МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ

(Machine Learning and Data Mining)

Н. Ю. Золотых

http://www.uic.unn.ru/~zny/ml

Лекция 5

Метод наименьших квадратов

Agenda

- Регрессионная функция
  - Метод наименьших квадратов
  - Метод максимального правдоподобия
- Линейная регрессия
- Оценка коэффициентов по выборке
- Переобучение
- Сокращение числа параметров и «усадка» коэффициентов
  - Выбор подмножества параметров
  - Гребневая регрессия
  - Лассо
  - Метод главных компонент
  - Частичные наименьшие квадраты

5.1. Регрессия

Гаусс (1794–1795), Лежандр (1805–1806) — метод наименьших квадратов

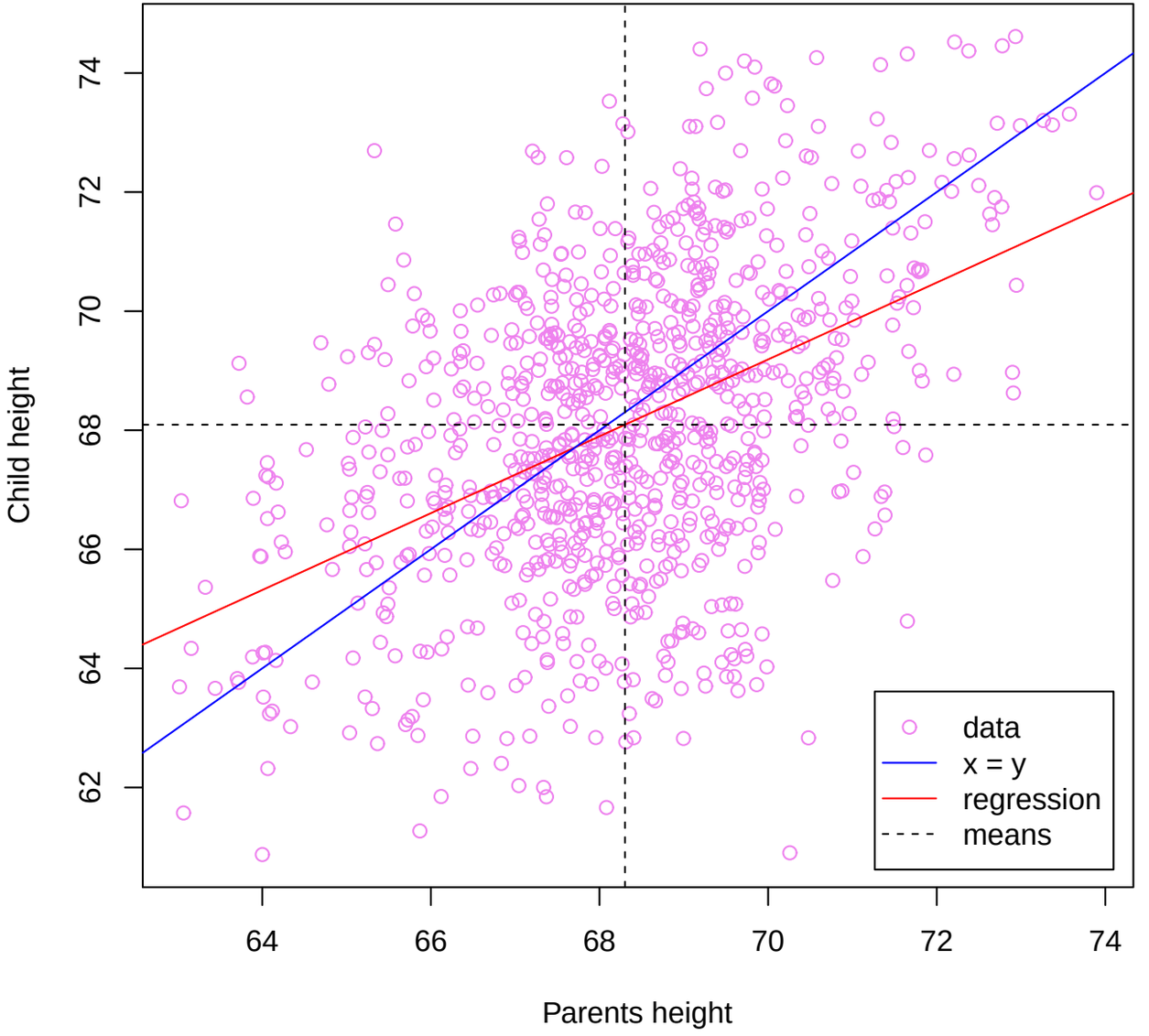
(Исследование траектории астероида Церера)

Фрэнсис Гальтона (1822–1911)

«Регрессия к середине в наследовании роста» (1885 г.)

«Где это возможно, считайте». Ф.Гальтон

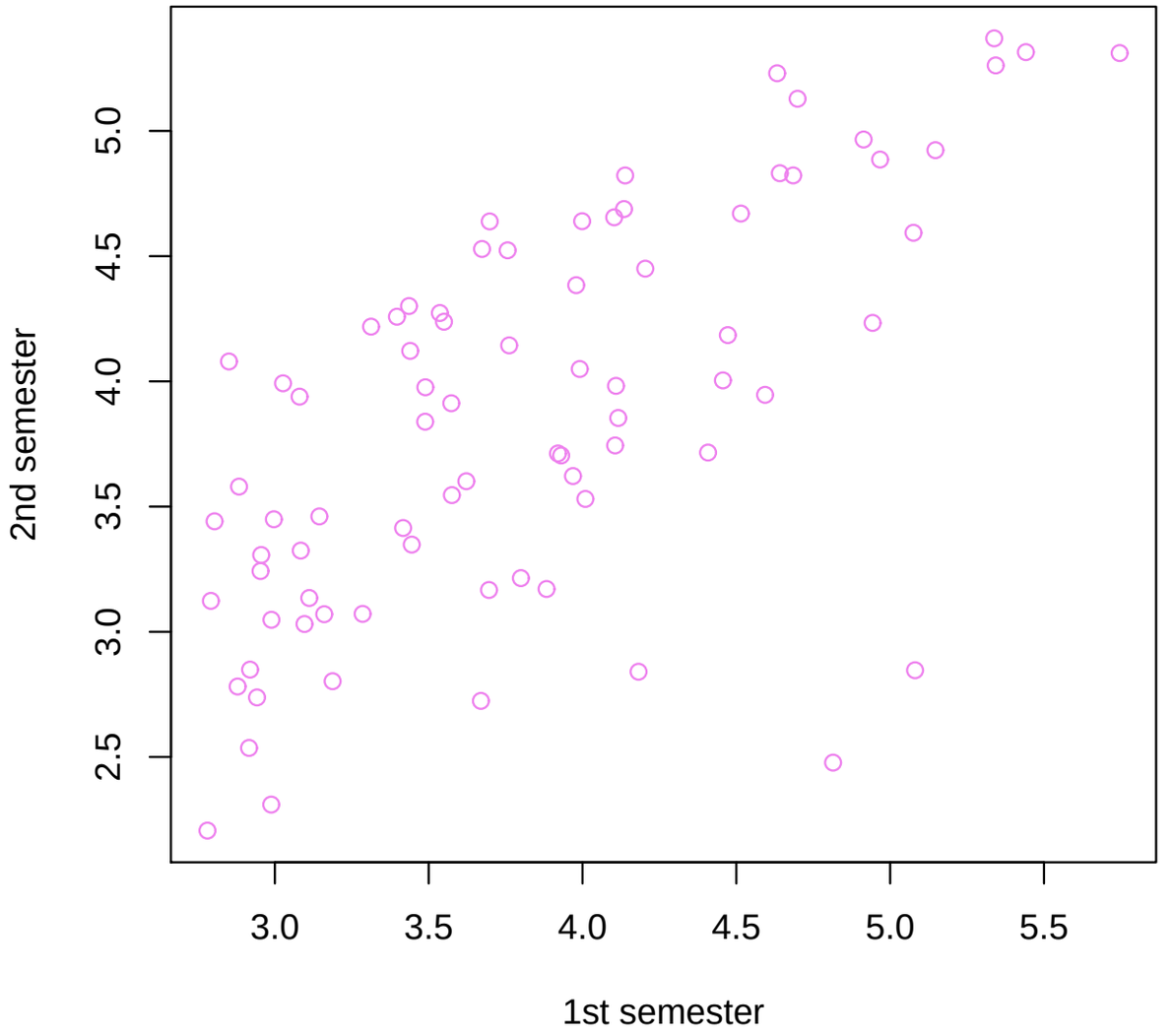
Зависимость роста ребенка от роста родителей в исследовании Ф. Гальтона



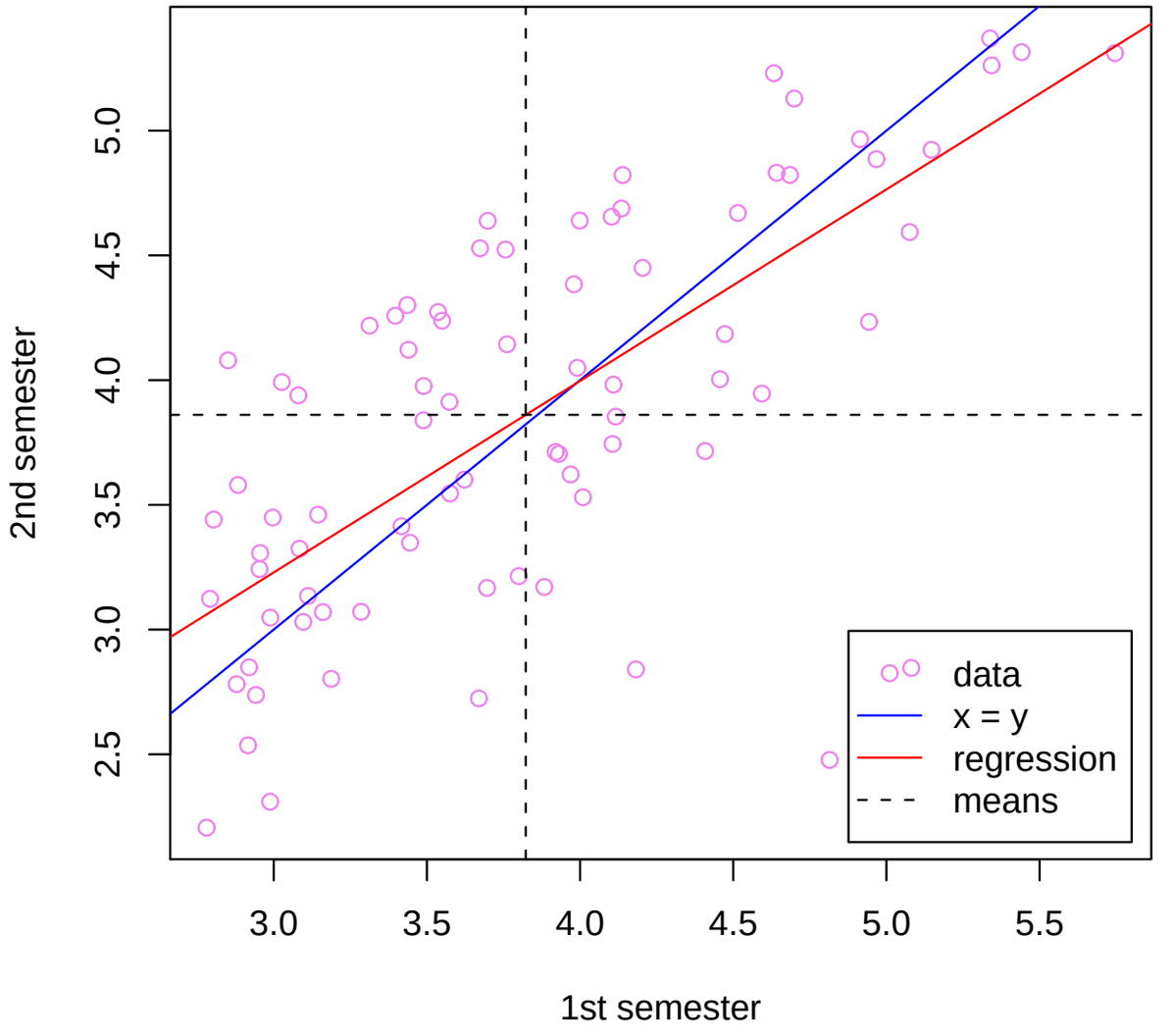
928 наблюдений $\texttt{ch} = 0.65\texttt{par} + 24 = 68.2 + 0.65 \times (\texttt{par} - 68.2)$

$x =$ средняя оценка по мат. анализу и алгебре в 1-м семестре

$y =$ средняя оценка по мат. анализу, алгебре и программированию во 2-м семестре



79 студентов

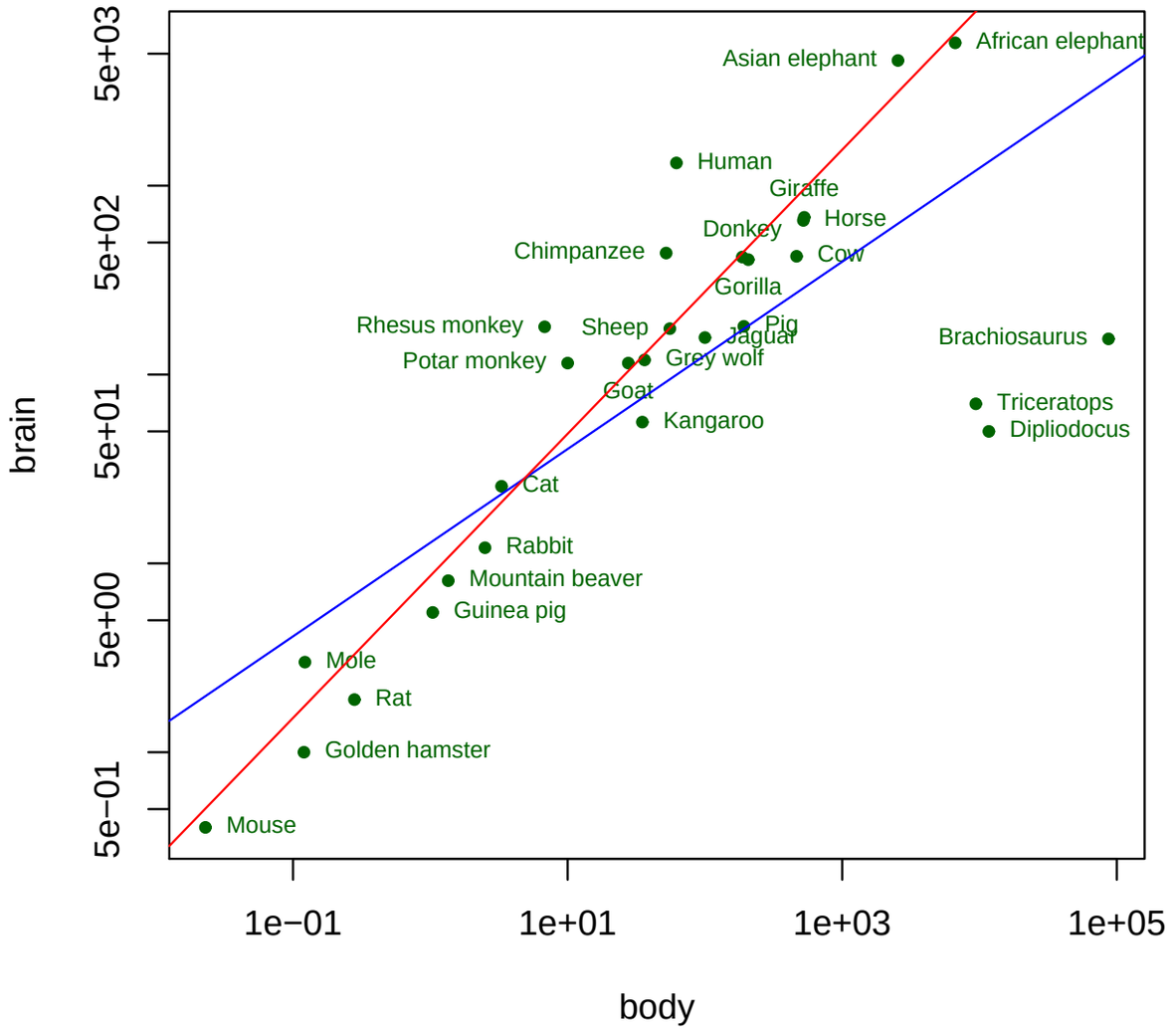


$\texttt{sem2} = 0.93 + 0.77 \times \texttt{sem1} \approx 3.86 + 0.77 \times (\texttt{sem1} - 3.82)$

3.82 — средняя оценка по всем студентам в 1-м семестре

3.86 — средняя оценка по всем студентам во 2-м семестре

Зависимость между массой тела и массой мозга животного (Найдите выбросы (outlayers).)



