Белорусский государственный университет
информатики и радиоэлектроники

Кафедра ЭВМ

# Отчет по лабораторной работе №2
# по курсу ХиУИ

Вариант #30

Выполнил:
студент группы 7М2432
Пантелеев В.В.

Проверил:
Марченко В.В.

**МИНСК 2017**

## Цель работы:

1) Изучить методику проведения регрессионного анализа при помощи языка программирования R.

# ВВЕДЕНИЕ

Регрессионный анализ — метод моделирования измеряемых данных и исследования их свойств. Данные состоят из пар значений зависимой переменной (переменной отклика) и независимой переменной (объясняющей переменной). Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной. Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные. Критерием качества приближения (целевой функцией) обычно является среднеквадратичная ошибка: сумма квадратов разности значений модели и зависимой переменной для всех значений независимой переменной в качестве аргумента. Регрессионный анализ — раздел математической статистики и машинного обучения. Предполагается, что зависимая переменная есть сумма значений некоторой модели и случайной величины. Относительно характера распределения этой величины делаются предположения, называемые гипотезой порождения данных. Для подтверждения или опровержения этой гипотезы выполняются статистические тесты, называемые анализом остатков. При этом предполагается, что независимая переменная не содержит ошибок. Регрессионный анализ используется для прогноза, анализа временных рядов, тестирования гипотез и выявления скрытых взаимосвязей в данных.

## Выполнение работы

**Входные данные:** n объектов, каждый из которых характеризуется двумя числовыми признаками: X, Y.

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

1. Построить модель линейной регрессии y = ax + b + ε, оценив оптимальные параметры a и b из условия минимизации суммы квадратов отклонений для заданных значений признаков {x, y}

2. Вычислить коэффициент детерминации для получившейся модели.

3. Визуализировать на одном графике точки (x[ i ] , y[ i ]) и прямую y = ax + b.

Все описанные выше задания требуется выполнить для двух наборов данных.

1. Смоделированные согласно модели y = ax + b + ε точки (x, y). В качестве ε нужно использовать гауссовский белый шум с нулевым математическим ожиданием и заданной дисперсией $\sigma^2$. Значения x[ i ] выбираются через равные промежутки на отрезке [0; 1].

2. Реальные статистические данные из заданного набора (выдаются преподавателем).

Исходные данные — n = 10000, a = -10, b = -10, σ = 1;

### Листинг программы:

```
analyse_regression <- function(x, y) {
  model <- lm(y ~ x)
  print(summary(model))
  dev.new()
  plot(x, y)
  abline(model)
}
fileName = "/home/panteleev/Documents/Data/iris.data.txt"
irisData <- read.table(fileName, header = FALSE, sep = ",")
head(irisData)
analyse_regression(irisData$V1, irisData$V2)
```

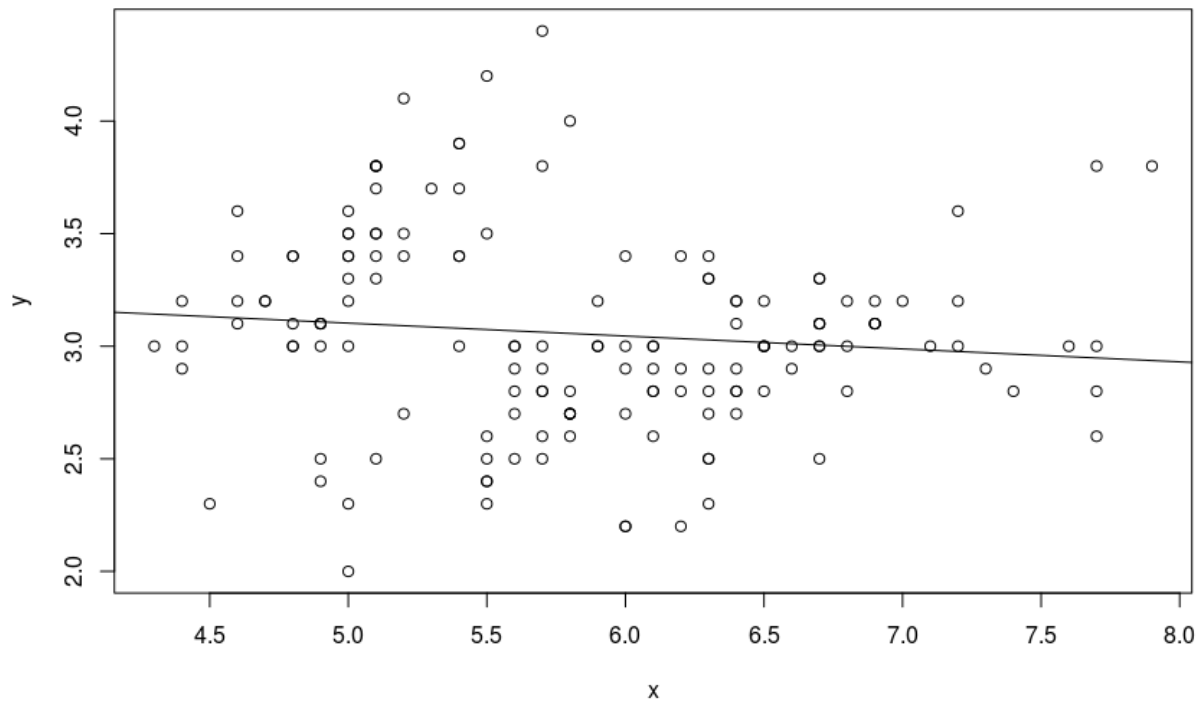
```
n <- 10000
a <- -10
b <- -10
s2 <- 1
x <- seq(0.0, 1.0, length=n)
y <- a * x + b + rnorm(n, 0, s2)
analyse_regression(x, y)
```

