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 группированой 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]:

- 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 данном курсовом проекте я буду моделировать функцию регрессии с аномальными наблюдениями, анализировать точность методов и находить для разных методов так называемый "breakpoint"– процент аномальных наблюдений, при котором увеличение количества наблюдений не повысит точность методов.

2 Теоретические сведения

Введем линейную регрессию:

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

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 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:

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

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

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

которая называется функцией линейной регрессии с выбросами. η_i-случайная величина из какого-то другого неизвестного нам распределения.
Для удобства далее обозначим, что $y_i = y_i^{\epsilon}$
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:

$$Ly_i = N_1(f(x_i;\theta),\sigma^2) \tag{6}$$

Логарифмическая функция правдоподобия выглядит так[2]:

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

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

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 медиала, принадлежащая к классу L-оценок:

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

где $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ - вариационный ряд выборки; $g(.)$-некоторая функция, $a{1}_{t=1}^{n}$- 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 цело, не превосходящее $\alpha \bullet n (0 \leq \alpha \leq \frac{1}{2})$. При определенном выборе α можем получить арифметическое среднее или выборочную медиану.

2.3 М-оценки

Швейцарский статистик П.Хьюбер преложил использовать М-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, $\phi(;\theta)$-некоторая функция, определяющая конкретный тип оценок и их точность.
Очевидно, что $\phi(;\theta) \equiv -\ln p(;\theta)$-обычная оценка макимального правдоподобия, построенная по модели без выбросов (1).
Рассмотрим теперь некоторые способы выбора $\phi(\bullet;\theta)$.

2.3.1 способы выбора $\phi(\bullet;\theta)$

Для начала определим:

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

Тогда существует такие методы[3]:

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

3 Численные эксперименты над моделью линейной регрессии с засорениями

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	$\sim N(100, 100)$

```
[ 99.98211212   3.54458638]
[ 99.89020336   3.84810157]
[ 99.98211212   3.54458638]
                    Robust linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 1000
Model:                            RLM   Df Residuals:                      998
Method:                          IRLS   Df Model:                            1
Norm:                          HuberT
Scale Est.:                       mad
Cov Type:                          H1
Date:                Thu, 07 Dec 2017
Time:                        15:02:12
No. Iterations:                    13
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         99.8902      0.147    679.415      0.000      99.602     100.178
x1             3.8481      0.051     75.965      0.000       3.749       3.947
==============================================================================
```

Рис. 1: На скриншоте видны выводы приближенных оценок: МНК, М-оценки

Получаем такие графики:

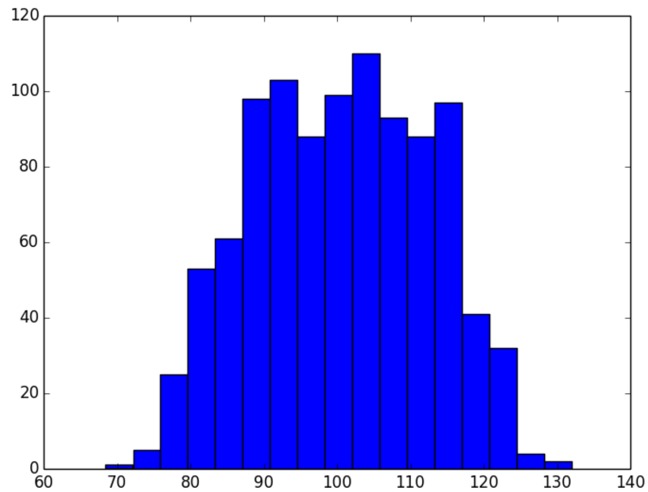

Рис. 2: Гистограмма вектора Y

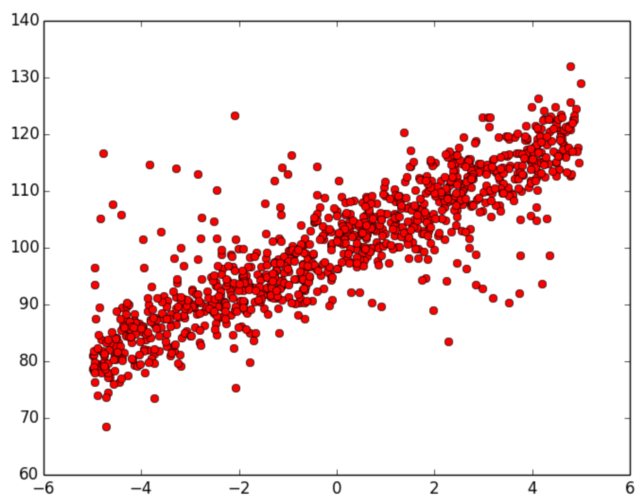

Рис. 3: Вывод вектора Y относительно вектора X

4 Поиск breakpoint 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
- На каждой итерации будем 20 раз моделировать выборку с 1000 и 3000 тысячами 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
- если полученная усредненная невязка при 1000 наблюдений меньше либо равно невязке при 3000 наблюдений, то заканчиваем цикл - нашли breakpoint
- иначе повышаем процент на 1 и повторяем цикл

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	9%
М-оценка с функцией Хьюбера	18%

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

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