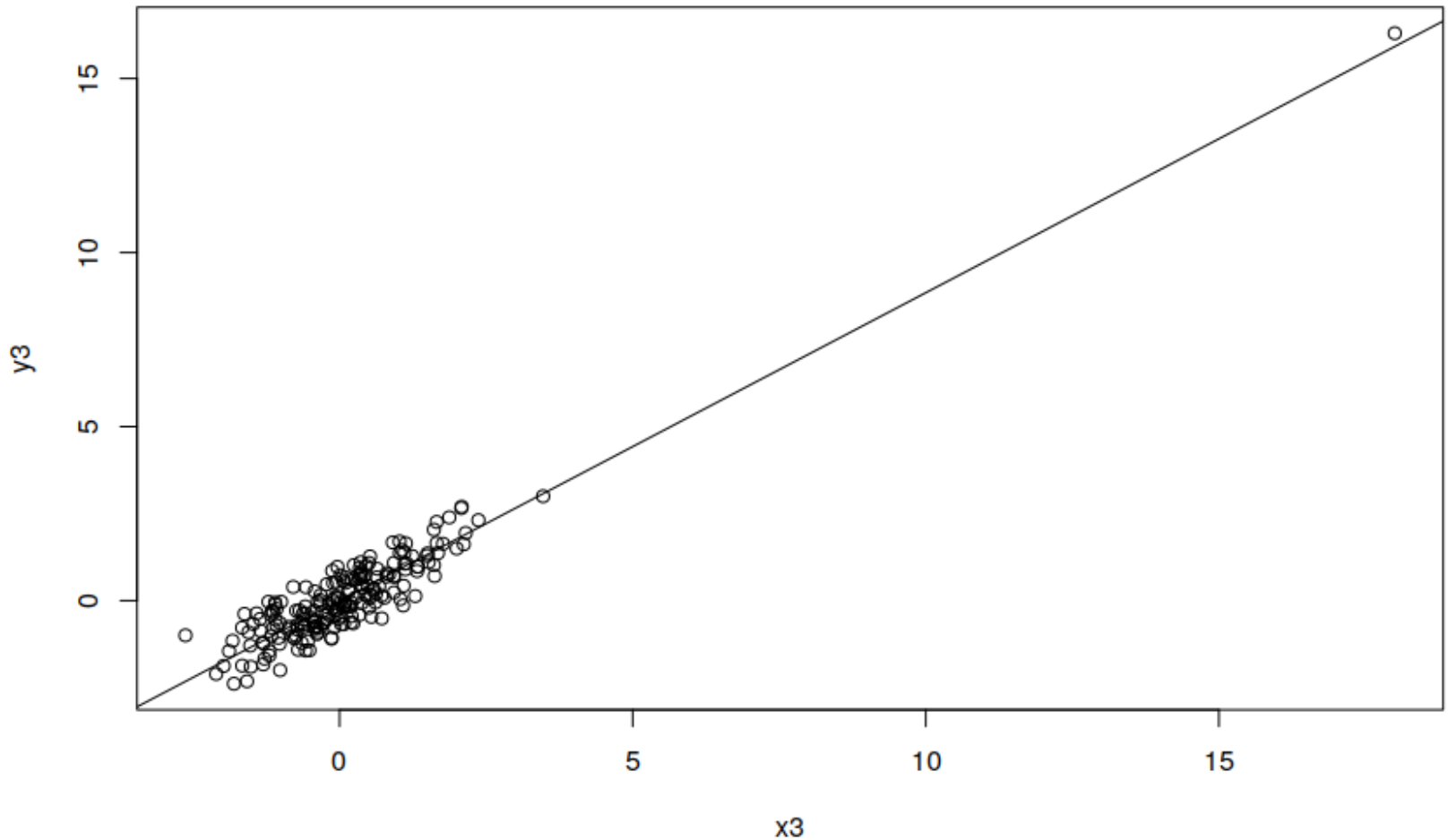
Теперь добавим одно единственное наблюдение ($x = 50, y = 25$):



$R^2 = 0.29$.

