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 Теоретические сведения

Введем линейную регрессию:

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

Или, в векторной форме:

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

Где y_i – i-е наблюдение из N наблюдений, $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$ регрессоры, $\{\beta_k, k = \overline{0, n}\}$– параметры регрессии, а ϵ_i – случайная ошибка i-го эксперемента, распределение которой подчиняется нормальному закону с нулевым 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:

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

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. $\widetilde{\epsilon}$ будем называть процентом аномальных наблюдений.
Теперь рассмотрим некоторые методы оценки параметров регрессии:

2.1 Метод Наименьших Квадратов

Предлоположим, что случайные ошибки подчиняются нормальному закону распределения 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:

$$L\{y_i^{\widetilde{\epsilon}}\} = N_1(f(x_i;\theta), \sigma^2) \tag{6}$$

Логарифмическая функция правдоподобия выглядит так[2]:

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

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

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

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

По формулам (5)-(7) видно, что метод никак не пригоден для модели регрессии с засорениями, в чем мы далее и убедимся.

2.2 L-оценки

Существует способ построения устойчивых оценок, при котором для описания искажений в выборке используются распределения с "хвостами" более тяжелыми, чем у гипотетического распределения. Например, используется распределение Лапласа, для которого МП-оценкой параметра "сдвига" является выборочная медиана, принадлежащая к классу L-оценок:

$$\hat{\theta}^L(x) = \sum_{t=1}^{n} a_t g(x_{(t)}), \tag{10}$$

где $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ - вариационный ряд выборки; $g(.)$-некоторая функция, $\{a_t\}_{t=1}^n$- весовые коэффициенты.
Чаще всего используется частный случай (10):

$$\hat{\theta}^L(x) = \frac{1}{n-2q} \sum_{t=q+1}^{n-q} x_{(t)}, \tag{11}$$

где q - наибольшее целое, не превосходящее $\alpha \cdot n (0 \leq \alpha \leq \frac{1}{2})$. При определенном выборе α можем получить арифметическое среднее или выборочную медиану.

2.3 M-оценки

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

$$\sum_{i=1}^{n} \phi(x_t; \theta) \to \min_{\theta \in \theta^*}, \tag{12}$$

где θ^*-замыкание θ, $\phi(\cdot;\theta)$-некоторая функция, определяющая конкретный тип оценок и их точность.
Очевидно, что $\phi(\cdot;\theta) \equiv -\ln p(\cdot;\theta)$-обычная оценка максимального правдоподобия, построенная по модели без выбросов (1).
Рассмотрим теперь некоторые способы выбора $\phi(\cdot;\theta)$.

2.3.1 способы выбора функции для решения экстремальной задачи в M-оценках

Для начала определим:

$$e_i = y_i^{\widetilde{\epsilon}} - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}) \tag{13}$$

Тогда существует такие методы[3]:

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

3 моделирование функции регрессии с аномальными наблюдениями

Для начала смоделируем функцию регрессии по методу (3). Для удобства моделируем регрессию с одним параметром x
Воспользуемся такими параметрами:

Переменная	значение
Размер выборки	1000
Процент выбросов	10
Параметры регрессии	$(100, 4)$
x_i	$\sim N(-5, 25)$
ϵ_i	$\sim N(0, 16)$
η_i	$\sim N(100, 100)$

Получаем такой график:

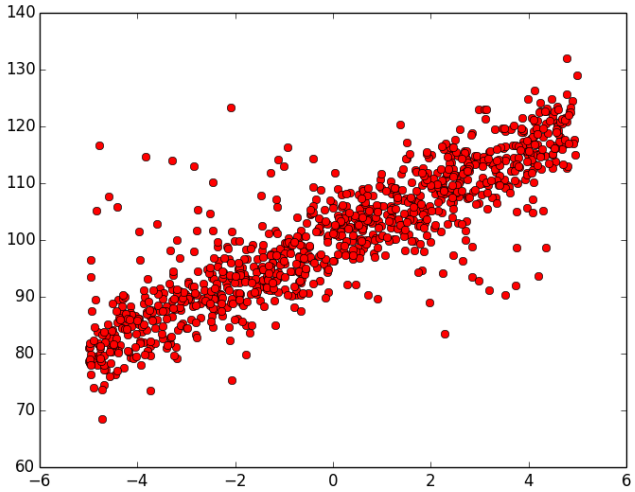

Рис. 1: Вывод графика рассеяния (Y,X)

4 Поиск breakdown point у МНК и М-оценок

Для поиска того процента загрязнений, при котором увеличение количества элементов выборки не повышает точности метода будем делать так:

- Организуем цикл по процентам загрязнений $\widetilde{\epsilon}$
- 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:
- после цикла делим на количество суммирования каждую из сумм невязок, т.е.