$\lg \texttt{brain} = \beta_0 + \beta_1 \lg \texttt{body}$

$\beta_0 = 0.94$, $\beta_1 = 0.75$, $\texttt{brain} = 8.6 \times (\texttt{body})^{3/4}$

Обучающая выборка

$$(x^{(1)}, y^{(1)}),\ (x^{(2)}, y^{(2)}),\ \ldots, (x^{(N)}, y^{(N)})$$

$$x^{(i)} \in \mathscr{X}, \qquad y^{(i)} \in \mathscr{Y} = \mathbf{R} \qquad (i = 1, 2, \ldots, N)$$

$$f^*(x^{(i)}) \approx y^{(i)} \qquad (i = 1, 2, \ldots, N)$$

Будем искать функцию $f^*(x)$ в виде (*линейная регрессионная модель*):

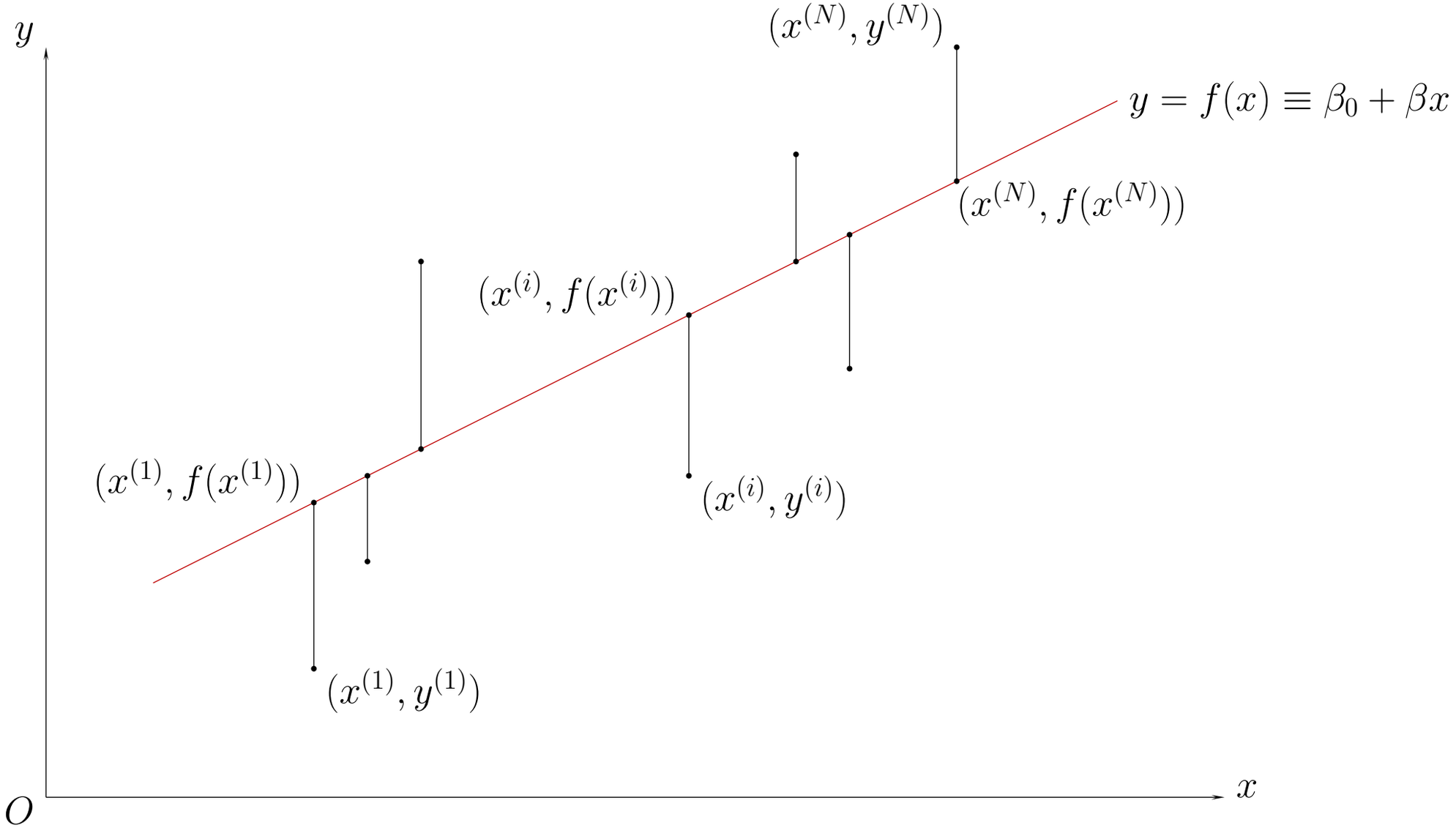
$$f(x) = f(x, \beta) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j$$

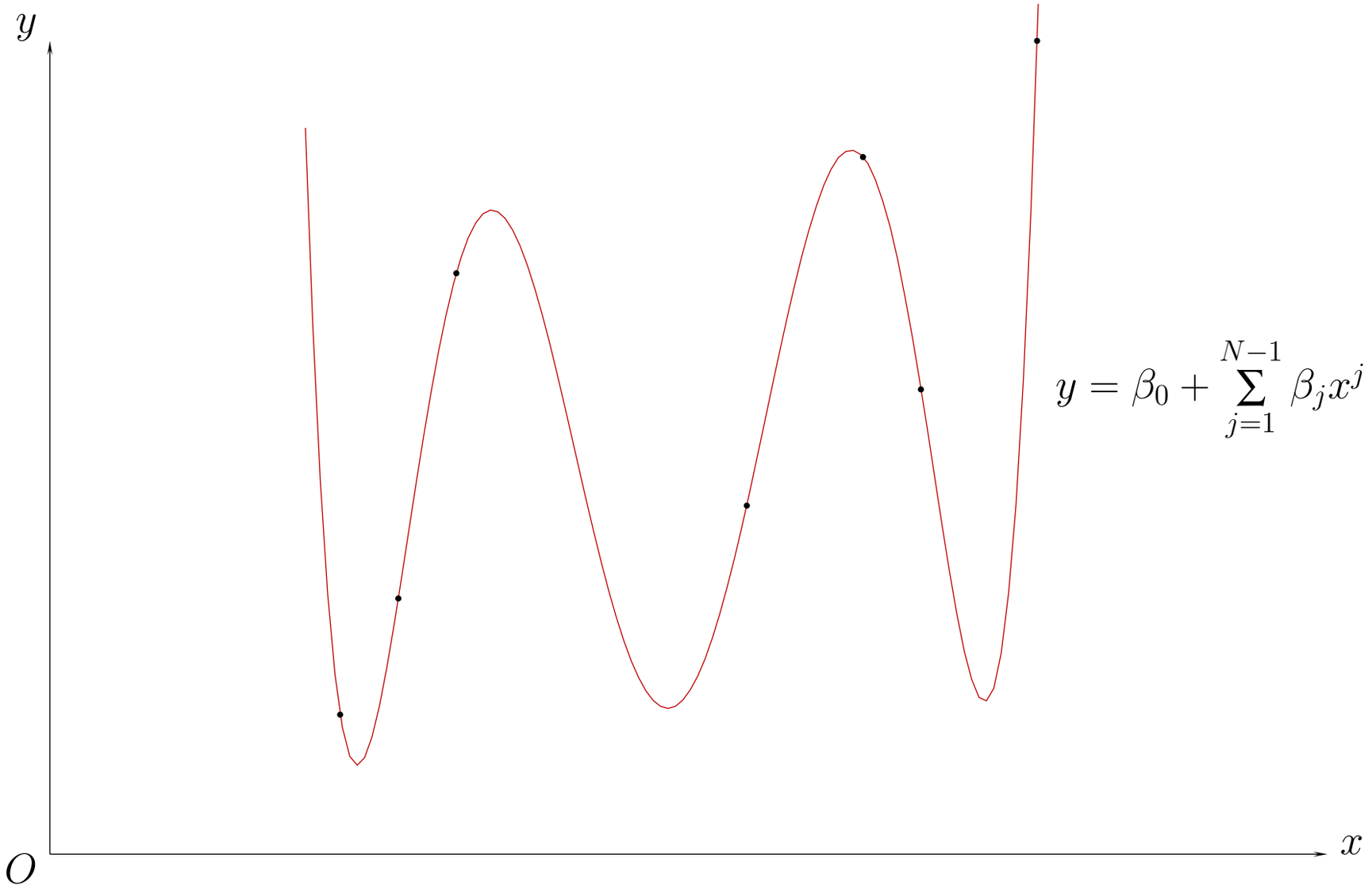
Метод наименьших квадратов — один из методов нахождения неизвестных параметров $\beta_0, \beta_1, \ldots, \beta_d$.

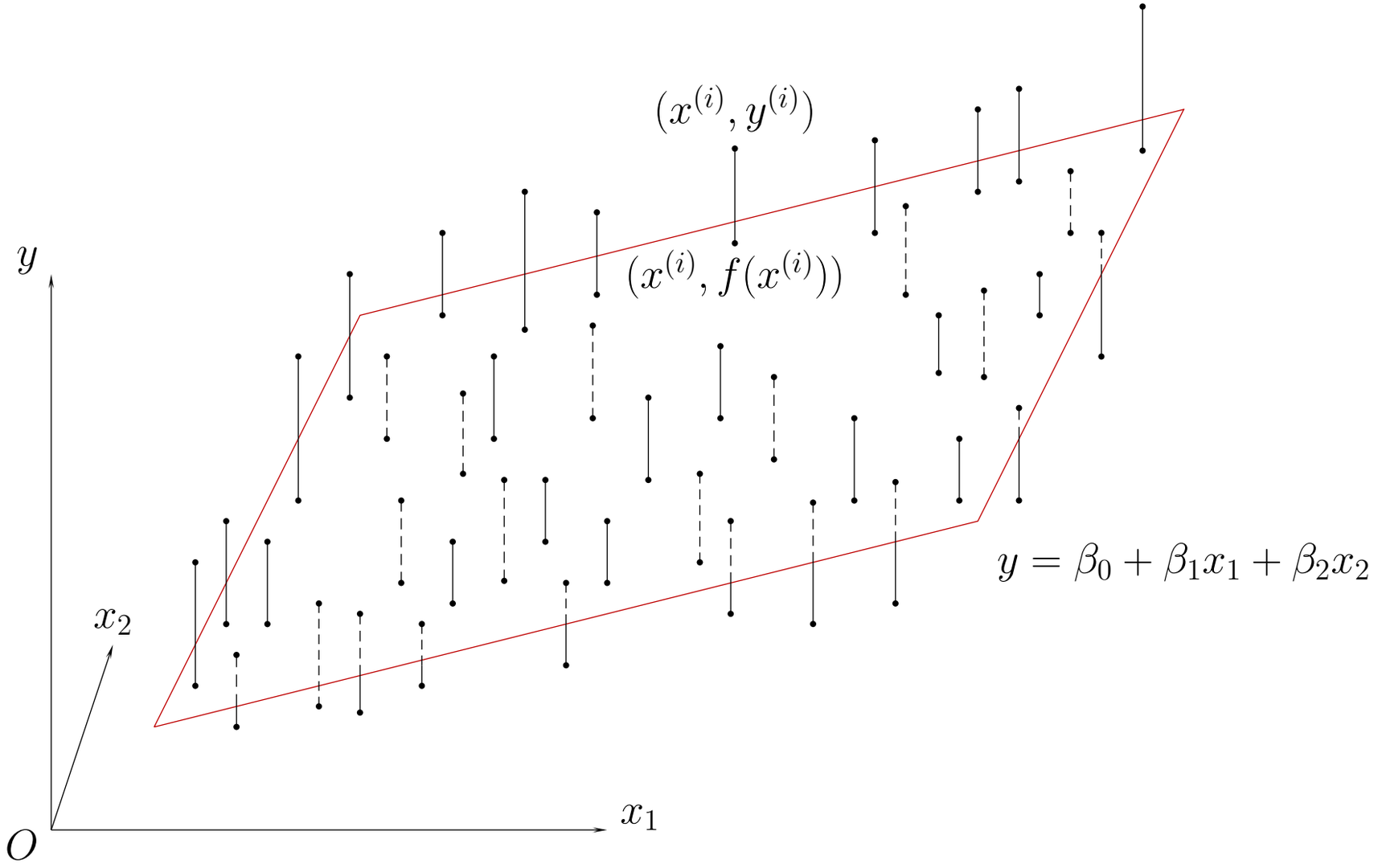
Ищем набор параметров $\widehat{\beta}$, доставляющих минимум *сумме квадратов невязок*, или *остаточной сумме квадратов* (*residual sum of squares*)

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} \Big(y^{(i)} - f(x^{(i)},\, \beta) \Big)^2.$$

Заметим, что $\dfrac{1}{N}\mathrm{RSS}(\beta)$ есть эмпирический риск для квадратичной функции потерь.


y
$(x^{(N)}, y^{(N)})$
$y = f(x) \equiv \beta_0 + \beta x$
$(x^{(N)}, f(x^{(N)}))$
$(x^{(i)}, f(x^{(i)}))$
$(x^{(1)}, f(x^{(1)}))$
$(x^{(i)}, y^{(i)})$
$(x^{(1)}, y^{(1)})$
O
x


y
O
x
$y = \beta_0 + \sum_{j=1}^{N-1} \beta_j x^j$


$(x^{(i)}, y^{(i)})$
$(x^{(i)}, f(x^{(i)}))$
y
x_2
O
x_1
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Функция $f^*(x)$ может быть функцией более общего вида:

$$f(x) = f(x, \beta) = \sum_{j=1}^{q} \beta_j h_j(x),$$

где β_j — неизвестные параметры, а $h_j(x)$ — заданные функции

Заметим, что параметры β_j входят *линейным* образом.

