МИНЕСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

КАФЕДРА МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ И АНАЛИЗА ДАННЫХ

Румянцев
Андрей Кириллович

"Робастные оценки параметров регрессии при наличии группирования выборки"

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

Минск, 2018

Постановка задачи

1. Модель линейной регрессии

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

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

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

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

Модель линейной регрессии с аномальными наблюдениями

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

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

2. Breakdown point

$$\widetilde{\delta_1^{\widetilde{\varepsilon}_i}} = \frac{1}{20}\sum_{k=1}^{20}(\sum_{i=0}^{n}(\beta_i - \hat{\beta}_{ki}^{(N_1)})^2)^{\frac{1}{2}},$$

$$\widetilde{\delta_2^{\widetilde{\varepsilon}_i}} = \frac{1}{20}\sum_{k=1}^{20}(\sum_{i=0}^{n}(\beta_i - \hat{\beta}_{ki}^{(N_2)})^2)^{\frac{1}{2}};$$

$$br = \left\{ \widetilde{\varepsilon}_i, \text{если } \widetilde{\delta_1^{\widetilde{\varepsilon}_i}} < \widetilde{\delta_2^{\widetilde{\varepsilon}_i}}; \right.$$

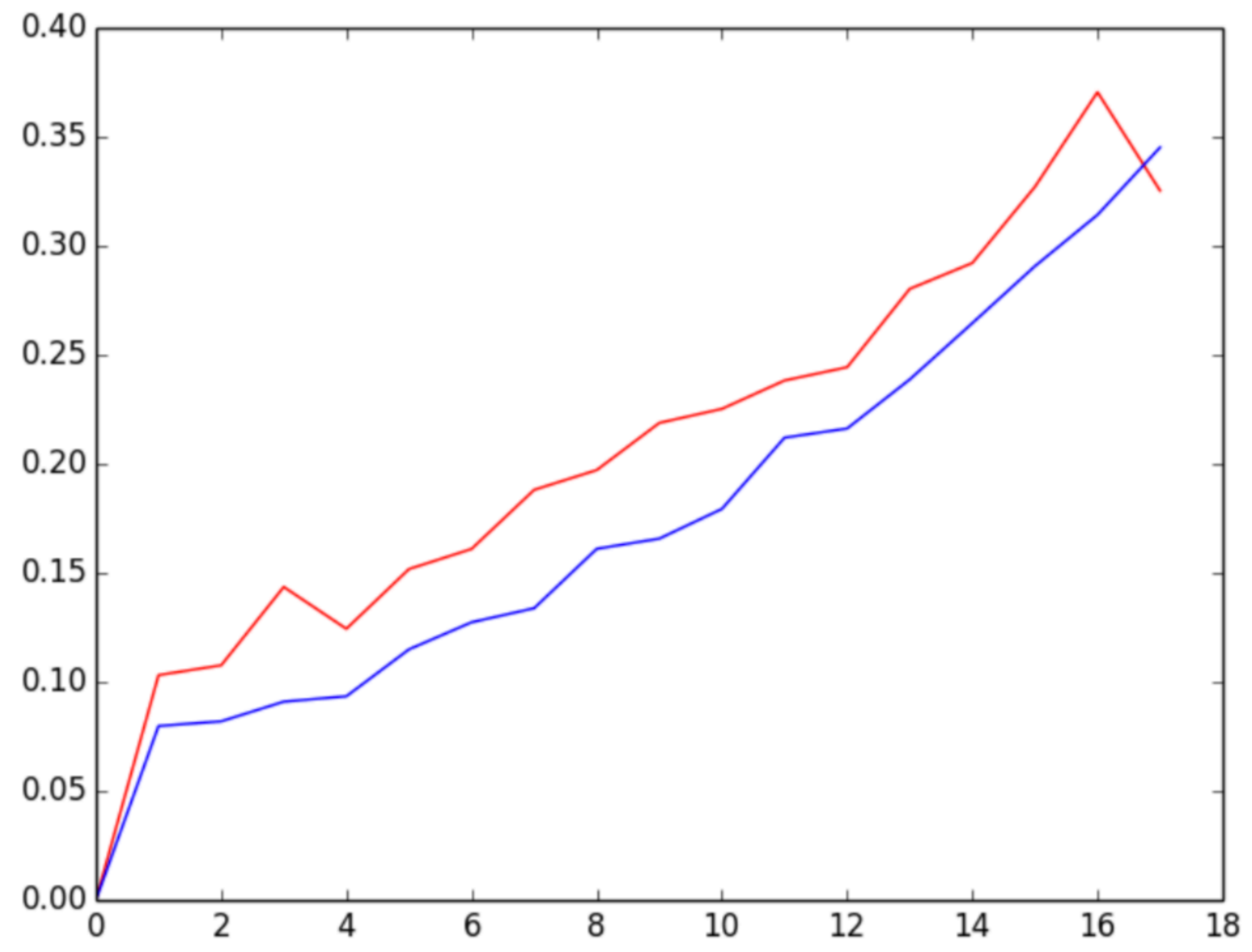

Рис. 3: График, на котором изображены $\widetilde{\delta}_1^{\widetilde{\varepsilon}_i}$ красным и $\widetilde{\delta}_2^{\widetilde{\varepsilon}_i}$синим относительно $\widetilde{\varepsilon}_i$ в случае М-оценок

