

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

КАФЕДРА МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ И АНАЛИЗА ДАННЫХ

Статистическое оценивание параметров линейной
регрессии с выбросами при наличии группирования
наблюдений

Курсовой проект

Румянцева Андрея Кирилловича
студента 4 курса, специальность
"прикладная математика"

Научный руководитель:
зав. кафедрой ММАД,
канд. физ.-мат. наук, доцент
Бодягин Игорь Александрович

Минск, 2018

Постановка задачи

- Изучить оценки наименьших квадратов
- Изучить М-Оценки
- Провести численные эксперименты по моделированию регрессионных данных с замещающими выбросами и построению изученных оценок
- Построить алгоритм робастного оценивания параметров линейной регрессии с выбросами при наличии группирования наблюдений .

1. Модель линейной регрессии

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} + \varepsilon_i, i = \overline{1, N},$$

$$y_i = f(x_i, \beta) + \varepsilon_i,$$

$$f(x_i, \beta) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in}$$

$$y_i = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_n \end{pmatrix} \times \begin{pmatrix} 1 \\ x_{i1} \\ \dots \\ x_{in} \end{pmatrix}^T + \varepsilon_i,$$

Модель линейной регрессии с аномальными наблюдениями


$$y_i^{\tilde{\varepsilon}} = (\xi_i)y_i + (1 - \xi_i)\eta_i$$

$$\begin{cases} p(\xi_i = 0) = \tilde{\varepsilon}, \\ p(\xi_i = 1) = 1 - \tilde{\varepsilon}. \end{cases}$$

Построение функции правдоподобия

$$y_i = f(x_i, \beta) + \varepsilon_i \sim \mathcal{N}(f(x_i, \beta), \sigma^2)$$

$$\mathcal{R} = (-\infty, a_1] \cup (a_1, a_2] \cup \dots \cup (a_{k-1}, +\infty)$$

$$\nu_0, \dots, \nu_{k-1}$$


$\mu_i = j$, если y_i отнесли к полуинтервалу ν_j .

3. Построение оценки параметров регрессии с помощью группирования выборки

Построение функции правдоподобия

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad \Phi(x) = \frac{1}{\sqrt{2}\sigma} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

функция распределения
стандартного нормального
закона

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \quad \Phi(x) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \right]$$

$$F(x) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2}\sigma}\right) \right]$$

Построение функции правдоподобия

$$P\{y_i \in \nu_j\} = F_{y_i}(a_{j+1}) - F_{y_i}(a_j) = \begin{cases} \frac{1}{2}(\operatorname{erf}(\frac{a_{j+1}-f(x_i,\beta)}{\sqrt{2}\sigma}) - \operatorname{erf}(\frac{a_j-f(x_i,\beta)}{\sqrt{2}\sigma})), & j = \overline{1, k-2} \\ \frac{1}{2}(1 + \operatorname{erf}(\frac{a_1-f(x_i,\beta)}{\sqrt{2}\sigma})), & j = 0 \\ \frac{1}{2}(1 + \operatorname{erf}(\frac{a_{k-1}-f(x_i,\beta)}{\sqrt{2}\sigma})), & j = k-1 \end{cases}$$

$$P(\mu_i = j) = P(y_i \in \nu_{\mu_i}).$$

Функция правдоподобия

$$\begin{aligned} l(\beta, \sigma^2, \nu_0, \dots, \nu_{k-1}) &= \ln\left(\prod_{i=1}^n P(\mu_i = j)\right) = \\ &= \sum_{i=1}^n \ln(P(\mu_i = j)). \end{aligned}$$