Но неизвестные параметры могут входить и нелинейным образом:

Например,

$$y = \beta_1 e^{\lambda_1 x} + \beta_2 e^{\lambda_2 x}.$$

β_1, β_2 — линейно, λ_1, λ_2 — нелинейно.

Если λ_1, λ_2 известны, то получаем линейную задачу (наименьших квадратов)

Если λ_1, λ_2 не известны — нелинейную задачу (наименьших квадратов)

Задача аппроксимации сплайнами тоже сводится к линейной задаче наименьших квадратов.

Почему метод наименьших *квадратов*?

Выведем (при некоторых дополнительных предположениях) метод наименьших квадратов из *принципа максимального правдоподобия*.

Y — с. в. с плотностью вероятности $p(y, \theta)$, где θ — вектор параметров.

N копий непрерывной случайной величины Y: $Y^{(1)}, Y^{(2)}, \ldots, Y^{(N)}$

(N независимых одинаково распределенных с.в. — испытания Бернулли)

N реализаций этих величин: $y^{(1)}, y^{(2)}, \ldots, y^{(N)}$

Плотность вероятности N-мерной с.в. $(Y^{(1)}, Y^{(2)}, \ldots, Y^{(N)})$:

$$L(\theta) = p(y^{(1)}, y^{(2)}, \ldots, y^{(N)}, \theta) = p(y^{(1)}, \theta) \cdot p(y^{(2)}, \theta) \cdot \cdots \cdot p(y^{(i)}, \theta)$$

$L(\theta)$ — *функция правдоподобия*

Логарифмическая функция правдоподобия:

$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{N} \ln p(y^{(i)}, \theta).$$

Принцип максимального правдоподобия предполагает, что наиболее разумные значения неизвестных параметров θ доставляют максимум функции $L(\theta)$ (и $\ell(\theta)$).

(Если Y — дискретная, то вместо $p(y^{(i)}, \theta)$ нужно рассмотреть $\mathsf{Pr}\left\{Y = y^{(i)}\right\}$)

Модель *с аддитивной случайной ошибкой*:

$$y = f^*(x) + E,$$

где E — случайная величина (ошибка), не зависящая от x, и $\mathsf{E}\, E = 0$.

$f^*(x) = \mathsf{E}\,(Y \,|\, X = x)$ — регрессионная функция.

Дополнительно предположим, что $E \sim N(0,\, \sigma^2) \Leftrightarrow$

$$p(y \,|\, x,\, \beta) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{1}{2}\cdot\frac{\left(y - f(x,\beta)\right)^2}{\sigma^2}}$$

Тогда $$\ell(\beta) = \sum_{i=1}^{N} \ln p(y^{(i)} \,|\, x,\, \beta) = -\frac{N}{2}\ln 2\pi - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - f(x^{(i)},\beta)\right)^2$$

В ней только последний член содержит вектор параметров β.

С точностью до множителя этот член равен $\mathrm{RSS}(\beta)$

Итак, при сделанных предположениях метод наименьших квадратов эквивалентен принципу максимального правдоподобия

Как найти минимум функции $\mathrm{RSS}(\beta)$?

Пусть

$$\mathbf{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_d^{(2)} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_1^{(N)} & x_2^{(N)} & \dots & x_d^{(N)} \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}$$

Тогда

$$\mathrm{RSS}(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) \to \min$$

Можно рассмотреть систему уравнений (относительно β)

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

$\widehat{\beta}$ называется *псевдорешением* этой системы (оно минимизирует норму невязки).

$\mathrm{RSS}(\beta)$ — квадратичная функция от $d+1$ неизвестных (параметров) $\beta_0, \beta_1, \dots, \beta_d$.

Дифференцируя, находим:

$$\frac{\partial\,\mathrm{RSS}}{\partial\beta} = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta), \qquad \frac{\partial^2\mathrm{RSS}}{\partial\beta\,\partial\beta^\top} = 2\mathbf{X}^\top\mathbf{X}.$$

Обозначим $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_d$ столбцы матрицы $\mathbf{X}$.

Если $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_d$ линейно независимы, то матрица $\mathbf{X}^\top\mathbf{X}$ невырождена и положительно определена, поэтому минимум функции $\mathrm{RSS}(\beta)$ достигается, когда первая производная по β обращается в ноль:

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta) = 0 \qquad \Leftrightarrow \qquad \mathbf{X}^\top\mathbf{X}\beta = \mathbf{X}^\top\mathbf{y}.$$

Это *нормальная система уравнений*, или *система нормальных уравнений*.

Единственным решением является вектор

$$\widehat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

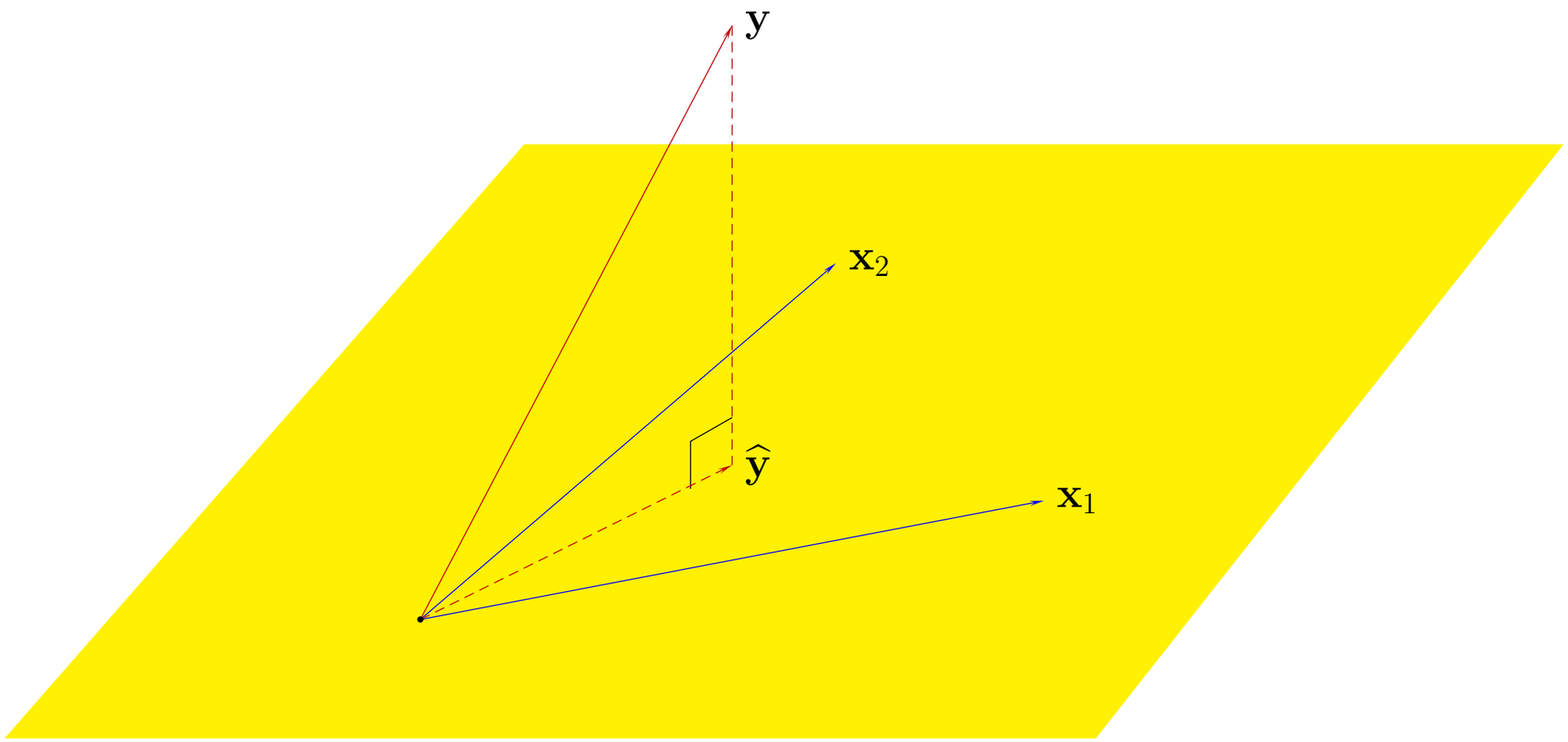
Итак, *псевдорешением* системы $\mathbf{X}\beta = \mathbf{y}$ является *решение* системы $\mathbf{X}^\top\mathbf{X}\beta = \mathbf{X}^\top\mathbf{y}$.

Матрица $\mathbf{X}^+ = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ называется *псевдообратной* (Мура–Пенроуза) к $\mathbf{X}$.

Входным значениям $x_1, x_2, \ldots, x_N$ будет соответствовать вектор выходных переменных

$$\widehat{\mathbf{y}} = (\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_d)^\top = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Пусть $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, тогда получаем $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$
$\widehat{\mathbf{y}}$ есть ортогональная проекция $\mathbf{y}$ на п/п-во, натянутое на $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_d$
$\mathbf{H}$ называется *матрицей проектирования*


$\mathbf{y}$
$\mathbf{x}_2$
$\widehat{\mathbf{y}}$
$\mathbf{x}_1$

Если столбцы матрицы $\mathbf{X}$ линейно зависимы, то $\widehat{\beta}$, на котором достигается минимум функции $\mathrm{RSS}(\beta)$, не единственен, однако, по-прежнему, $\widehat{\mathbf{y}}$ является ортогональной проекцией вектора $\mathbf{y}$ на линейную оболочку векторов $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_d$.

5.2. Проверка значимости и доверительные интервалы для коэффициентов (регрессионный анализ)

$\mathbf{y}$, E, $\widehat{\beta}$ случайны; $\mathbf{X}$, β детерминированы; $\mathsf{E}\,E = 0$, $\mathsf{Cov}\,E = \sigma^2\mathbf{I}$.

$$\mathbf{y} = \mathbf{X}\beta + E, \qquad \widehat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\beta + E).$$

$$\widehat{\beta} - \beta = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\beta + E) - \beta = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E$$

Несмещенность:

$$\mathsf{E}\,(\widehat{\beta} - \beta) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathsf{E}\,E = 0 \qquad \Rightarrow \qquad \mathsf{E}\,\widehat{\beta} = \beta.$$

Матрица ковариации:

$$\mathsf{Cov}\,\widehat{\beta} = \mathsf{E}\,\big((\widehat{\beta} - \beta)(\widehat{\beta} - \beta)^\top\big) = \mathsf{E}\,\Big(\big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E\big)\big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E\big)^\top\Big)$$

$$= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \cdot \mathsf{E}\,(EE^\top) \cdot \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \cdot \sigma^2\mathbf{I} \cdot \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$$

$$= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

Несмещенной оценкой для σ^2 является *остаточная дисперсия* $\widehat{\sigma}^2 = \mathrm{RSS}\,/(N - d - 1)$:

$$\mathsf{E}\,\widehat{\sigma}^2 = \mathsf{E}\,\big(\mathrm{RSS}\,/(N - d - 1)\big) = \mathsf{E}\,\big(\mathbf{y}^\top(\mathbf{I} - \mathbf{H})\mathbf{y}/(N - d - 1)\big) = \sigma^2.$$

Дополнительные предположения:

Пусть ошибки $E^{(i)}$ распределены по нормальному закону:

$$E \sim N(0,\, \sigma \mathbf{I}).$$

В этом случае из некоррелированности случайных величин $E^{(i)}$ следует их независимость.

Теперь можно показать (faciat, qui potest), что

$$\widehat{\beta} \sim N\big(\beta,\ (\mathbf{X}^\top \mathbf{X})^{-1}\sigma^2\big) \qquad \text{и} \qquad (N-d-1)\,\widehat{\sigma}^2 \sim \sigma^2 \chi^2_{N-d-1}.$$

Эти свойства будем использовать для построения статистических тестов и доверительных интервалов для β_j.

Проверка значимости модели.

Гипотеза H_0: $\beta_1 = \cdots = \beta_d = 0$

Альтернативная гипотеза H_1: найдется j, для которого $\beta_j \neq 0$

Если гипотеза H_0 верна, то решением задачи наименьших квадратов будет

$$\beta_0 = \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)}.$$

В этом случае остаточная сумма квадратов (называемая в даном случае *полной суммой квадратов относительно среднего*) равна

$$\mathrm{TSS} = \sum_{i=1}^{N}(y^{(i)} - \beta_0)^2 = \sum_{i=1}^{d}(y^{(i)} - \overline{y})^2$$

Если принимается гипотеза H_1, то говорят, что модель *статистически значима*.

Можно показать, что

$$\sum_{i=1}^{d}(y^{(i)}-\overline{y})^2=\sum_{i=1}^{d}(y^{(i)}-\widehat{y}^{(i)})^2+\sum_{i=1}^{d}(\widehat{y}-\overline{y})^2$$
$$\text{TSS} \qquad\qquad \text{RSS} \qquad\qquad \text{ESS}$$

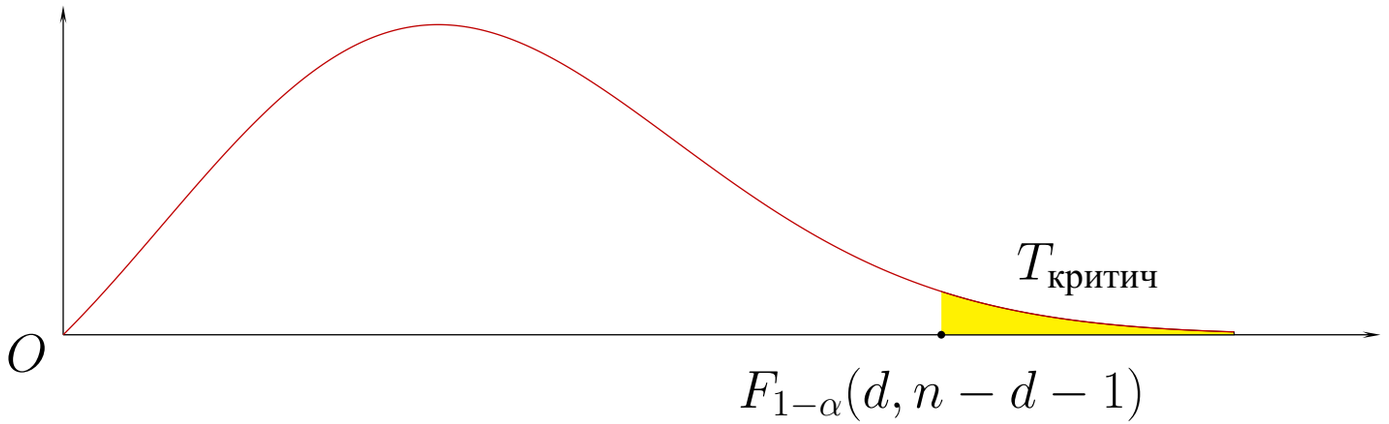
$\text{TSS}=\sum\limits_{i=1}^{N}(y^{(i)}-\overline{y})^2$ — *полная сумма квадратов* (total sum of squares)

$\text{ESS}=\sum\limits_{i=1}^{N}(\widehat{y}^{(i)}-\overline{y})^2$ — *сумма квадратов, обусловленная регрессией* (explained s. of sq.)

