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"

Курсовая работа

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

Минск, 2018

Содержание

1 Введение

Существует несколько подходов для оценки параметров регрессии, но далеко не все устойчивы к возникновениям аномальных наблюдений. В реальной жизни аномальные наблюдения возникают постоянно, поэтому большинство методов просто неприменимо.
В прошлом веке в работах Хьюбера была заложена теория робастного оценивания.
Были предложены следующие робастные оценки[1]:

- M-Оценки
- R-Оценки
- L-Оценки

M-оценки – некоторое подобие оценок максимального правдоподобия (ММП-оценки - частный случай), L-оценки строятся на основе линейных комбинаций порядковых статистик, R-оценки – на основе ранговых статистик. В данном курсовом работе я буду моделировать функцию регрессии с аномальными наблюдениями, анализировать точность методов и находить для разных методов так называемый "breakdown point"– процент аномальных наблюдений, при котором увеличение количества наблюдений не повысит точность методов.
Также будет предложен новый способ оценивания параметров регрессии, где используется группирование выборки.

2 Модель функции регрессии с аномальными наблюдениями и оценки ее параметров

Введем модель линейной регрессию:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i, i = \overline{1, N} \tag{1}$$
$$y_i = f(x_i, \beta) + \varepsilon_i,$$
$$f(x_i, \beta) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}$$

Или, в векторной форме:

$$y_i = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_n \end{pmatrix} \times \begin{pmatrix} 1 \\ x_{i1} \\ \dots \\ x_{in} \end{pmatrix}^T + \varepsilon_i, \tag{2}$$

где y_i – i-е наблюдение из N наблюдений(N-объем выборки), $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$ регрессоры, $\{\beta_k, k = \overline{0, n}\}$– параметры регрессии, а ε_i – случайная ошибка i-го эксперемента, распределение которой подчиняется нормальному закону с нулевым математическим ожиданием и дисперсией σ^2.
В нашей задаче считаем параметры $\{\beta_k, k = \overline{0, n}\}$ неизвестными, их нам и требуется найти.
Но мы будем рассматривать не линейную регрессию, заданную формулами (1)-(2), а линейную регрессию с аномальными наблюдениями вида:

$$y_i^{\widetilde{\varepsilon}} = (\xi_i) y_i + (1 - \xi_i) \eta_i, \tag{3}$$

где ξ_i принимает значение, равное 1, с вероятностью $1 - \widetilde{\varepsilon}$ и значение, равное 0, с вероятностью $\widetilde{\varepsilon}$, т.е.:

$$\begin{cases} p(\xi_i = 0) = \widetilde{\varepsilon}, \\ p(\xi_i = 1) = 1 - \widetilde{\varepsilon}. \end{cases}, \tag{4}$$

которая называется функцией линейной регрессии с выбросами. η_i-случайная величина из некоторого вообще говоря неизвестного распределения. Переменную $\widetilde{\varepsilon}$ будем называть процентом аномальных наблюдений. Величины ξ_i, x_i и η_i являются независимыми
Теперь рассмотрим некоторые методы оценки параметров регрессии:

2.1 Метод наименьших квадратов

Предлоположим, что случайные ошибки подчиняются нормальному закону распределения вероятностей:

$$L\{\varepsilon_i\} = N_1(0, \sigma^2), i = \overline{1, n}. \tag{5}$$

Строим логарифмическую функцию правдоподобия. В силу (1) и (2) имеем:

$$L\{y_i\} = N_1(f(x_i; \beta), \sigma^2). \tag{6}$$

Логарифмическая функция правдоподобия выглядит так[2]:

$$l(\beta) = \ln \prod_{i=1}^{n} (\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - f(x_i;\beta))^2}{2\sigma^2}}) = -\frac{1}{2} n \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} R^2(\beta), \tag{7}$$

$$R^2(\beta) = \sum_{i=1}^{n} (\delta y_i)^2 = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 \geq 0. \tag{8}$$

Тогда оценка максимального правдоподобия из формул (4)-(5) такова:

$$\hat{\beta} = arg \min_{\beta} R^2(\beta). \tag{9}$$

2.2 М-оценки

Швейцарский статистик П.Хьюбер предложил использовать М-оценки [2], которые являются решениями экстремальных задач вида:

$$\sum_{i=1}^{n} \phi(x_t; \beta) \to \min_{\beta}, \tag{10}$$

где $\phi(\cdot;\beta)$-некоторая функция, определяющая конкретный тип оценок и их точность.
Очевидно, что $\phi(\cdot;\beta) \equiv -\ln p(\cdot;\beta)$ дает обычную оценку максимального правдоподобия, построенная по модели без выбросов (1).
Рассмотрим теперь некоторые способы выбора $\phi(\cdot;\beta)$.