вычисляем такие значения:

$$\widetilde{e_1^{\widetilde{\epsilon}_i}} = \frac{1}{20}\sum_{k=1}^{20}(\sum_{i=0}^{n}(\beta_i - \hat{\beta_{N_1 ki}})^2)^{\frac{1}{2}} \tag{14}$$

$$\widetilde{e_2^{\widetilde{\epsilon}_i}} = \frac{1}{20}\sum_{k=1}^{20}(\sum_{i=0}^{n}(\beta_i - \hat{\beta_{N_2 ki}})^2)^{\frac{1}{2}} \tag{15}$$

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

$$br_p = \left\{\widetilde{\epsilon}_i, if\ \widetilde{e_1^{\widetilde{\epsilon}_i}} < \widetilde{e_2^{\widetilde{\epsilon}_i}}\right. \tag{16}$$

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

Такие тесты проведем для МНК и 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 Результаты программы

Метод	breakpoint
МНК	7%
М-оценка с функцией Хьюбера	23%

Итак, видим, что М-оценки значительно устойчивее к выбросам чем МНК.

Консольный вывод программы, где показываются усредненные невязки для разных методов при $N_1 = 1000$ и $N_2 = 3000$ наблюдений соответственно:

```
[MacBook-Pro-Andrew:Курсовая работа akrum$ python breakpointSearcher.py
 /Library/Python/2.7/site-packages/statsmodels/compat/pandas.py:56: Futu
[e the pandas.tseries module instead.
   from pandas.core import datetools
 Going to perform test with percentage 1.000000%....
 Disperancies: 0.135753, 0.062673
 Going to perform test with percentage 2.000000%....
 Disperancies: 0.116544, 0.079605
 Going to perform test with percentage 3.000000%....
 Disperancies: 0.090084, 0.067697
 Going to perform test with percentage 4.000000%....
 Disperancies: 0.146603, 0.098462
 Going to perform test with percentage 5.000000%....
 Disperancies: 0.165968, 0.117375
 Going to perform test with percentage 6.000000%....
 Disperancies: 0.121430, 0.106455
 Going to perform test with percentage 7.000000%....
 Disperancies: 0.203776, 0.139066
 Going to perform test with percentage 8.000000%....
 Disperancies: 0.162438, 0.147290
 Going to perform test with percentage 9.000000%....
 Disperancies: 0.207670, 0.168318
 Going to perform test with percentage 10.000000%....
 Disperancies: 0.252236, 0.190055
 Going to perform test with percentage 11.000000%....
 Disperancies: 0.221233, 0.209581
 Going to perform test with percentage 12.000000%....
 Disperancies: 0.271572, 0.224439
 Going to perform test with percentage 13.000000%....
 Disperancies: 0.295273, 0.241331
 Going to perform test with percentage 14.000000%....
 Disperancies: 0.311751, 0.259758
 Going to perform test with percentage 15.000000%....
 Disperancies: 0.309188, 0.263077
 Going to perform test with percentage 16.000000%....
 Disperancies: 0.340398, 0.328303
 Going to perform test with percentage 17.000000%....
 Disperancies: 0.344174, 0.328521
 Going to perform test with percentage 18.000000%....
 Disperancies: 0.395530, 0.358207
 Going to perform test with percentage 19.000000%....
 Disperancies: 0.414452, 0.379166
 Going to perform test with percentage 20.000000%....
 Disperancies: 0.426133, 0.392336
 Going to perform test with percentage 21.000000%....
 Disperancies: 0.464612, 0.417694
 Going to perform test with percentage 22.000000%....
 Disperancies: 0.497176, 0.458054
 Going to perform test with percentage 23.000000%....
 Disperancies: 0.492249, 0.494707
 Breakpoint for this approximation model is:23.000000%
 MacBook-Pro-Andrew:Курсовая работа akrum$
```

Рис. 2: Невязки М-оценок с функцией Хьюбера

```
[MacBook-Pro-Andrew:Курсовая работа akrum$ python breakpointSearcher.py
 /Library/Python/2.7/site-packages/statsmodels/compat/pandas.py:56: Futu
 e the pandas.tseries module instead.
   from pandas.core import datetools
 Going to perform test with percentage 1.000000%....
 Disperancies: 0.122232, 0.098061
 Going to perform test with percentage 2.000000%....
 Disperancies: 0.134754, 0.102790
 Going to perform test with percentage 3.000000%....
 Disperancies: 0.215045, 0.134158
 Going to perform test with percentage 4.000000%....
 Disperancies: 0.246705, 0.167242
 Going to perform test with percentage 5.000000%....
 Disperancies: 0.260193, 0.210426
 Going to perform test with percentage 6.000000%....
 Disperancies: 0.268767, 0.264489
 Going to perform test with percentage 7.000000%....
 Disperancies: 0.303335, 0.316891
 Breakpoint for this approximation model is:7.000000%
 MacBook-Pro-Andrew:Курсовая работа akrum$
```

Рис. 3: Невязки МНК

6 Заключение

По консольному выводу можно заметить, как растет невязка у обоих методов и при этом уменьшается расхождение невязок при увеличении объема выборки, когда процент аномальных наблюдений растет.

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