Если гипотеза H_0 верна, то

$$\widehat{F}=\frac{\text{ESS}}{d\sigma^2}\Big/\frac{\text{RSS}}{(n-d-1)\sigma^2}\sim F(d,\,n-d-1)$$

Если модель значима, то ESS большая и критическая область справа:



Коэффициент детерминации (взгляд с другой стороны)

Коэффициент детерминации, или *коэффициент регрессии Пирсона*

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

— доля объясняемого регрессией разброса относительно среднего.

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

Если R^2 близок к 1, то RSS намного больше TSS.

Очевидно, что $R = \sqrt{R^2}$ по модулю равно выборочной корреляции между $y^{(i)}$ и $\widehat{y}^{(i)}$.

Если $d = 1$, то R равен выборочной корреляции между $x^{(i)}$ и $y^{(i)}$.

Если гипотеза H_0 верна, то

$$\widehat{F} = \frac{R^2/d}{(1-R^2)/(N-d-1)} \sim F(d,\, n-d-1).$$

Недостаток R^2: при добавлении в модель новых признаков R^2 не изменяется. (Если d велико и близко к N, то может быть, что $\mathrm{RSS} \approx 0$ и $R^2 \approx 1$.) В этом случае используется скорректированный («подправленный») коэффициент

$$R^2_{\mathrm{a}} = R^2 - \frac{1-R^2}{N-d-1}.$$

Упражнение 5.1 Доказать, что $\mathrm{TSS} = \mathrm{RSS} + \mathrm{ESS}$. Это можно доказать непосредственное, а можно предварительно доказать, что $\mathbf{y} - \widehat{\mathbf{y}} \perp \widehat{\mathbf{y}} - \overline{\mathbf{y}}$, где $\overline{\mathbf{y}}$ — вектор, составленный из $\overline{y}$, и воспользоваться этим.

Упражнение 5.2 Разности

$$e^{(i)} = y^{(i)} - \widehat{y}^{(i)} = y^{(i)} - \sum_{j=1}^{d} \widehat{\beta}_j x_j^{(i)}$$

называются *остатками*. Доказать, что

$$\sum_{i=1}^{N} e^{(i)} = 0.$$

Упражнение 5.3 Доказать, что

$$\overline{y} = \beta_0 + \sum_{j=1}^{d} \widehat{\beta}_j \overline{x}_j.$$

Упражнение 5.4 Доказать, что

$$\sum_{i=1}^{N} y^{(i)} = \sum_{i=1}^{N} \widehat{y}^{(i)}.$$

Проверка значимости одного коэффициента.

Гипотеза H_0: $\beta_j = 0$ (j фиксировано):

использование переменной x_j не улучшает предсказание по сравнению с предсказанием, полученным на основе только остальных $d-1$ переменных.

Для проверки этой гипотезы (против гипотезы $\beta_j \neq 0$) рассмотрим *стандартный коэффициент* (z-score)

$$t_j = \frac{\widehat{\beta}_j}{\operatorname{se} \beta_j}, \qquad (\star)$$

где

$$\operatorname{se} \beta_j = \widehat{\sigma}\sqrt{v_j}$$

— *стандартная ошибка* коэффициента β_j, а v_j — j-й диаг. элемент матрицы $(\mathbf{X}^\top \mathbf{X})^{-1}$.

В предположении $\beta_j = 0$ коэффициент t_j имеет t-распределение Стьюдента t_{N-d-1}.

Если $|t_j|$ «велико», то гипотезу H_0 следует отбросить.

Если гипотеза H_0 отбрасывается, то говорят, что коэффициент $\widehat{\beta}_j$ *статистически значим*.

Можно проверить гипотезу $\beta_j = \beta'_j$ (относительно односторонней или двусторонней альтернативы), где β'_j — некоторое заданное значение.

Статистика критерия имеет в этом случае вид

$$t'_j = \frac{\widehat{\beta}_j - \beta'_j}{\operatorname{se} \beta_j}.$$

Коэффициент t'_j имеет распределение t_{N-d-1}.

Проверка гипотезы зависит от вида альтернативной гипотезы и происходит обычным образом.

Проверка значимости группы коэффициентов. Гипотеза о равенстве нулю группы коэффициентов (против гипотезы, что по крайней мере один из коэффициентов не равен нулю): переменные этой группы не улучшают предсказание по отношению к предсказанию, полученному без этих переменных.

Будем использовать статистику

$$F = \frac{(\mathrm{RSS}_2 - \mathrm{RSS}_1)/(d_1 - d_2)}{\mathrm{RSS}_1/(N - d_1 - 1)},$$

где RSS_1 — остаточная сумма квадратов «бóльшей» модели с $d_1 + 1$ параметрами, а RSS_2 — остаточная сумма квадратов «вложенной» модели с $d_2 + 1$ параметрами,

(«вложенная» модель получается из «бóльшей» обнулением $d_1 - d_2$ параметров).

В предположении, что E имеет нормальное распределение, статистика F имеет $F(d_1 - d_2,\, N - d_1 - 1)$ распределение Фишера.

Если отбрасывается один коэффициент, то F равен квадрату стандартного коэффициента t_j из $(\star)$.

Доверительные интервалы. Для β_j доверительным интервалом является

$$\left(\widehat{\beta}_j - z^{(1-\alpha)}\widehat{\sigma}\sqrt{v_j},\ \ \widehat{\beta}_j + z^{(1-\alpha)}\widehat{\sigma}\sqrt{v_j}\right),$$

где $z^{(1-\alpha)}$ есть $(1-\alpha)$-процентиль для нормального распределения:

$$\begin{aligned} z^{(1-0.1)} &= 1.645, \\ z^{(1-0.05)} &= 1.96, \\ z^{(1-0.01)} &= 2.58, \quad \text{и т. д.} \end{aligned}$$

(v_j — j-й диаг. элемент в $(\mathbf{X}^\top\mathbf{X})^{-1}$, $\operatorname{se}\widehat{\beta}_j = \widehat{\sigma}\sqrt{v_j}$ — стандартная ошибка для $\widehat{\beta}_j$).

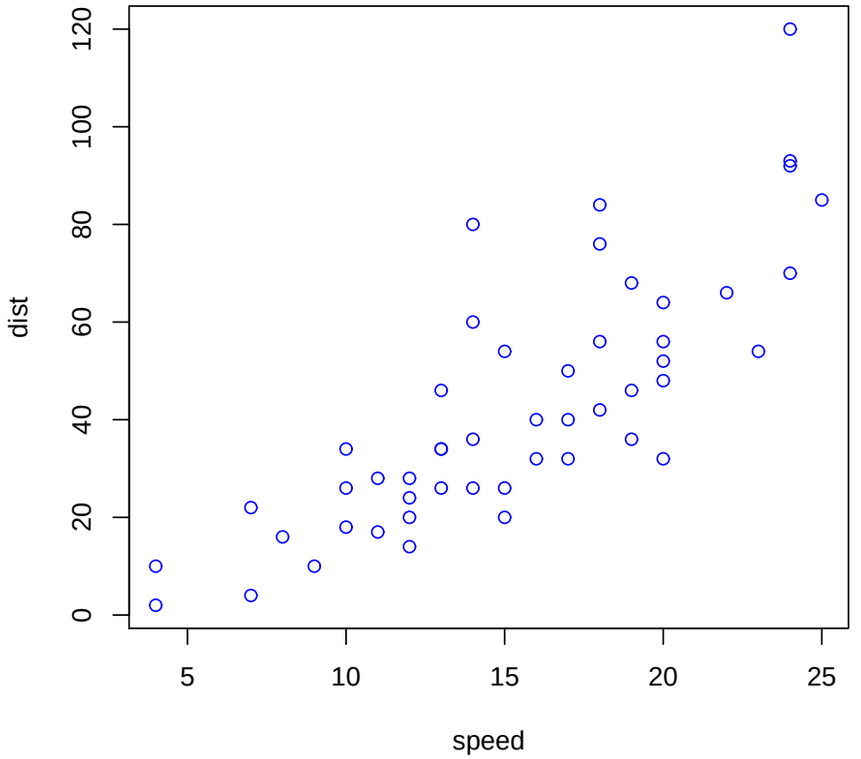
Таким образом, интервал $\widehat{\beta} \pm 2 \cdot \operatorname{se}\widehat{\beta}$ соответствует мере доверия около 95%.

5.2.1. Пример

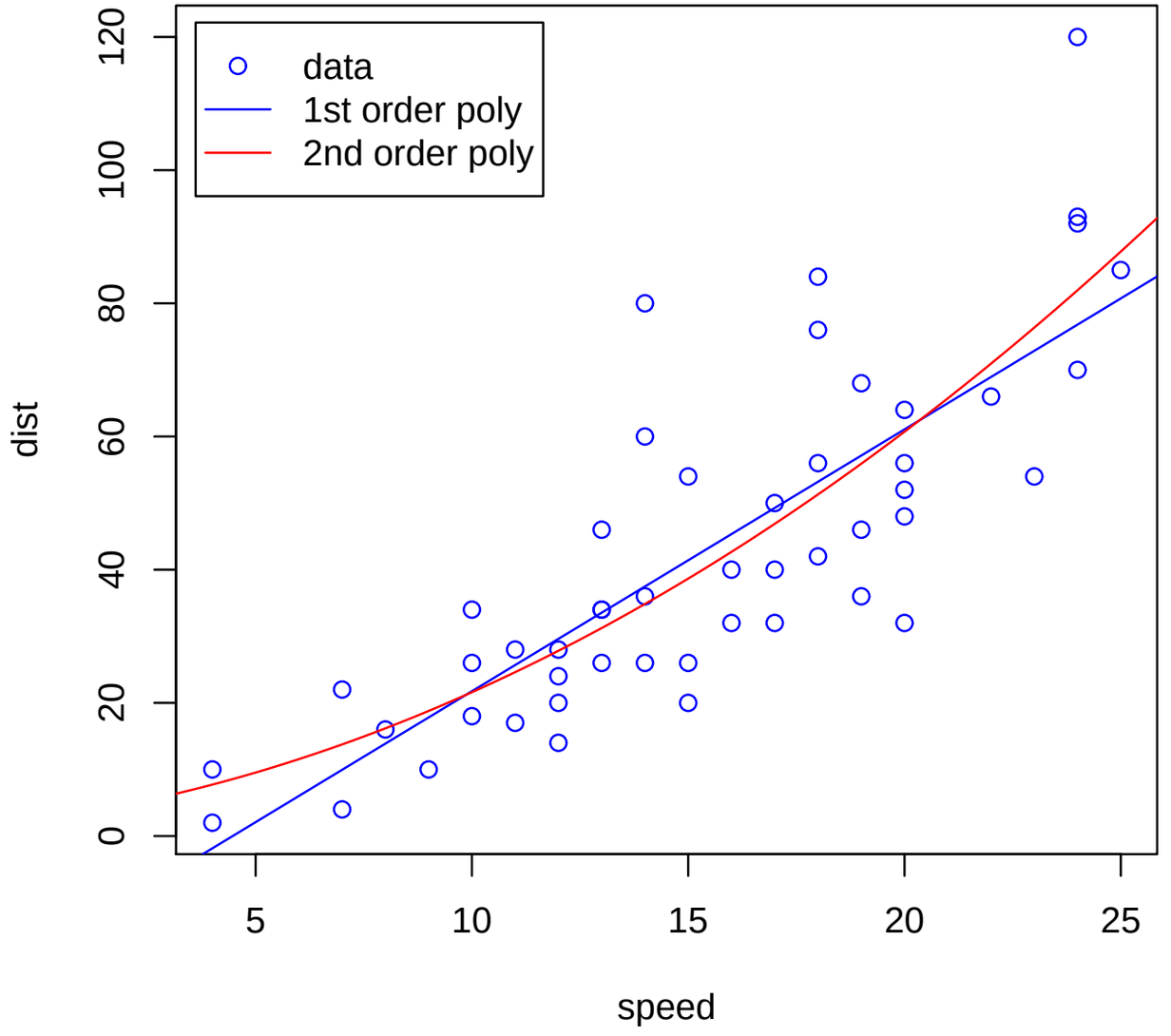
Зависимость длина тормозного пути (dist) от начальной скорости автомобиля (speed) [*Ezekiel M.* Methods of Correlation Analysis. Wiley. 1930], $N = 50$.

В качестве модели рассмотрим

$$\text{dist} = \beta_0 + \beta_1 \times \text{speed}.$$



Найдены значения $\widehat{\beta}_0 = -17.579$, $\widehat{\beta}_1 = 3.932$.



Синий график: регрессия в виде $\text{dist} = \beta_0 + \beta_1 \times \text{speed}$.

