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"

Курсовой проект

Допущен к защите
«___» ________ 2017 г
Агеева Елена Сергеевна

Научный руководитель:
Агеева Елена Сергеевна

Минск, 2017

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 проекте 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.

2 Модель функции регрессии с аномальными наблюде и оценки ее параметров

Введем линейную регрессию:

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

Или, в векторной форме:

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

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 экспереме 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:

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

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

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

которая называется функцией линейной регрессии с выбросами. η_i-случайная величина из какого-то другого неизвестного нам распределения. Переменную $\widetilde{\epsilon}$ будем называть процентом аномальных наблюдений.
Теперь рассмотрим некоторые методы оценки параметров регрессии:

2.1 Метод Наименьших Квадратов

Предлоположим, что случайные ошибки подчиняются нормальному закону распределен вероятностей:

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

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 преложил 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 $\phi(\cdot; \beta) \equiv -\ln p(\cdot; \beta)$-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:

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

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
Процент выбросов $\widetilde{\epsilon}$	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:

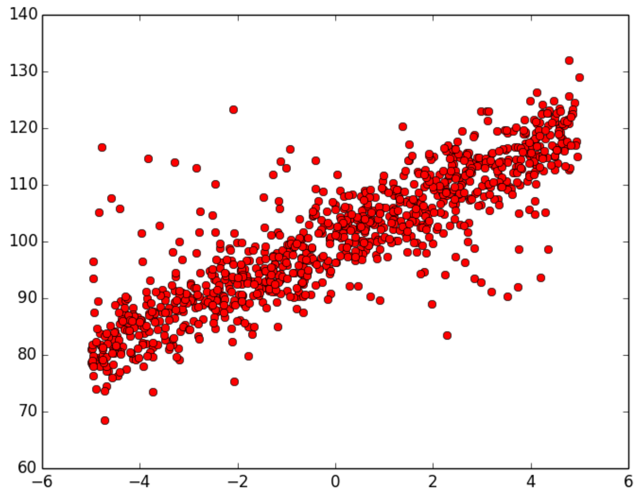

Рис. 1: Вывод графика рассеяния (y_i, x_i)

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 $\widetilde{\epsilon_i}$ от $\widetilde{\epsilon_0} = 0$ до $\widetilde{\epsilon_{100}} = 100$, увеличивая каждый раз $\widetilde{\epsilon_i}$ 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:

$$\widetilde{\delta_1^{\widetilde{\epsilon_i}}} = \frac{1}{20}\sum_{k=1}^{20}(\sum_{i=0}^{n}(\beta_i - \hat{\beta_{N_1ki}})^2)^{\frac{1}{2}} \tag{12}$$

$$\widetilde{\delta_2^{\widetilde{\epsilon_i}}} = \frac{1}{20}\sum_{k=1}^{20}(\sum_{i=0}^{n}(\beta_i - \hat{\beta_{N_2ki}})^2)^{\frac{1}{2}} \tag{13}$$

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

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

- иначе повышаем процент на 1 и повторяем цикл: $\widetilde{\epsilon_{i+1}} = \widetilde{\epsilon_i} + 1$

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.

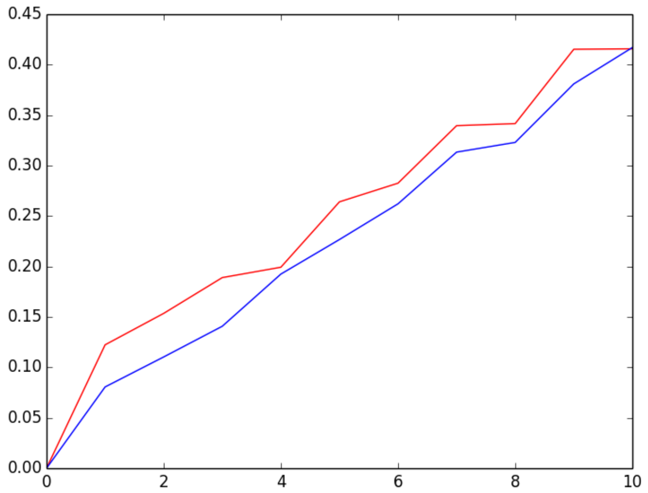

Рис. 2: График, на котором изображены $\widetilde{\delta_1^{\widetilde{\epsilon_i}}}$ красным и $\widetilde{\delta_2^{\widetilde{\epsilon_i}}}$синим относительно $\widetilde{\epsilon_i}$ в случае МНК

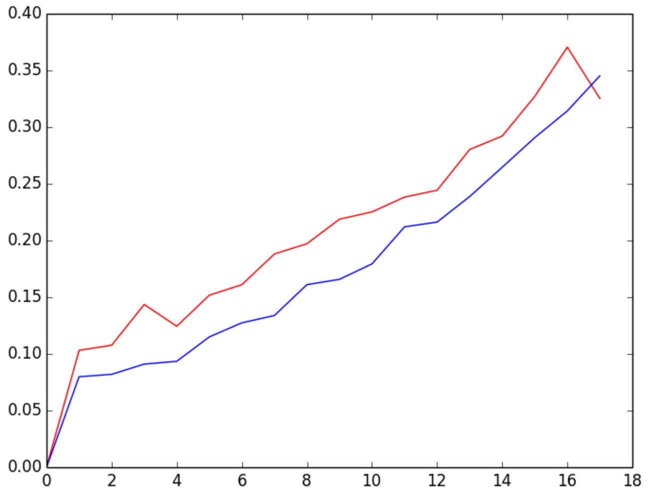

Рис. 3: График, на котором изображены $\widetilde{\delta_1^{\widetilde{\epsilon_i}}}$ красным и $\widetilde{\delta_2^{\widetilde{\epsilon_i}}}$синим относительно $\widetilde{\epsilon_i}$ в случае М-оценок

Замечания:

- Мы могли бы моделировать не 20 раз , а значительно больше, тем самым мы уменьшаем зависимость результата работы метода он моделируемой выборки.

- Аналогично можно заключить и для размера выборок(отношение моделируемых количеств можно значительно увеличить)

5 Заключение

Была рассмотрена модель функции регрессии с аномальными наблюдениями. Были построены МНК оценки и М-оценки, а также была оценена их устройчивость при разных количествах выбросов.
Численные эксперименты показали, что МНК оценки более чувствительны к наличию в выборке аномальных наблюдений чем М-оценки.

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