Производная функции правдоподобия

$$\frac{\delta l}{\delta \beta} = \frac{\delta \sum_{i=1}^n \ln(P(\mu_i = j))}{\delta \beta} = \frac{\delta \sum_{i=1}^n \ln P(y_i \in \nu_{\mu_i})}{\delta \beta} = \quad (30)$$

$$= \frac{\delta \sum_{i=1}^n \ln\left(\frac{1}{2}\left(\operatorname{erf}\left(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma}\right)\right)\right)}{\delta \beta} = \quad (31)$$

$$= \sum_{i=1}^n \left((1 - (\delta_{\mu_i 0} + \delta_{\mu_i k-1})) \frac{(\operatorname{erf}'(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma}) - \operatorname{erf}'(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma}))}{(\operatorname{erf}(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma}) - \operatorname{erf}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma}))} + \right. \quad (32)$$

$$\left. + (\delta_{\mu_i 0} + \delta_{\mu_i k-1}) \frac{\operatorname{erf}'(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma})}{(1 + \operatorname{erf}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma}))} \right) (-1) \frac{\delta f(x_i, \beta)}{\delta \beta} =$$

$$= - \sum_{i=1}^n \begin{pmatrix} 1 \\ x_{i1} \\ \dots \\ x_{in} \end{pmatrix} \times \left((1 - (\delta_{\mu_i 0} + \delta_{\mu_i k-1})) \frac{(\operatorname{erf}'(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma}) - \operatorname{erf}'(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma}))}{(\operatorname{erf}(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma}) - \operatorname{erf}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma}))} + \right. \quad (33)$$

$$\left. + (\delta_{\mu_i 0} + \delta_{\mu_i k-1}) \frac{\operatorname{erf}'(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma})}{(1 + \operatorname{erf}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma}))} \right),$$

Приближение функции erf

$$(\operatorname{erf} x)^2 \approx 1 - \exp\left(-x^2 \frac{\frac{4}{\pi} + ax^2}{1 + ax^2}\right),$$
$$a = \frac{8}{3\pi} \frac{3 - \pi}{\pi - 4}.$$

$$\operatorname{erf}'(x) = \exp\left(-x^2 \frac{\frac{4}{\pi} + ax^2}{1 + ax^2}\right) \frac{-2x \frac{\frac{4}{\pi} + ax^2}{1 + ax^2} + (2ax^3) \frac{\frac{4}{\pi} + ax^2}{1 + ax^2} - \frac{2ax^3}{1 + ax^2}}{2\sqrt{1 - \exp\left(-x^2 \frac{\frac{4}{\pi} + ax^2}{1 + ax^2}\right)}}.$$

Метод секущих

$$\beta^{(k+1)} = \beta^{(k)} + \Delta\beta^{(k)}.$$

$$\frac{\delta}{\delta\beta} \frac{\delta l(\beta^{(k)})}{\delta\beta} \Delta\beta^{(k)} = - \frac{\delta l(\beta^{(k)})}{\delta\beta}.$$

$$\frac{\delta}{\delta\beta_j} \frac{\delta l(\beta_1^{(k)}, \dots, \beta_n^{(k)})}{\delta\beta} \approx \frac{\frac{\delta l(\beta_1^{(k)}, \dots, \beta_j^{(k)}, \dots, \beta_n^{(k)})}{\delta\beta}(\beta^{(k)}) - \frac{\delta l(\beta_1^{(k)}, \dots, \beta_j^{(k-1)}, \dots, \beta_n^{(k)})}{\delta\beta}(\beta^{(k)})}{\beta_j^{(k)} - \beta_j^{(k-1)}}.$$

Переклассификация

$$\check{\mu}_i = \arg \max_j \sum_{k \in V_i, k \neq i} \delta_{\check{\mu}_k j} , \quad (32)$$

где V_i множество индексов l первых K векторов x_l , отсортированных по возрастанию расстояния до вектора x_i .

Компьютерные эксперименты

Моделирование функции регрессии

Параметры программы	
Переменная	значение
Размер выборки N	1000
Доля выбросов $\tilde{\varepsilon}$	0.1
Параметры регрессии β	(100, 4)
Регрессоры x_i	$\sim U(-5, 5)$
ε_i	$\sim N(0, 16)$
η_i	$\sim N(100, 100)$