Красный график: регрессия в виде $\text{dist} = \beta_0 + \beta_1 \times \text{speed} + \beta_2 \times \text{speed}^2$.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  \pvalue: 1.49e-12

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.47014   14.81716   0.167    0.868
speed        0.91329    2.03422   0.449    0.656
I(speed^2)   0.09996    0.06597   1.515    0.136

Residual standard error: 15.18 on 47 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6532
F-statistic: 47.14 on 2 and 47 DF,  \pvalue: 5.852e-12
```

Для dist $= \beta_0 + \beta_1 \times$ speed

Коэффициент детерминации и подправленный коэфициент детерминации:

$$r^2 = 0.6511, \qquad r^2_{\text{a}} = 0.6438.$$

Таким образом, регрессия объясняет около 65% изменчивости данных.

Стандартные ошибки и стандартные коэффициенты:

$$\text{se}\,\beta_0 = 6.7584, \quad \text{se}\,\beta_1 = 0.4155, \quad t_0 = -2.601, \quad t_1 = 9.464.$$

Задача имеет $N - d - 1 = 48$ степеней свободы.

Сравнивая значения t_0 и t_1 с квантилями t-распределения Стьюдента с 48 степенями свободы, получаем, что для гипотезы $\beta_0 = 0$ значение p-valueменьше 0.0123, а для гипотезы $\beta_1 = 0$ оно равно 1.49×10^{-12}.

Если считать, что уровень надежности равен, например, $\alpha = 0.01$, то гипотезу $\beta_0 = 0$ принимаем, гипотезу $\beta_1 = 0$ отвергаем.

Остаточная стандартная ошибка равна $\widehat{\sigma} = 15.38$.

Значение F-статистики $F = 89.57$ сравниваем с квантилями F-распределения $F_{1,48}$.

Получаем p-value, равное 1.490×10^{-12} — для указанного уровня надежности модель статистически значима.

Для dist $= \beta_0 + \beta_1 \times$ speed $+ \beta_2 \times$ speed^2:

Найдены значения $\widehat{\beta}_0 = 2.47014$, $\widehat{\beta}_1 = 0.91329$, $\widehat{\beta}_1 = 0.09996$.

Имеем

$$r^2 = 0.6673, \qquad r^2_{\text{a}} = 0.6532,$$

$$\text{se}\,\beta_0 = 14.81716, \quad \text{se}\,\beta_1 = 2.03422, \quad \text{se}\,\beta_2 = 0.06597,$$

$$t_0 = 0.167, \quad t_1 = 0.449, \quad t_2 = 1.515.$$

Задача имеет $N - d - 1 = 47$ степеней свободы.

Значение p-valueдля гипотезы $\beta_0 = 0$ меньше 0.868.

Для гипотезы $\beta_1 = 0$ оно равно 0.656,

а для гипотезы $\beta_2 = 0$ равно 0.136.

Таким образом, при уровне надежности $\alpha = 0.01$ гипотезу $\beta_2 = 0$ принимаем.

Коэффициент β_2 не является статистически значимым.

Остаточная стандартная ошибка равна $\widehat{\sigma} = 15.18$.

Значение F-статистики $F = 47.14$ нужно сравнить с квантилями F-распределения Фишера $F_{2,47}$.

Это сравнение приводит к p-value, равному 5.852×10^{-12}.

Таким образом, для указанного уровня надежности модель статистически значима.

Сравниваем обе модели (1-я квадратичная; 2-я линейная)

$$F = \frac{(\mathrm{RSS}_2 - \mathrm{RSS}_1)/(d_1 - d_2)}{\mathrm{RSS}_1/(N - d_1 - 1)} = \frac{(11353.52 - 10824.72)/(2-1)}{10824.72/(50-2-1)} = 2.296027$$

Заметим, что $\sqrt{F} = 1.515265$ совпадает с t-статистикой для β_2

Вычисляем p-value$= 0.13640241 - \beta_2$ незначим.

5.2.2. Пример. Boston

Все данные:

$N = 506$, $d = 13$ $\mathrm{RSS} = 11078.78$ $\mathrm{RSS}/N = 21.90$

```
Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
CRIM        -1.080e-01  3.286e-02  -3.287 0.001087 **
ZN           4.642e-02  1.373e-02   3.382 0.000778 ***
INDUS        2.056e-02  6.150e-02   0.334 0.738288
CHAS         2.687e+00  8.616e-01   3.118 0.001925 **
NOX         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
RM           3.810e+00  4.179e-01   9.116  < 2e-16 ***
AGE          6.922e-04  1.321e-02   0.052 0.958229
DIS         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
RAD          3.060e-01  6.635e-02   4.613 5.07e-06 ***
TAX         -1.233e-02  3.760e-03  -3.280 0.001112 **
```

```
PTRAT     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
B          9.312e-03  2.686e-03   3.467 0.000573 ***
LSTAT     -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared: 0.7406,     Adjusted R-squared: 0.7338
F-statistic: 108.1 on 13 and 492 DF,  \pvalue: < 2.2e-16
```

Выделим обучающую ($N = 405$) и тестовую ($N_{\text{test}} = 101$) выборки

на обучающей $\text{RSS} = 8655.383$, $\text{RSS}/N = 21.37132$

на тестовой $\text{RSS}_{\text{test}} = 2522.615$, $R \approx \text{RSS}_{\text{test}}/N_{\text{test}} = 24.97638$, $\text{se}\, R = 0.4997638$.

$$(\text{se}\, R)^2 = \frac{1}{N_{\text{test}}} \mathsf{Var}(Y - \widehat{Y}) = \frac{1}{N_{\text{test}}} \left(\frac{1}{N_{\text{test}} - 1} \sum_{i=1}^{N_{\text{test}}} (y^{(i)} - \widehat{y}^{(i)})^2 \right)$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.353120   5.775840   5.601 4.01e-08 ***
CRIM         -0.116261   0.034297  -3.390  0.00077 ***
ZN            0.049895   0.015084   3.308  0.00103 **
INDUS         0.032491   0.073167   0.444  0.65724
CHAS          2.054215   1.059339   1.939  0.05320 .
NOX         -17.501306   4.194206  -4.173 3.71e-05 ***
RM            4.261255   0.500527   8.514 3.67e-16 ***
AGE          -0.003699   0.015249  -0.243  0.80845
DIS          -1.389752   0.216254  -6.426 3.80e-10 ***
RAD           0.310786   0.072337   4.296 2.19e-05 ***
TAX          -0.013453   0.004167  -3.229  0.00135 **
PTRATIO      -0.913321   0.148451  -6.152 1.89e-09 ***
B             0.008908   0.002929   3.042  0.00251 **
LSTAT        -0.475633   0.061656  -7.714 1.02e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.705 on 391 degrees of freedom
Multiple R-squared:  0.7432,    Adjusted R-squared:  0.7346
F-statistic: 87.03 on 13 and 391 DF,  \pvalue: < 2.2e-16
```

Возьмем только значимые (***) переменные: CRIM + NOX + RM + DIS + RAD + PTRATIO + LSTAT

