Белорусский Государственный Университет
Информатики и Радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра ЭВМ

Лабораторная работа №2

Тема «Регрессионный анализ»

Выполнил:

Студент группы 7М2431

Зайцев Ю.В.

Проверил:

Марченко В.В.

Минск, 2017

Задание:

Входные данные: n объектов, каждый из которых характеризуется двумя числовыми признаками: $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^n$.

Требуется исследовать регрессионную зависимость признака y от признака x. Для каждого набора данных необходимо выполнить следующие задания:

1. Построить модель линейной регрессии $y = ax + b + \varepsilon$, оценив оптимальные параметры a и b из условия минимизации суммы квадратов отклонения для заданных значений признаков $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^n$.
2. Вычислить коэффициент детерминации для получившейся модели.
3. Визуализировать на одном графике точки (x_i, y_i) и прямую y = ax + b.

Исходные данные:

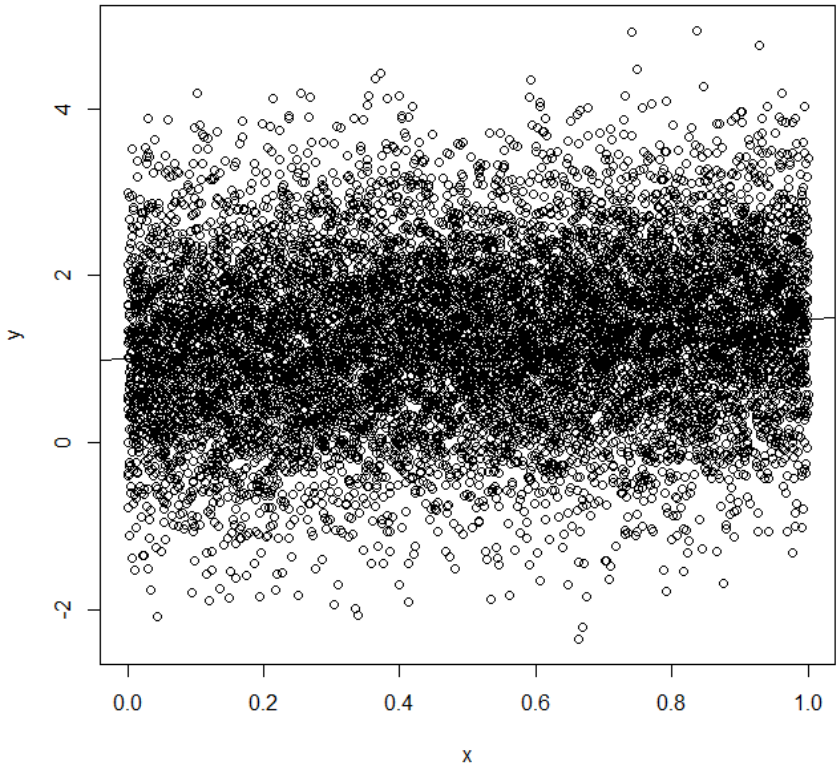
N	a	b	σ^2
10000	0.5	1	1

Где N – это количество точек, a и b – коэффициенты в линейной функции y = ax + b + ε, а σ^2 – дисперсия гауссовского белого шума ε. Сами значения x задаются в виде равномерной сетки на отрезке [0; 1].

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:

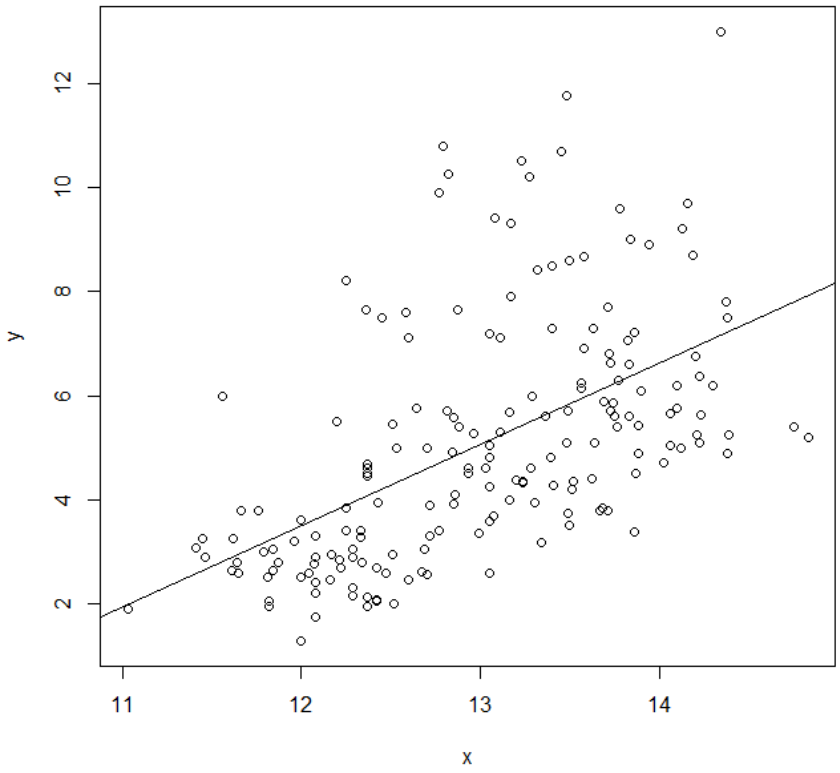
1. Смоделированные данные:




```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0538 -0.6733 -0.0131  0.6864  4.4301

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.01629    0.02015   50.44   <2e-16 ***
x            0.47271    0.03490   13.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.008 on 9998 degrees of freedom
Multiple R-squared:  0.01802,   Adjusted R-squared:  0.01792
F-statistic: 183.5 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Коэффициент детерминации = 0.018

2. Реальные данные:




```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0189 -1.3322 -0.4905  0.6174  6.0705

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.2257     2.3483  -6.484 8.72e-10 ***
x             1.5602     0.1803   8.654 3.06e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.947 on 176 degrees of freedom
Multiple R-squared:  0.2985,    Adjusted R-squared:  0.2945
F-statistic:  74.9 on 1 and 176 DF,  p-value: 3.056e-15
```

Коэффициент детерминации = 0.295

Листинг программы:

```
analyse_regression <- function(x, y) {
  model <- lm(y ~ x)
  print(summary(model))
  dev.new()
  plot(x, y)
  abline(model)
}

dat <- read.table("wine.csv", sep=",")
analyse_regression(dat$V2, dat$V11)
n <- 10000
a <- 0.5
b <- 1
s2 <- 1
x <- seq(0.0, 1.0, length=n)
y <- a * x + b + rnorm(n, 0, s2)
analyse_regression(x, y)
```