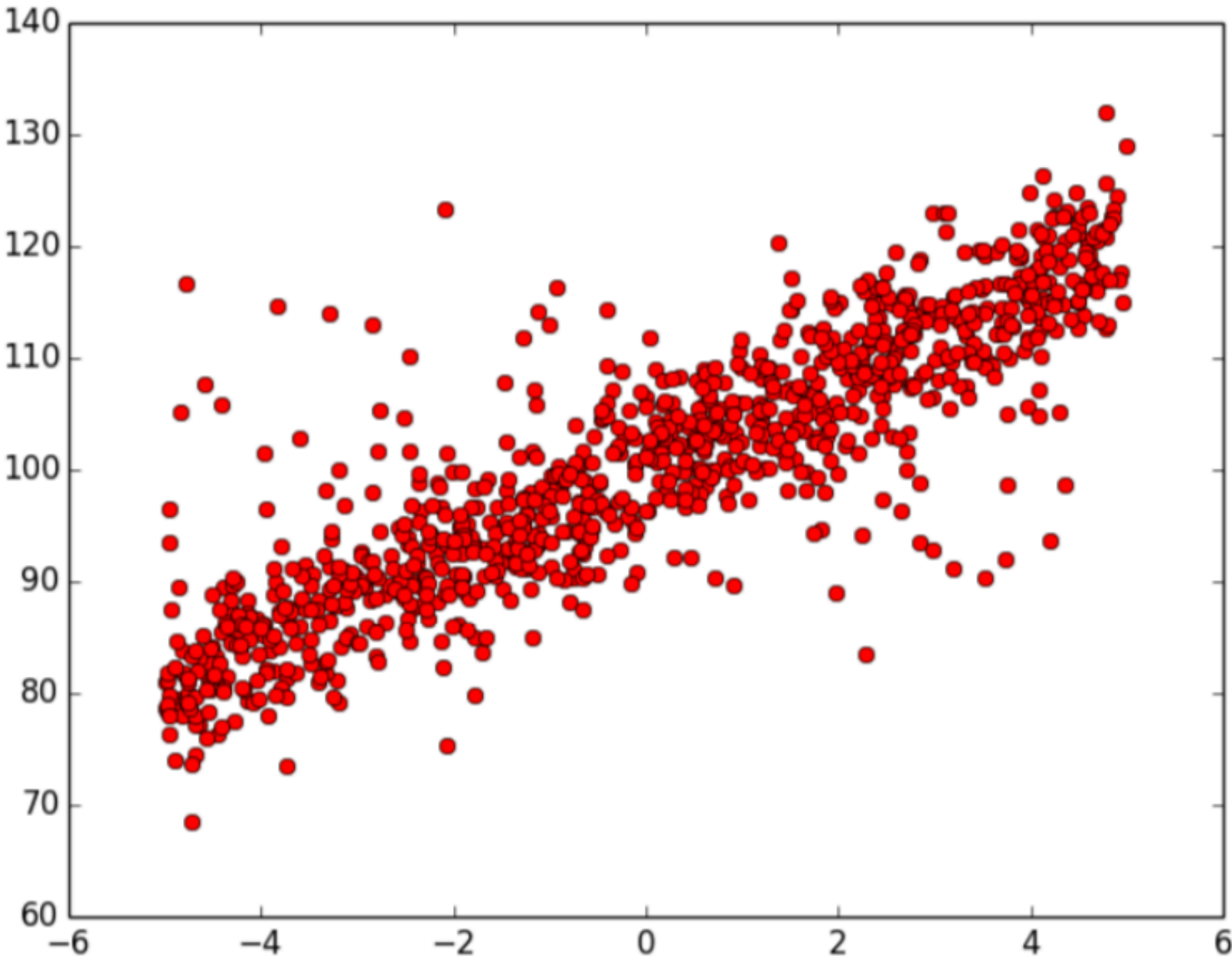


Рис. 1: Вывод графика рассеяния (y_i, x_i)

Эксперименты с оценками

Параметры программы	
Переменная	значение
Размер выборки N	100
Доля выбросов $\tilde{\varepsilon}$	0.8
Параметры регрессии β	(90, 4)
Регрессоры x_i	$\sim U(-5, 5)$
ε_i	$\sim N(0, 16)$
η_i	$\sim N(100, 100)$
Величина K из пункта 2.3	10

