

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:
Бодягин Игорь Александрович

Минск, 2017

Содержание

1	Введение	2
2	Модель функции регрессии с аномальными наблюдениями и оценки ее параметров	3
2.1	Метод Наименьших Квадратов	3
2.2	М-оценки	4
2.2.1	Способы выбора функции для решения экстремальной задачи в М-оценках	4
3	Моделирование функции регрессии с аномальными наблюдениями	5
4	Поиск breakdown point у МНК и М-оценок	6
4.1	Результаты программы	6
5	Построение оценки параметров регрессии с помощью группирования выборки	8
5.1	Построение функции правдоподобия	8
5.2	Метод секущих	9
5.3	Переклассификация выборки	10
6	Заключение	10
	Список Литературы	11

1 Введение

Существует несколько подходов для оценки параметров регрессии, но далеко не все устойчивы к возникновению 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]:

- М-Оценки
- R-Оценки
- L-Оценки

М-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 данной курсовой 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 линейную регрессию:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \epsilon_i, i = \overline{1, N} \\ y_i &= f(x_i, \beta) + \epsilon_i, \\ f(x_i, \beta) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} \end{aligned} \quad (1)$$

Или, в векторной форме:

$$y_i = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_n \end{pmatrix} \times \begin{pmatrix} 1 \\ x_{i1} \\ \dots \\ x_{in} \end{pmatrix}^T + \epsilon_i, \quad (2)$$

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 эксперимента, распределение которой подчиняется нормальному закону с нулевым 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:

$$y_i^{\tilde{\epsilon}} = (\xi_i) y_i + (1 - \xi_i) \eta_i, \quad (3)$$

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

$$\begin{cases} p(\xi_i = 0) = \tilde{\epsilon} \\ p(\xi_i = 1) = 1 - \tilde{\epsilon} \end{cases}, \quad (4)$$

которая называется функцией линейной регрессии с выбросами. η_i – случайная величина из какого-то другого неизвестного нам распределения. Переменную $\tilde{\epsilon}$ будем называть процентом аномальных наблюдений.

Теперь рассмотрим некоторые методы оценки параметров регрессии:

2.1 Метод Наименьших Квадратов

Предположим, что случайные ошибки подчиняются нормальному закону распределения вероятностей:

$$L\{\epsilon_i\} = N_1(0, \sigma^2), i = \overline{1, n} \quad (5)$$

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

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

Логарифмическая функция правдоподобия выглядит так[2]:

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

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

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

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

2.2 М-оценки

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

$$\sum_{i=1}^n \phi(x_i; \beta) \rightarrow \min_{\beta}, \quad (10)$$

где $\phi(\cdot; \beta)$ -некоторая функция, определяющая конкретный тип оценок и их точность. Очевидно, что $\phi(\cdot; \beta) \equiv -\ln p(\cdot; \beta)$ -обычная оценка 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:

$$u_i = y_i^{\epsilon} - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in}) \quad (11)$$

Тогда существует такие методы[3]:

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

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
Процент выбросов $\tilde{\epsilon}$	10
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 загрязнений, при котором увеличение количества элементов выборки не повышает точности метода будем делать так:

- Организуем цикл по процентам загрязнений $\tilde{\epsilon}_i$ от $\tilde{\epsilon}_0 = 0$ до $\tilde{\epsilon}_{100} = 100$, увеличивая каждый раз $\tilde{\epsilon}_i$ 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:

$$\widetilde{\delta}_1^{\tilde{\epsilon}_i} = \frac{1}{20} \sum_{k=1}^{20} \left(\sum_{i=0}^n (\beta_i - \hat{\beta}_{N_1 ki})^2 \right)^{\frac{1}{2}} \quad (12)$$

$$\widetilde{\delta}_2^{\tilde{\epsilon}_i} = \frac{1}{20} \sum_{k=1}^{20} \left(\sum_{i=0}^n (\beta_i - \hat{\beta}_{N_2 ki})^2 \right)^{\frac{1}{2}} \quad (13)$$

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

$$br = \left\{ \tilde{\epsilon}_i, \text{ если } \widetilde{\delta}_1^{\tilde{\epsilon}_i} < \widetilde{\delta}_2^{\tilde{\epsilon}_i} \right. \quad (14)$$

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

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.



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



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

Замечания:

- Мы могли бы моделировать не 20 раз, а значительно больше, тем самым мы уменьшаем зависимость результата работы метода от моделируемой выборки.
- Аналогично можно заключить и для размера выборок (отношение моделируемых количеств можно значительно увеличить)

5 Построение оценки параметров регрессии с помощью группирования выборки

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

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

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

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

Обозначим полученные интервалы: ν_0, \dots, ν_{k-1} .