**Результат выполнения программы:**

Часть загруженных данных из файла *iris.data.txt:*

```
  V1  V2  V3  V4          V5
1 5.1 3.5 1.4 0.2 Iris-setosa
2 4.9 3.0 1.4 0.2 Iris-setosa
3 4.7 3.2 1.3 0.2 Iris-setosa
4 4.6 3.1 1.5 0.2 Iris-setosa
5 5.0 3.6 1.4 0.2 Iris-setosa
6 5.4 3.9 1.7 0.4 Iris-setosa
```

*Визуализация данных из файла iris.data.txt:*



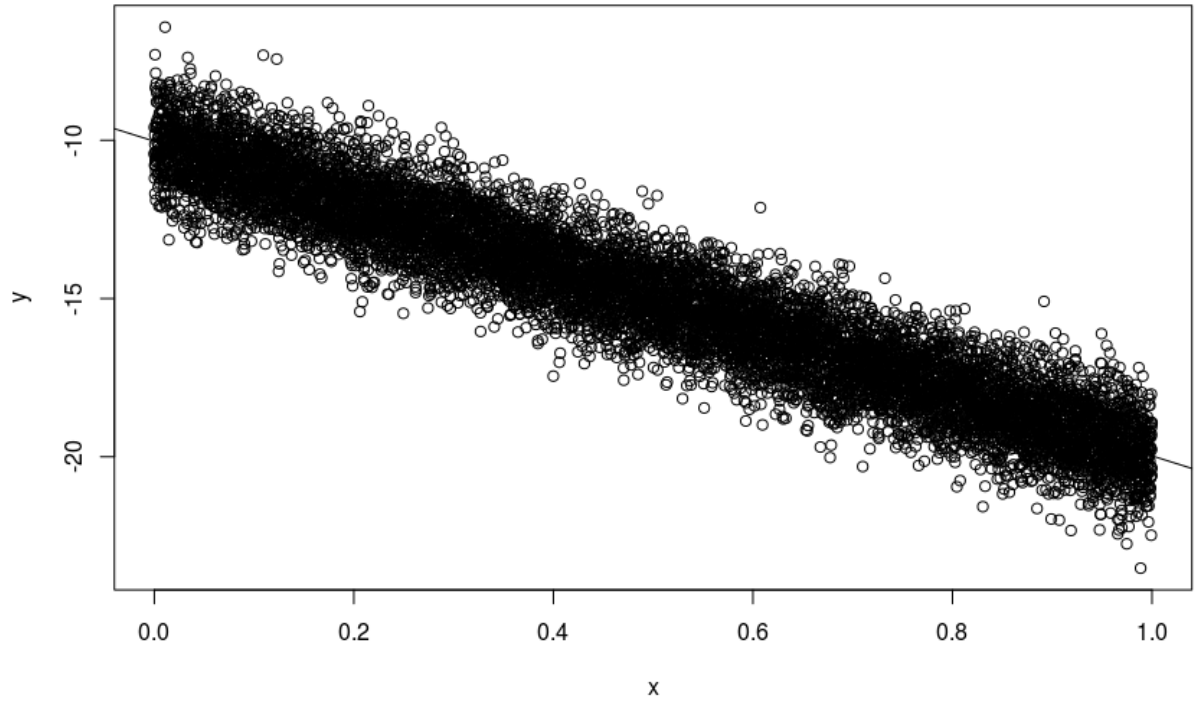
Результат регрессионного анализа данных из файла *iris.data.txt:*

```
Residuals:
      Min       1Q   Median       3Q      Max
-1.10230 -0.23930 -0.01639  0.27414  1.33779

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.38864    0.25248  13.421   <2e-16 ***
x           -0.05727    0.04278  -1.339    0.183
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4324 on 148 degrees of freedom
Multiple R-squared:  0.01196,   Adjusted R-squared:  0.005286
F-statistic: 1.792 on 1 and 148 DF,  p-value: 0.1828
```

Часть сгенерированных данных*:*

```
[x] 0.00000000 0.00010001 0.00020002 0.00030003 0.00040004 0.00050005

[y]  -9.591310 -10.444093 -10.407709 -11.197622 -10.788755  -9.831493
```

Визуализация сгенерированных данных:




Результат регрессионного анализа сгенерированных данных *:*

```
Residuals:
     Min       1Q   Median       3Q      Max
-3.6616 -0.6858 -0.0134  0.6918  3.9446

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.03186    0.02011  -498.8   <2e-16 ***
x            -9.94138    0.03484  -285.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.006 on 9998 degrees of freedom
Multiple R-squared:  0.8907,    Adjusted R-squared:  0.8906
F-statistic: 8.144e+04 on 1 and 9998 DF,  p-value: < 2.2e-16
```

## Выводы

Для данных из файла *iris.data.txt,* мы получили следующий вектор коэфицентов в уровнении регрессии - `-0.05727 0.04278 -1.339 0.183` , для сгенерированных данных мы получили следующий вектор коэфицентов в уровнении регрессии `-9.94138 0.03484 -285.4 <2e-16.`

Коэфицент детерминации для модели данных из файла — 0.01, для сгенерированных данных коэфицент равен — 0.89. Из этого можно сказать что практически все точки из сгенерированного набора лежат на линии регрессии, чего нельзя сказать о данных импортированных из файла.