2.2.1 Способы выбора функции для решения экстремальной задачи в М-оценках

Для начала определим:

$$u_i = y_i^{\tilde{\varepsilon}} - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}). \tag{11}$$

Тогда существует такие методы[3]:

Способы выбора $\phi(\cdot;\beta)$	
Метод	Целевая функция
Метод Наименьших Квадратов	$\phi(\cdot;\beta)_{OLS} = u^2$
Хьюбера	$\phi(\cdot;\beta)_H = \begin{cases} \frac{1}{2}u^2, \lvert u \rvert \leq k, \\ k\lvert u \rvert - \frac{1}{2}k^2, \lvert u \rvert > k \end{cases}$
Биквадратный	$\phi(\cdot;\beta)_B = \begin{cases} \frac{k^2}{6}(1 - [1 - (\frac{u}{k})^2]^3), \lvert u \rvert \leq k \\ \frac{k^2}{6}, \lvert u \rvert > k \end{cases}$

3 Моделирование функции регрессии с аномальными наблюдениями

Для начала смоделируем функцию регрессии по методу (3). Для удобства моделируем регрессию с одномерными регрессорами $x_i, i = \overline{1, N}$.
Воспользуемся такими параметрами:

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

,

$U(-5, 5)$ - равномерное распределение на отрезке $[-5, 5]$.
Получаем такой график:

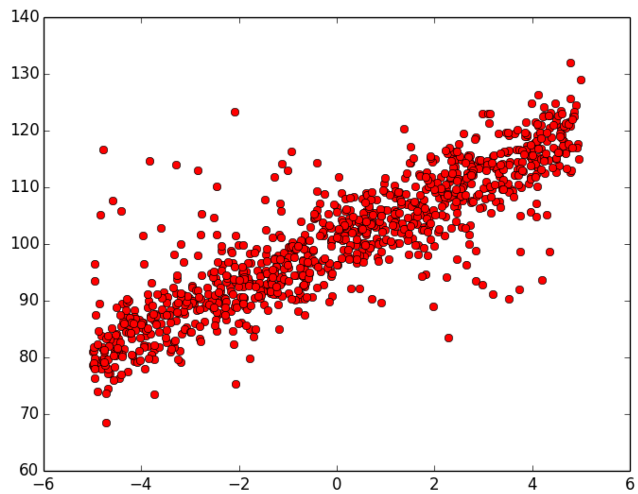

Рис. 1: Вывод графика рассеяния (y_i, x_i)

4 Поиск breakdown point у МНК и М-оценок

Будем пользоваться той же моделью, как и в пункте 3. Для поиска того процента аномальных наблюдений, при котором увеличение количества элементов выборки не повышает точности метода будем делать так:

- Организуем цикл по процентам загрязнений $\widetilde{\varepsilon}_i$ от $\widetilde{\varepsilon}_0 = 0$ до $\widetilde{\varepsilon}_{100} = 100$, увеличивая каждый раз $\widetilde{\varepsilon}_i$ на 1;

- На каждой итерации будем 20 раз моделировать выборку с $N_1 = 1000$ и $N_2 = 3000$ наблюдений. На каждой такой итерации суммируем невязку с точными значениями параметров для каждого количества элементов, а потом находим среднее, поделив на количество суммирования, т.е. посчитаем усредненную невязку:

$$\widetilde{\delta_1^{\widetilde{\varepsilon}_i}} = \frac{1}{20}\sum_{k=1}^{20}(\sum_{i=0}^{n}(\beta_i - \hat{\beta}_{ki}^{(N_1)})^2)^{\frac{1}{2}}, \tag{12}$$

$$\widetilde{\delta_2^{\widetilde{\varepsilon}_i}} = \frac{1}{20}\sum_{k=1}^{20}(\sum_{i=0}^{n}(\beta_i - \hat{\beta}_{ki}^{(N_2)})^2)^{\frac{1}{2}}; \tag{13}$$

- если полученная усредненная невязка при 1000 наблюдений меньше либо равна невязке при 3000 наблюдений, то заканчиваем цикл - нашли breakdown point, т.е.:

$$br = \begin{cases}\widetilde{\varepsilon}_i, \text{если } \widetilde{\delta_1^{\widetilde{\varepsilon}_i}} < \widetilde{\delta_2^{\widetilde{\varepsilon}_i}};\end{cases} \tag{14}$$

- иначе повышаем процент на 1 и повторяем цикл: $\widetilde{\varepsilon}_{i+1} = \widetilde{\varepsilon}_i + 0.01$

Такие тесты проведем для МНК и М-оценок.