На обучающей выборке $\mathrm{RSS} = 9437.129$, $\mathrm{RSS}/N = 23.30155$

На тестовой выборке $\mathrm{RSS} = 2798.908$, $\mathrm{RSS}/N = 27.71196$ $\mathrm{se} = 0.5264215$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.55914    5.71594   6.221 1.25e-09 ***
CRIM         -0.11438    0.03527  -3.243  0.00128 **
NOX         -20.55633    3.86533  -5.318 1.75e-07 ***
RM            4.59477    0.48776   9.420  < 2e-16 ***
DIS          -1.06844    0.17807  -6.000 4.44e-09 ***
RAD           0.11284    0.04673   2.415  0.01619 *
PTRATIO      -1.13134    0.14049  -8.053 9.54e-15 ***
LSTAT        -0.51548    0.05926  -8.699  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.876 on 397 degrees of freedom
Multiple R-squared:   0.72,     Adjusted R-squared:  0.715
F-statistic: 145.8 on 7 and 397 DF,  \pvalue: < 2.2e-16
```

$$F = \frac{(\mathrm{RSS}_2 - \mathrm{RSS}_1)/(d_1 - d_2)}{\mathrm{RSS}_1/(N - d_1 - 1)} = \frac{(9437.129 - 8655.383)/(13 - 7)}{8655.383/(405 - 13 - 1)} = 5.885792$$

Распределение $F(d_1 - d_2, N - d_1 - 1) = F(6, 391)$.

Находим p-value: $6.783518e - 06$, т. е. гипотеза о незначимости 6 переменных не принимается (для разумного уровня значимости α)

5.2.3. Анализ остатков

Ранее мы сделали предположение, что остатки должны быть независимы (некоррелированы) и иметь нормальное распределение.

Многие рассмотренные выше тесты достаточно устойчивы по отношению к отклонениям от таких предположений. Однако перед тем, как исследовать статистичекую значимость модели и строить доверительные интервалы, рекомендуется провести анализ остатков.

«Визуальные» способы:

- Гистограмма остатков — должна быть близка к нормальному распределению;
- график остатков — без зависимостей и нелинейностей;
- *QQ-график* (график «квантиль–квантиль»)

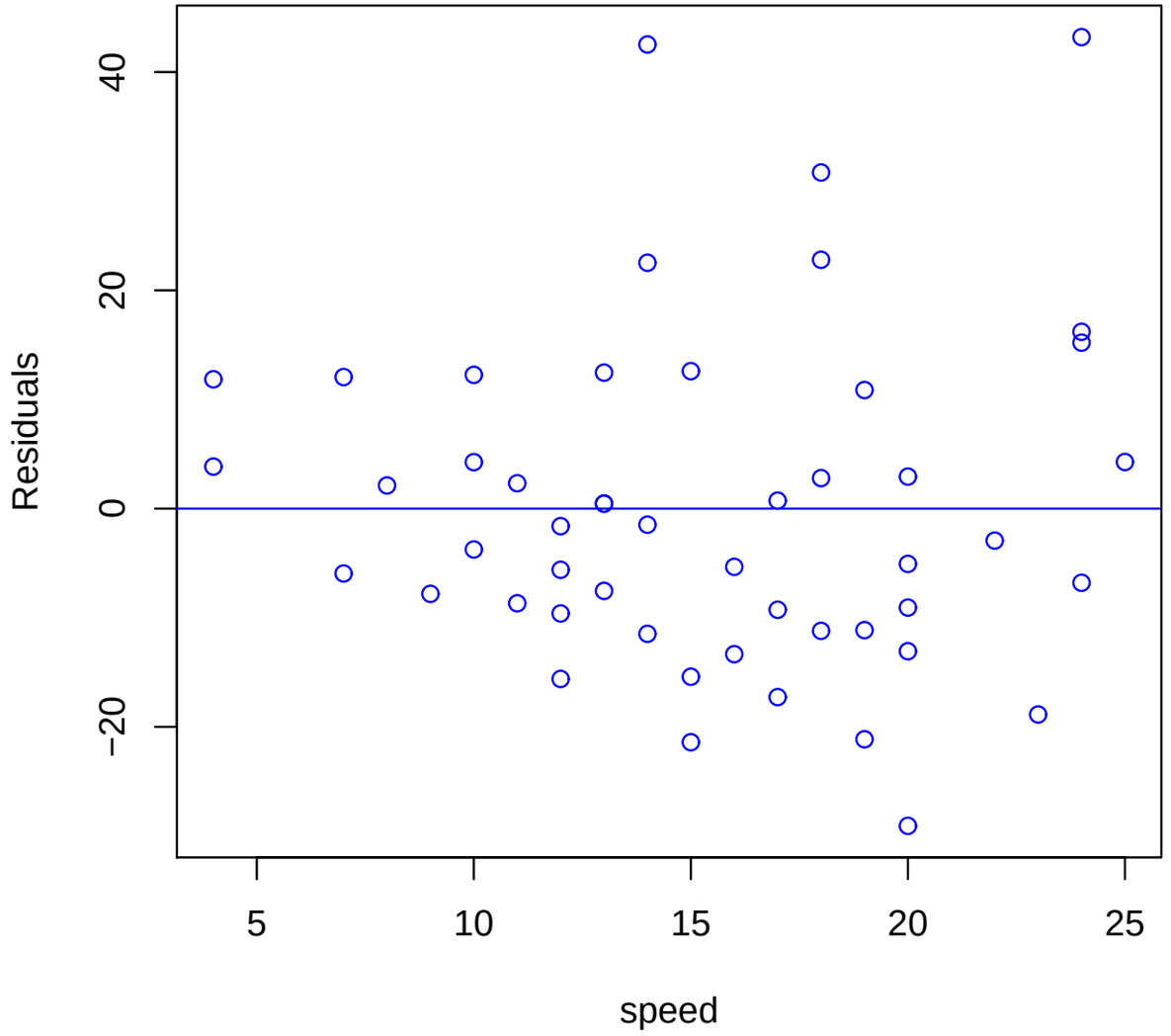
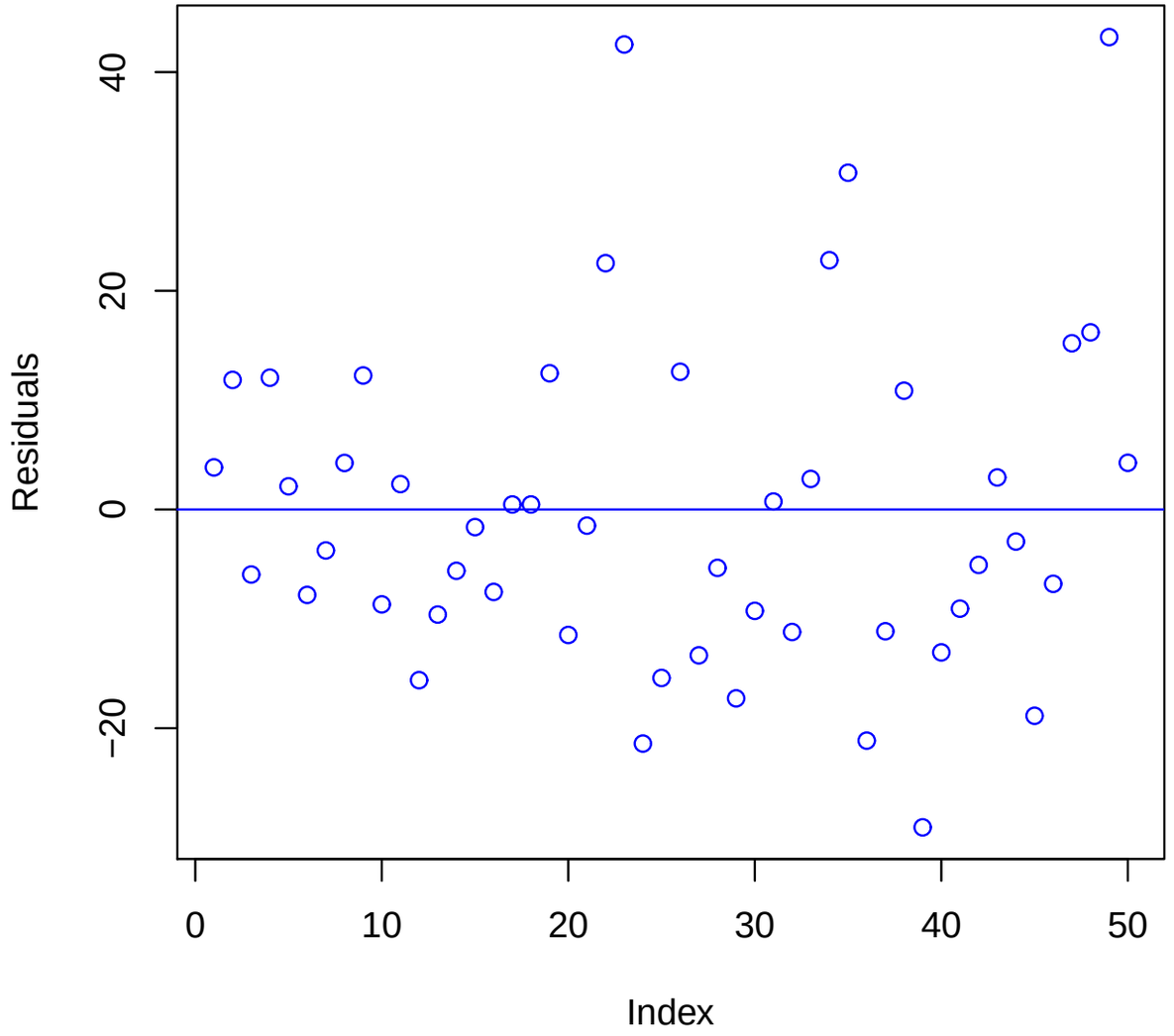
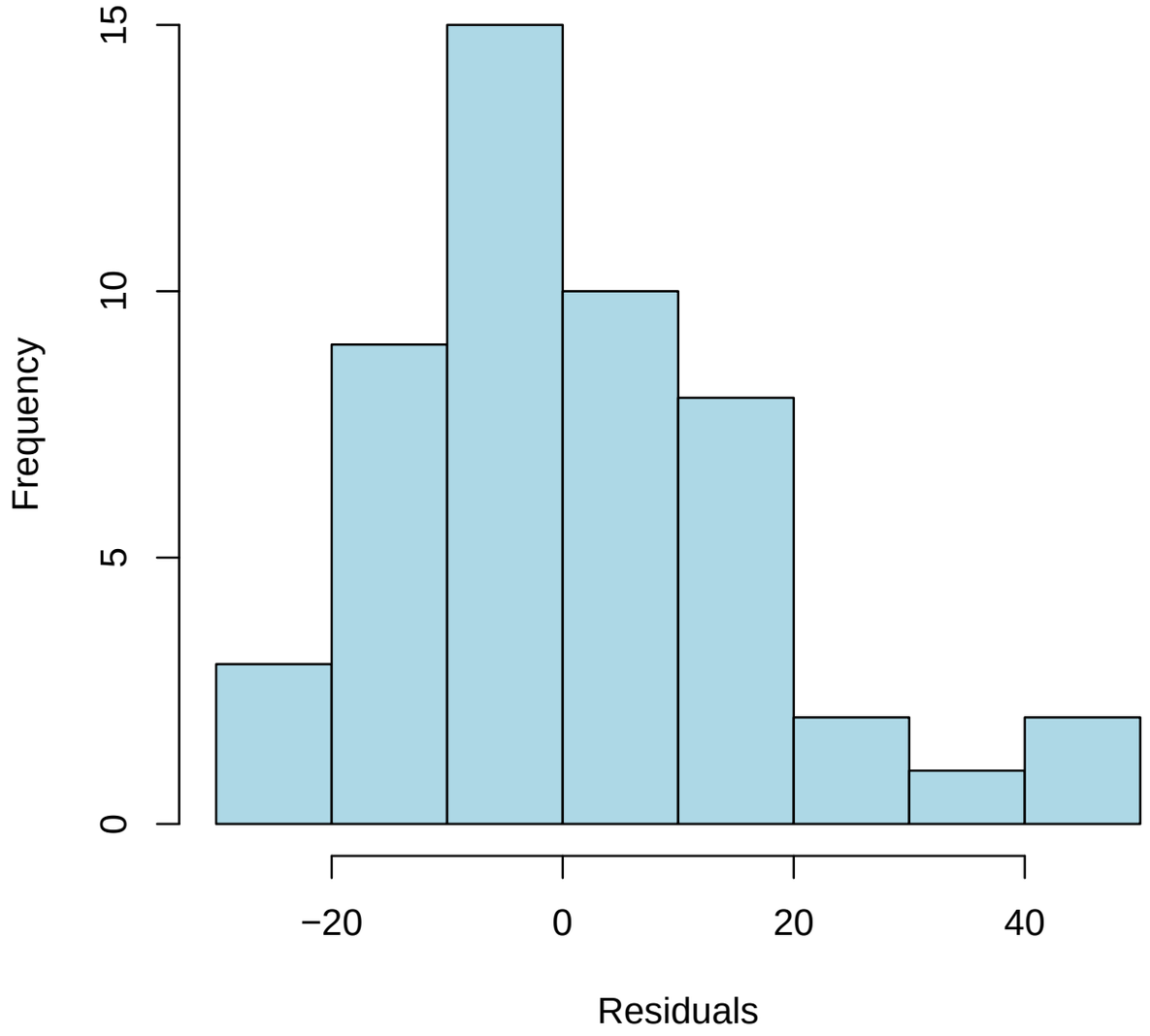