Эксперименты с оценками

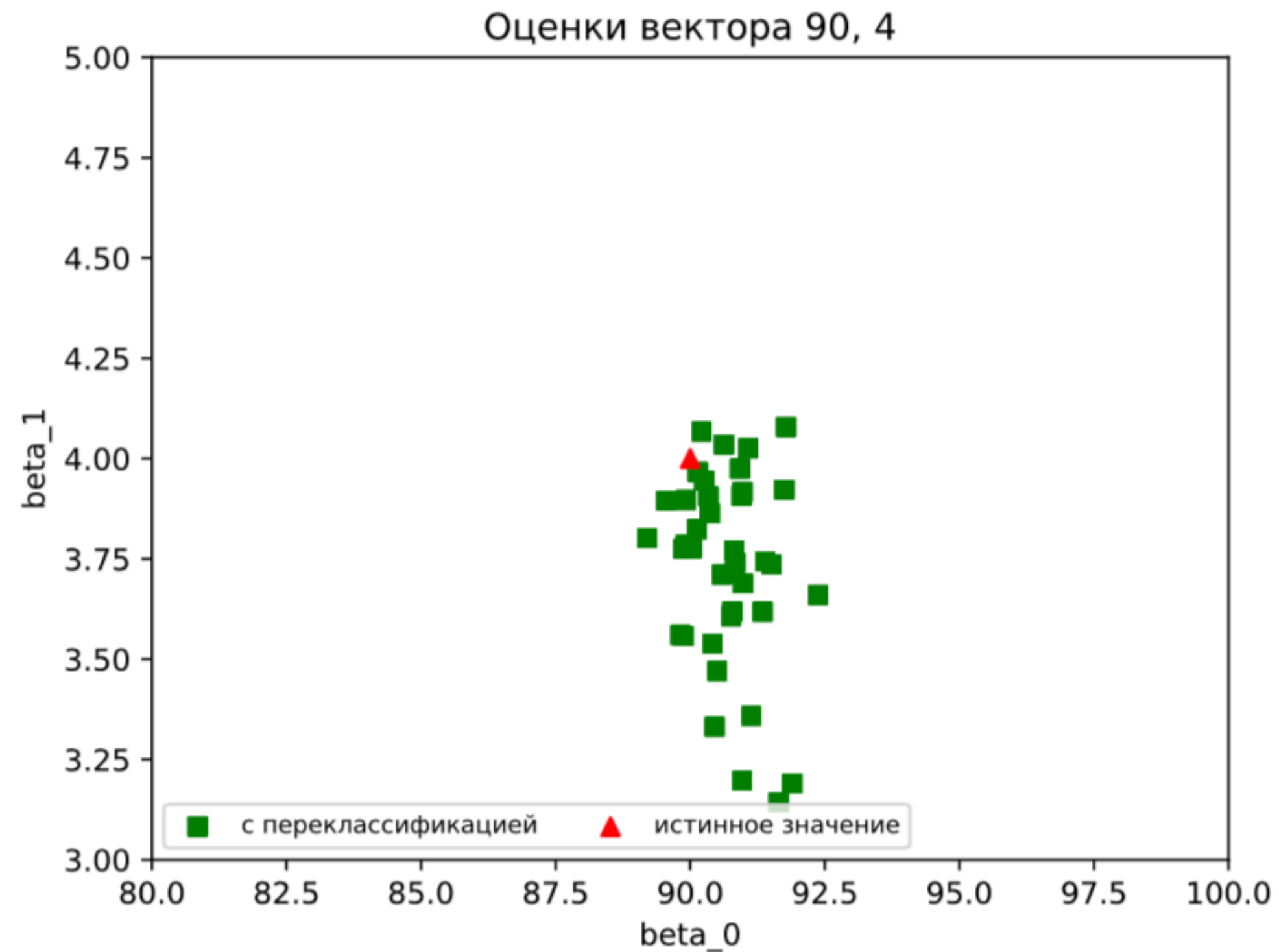


Рис. 2: Вывод графика рассеяния $(\hat{\beta}_0, \hat{\beta}_1)$

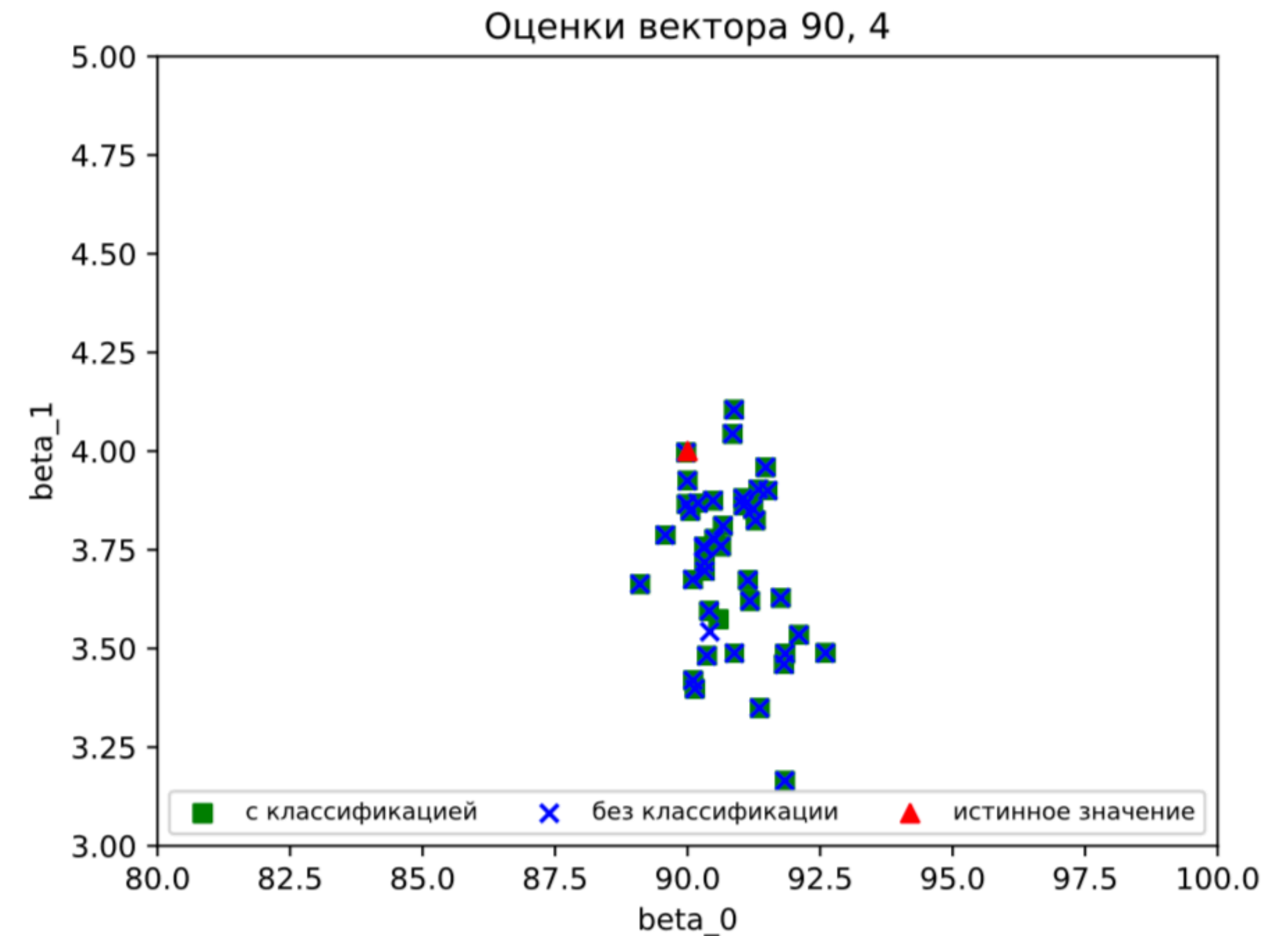


Рис. 3: Вывод графика рассеяния $(\hat{\beta}_0, \hat{\beta}_1)$

Эксперименты с оценками

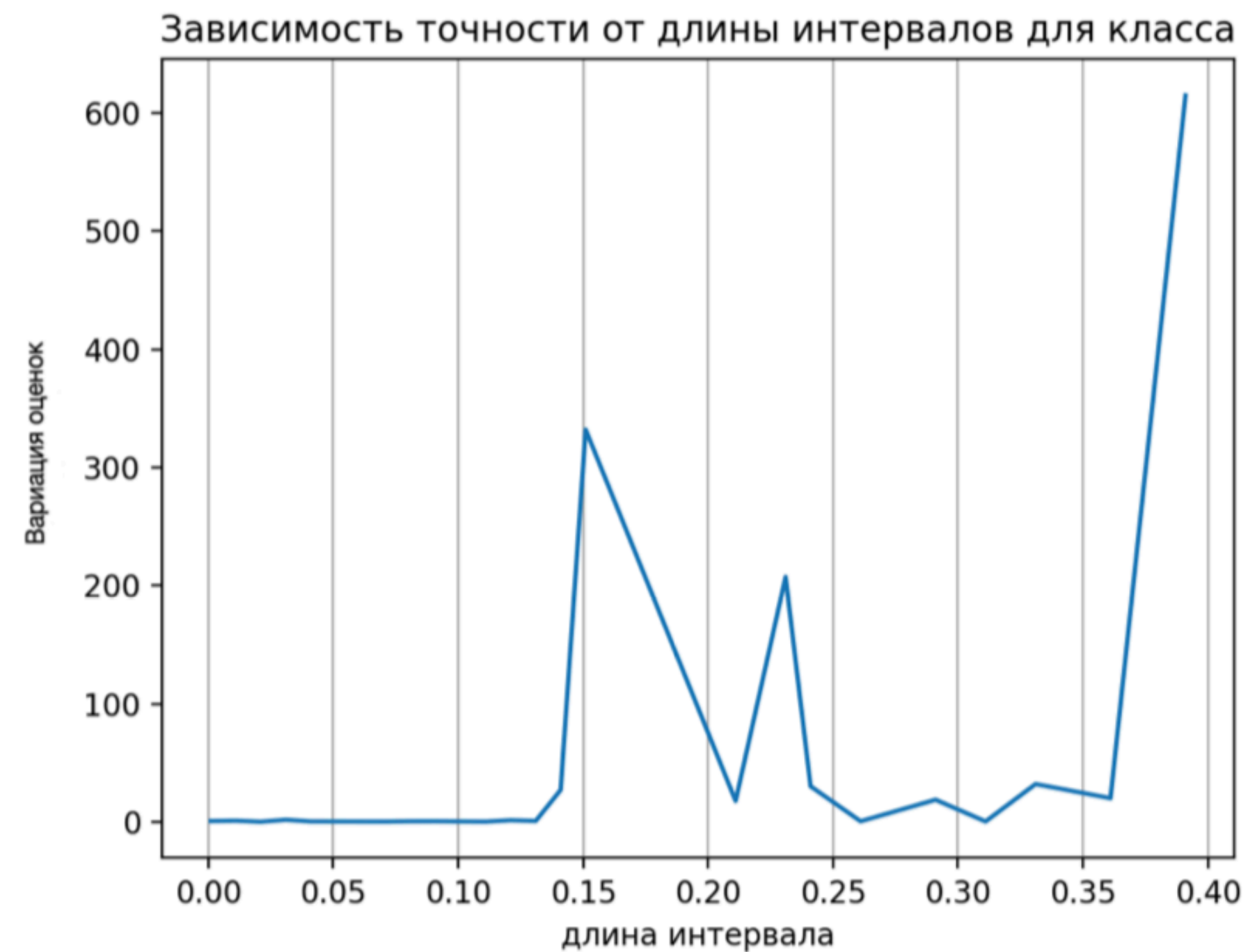