4.1 Результаты программы

Найденные breakdown point для МНК и М-оценок	
Метод	breakpoint
МНК	10%
М-оценка с функцией Хьюбера	17%

Итак, видим, что М-оценки значительно устойчивее к выбросам чем МНК, т.е.при увеличении доли аномальных наблюдений, с увеличением объема выборки, оценки не сильно изменяются.

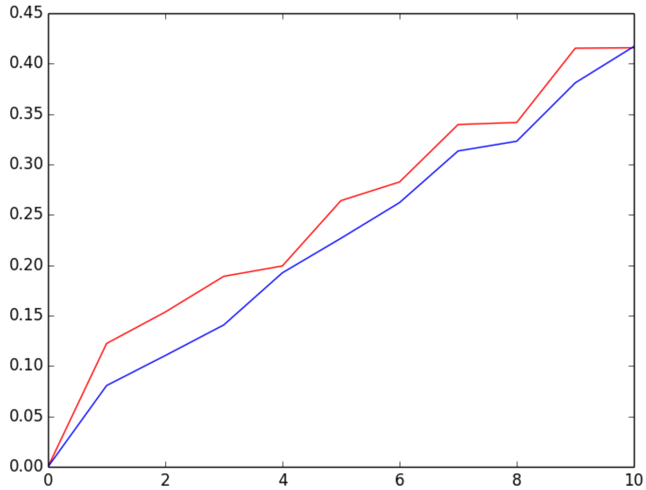

Рис. 2: График, на котором изображены $\widetilde{\delta_1^{\widetilde{\varepsilon}_i}}$ красным и $\widetilde{\delta_2^{\widetilde{\varepsilon}_i}}$синим относительно $\widetilde{\varepsilon}_i$ в случае МНК

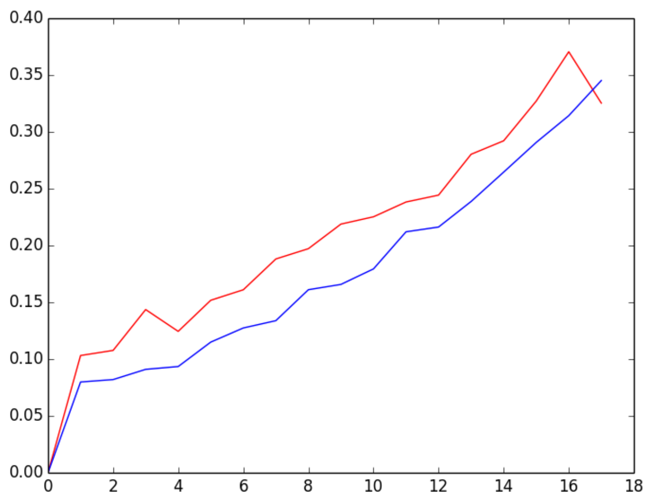

Рис. 3: График, на котором изображены $\widetilde{\delta_1^{\widetilde{\varepsilon}_i}}$ красным и $\widetilde{\delta_2^{\widetilde{\varepsilon}_i}}$синим относительно $\widetilde{\varepsilon}_i$ в случае М-оценок

Замечания:

- Можно моделировать не 20 раз , а значительно больше, тем самым уменьшая зависимость результата работы метода от моделируемой выборки.

- Аналогично можно заключить и для размера выборок(отношение моделируемых количеств можно значительно увеличить), однако даже для таких объемов результаты являются достаточно показательными.