Тогда:

$$P\{y_i \in \nu_j\} = \begin{cases} \frac{1}{2}(\operatorname{erf}(\frac{a_{j+1}-f(x_i, \beta)}{\sqrt{2}\sigma^2}) - \operatorname{erf}(\frac{a_j-f(x_i, \beta)}{\sqrt{2}\sigma^2})), & j = \overline{1, k-2} \\ \frac{1}{2}(1 + \operatorname{erf}(\frac{a_1-f(x_i, \beta)}{\sqrt{2}\sigma^2})), & j = 0 \\ \frac{1}{2}(1 + \operatorname{erf}(\frac{a_{k-1}-f(x_i, \beta)}{\sqrt{2}\sigma^2})), & j = k-1 \end{cases}, \quad (17)$$

где

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \quad (18)$$

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

$$\mu_i = j, \text{ если } y_i \text{ отнесли к полуинтервалу } \nu_j \quad (19)$$

Понятно, что:

$$P(\mu_i = j | y_i \in \nu_j) = P(y_i \in \nu_{\mu_i}) \quad (20)$$

5.1 Построение функции правдоподобия

Составим функцию правдоподобия:

$$l(\beta, \sigma^2, \nu_0, \dots, \nu_{k-1}) = \ln\left(\prod_{i=1}^n P(\mu_i = j | y_i \in \nu_j)\right) = \quad (21)$$

$$= \sum_{i=1}^n \ln(P(\mu_i = j | y_i \in \nu_j)) \quad (22)$$

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

$$(\operatorname{erf}x)^2 \approx 1 - \exp(-x^2 \frac{\frac{4}{\pi} + ax^2}{1 + ax^2}), \quad (23)$$

$$a = \frac{8}{3\pi} \frac{3 - \pi}{\pi - 4} \quad (24)$$

Оно считается достаточно точным для x близких к 0 и к ∞ [7].

Найдем сразу производную для этого приближения:

$$\text{erf}'(x) = \exp(-x^2 \frac{4}{\pi} + ax^2) \frac{-2x \frac{4}{\pi} + ax^2}{1 + ax^2} + (2ax^3) \frac{\frac{4}{\pi} + ax^2}{1 + ax^2} - \frac{2ax^3}{1 + ax^2} \quad (25)$$

$$2\sqrt{1 - \exp(-x^2 \frac{4}{\pi} + ax^2)}$$

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

$$\frac{\delta l}{\delta \beta} = \frac{\delta \sum_{i=1}^n \ln(P(\mu_i = j | y_i \in \nu_j))}{\delta \beta} = \frac{\delta \sum_{i=1}^n \ln P(y_i \in \nu_{\mu_i})}{\delta \beta} = \quad (26)$$

$$= \frac{\delta \sum_{i=1}^n \ln(\frac{1}{2}(\text{erf}(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma^2}) - \text{erf}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma^2})))}{\delta \beta} = \quad (27)$$

$$= \sum_{i=1}^n \left((1 - (\delta_{\mu_i 0} + \delta_{\mu_i k-1})) \frac{(\text{erf}'(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma^2}) - \text{erf}'(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma^2}))}{(\text{erf}(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma^2}) - \text{erf}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma^2}))} + \right. \quad (28)$$

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

$$= - \sum_{i=1}^n \begin{pmatrix} 1 \\ x_{i1} \\ \dots \\ x_{in} \end{pmatrix} \times \left((1 - (\delta_{\mu_i 0} + \delta_{\mu_i k-1})) \frac{(\text{erf}'(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma^2}) - \text{erf}'(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma^2}))}{(\text{erf}(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma^2}) - \text{erf}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma^2}))} + \right. \quad (29)$$

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

где δ_{ij} - символ Кронекера.

Итак, выражение (29) и будем использовать для метода дихотомии, приближая $\text{erf}'(x)$ 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 $\epsilon^{(k)} = \beta^* - \beta^{(k)}$. Тогда для его определения имеем:

$$\frac{\delta l(\beta^{(k)} + \epsilon^{(k)})}{\delta \beta} = 0 \quad (30)$$

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

$$\frac{\delta}{\delta \beta} \frac{\delta l(\beta^{(k)})}{\delta \beta} \Delta \beta^{(k)} = - \frac{\delta l(\beta^{(k)})}{\delta \beta} \quad (31)$$

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:

$$\beta^{(k+1)} = \beta^{(k)} + \Delta\beta^{(k)} \quad (32)$$

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

$$\frac{\delta}{\delta\beta_j} \frac{\delta l(\beta_1^{(k)}, \dots, \beta_n^{(k)})}{\delta\beta}(\beta^{(k)}) \approx \frac{\frac{\delta l(\beta_1^{(k)}, \dots, \beta_j^{(k)}, \dots, \beta_n^{(k)})}{\delta\beta}(\beta^{(k)}) - \frac{\delta l(\beta_1^{(k)}, \dots, \beta_j^{(k-1)}, \dots, \beta_n^{(k)})}{\delta\beta}(\beta^{(k)})}{\beta_j^{(k)} - \beta_j^{(k-1)}} \quad (33)$$

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 \tilde{y}_i и класс $\mu_i: (x_i, \tilde{y}_i, \mu_i)$. Теперь попытаемся переклассифицировать выборку. Для этого будем строить новую выборку такого же объема N . Будем идти по каждому элементу $(x_i, \tilde{y}_i, \mu_i)$ выборки и для этого наблюдения построим новое:

$$(x_i, \tilde{y}_i, \check{\mu}_i), \quad (34)$$

где $\check{\mu}_i$ максимально встречающийся класс близлежащих соседей:

$$\check{\mu}_i = \arg \max_j \sum_{|x_k - x_i| \leq \Delta, k \neq i} \delta_{\check{\mu}_k j}, \quad (35)$$

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 $\check{\mu}_i$ вместо μ_i . Аналогично пунктам 5.1-5.2 максимизируем ее и найдем новую оценку параметров $\hat{\beta}$.

6 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*