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
Ассистент кафедры математического моделирования и анализа данных ФПМИ, кандидат физико-математических наук,
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

На данном этапе будем работать с линейной регрессией:

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

Где y_i – i-е наблюдение из N наблюдений, x_i регрессоры, $\{\alpha, \beta_k, k = \overline{1,n}\}$– параметры регрессии, а ϵ_i – ошибка, распределение которой подчиняется нормальному закону с нулевым ожиданием и дисперсией σ^2.
В нашей задаче считаем параметры $\{\alpha, \beta_k, k = \overline{1,n}\}$ неизвестными, их нам и требуется найти.
Теперь рассмотрим некоторые методы оценки параметров регрессии:

2.1 Метод Наименьших Квадратов

2.2 M-оценки

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]