5 Построение оценки параметров регресии с помощью группирования выборки

Будем работать с моделью регрессии (3), предполагая что имеем регрессию без выбросов (1). Каждый y_i принадлежит нормальному распределению:

$$y_i = f(x_i, \beta) + \varepsilon_i \sim \mathcal{N}(f(x_i, \beta), \sigma^2). \tag{15}$$

Будем строить оценки таким образом: разделим множество значений функции регресси, т.е множество $\mathcal{R}$ на k полуинтервалов:

$$\mathcal{R} = (-\infty, a_1]U(a_1, a_2]U \dots U(a_{k-1}, +\infty). \tag{16}$$

Обозначим полученные интервалы: $\nu_0, \dots, \nu_{k-1}$.
Функцию распределения нормального закона с параметрами μ, σ^2 можно представить как:

$$F(x) = \Phi(\frac{x-\mu}{\sigma}), \tag{17}$$

где $\Phi(x)$ функция распределения стандартного нормального закона, а $\sigma = \sqrt{\sigma^2}$:

$$\Phi(x) = \frac{1}{\sqrt{2}\sigma} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt. \tag{18}$$

Обозначим:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt. \tag{19}$$

Тогда:

$$\Phi(x) = \frac{1}{2}\Big[1 + \operatorname{erf}\Big(\frac{x}{\sqrt{2}}\Big)\Big]. \tag{20}$$

Поэтому:

$$F(x) = \frac{1}{2}\Big[1 + \operatorname{erf}\Big(\frac{x-\mu}{\sqrt{2}\sigma}\Big)\Big]. \tag{21}$$

Тогда:

$$P\{y_i \in \nu_j\} = \begin{cases} \frac{1}{2}(\operatorname{erf}(\frac{a_{j+1}-f(x_i,\beta)}{\sqrt{2}\sigma}) - \operatorname{erf}(\frac{a_j - f(x_i,\beta)}{\sqrt{2}\sigma})), \ j = \overline{1, k-2} \\ \frac{1}{2}(1 + \operatorname{erf}(\frac{a_1 - f(x_i,\beta)}{\sqrt{2}\sigma})), \ j = 0 \\ \frac{1}{2}(1 + \operatorname{erf}(\frac{a_{k-1} - f(x_i,\beta)}{\sqrt{2}\sigma})), \ j = k-1 \end{cases}. \tag{22}$$

Тогда для каждого y_i будем иметь полуинтервал, в который он попал μ_i.

$$\mu_i = j, \text{если } y_i \text{ отнесли к полуинтервалу } \nu_j. \tag{23}$$

Понятно, что:

$$P(\mu_i = j | y_i \in \nu_j) = P(y_i \in \nu_{\mu_i}). \tag{24}$$

5.1 Построение функции правдоподобия

Составим функцию правдоподобия:

$$l(\beta, \sigma^2, \nu_0, \ldots, \nu_{k-1}) = \ln(\prod_{i=1}^{n} P(\mu_i = j | y_i \in \nu_j)) = \tag{25}$$

$$= \sum_{i=1}^{n} \ln(P(\mu_i = j | y_i \in \nu_j)). \tag{26}$$

Известно приближение для функции $\mathrm{erf}(x)$:

$$(\mathrm{erf} x)^2 \approx 1 - \exp(-x^2 \frac{\frac{4}{\pi} + ax^2}{1 + ax^2}), \tag{27}$$

$$a = \frac{8}{3\pi} \frac{3 - \pi}{\pi - 4}. \tag{28}$$

Оно считается достаточно точным для x близких к 0 и к ∞ [7].
Найдем сразу производную для этого приближения:

$$\mathrm{erf}'(x) = \exp(-x^2 \frac{\frac{4}{\pi} + ax^2}{1 + ax^2}) \frac{-2x \frac{\frac{4}{\pi} + ax^2}{1 + ax^2} + (2ax^3) \frac{\frac{4}{\pi} + ax^2}{1 + ax^2} - \frac{2ax^3}{1 + ax^2}}{2\sqrt{1 - \exp(-x^2 \frac{\frac{4}{\pi} + ax^2}{1 + ax^2})}}. \tag{29}$$

Будем максимизировать функцию L. Для этого будем искать нули ее производной с помощью вычислительных методов(будем использовать метод дихотомии).

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

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

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

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

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

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

где δ_{ij} - символ Кронекера.
Итак, выражение (29) и будем использовать для метода дихотомии, приближая $\mathrm{erf}'(x)$ с помощью выражения (25).

