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]:

- 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 "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 регрессию 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.:

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

которая называется функцией линейной регрессии с выбросамиη_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:

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

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

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

Логарифмическая функция правдоподобия выглядит так[2]:

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

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

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

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

По формулам (4-6) никак не пригоден для модели регрессии с засорениями, в чем мы далее и убедимся.

2.2 М-оценки

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

$$\sum_{i=1}^{n} \phi(x_t; \theta) \to \min_{\theta \in \theta^*}, \quad (10)$$

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).

2.3 L-оценки

3 Моделирование регрессии на языке Python

Подключим необходимые библиотеки:

```
import numpy as np
import matplotlib.pyplot as plt
from random import random
import pylab
import scipy
from outliers import smirnov_grubbs as grubbs
from matplotlib.backends.backend_pdf import PdfPages
from statsmodels.robust.scale import mad
import theano
import theano.tensor as T
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

Заведем константы для моделирования: количество наблюдейний, процент аномальных наблюдений, и параметры регрессии, использующиеся в моделировании:

```
SAMPLE_QUINTITY=100
OUTLIER_PERCENTAGE = 10.0
regressionParameters = np.matrix([100,4]).T
```

Проинициализируем результирующий вектор y:

```
y_points = np.zeros(shape = SAMPLE_QUINTITY)
```

Теперь моделируем y:

```
x_points = np.zeros(shape=[SAMPLE_QUINTITY,len(regressionParameters)])
y_points = np.zeros(shape = SAMPLE_QUINTITY)
# plt.plot(x_points,y_points,'ro')
# # plt.hist(y_points,bins="auto")
# plt.show()
for i in range(0,SAMPLE_QUINTITY):
    if random()>OUTLIER_PERCENTAGE/100:
        x_points[i] = np.append(np.ones(1),np.random.uniform(-5,5,size = len(regressionP
        # print(x_points[i])
        y_points[i]=(x_points[i]*regressionParameters)+np.random.normal(0,4)
    else:
        x_points[i] = np.append(np.ones(1),np.random.uniform(-5,5,size = len(regressionP
        y_points[i]=np.random.normal(100,10, size=1)
plt.plot(x_points.T[1],y_points,'ro')
plt.show()
```

Программа выводит такой график:

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