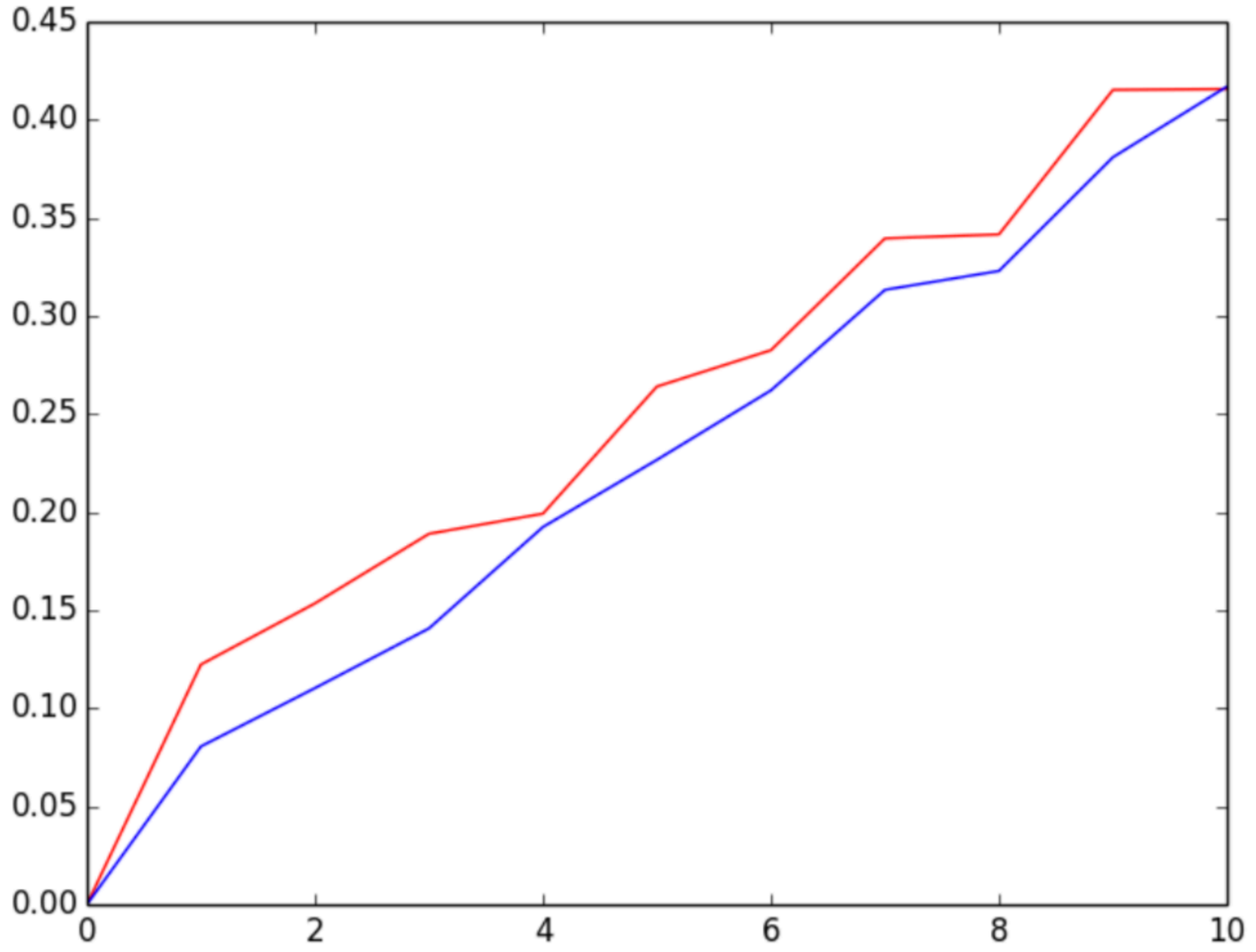

Рис. 2: График, на котором изображены $\widetilde{\delta}_1^{\widetilde{\varepsilon}_i}$ красным и $\widetilde{\delta}_2^{\widetilde{\varepsilon}_i}$синим относительно $\widetilde{\varepsilon}_i$ в случае МНК

3. Построение оценки параметров регрессии с помощью группирования выборки

Построение функции правдоподобия

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

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

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

Построение функции правдоподобия

$$F(x) = \Phi(\frac{x - \mu}{\sigma}),$$

$$\Phi(x) = \frac{1}{\sqrt{2}\sigma} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt.$$

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt.$$

$$\Phi(x) = \frac{1}{2}\Big[1 + \operatorname{erf}\Big(\frac{x}{\sqrt{2}}\Big)\Big]$$

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

Построение функции правдоподобия

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

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

Функция правдоподобия

$$l(\beta, \sigma^2, \nu_0, \ldots, \nu_{k-1}) = \ln(\prod_{i=1}^{n} P(\mu_i = j)) =$$

$$= \sum_{i=1}^{n} \ln(P(\mu_i = j)).$$

Производная функции правдоподобия

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

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

$$= \sum_{i=1}^{n} \Bigg((1 - (\delta_{\mu_i 0} + \delta_{\mu_i k-1})) \frac{(\text{erf'}(\frac{a_{\mu_i+1} - f(x_i,\beta)}{\sqrt{2}\sigma}) - \text{erf'}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))}{(\text{erf}(\frac{a_{\mu_i+1} - f(x_i,\beta)}{\sqrt{2}\sigma}) - \text{erf}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))} + \tag{32}$$

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

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

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

Приближение функции erf

$$(\mathrm{erf} x)^2 \approx 1 - \exp(-x^2 \frac{\frac{4}{\pi} + ax^2}{1 + ax^2}),$$

$$a = \frac{8}{3\pi} \frac{3 - \pi}{\pi - 4}.$$

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

Метод секущих

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

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

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

Переклассификация

$$\check{\mu}_i = \arg\max_j \sum_{|x_k - x_i| \le \Delta,\ k \ne i} \delta_{\check{\mu}_k j},$$

Поведение производной функции на точном решении при разных долях аномальных наблюдений