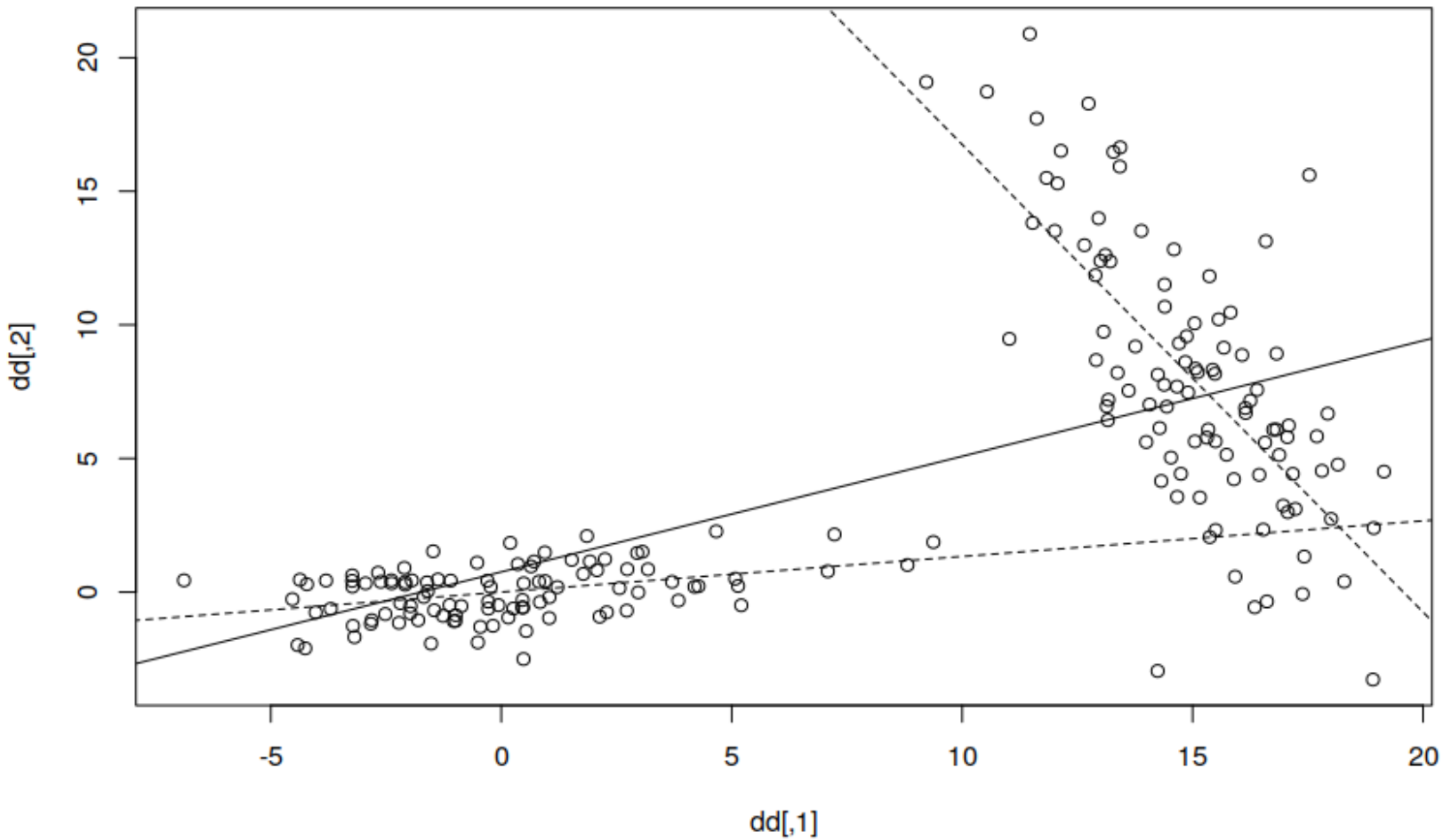
Выбросы

Иногда выбросы почти не влияют на регрессию:



$R^2 = 0.88$.