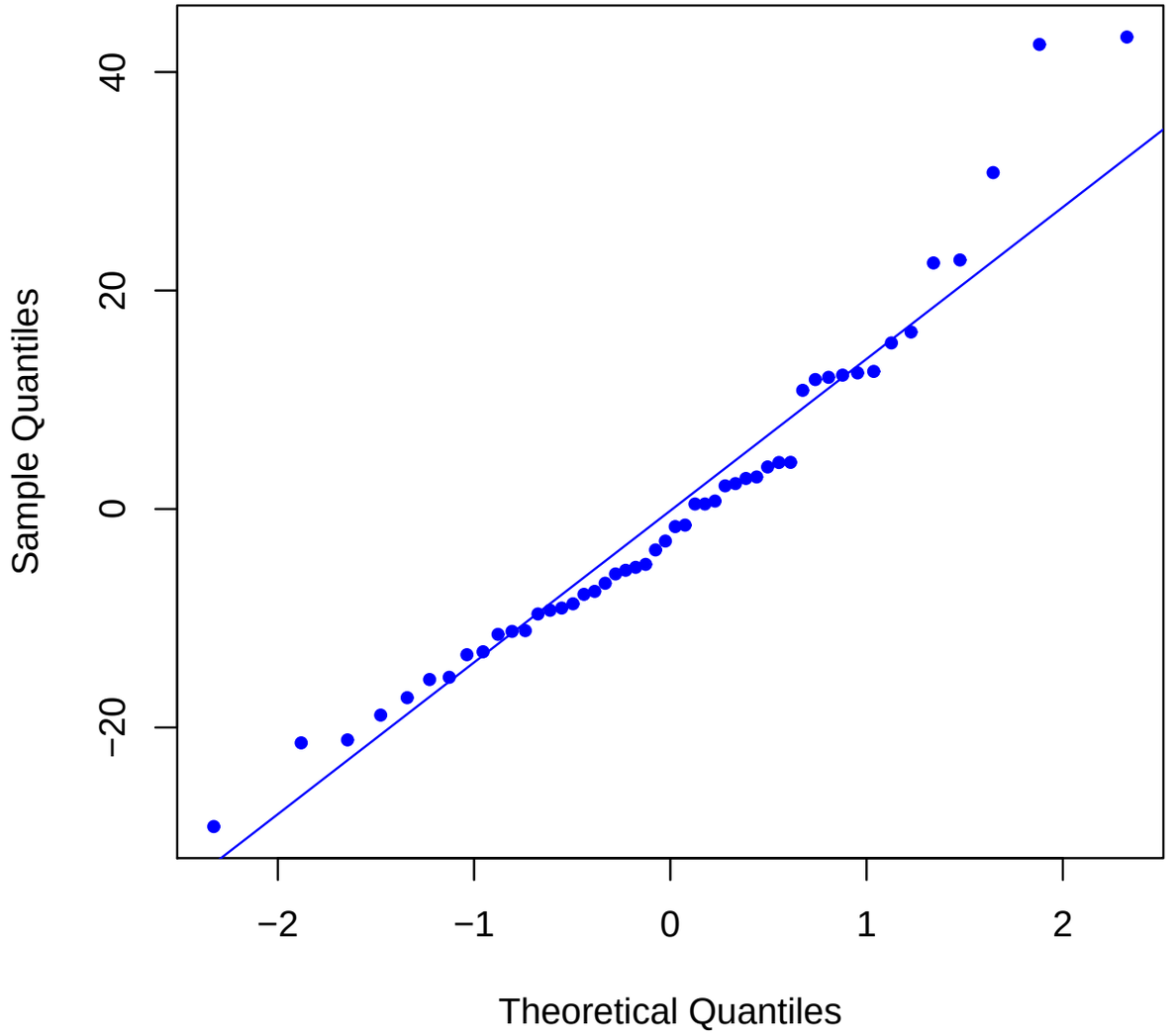


График остатков в задаче восстановления зависимости длины тормозного пути от начальной скорости.


Residuals
40
20
0
−20
0
10
20
30
40
50
Index



Гистограмма остатков в задаче восстановления зависимости вида $\text{dist} = \beta_0 + \beta_1 \times \text{speed}$ длины тормозного пути от начальной скорости.



QQ-график в задаче восстановления зависимости длины тормозного пути от начальной скорости.

Теоретическая проверка нормальности остатков:

- общие критерии согласия: χ^2, критерий Колмогорова–Смирнова и т. п.
- специализированные, например, *критерий Шапиро–Уилка*

Нуль-гипотеза в критерии Шапиро–Уилка заключается в том, что статистически наблюдаемая случайная величина имеет нормальное распределение.

В задаче с определением длины тормозного пути автомобиля статистика Шапиро–Уилка оказалось равной $W = 0.9451$.

Соответствующее p-valueравно 0.02153. При уровне значимости $\alpha = 0.01$ гипотезу о нормальности распределения остатков принимаем.

Критерий Уайта

Проверка на гомоскедастичность (постоянство дисперсии) остатков.

Гипотеза H_0: $\sigma_1 = \sigma_2 = \cdots = \sigma_N$

Гипотеза H_1: H_0 не выполнено

К исходной модели применяется метод наименьших квадратов и находятся остатки $e_1, e_2, \ldots, e_N$.

Затем строится модель, описывающая квадраты этих остатков через исходные переменные, их квадраты и попарные произведения:

$$e^2 = \gamma_0 + \sum_{j=1}^{d} \gamma_j x_j + \sum_{j \le j'} \gamma_{jj'} x_j x_{j'}.$$

При гипотезе H_0 величина nr^2 асимптотически имеет распределение $\chi^2(d' - 1)$, где

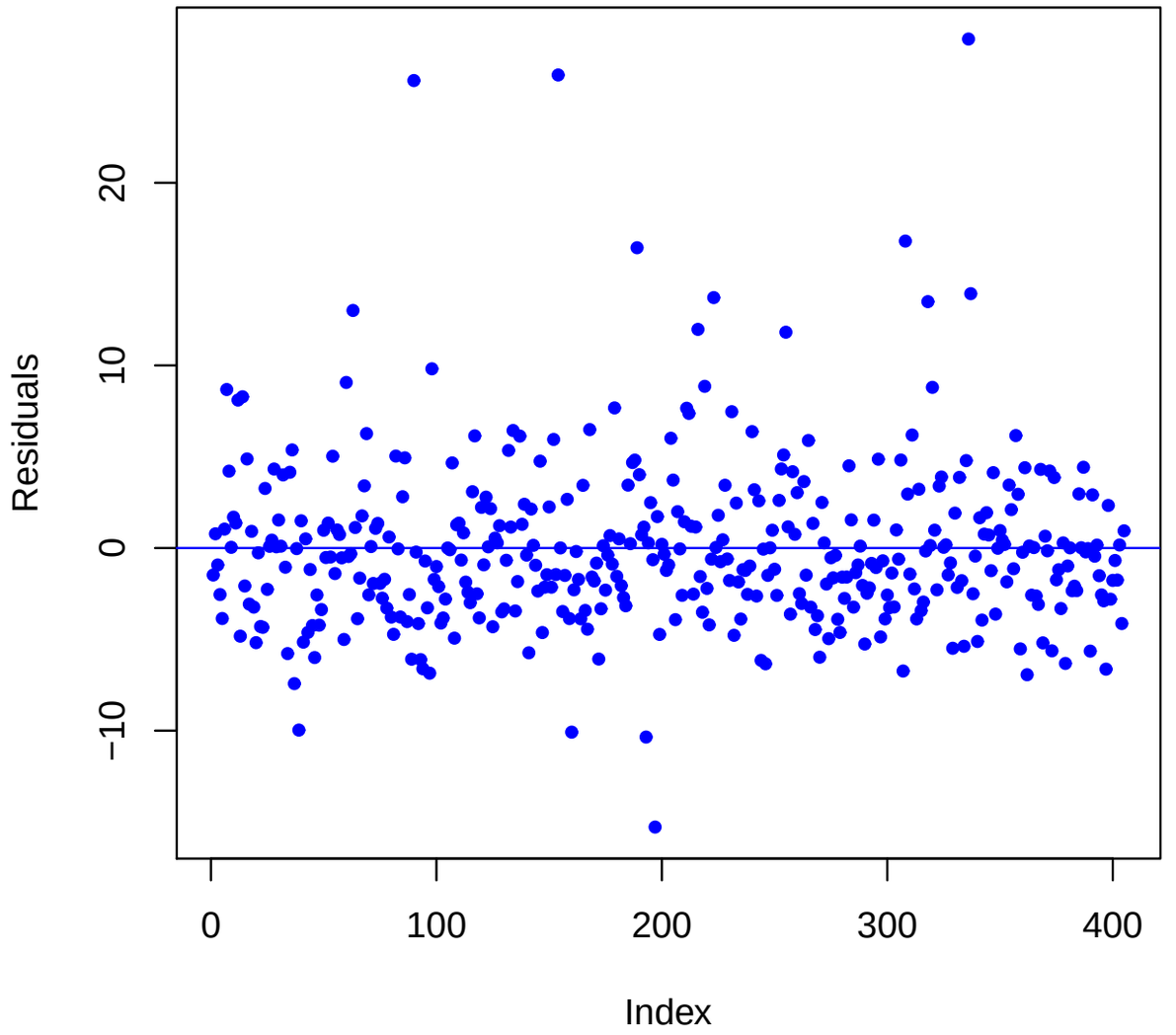
$d' = 1 + d + \dfrac{d(d+1)}{2}$ — число переменных новой модели,

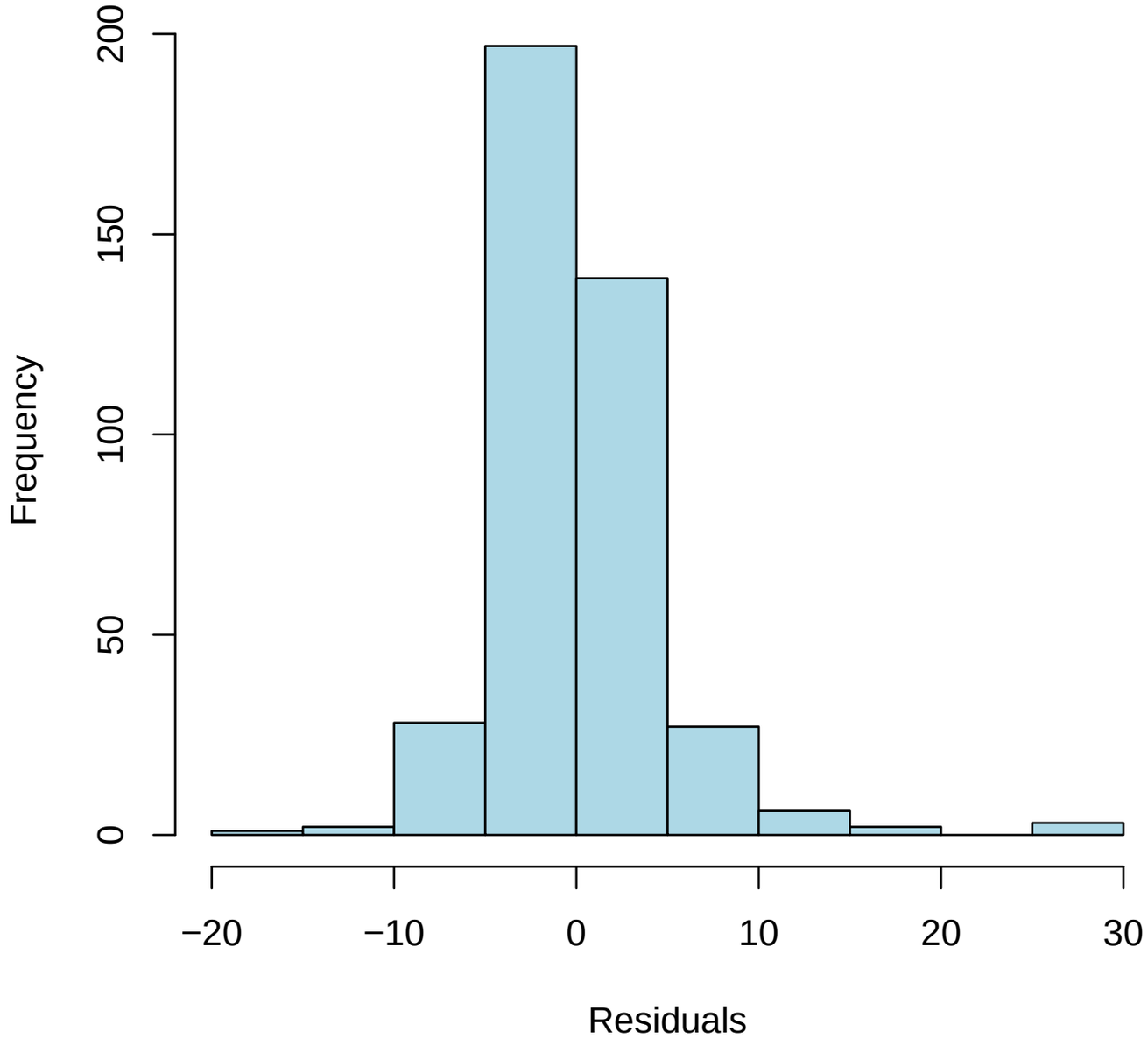
r^2 — коэффициент детерминации новой модели.

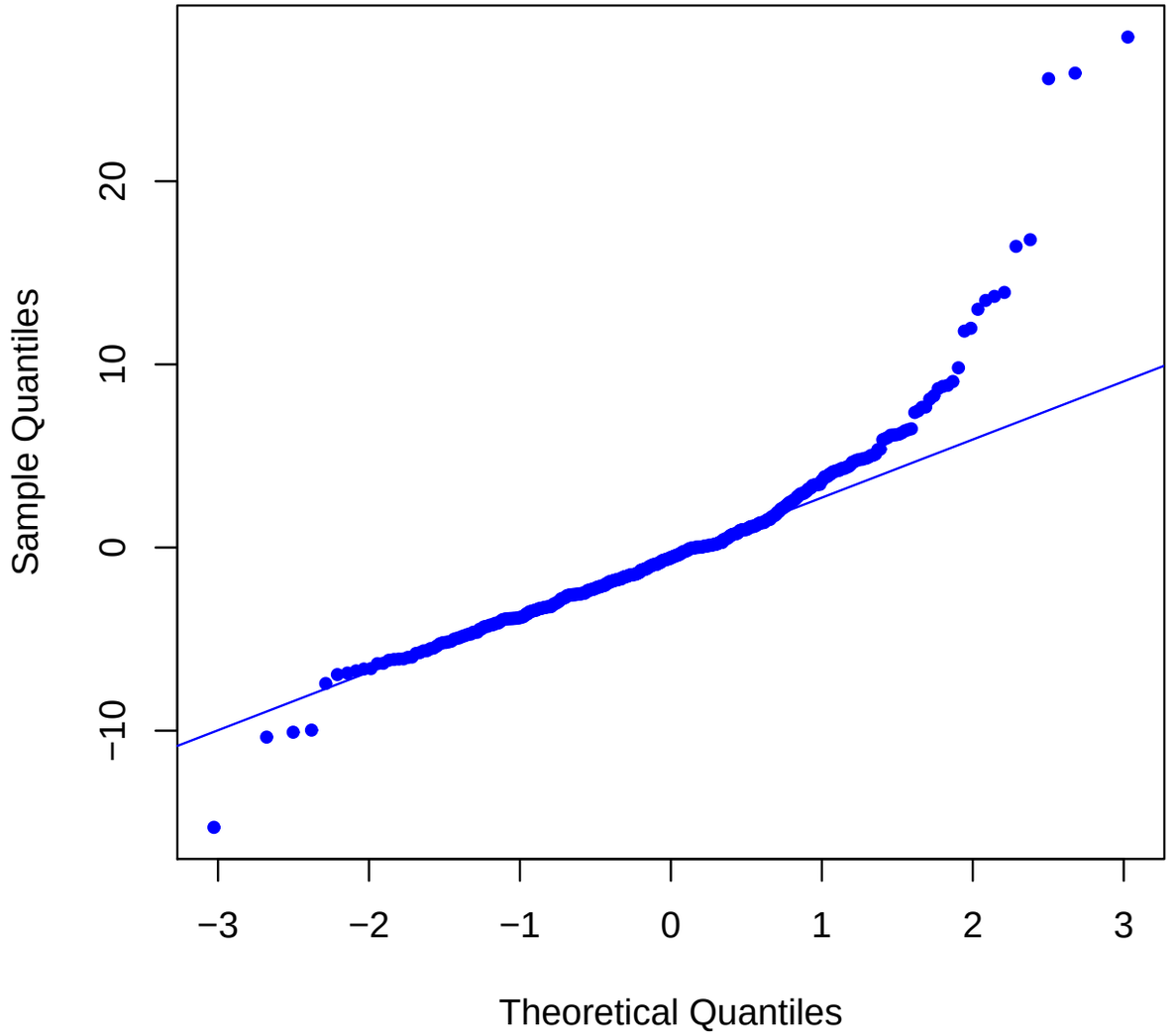
Анализ остатков для линейной модели dist $= \beta_0 + \beta_1 \times$ speed:

