Санкт-Петербургский политехнический университет

Институт прикладной математики и механики

Кафедра прикладной математики и информатики

**Отчет 6**

**«Математическая статистика»**

Группа 3630102/70301

Выполнил студент Ли Жуйци

Преподаватель Баженов А.Н.

Санкт-Петербург

2020

Оглавление

[Постановка задачи 2](#_Toc38231920)

[теория ..](#теория)................................................................................................................................................................3

[Оценки коэффициентов линейной регрессии](#оценки) ...........................................................................5

[обсуждение](#обсуждение) ....................................................................................................................................6

[Список литературы ........................................................................................................................................6](#список)

# 

# Постановка задачи

# Сгенерировать выборку объёмом 100 элементов для нормального распределения N(x;0;1) . По сгенерированной выборке оценить параметры μ и σ нормального закона методом максимального правдоподобия. В качестве основной гипотезы H0 будем считать, что сгенерированное распределение имеет вид N(x;μ ˆ ;σ ˆ) . Проверить основную гипотезу, используя критерий согласия 2 χ . В качестве уровня значимости взять α = 0,05 . Привести таблицу вычислений 2 χ .

### 2 [Теория](#теория)

### 2.1 Робастные оценки коэффициентов линейной регрессии

### （稳健回归）

# Робастность оценок коэффициентов линейной регрессии (т.е. их устойчивость по отношению к наличию в данных редких, но больших по величине

# выбросов) может быть обеспечена различными способами. Одним из них

# является использование метода наименьших модулей вместо метода наименьших квадратов:

# → min 𝛽0, 𝛽1 . (1)

# 

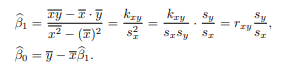
# Напомним, что использование метода наименьших модулей в задаче оценивания параметра сдвига распределений приводит к оценке в виде выборочной медианы, обладающей робастными свойствами. В отличие от этого случая и от задач метода наименьших квадратов, на практике задача (1) решается численно. Соответствующие процедуры представлены в некоторых современных пакетах программ по статистическому анализу. Здесь мы рассмотрим простейшую в вычистлительном отношении робастную альтернативу оценкам коэффициентов линейной регрессии по МНК. Для этого сначала запишем выражения для оценок (2) и (3)

# 

# 

(2) (3)

# в другом виде:



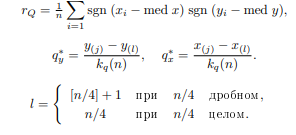
(4)

В формулах (4) заменим выборочные средние 𝑥 и 𝑦 соответственно на робастные выборочные медианы med 𝑥 и med 𝑦, среднеквадратические отклонения 𝑠𝑥 и 𝑠𝑦 на робастные нормированные интерквартильные широты 𝑞 \* 𝑥 и 𝑞 \* 𝑦 , выборочный коэффициент корреляции 𝑟𝑥𝑦 — на знаковый коэффициент корреляции 𝑟𝑄:



(5)

𝛽̂︀0𝑅 = med 𝑦 − 𝛽̂︀1𝑅 med 𝑥, (6)

 (7)

(8)

j = n -l +1

1 при z>0

sgn z= 0 при z=0

-1 при z<0

Уравнение регрессии здесь имеет вид



(9)

Статистики выборочной медианы и интерквартильной широты обладают робастными свойствами в силу того, что основаны на центральных порядковых статистиках, малочувствительных к большим по величине выбросам в данных. Статистика выборочного знакового коэффициента корреляции робастна, так как знаковая функция sgn 𝑧 чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка прямой регрессии (22) обладает очевидными робастными свойствами устойчивости к выбросам по координате 𝑦, но она довольно груба

**2.2 Метод максимального правдоподобия**

Пусть есть выборка X1, ... , Xn {\displaystyle X\_{1},\ldots ,X\_{n}} из распределения{\displaystyle \mathbb {P} \_{\theta }} P0 , где {\displaystyle \theta \in \Theta } — неизвестные параметры. Пусть {\displaystyle L(\mathbf {x} \mid \theta )\colon \Theta \to \mathbb {R} }L(x | ) : —> R функция правдоподобия, где x Точечная оценка

 {\displaystyle {\hat {\theta }}\_{\mathrm {M\Pi } }={\hat {\theta }}\_{\mathrm {M\Pi } }(X\_{1},\ldots ,X\_{n})=\mathop {\rm {argmax}} \limits \_{\theta \in \Theta }L(X\_{1},\ldots ,X\_{n}\mid \theta )}

называется оце́нкой максима́льного правдоподо́бия параметра  {\displaystyle \theta }. Таким образом оценка максимального правдоподобия — это такая оценка, которая максимизирует функцию правдоподобия при фиксированной реализации выборки.