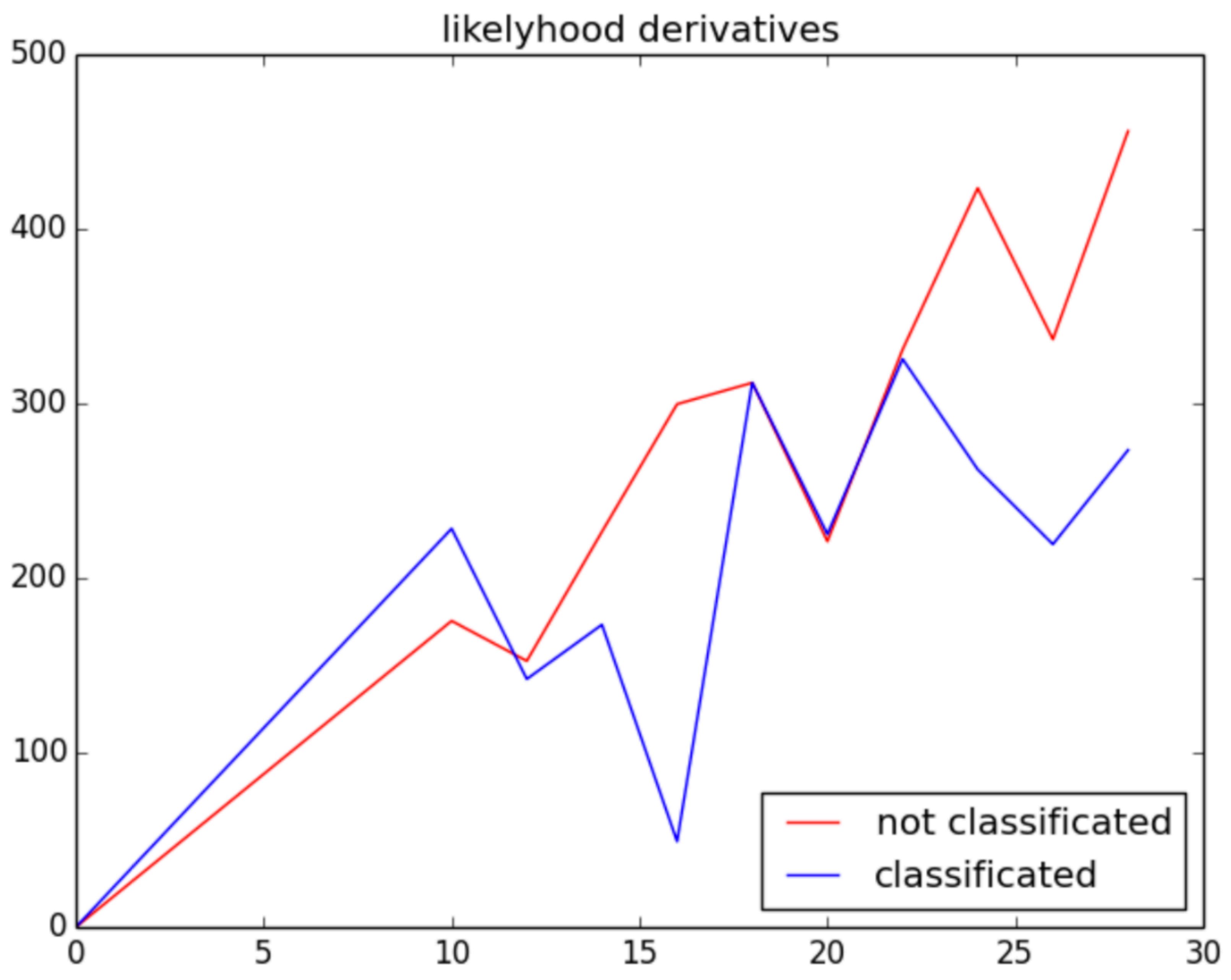

Рис. 4: При объеме $N = 500$

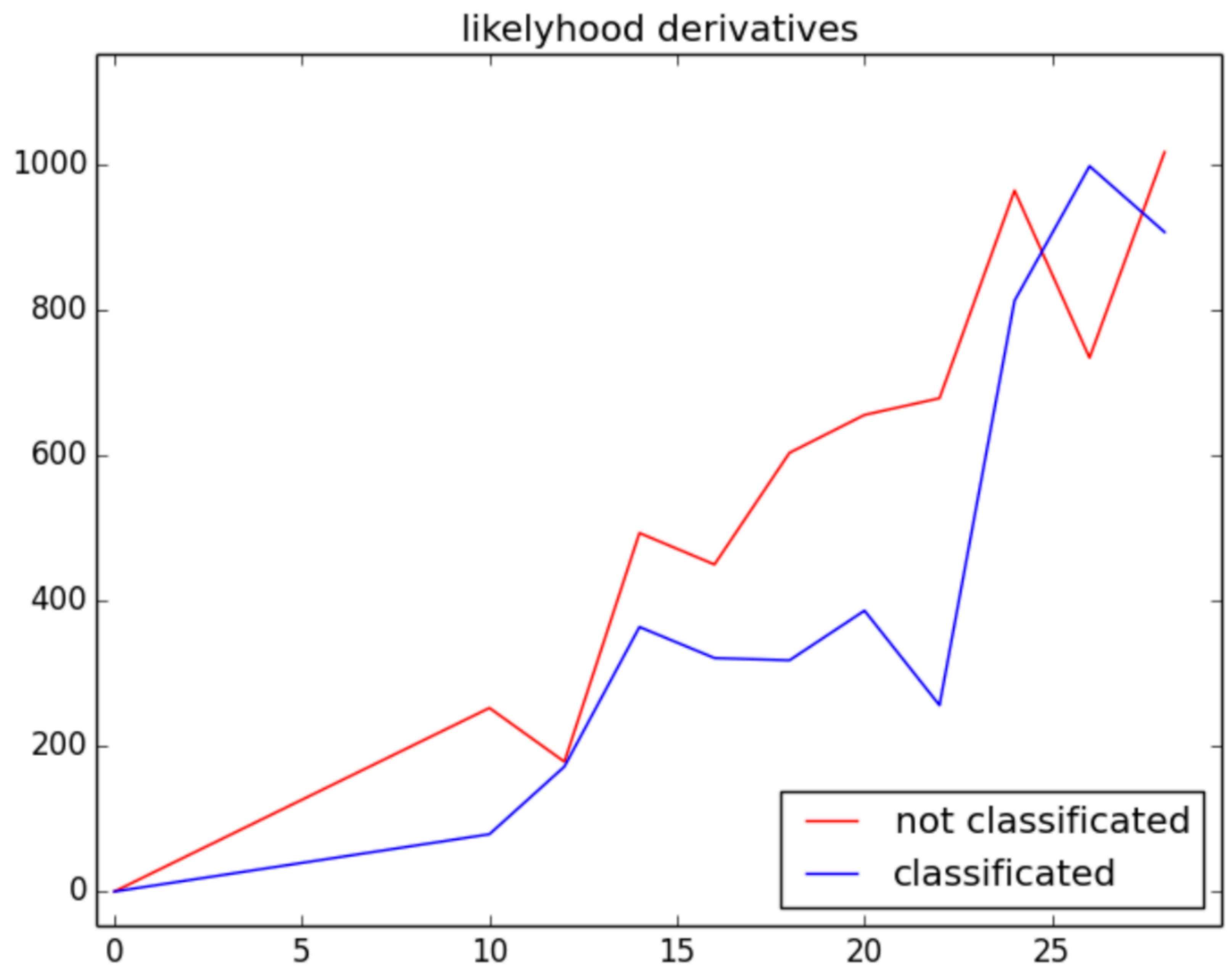

Рис. 5: При объеме $N = 1000$

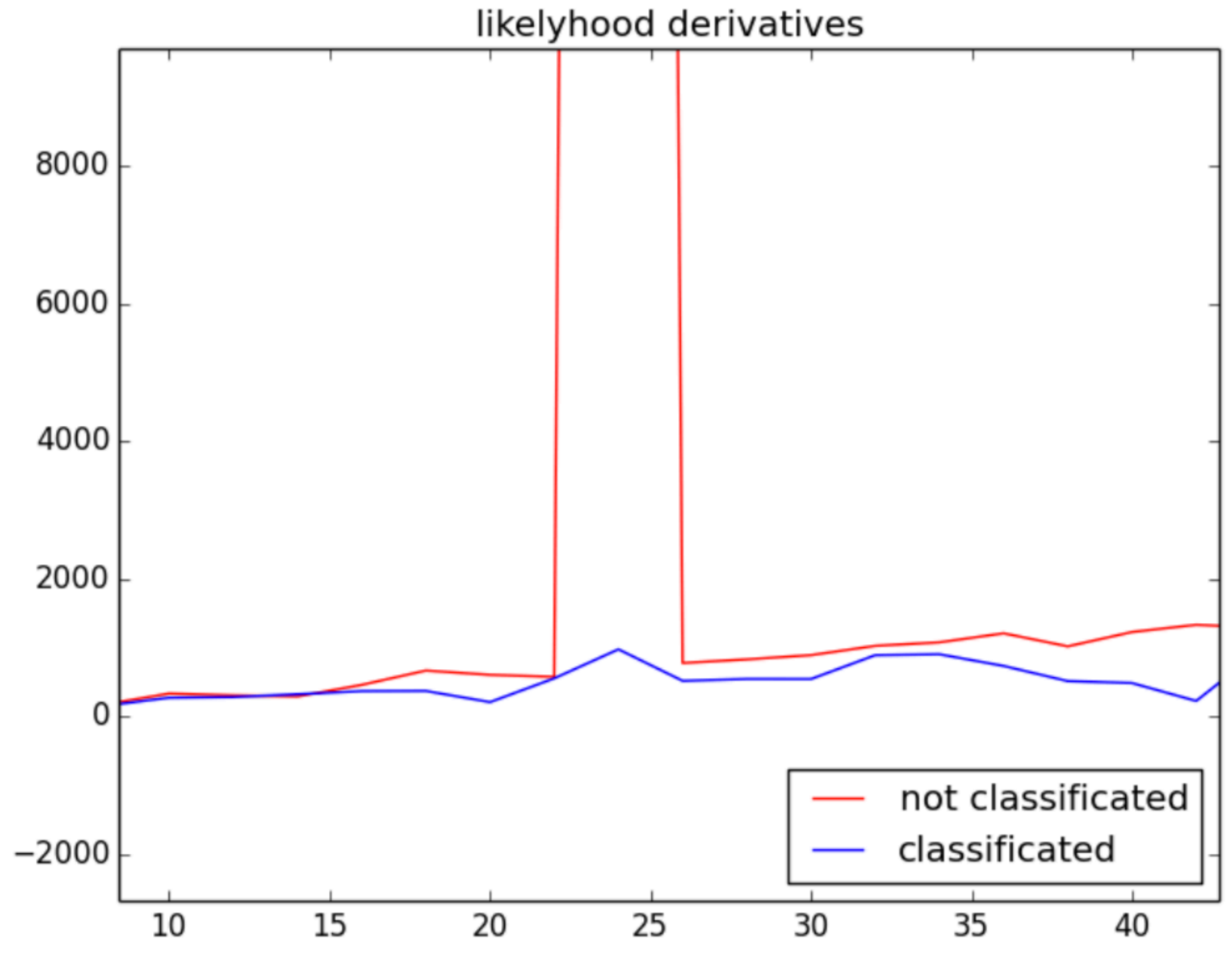

Рис. 6: При объеме $N = 1000$

Значение производной функции на точном решении при разных доли аномальных наблюдений, равной 0.08

$$\frac{\delta l(\hat{\beta})}{\delta \beta} = [6777.80925977 \ 11935.60045093]^T, \ \text{при} \ \hat{\beta} = [1, 1]^T,$$

$$\frac{\delta l(\hat{\beta})}{\delta \beta} = [-49.05706716 \ 283.92412386]^T, \ \text{при} \ \hat{\beta} = [90, 4]^T,$$

$$\frac{\delta l(\hat{\beta})}{\delta \beta} = [-3129.555067 \ -11908.91415502]^T, \ \text{при} \ \hat{\beta} = [170, 8]^T.$$

Заключение

I. Были описаны некоторые методы робастного оценивания параметров регрессии.
II. В курсовой работе был предложен еще один способ оценки “меры робастности” методов: “breakdown point”, который был далее использован на описанных методах оценивания — М-оценках и МНК.
III. Был предложен еще один способ оценивания параметров регрессии для модели регрессии с аномальными наблюдениями при наличии группирования выборки.
IV. С проведением вычислительного эксперимента было исследовано поведение производной функции правдопобия на точном решении для метода, описанного в пункте 3.

Спасибо за внимание