Белорусский Государственный Университет

Информатики и Радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра ЭВМ

Лабораторная работа №1

Тема «Корреляционный анализ»

Выполнил: Проверил:

Студент группы 7М2431 Марченко В.В.

Зайцев Ю.В.

Минск, 2017

Задание:

Входные данные: n объектов, каждый из которых характеризуется

двумя числовыми признаками: и .

Требуется исследовать степень взаимосвязи между двумя признаками некоторых объектов. Для каждого набора данных необходимо выполнить следующие задания:

1. Визуализировать данные на плоскости в виде точек с координатами .

2. Статистически оценить коэффициент корреляции Пирсона между признаками *x* и *y*.

3. Проверить статистическую гипотезу о некоррелированности признаков *x* и *y* на уровне значимости 0,05.

Исходные данные:

|  |  |  |
| --- | --- | --- |
| N | a | R |
| 10000 | (-1, 1) |  |

Где N – это количество точек, a – значения вектора математических ожиданий, R – корреляционная матрицы для моделируемой выборки из гауссовских случайных векторов.

Название файла: wine.csv

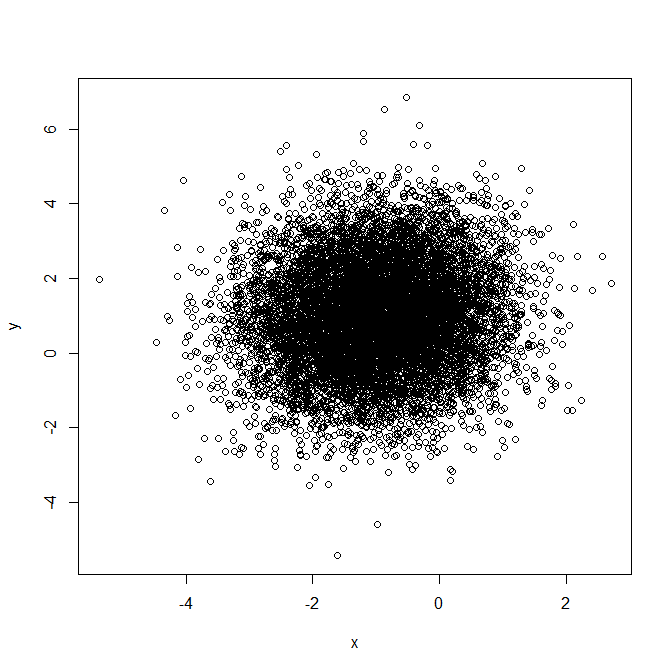
Ссылка: <http://archive.ics.uci.edu/ml/datasets/Wine>

Первый признак: alcohol (столбец № 2)

Второй признак: color-intensity (столбец № 11)

Результаты:

1. Смоделированные данные:



data: x and y

t = 7.9895, df = 9998, p-value = 1.505e-15

alternative hypothesis: true correlation is not equal to 0

5 percent confidence interval:

0.07902570 0.08027207

sample estimates:

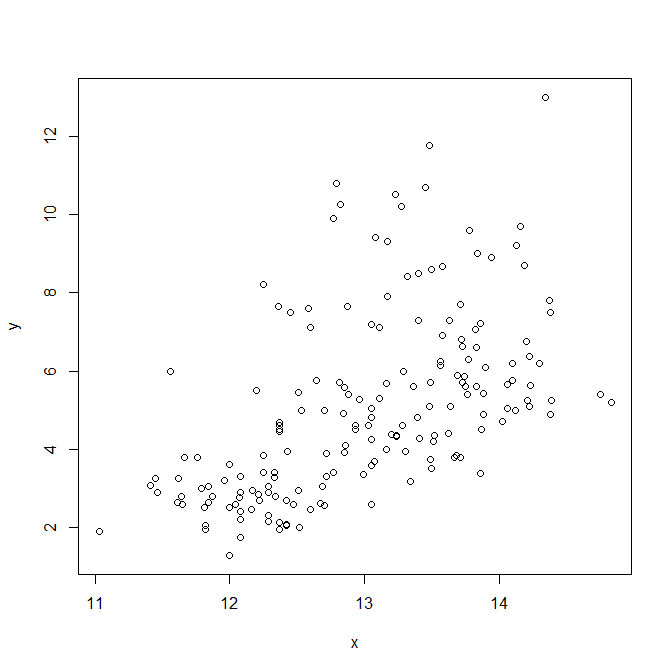
cor

0.07964892

Число Стьюдента для уровня значимости 0,05 и степеней свободы >100 равно 1,96.

Т.к. |t| > 1,96, то гипотеза о некоррелированности отвергается.

1. Реальные данные:



data: x and y

t = 8.6542, df = 176, p-value = 3.056e-15

alternative hypothesis: true correlation is not equal to 0

5 percent confidence interval:

0.5430304 0.5496808

sample estimates:

cor

0.5463642

Число стьюдента для уровня значимости 0.05 и степеней свободы 200

равно 1.97.

Т.к. |t| > 1,971, то гипотеза о некоррелированности отвергается.

Листинг программы:

require(MASS)

analyse\_cor <- function(x, y) {

print(cor.test(x, y))

dev.new()

plot(x, y)

}

dat <- read.table("wine.csv", sep=",")

analyse\_cor(dat$V2, dat$V11)

n <- 10000

a <- c(-1, 1)

r <- cbind(c(1, 0.1), c(0.1, 2))

dat <- mvrnorm(n, a, r)

analyse\_cor(dat[,1], dat[,2])