```
bptest(dist ~ speed, data=cars) # BP = 3.2149, df = 1, \pvalue = 0.07297
White = 3.2157, df = 2, \pvalue = 0.2003
```

Boston




Frequency
0
50
100
150
200
−20
−10
0
10
20
30
Residuals

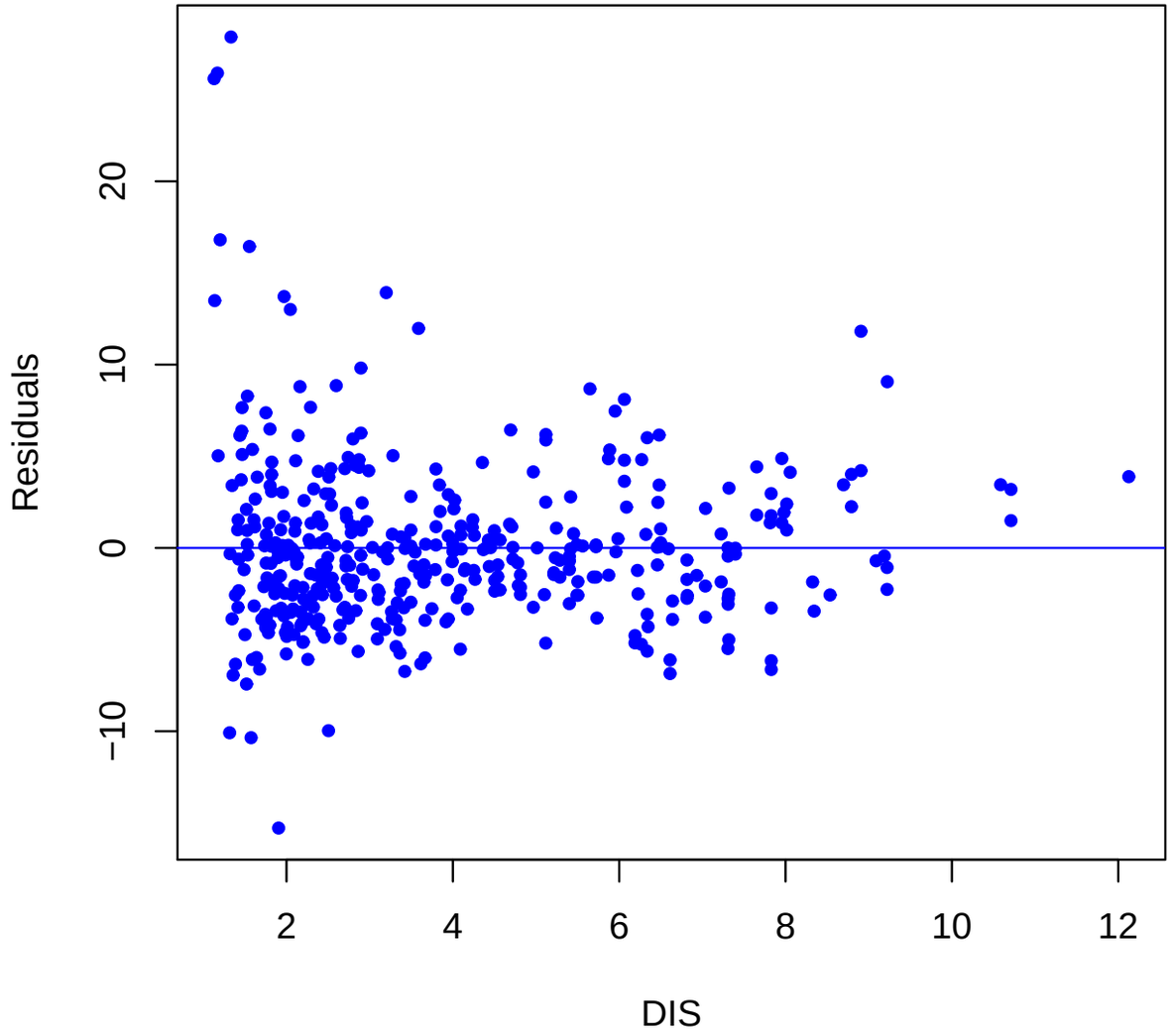


Критерий Шапиро-Уилка W = 0.8739, p-value< 2.2e-16 Отвергаем

Тест на гомоскедастичность

Тест Бройша–Пагана = 51.0623, df = 13, p-value= 1.958e-06

Критерий Уайта: White-test = 285.8917, df = 103, p-value< 2.2e-16

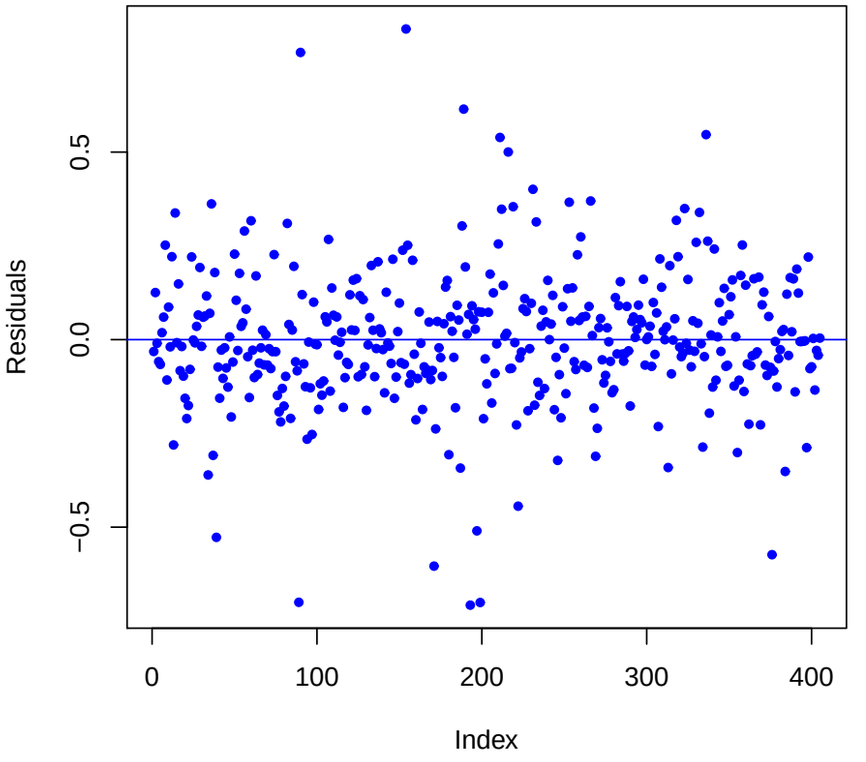


В оригинальной статье [Harrison and Rubinfeld, 1978] рассматривается модель

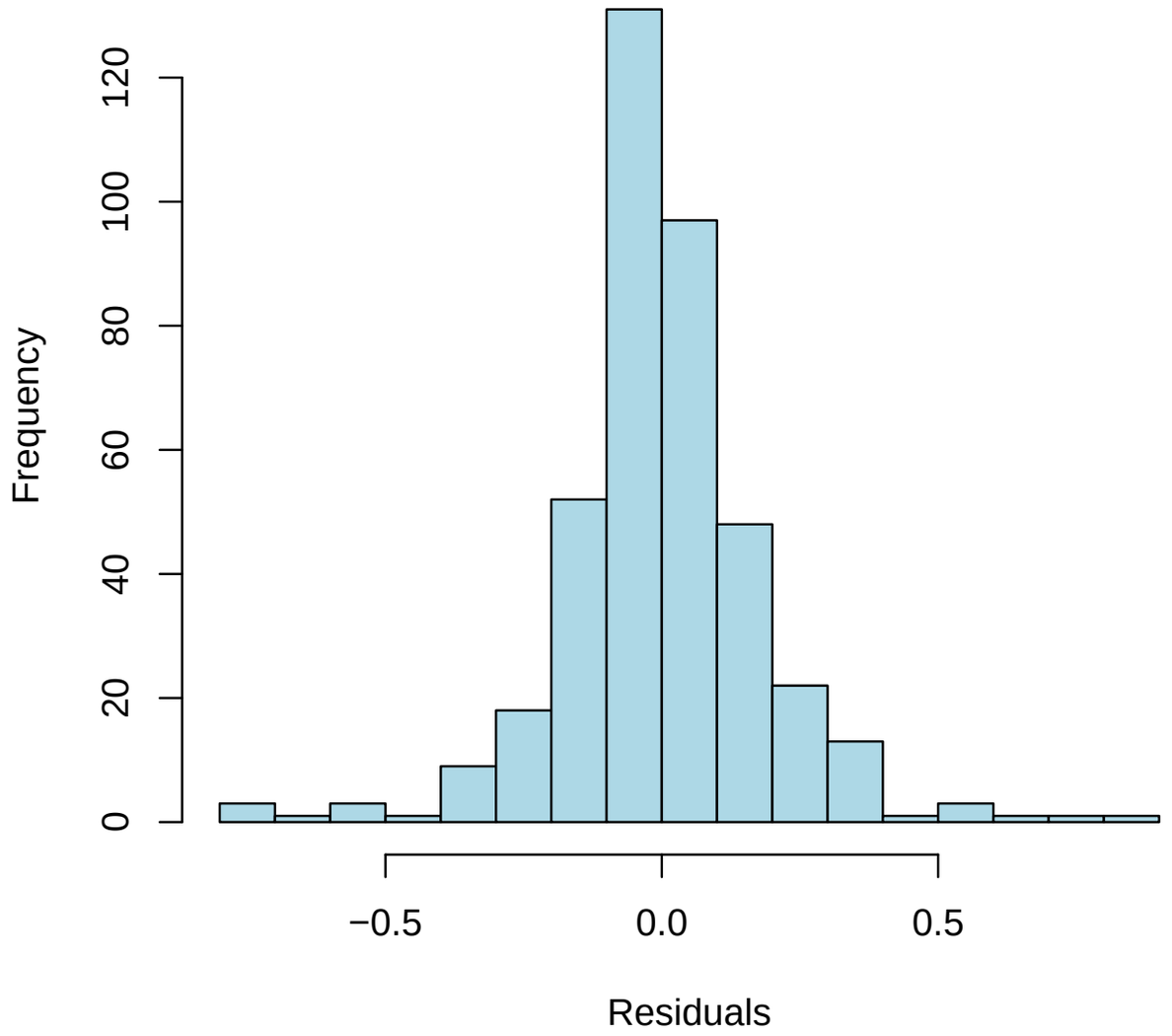
$$\ln \texttt{MEDV} = \beta_0 + \beta_1 \texttt{CRIM} + \beta_2 \texttt{ZN} + \beta_3 \texttt{INDUS} + \beta_4 \texttt{CHAS} + \beta_5 \texttt{NOX}^2 + \beta_6 \texttt{RM}^2 + \beta_7 \texttt{AGE} + \beta_8 \ln \texttt{DIS} +$$

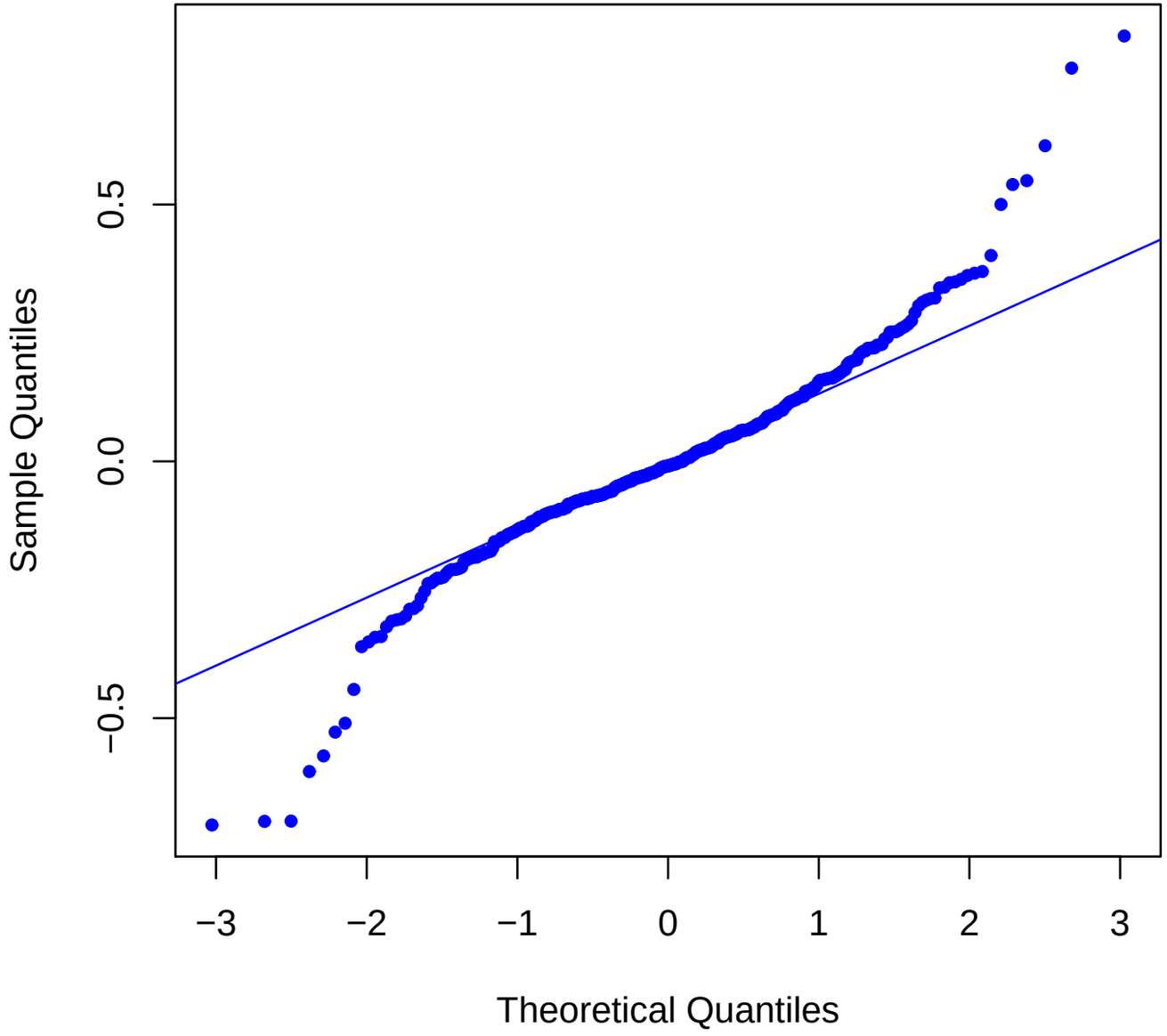
$$+ \beta_9 \ln \texttt{RAD} + \beta_{10} \texttt{TAX} + \beta_{11} \texttt{PTRAT} + \beta_{12} \texttt{B} + \beta_{13} \ln \texttt{LSTAT} + E.$$

$$\text{RSS}/Ntrain = 16.05832, \quad \text{RSS}/Ntest = 14.77086$$



Но также не проходит тест на нормальность. Shapiro-Wilk normality test W = 0.9439, p-value= 2.989e-11

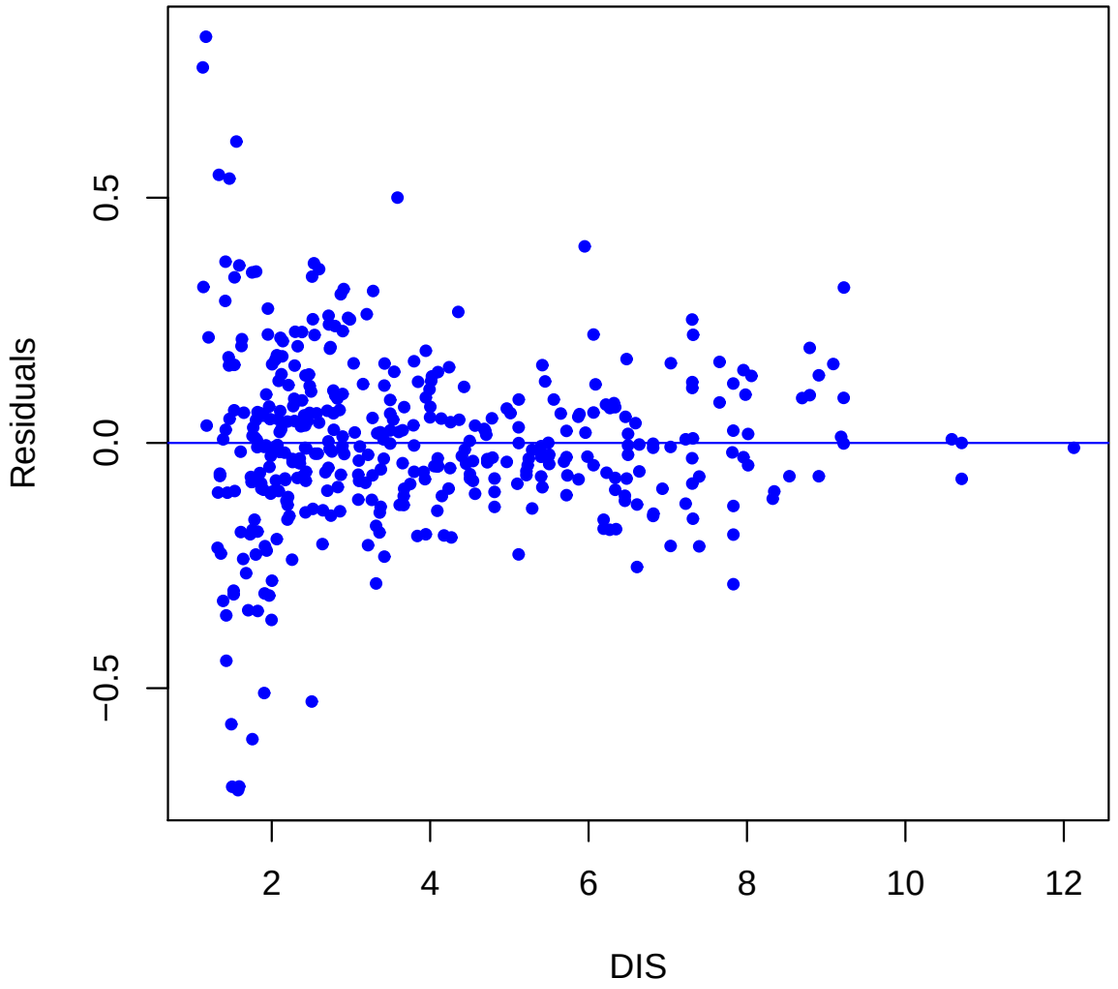



Sample Quantiles
0.5
0.0
−0.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Тест на гомоскедастичность

Тест Бройша–Пагана = 76.1122, df = 13, p-value= 5.905e-11

Критерий Уайта =



Для выявления корреляции между остатками (точнее: сериальной корреляции) часто используется *критерий Дарбина–Уотсона*. Статистика Дарбина–Уотсона вычисляется по формуле

$$D = \frac{\sum\limits_{i=1}^{N-1} \left(e^{(i+1)} - e^{(i)}\right)^2}{\sum\limits_{i=1}^{N} {e^{(i)}}^2}.$$

Упражнение 5.5 Доказать, что $0 \leq D \leq 4$.

Если $D < D_{\mathrm{L}}(\alpha)$ или $D > 4 - D_{\mathrm{L}}(\alpha)$, то с достоверностью α принимается гипотеза о наличии отрицательной или соответственно положительной корреляции остатков.

Если $D_{\mathrm{L}}(\alpha) < D < D_{\mathrm{U}}(\alpha)$ или $4 - D_{\mathrm{U}}(\alpha) < D < 4 - D_{\mathrm{L}}(\alpha)$, то критерий не позволяет принять решение о наличии или отсутствии корреляции остатков.

Если $D_{\mathrm{U}}(\alpha) < D < 4 - D_{\mathrm{U}}(\alpha)$, то гипотеза о наличии корреляции остатков отклоняется.

Таблицы критических значений $D_{\mathrm{L}}(\alpha)$ и $D_{\mathrm{U}}(\alpha)$ для различных α, N, d приведены, например, в [Дрейпер, Смит Т. 1, С. 211].

5.3. Подготовка данных

В линейной регрессионной модели в качестве признаков x_j могут выступать:

- исходные количественные признаки
- функции от исходных количественных признаков, например, x_1^2, $\ln x_2$
- функции от нескольких исходных признаков, например, $x_3 = x_1 \cdot x_2$
- бинарные признаки
- бинаризованные категориальные признаки

$$y = f(x) \equiv \sum_{j=1}^{s} \beta_j h_j(x)$$

Например, $$y = f(x) \equiv \beta_1 e^{\lambda_1 x} + \beta_2 e^{\lambda_2 x}.$$

Если λ_1, λ_2 известны, то получаем линейную регрессию

Если λ_1, λ_2 не известны, то получаем нелинейную регрессию

5.3.1. Бинаризация категориальных признаков

(one hot encoding)

Категориальный признак House, принимающий значения

$$\mathsf{Block}, \quad \mathsf{Brick}, \quad \mathsf{Monolithic}, \quad \mathsf{Panel}, \quad \mathsf{Wooden},$$

заменяем на 5 бинарных признаков:

$$x_{\mathsf{Block}}, \quad x_{\mathsf{Brick}}, \quad x_{\mathsf{Monolithic}}, \quad x_{\mathsf{Panel}}, \quad x_{\mathsf{Wooden}}$$

$$\mathsf{House} = \mathsf{Block} \quad \Leftrightarrow \quad x_{\mathsf{Block}} = 1,\ x_{\mathsf{Brick}} = 0,\ x_{\mathsf{Monolithic}} = 0,\ x_{\mathsf{Panel}} = 0,\ x_{\mathsf{Wooden}} = 0$$

$$\mathsf{House} = \mathsf{Brick} \quad \Leftrightarrow \quad x_{\mathsf{Block}} = 0,\ x_{\mathsf{Brick}} = 1,\ x_{\mathsf{Monolithic}} = 0,\ x_{\mathsf{Panel}} = 0,\ x_{\mathsf{Wooden}} = 0$$

и т. д.

Нужны только 4 бинарных признака!

Price	цена квартиры (тыс. руб.)
Date	№ дня, в который квартира выставлена на продажу
Lat	географическая широта объекта недвижимости
Lng	географическая долгота объекта недвижимости
Housing	тип недвижимости (0 — вторичное жилье, 1 — новостройка)
Floors	количество этажей в доме
House	тип дома
Rooms	количество комнат
Floor	№ этажа
Area	площадь квартиры (м^2)

$$y = \beta_0 + \sum_{j=1}^{12} \beta_j x_j$$

$$\beta_0 = -1.398 \times 10^5,\ \beta_1 = \beta_{\mathsf{Date}} = 3.634 \times 10^{-1},\ \beta_2 = \beta_{\mathsf{Lat}} = 7.540 \times 10^2,\ \ldots$$

$$\widehat{R}_{\text{train}} = \frac{1}{N_{\text{train}}} \text{RSS}_{\text{train}} = 495194.6, \qquad \widehat{R}_{\text{test}} = \frac{1}{N_{\text{test}}} \text{RSS}_{\text{test}} = 547395.1$$

5.3.2. Центрирование и нормализация данных

Центрирование:

- Центрирование $x^{(i)}$:
  каждое $x_j^{(i)}$ заменяем на $x_j^{(i)} - \overline{\mathbf{x}}_j$ $(i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, d)$, где

$$\overline{\mathbf{x}}_j = \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)}$$

Приведение к единичному стандартному отклонению:

- Приведение к нулевому среднему и единичному стандартному отклонению:
  каждое $x_j^{(i)}$ заменяем на $\dfrac{x_j^{(i)} - \overline{\mathbf{x}}_j}{s_j}$ $(i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, d)$, где

$$\overline{\mathbf{x}}_j = \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)}, \qquad s_j = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_j^{(i)} - \overline{\mathbf{x}}_j)^2}$$

Приведение к отрезку $[0, 1]$:

- каждое $x_j^{(i)}$ заменяем на

$$\frac{x_j^{(i)} - \min_i \overline{\mathbf{x}}_j}{\max_i \overline{\mathbf{x}}_j - \min_i \overline{\mathbf{x}}_j} \quad (i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, d),$$

где

$$\min_i \overline{\mathbf{x}}_j = \min\left\{x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(N)}\right\}, \quad \max_i \overline{\mathbf{x}}_j = \max\left\{x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(N)}\right\}$$

- Центрирование y:
  каждое $y^{(i)}$ заменяем на $y^{(i)} - \overline{\mathbf{y}}$ $(i = 1, 2, \ldots, N)$

$$\overline{\mathbf{y}} = \frac{1}{N} \sum_{i=1}^{N} y^{(i)}$$

Если центрирование y проведено, то далее используем МНК без β_0:

$$y = f(x) \equiv \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_d x_d,$$

а потом кладем $\beta_0 = \overline{\mathbf{y}}$