Часто вместо функции логарифимическую L правдоподобия {\displaystyle L}L используют функциюправдоподобия l = lnL  {\displaystyle l=\ln L}Так как функция x->lnx,x>0 монотонно возрастает{\displaystyle x\to \ln x,\;x>0}  на всей области определения,максимум  любой функции {\displaystyle L(\theta )}L()является максимумом функции {\displaystyle \ln L(\theta )}lnL(), и наоборот. Таким образом,

{\displaystyle {\hat {\theta }}\_{\mathrm {M\Pi } }=\mathop {\rm {argmax}} \limits \_{\theta \in \Theta }l(X\_{1},\ldots ,X\_{n}\mid \theta )} 

Если функция правдоподобия дифференцируема, то необходимое условие экстремума — равенство нулю её градиента

{\displaystyle g(\theta )={\frac {\partial l(\mathbf {x} ,\theta \_{0})}{\partial \theta }}=0}

Достаточное условие экстремума может быть сформулировано как отрицательная определённость гессиана — матрицы вторых производных:



{\displaystyle H={\frac {\partial ^{2}l(\mathbf {x} ,\theta \_{0})}{\partial \theta \partial \theta ^{T}}}}

Важное значение для оценки свойств оценок метода максимального правдоподобия играет так называемая информационная матрица , равная по определению:



{\displaystyle I(\theta )=E[g(\theta )g(\theta )^{T}]}

В оптимальной точке информационная матрица совпадает с математическим ожиданием гессиана, взятым со знаком минус:

{\displaystyle I=-E(H\_{0})}  *I=-E(H0)*

i

**2.3 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат**

# Теорема К.Пирсона. Статистика критерия 𝜒 2 асимптотически распределена по закону 𝜒 2 с 𝑘 − 1 степенями свободы. Это означает, что независимо от вида проверяемого распределения, т.е. функции 𝐹(𝑥), выборочная функция распределения статистики 𝜒 2 при 15 𝑛 → ∞ стремится к функции распределения случайной величины с плотностью вероятности

# 0 𝑥 ≤ 0;

# 𝑓𝑘−1(𝑥) =

# 

# x>0

# Правило проверки гипотезы о законе распределения по методу 𝜒 2 .

# 1. Выбираем уровень значимости 𝛼.

# 2. По таблице находим квантиль (𝑘 − 1) распределения хи-квадрат с 𝑘 − 1 степенями свободы порядка 1 − 𝛼.

# 3. С помощью гипотетической функции распределения 𝐹(𝑥) вычисляем вероятности 𝑝𝑖 = 𝑃 (𝑋 ∈ ∆𝑖), 𝑖 = 1, ... , 𝑘.

# 4. Находим частоты 𝑛𝑖 попадания элементов выборки в подмножества ∆𝑖 , 𝑖 = 1, ... , 𝑘.

# 5. Вычисляем выборочное значение статистики критерия 𝜒 2 :

# = .

# 6. Сравниваем и квантиль (𝑘 − 1).

# а) Если 𝜒 2 В < (𝑘 − 1), то гипотеза 𝐻0 на данном этапе проверки принимается.

# б) Если 𝜒 2 В ≥ (𝑘 − 1)то гипотеза 𝐻0 отвергается, выбирается одно из альтернативных распределений, и процедура проверки повторяется.

# Результат

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| i | Границы |  |  |  |  |  |
| 1 | −∞ -1.2387 | 6 | 0.0517 | 5.1694 | 0.8306 | 0.1335 |
| 2 | -1.4496 0.1126 | 44 | 0.4411 | 44.1092 | -0.1092 | 0.0003 |
| 3 | 0.1126 −∞ | 50 | 0.5072 | 50.7214 | -0.7214 | 0.0103 |
|  |  | 100 | 1.0000 | 100.0000 | 0.0000 |  |

# Количество промежутков k=3

# Уровень значимости 𝛼 =0.1440

# Тогда квантиль

# [Обсуждение](#обсуждение)

# 1： Критерий наименьших квадратов точнее оценивает коэффициенты линейной регрессии на выборке без возмущений.

# 2： Критерий наименьших модулей точнее оценивает коэффициенты линейной регрессии на выборке с возмущениями.

# 3： Критерий наименьших модулей устойчив к редким выбросам

# Список литературы

# 1: линейная регрессия <https://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%B0%D1%8F_%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F>

# 2：Метод наименьших квадратов

# <https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%BD%D0%B0%D0%B8%D0%BC%D0%B5%D0%BD%D1%8C%D1%88%D0%B8%D1%85_%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D0%BE%D0%B2>