Рис. 4: Зависимость точности от длины интервала



Рис. 5: Зависимость точности от размера выборки

Эксперименты с оценками

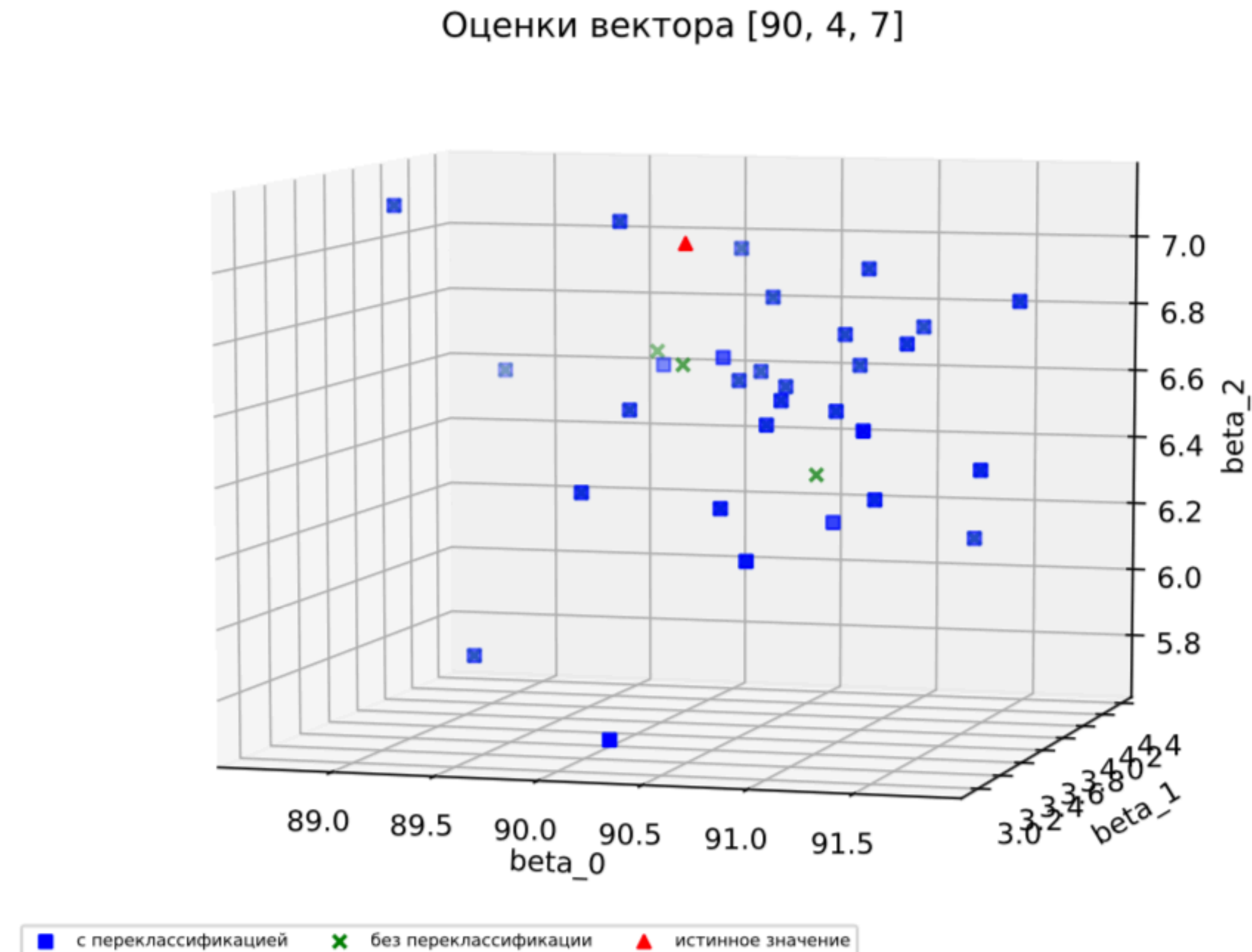


Рис. 6: Вывод графика рассеяния $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$

Заключение

В ходе выполнения курсового проекта были получены следующие результаты:

- рассмотрена математическая модель линейной регрессии с выбросами при наличии группирования наблюдений;
- описаны основные методы оценивания параметров линейной регрессии при наличии выбросов: оценки МНК, М-оценки;
- построены оценки параметров линейной регрессии при наличии группирования наблюдений по методу максимального правдоподобия;
- проведены компьютерные эксперименты в которых построенные оценки применялись к модельным данным;
- результаты экспериментов показали, что построенные оценки могут быть состоятельными.

Спасибо за внимание.