5.2 Метод секущих

Так как мы не можем привести систему $\frac{\delta l}{\delta \beta} = 0$ к виду, удобному для итерации, то нам придется искать ее нули с помощью метода Ньютона. Введем вектор ошибки $\breve{\varepsilon}^{(k)} = \beta^* - \beta^{(k)}$. Тогда для его определения имеем:

$$\frac{\delta l(\beta^{(k)} + \breve{\varepsilon}^{(k)})}{\delta \beta} = 0. \tag{34}$$

Разлагая левую часть по формуле Тейлора и ограничиваясь лишь линейными членами[8], будем иметь систему:

$$\frac{\delta}{\delta \beta} \frac{\delta l(\beta^{(k)}}{\delta \beta} \Delta \beta^{(k)} = -\frac{\delta l(\beta^{(k)})}{\delta \beta}. \tag{35}$$

Если матрица $\frac{\delta}{\delta \beta} \frac{\delta l(\beta^{(k)}}{\delta \beta}$(а в нашем случае она диагональная), то из этой системы можно единственным образом найти $\Delta \beta^{(k)}$ и построить приближение:

$$\beta^{(k+1)} = \beta^{(k)} + \Delta \beta^{(k)}. \tag{36}$$

Так как для второй производной l получится довольно сложное выражение, то будем приближать ее с помощью выражения:

$$\frac{\delta}{\delta \beta_j} \frac{\delta l(\beta_1^{(k)}, \ldots, \beta_n^{(k)})}{\delta \beta} \approx \frac{\frac{\delta l(\beta_1^{(k)}, \ldots, \beta_j^{(k)}, \ldots, \beta_n^{(k)})(\beta^{(k)}}{\delta \beta} - \frac{\delta l(\beta_1^{(k)}, \ldots, \beta_j^{(k-1)}, \ldots, \beta_n^{(k)})(\beta^{(k)}}{\delta \beta}}{\beta_j^{(k)} - \beta_j^{(k-1)}}. \tag{37}$$

Теперь имеем нули производной функции l, а также ее значения на границе отрезка $[a, b]$. Переберем эти значения и таким образом найдем значение вектора $\hat{\beta}$, где она достигает своего максимального значения.

5.3 Переклассификация выборки

На данном этапе для каждого x_i, значение функции регрессии $y_i^{\tilde{\varepsilon}}$ и класс μ_i: $(x_i, y_i^{\tilde{\varepsilon}}, \mu_i)$. Теперь попытаемся переклассифицировать выборку. Для этом будем строить новую выборку такого же объема N. Будем идти по каждому элементу $(x_i, y_i^{\tilde{\varepsilon}}, \mu_i)$ выборки и для этого наблюдения построим новое:

$$(x_i, y_i^{\tilde{\tilde{\varepsilon}}}, \breve{\mu}_i), \tag{38}$$

где $\breve{\mu}_i$ максимально встречающийся класс близлежайших соседей:

$$\breve{\mu}_i = \arg\max_j \sum_{|x_k - x_i| \le \Delta,\ k \ne i} \delta_{\breve{\mu}_k j}, \tag{39}$$

где Δ параметр, задающий уровень близости. Чем он выше, тем больше используется соседей для коррекции класса нашего наблюдения.

Итак, переклассифицировав выборку, применим к ней функцию правдоподобия из уравнений (21-22), только используя теперь новые классы $\breve{\mu}_i$ вместо μ_i. Аналогично пунктам 5.1-5.2 максимизируем ее и найдем новую оценку параметров $\hat{\beta}$.

6 Численные эксперименты

С помощью численных экспериментов можно показать, как себя ведет производная функции правдоподобия при переклассификации выборки.

7 Заключение

Список литературы

[1] Хьюбер Дж П., *Робастность в статистике:пер. с англ.*. М.:Мир,1984-304с

[2] Харин Ю.С., Зуев Н.М., Жук Е.Е, *Теория вероятностей, математическая и прикладная статистика: учебник* Минск: БГУ, 2011.-463с

[3] John Fox & Sanford Weisberg, *Robust Regression*, October 8, 2013

[4] А.В. Омельченко, *Робастное оценивание параметров полиномиальной регрессии второго порядка,* Харьковский национальный университет радиоэлектроники, Украина, 2009

[5] Özlem Gürünlü Alma, *Comparison of Robust Regression Methods in Linear Regression*, Int. J. Contemp. Math. Sciences, Vol. 6, 2011, no. 9, 409 - 421

[6] Özlem Gürünlü Alma, *Comparison of Robust Regression Methods in Linear Regression*, Int. J. Contemp. Math. Sciences, Vol. 6, 2011, no. 9, 409 - 421

[7] Sergei Winitzki, *A handy approximation for the error function and its inverse*

[8] Мандрик П.А., Репников В.И., Фалейчик Б.В., *Численные методы*