Выбросы



Математическая модель

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n,$$

где все ε_i имеют нормальное распределение $N_{0,\sigma}$.

$$\varepsilon_i \neq \hat{\varepsilon}_i := y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Математическая модель: МНК

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n,$$

где все ε_i имеют нормальное распределение N_{0,σ^2}.
Тогда y_i имеют распределение $N_{\beta_0+\beta_1 x_i,\sigma^2}$.
Применение метода наибольшего правдоподобия для нахождения параметров влечет за собой нахождения максимума

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right),$$

что тоже самое что и минимизация

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Математическая модель: МНК для множественной регрессии

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

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

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \qquad (1)$$

Тогда y имеет распределение

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right)$$

Математическая модель: МНК для множественной регрессии

Минимизируем

$$(y - X\beta)^T(y - X\beta) = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta$$

$$= y^T y - 2y^T X\beta + \beta^T X^T X\beta.$$

используя

$$\frac{\partial}{\partial \beta} = \begin{bmatrix} \frac{\partial}{\partial \beta_0} \\ \frac{\partial}{\partial \beta_1} \\ \frac{\partial}{\partial \beta_2} \\ \vdots \\ \frac{\partial}{\partial \beta_k} \end{bmatrix} \tag{2}$$

Математическая модель: МНК для множественной регрессии

$$\frac{\partial(y^T y - 2y^T X\beta + \beta^T X^T X\beta)}{\partial\beta} = -2X^T y + 2X^T X\beta.$$

Приравняв к нулю, получим оценки для параметров:

$$X^T X\hat{\beta} = X^T y$$

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

Аналогично можно получить и оценку для дисперсии ε_i:

$$\sigma^2_{\text{МНП}} = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

$$\sigma^2 = \frac{1}{n-p}(y - X\hat{\beta})^T(y - X\hat{\beta}),$$

где $p = \text{rank}X = k + 1$.

Математическая модель: коэффициент детерминации для множественной регрессии

$$\hat{y} = X\hat{\beta}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Может использоваться для сравнения моделей, если:

1. Все модели используют одну и ту же "ответную переменную" y: нельзя сравнивать модели, предсказывающие y и $\log(y)$;
2. Все модели содержат одинаковое число параметров;
3. Все модели включают константный член β_0.

Проверка гипотез и доверительные интервалы

Типичные гипотезы, которые будут нас интересовать, таковы:

1. Тест на значимость параметра $H_0 : \beta_j = 0$ против $H_a : \beta_j \neq 0$
2. Тест для подвекторов $\beta_* = (\beta_1, \ldots, \beta_r)$ $H_0 : \beta_* = 0$ против $H_a : \beta_* \neq 0$
3. Тест на равенство $H_0 : \beta_j = \beta_r$ против $H_a : \beta_j \neq \beta_r$

Все они могут быть представлены в общем виде:

$$H_0 : C\beta = d \quad \text{против} \quad H_a : C\beta \neq d,$$

где C — матрица rxp для некоторого r, в зависимости от гипотезы.

F-тест

Пусть $SSE := \hat{\varepsilon}^T\hat{\varepsilon} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2$, а SSE_{H_0} — тоже самое, но с ограничением $C\beta = d$.
Тогда

$$\frac{\Delta SSE}{SSE} = \frac{SSE_{H_0} - SSE}{SSE} \geqslant 0$$

описывает величину разницы между ошибками при и без гипотезы H_0.
Для того, чтобы это отношение имело известное распределение Фишера (с параметрами r и $n - p$) обычно используется статистика

$$F = \frac{n - p}{r}\frac{\Delta SSE}{SSE}.$$

F-тест

Для нахождения распределения статистики F при условии гипотезы H_0 нам потребуется показать что:

1. Оценка β при условии выполнения H_0 имеет вид: $\hat{\beta}^R = \hat{\beta} - (X^T X)^{-1} C^T (C(X^T X)^{-1} C^T)^{-1} (C\hat{\beta} - d)$;
2. $\Delta SSE = (C\hat{\beta} - d)^T (C(X^T X)^{-1} C^T)^{-1} (C\hat{\beta} - d)$;
3. При условии выполнения H_0 имеем $\frac{1}{\sigma^2}\Delta SSE \sim \chi^2_r$;
4. ΔSSE и SSE независимы.