**Лабораторная работа №2**

**«Прогнозирование. Линейная однофакторная регрессионная модель. Множественная линейная регрессия.»**

**Цель работы**

Исследовать возможности языка R для построения линейной однофакторной регрессионной модели и множественной линейной регрессий;

**Ход работы**

Для выполнения лабораторной работы был выбран вариант 10 по номеру в списке группы, а так же тестовые данные X5, Y5. Весь программный код находится в приложении А.

**Задание 1.**

Используя тестовые данные для решения задач по теме «Однофакторная линейная регрессия» или данные, смоделированные Вами для двух признаков, выполните следующее:

1. Постройте на графике облако точек;

2. Найдите коэффициенты линейной регрессии;

3. Совместите на графике линию регрессии с облаком точек;

4. Оцените визуально характер зависимости признаков.

5. Сделайте выводы.

Был построен график с облаком точек.

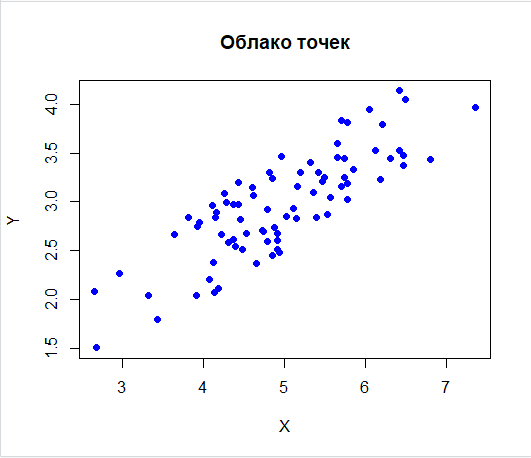


Рисунок 1 – График облака точек

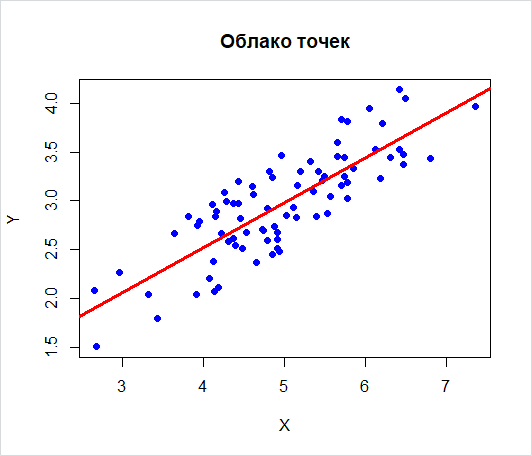
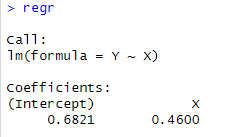


Рисунок 2 – Совмещенный график с линией регрессии

  
Рисунок 3 – Коэффициенты регрессии

Полученное уравнение линейной регрессии, в соответствии с коэффициентами имеет вид: y = 0.46x + 0.6821. Выборочный коэффициент корреляции полученный инструкцией cor(x, y) равен 0.8234324, что говорит об хорошей корреляционной зависимости.

**Задание 2.**

Подберите реальные данные для задачи однофакторной линейной регрессии (можно использовать данные из приложения А). Выполните следующее:

1. Постройте на графике облако точек.

2. Найдите уравнение регрессии. (Укажите численные значения коэффициентов регрессии)

3. Совместите уравнение регрессии с облаком точек.

4. Сделайте выводы относительно зависимости исследуемых признаков.

5. Выберите произвольно несколько значений независимого признака x и вычислите ожидаемые (согласно полученному уравнению регрессии) значения признака y.

6. Покажите на графике точки, соответствующие сделанному прогнозу.

7. Оформите отчёт. Включите в отчёт построенный график, запишите численно коэффициенты регрессии, коэффициент корреляции, значение критерия Стьюдента, расчётные значения прогнозируемого признака. Обоснуйте выводы относительно зависимости признаков и качества прогноза.

Для исследования регрессии были выбраны две колонки со значениями: fLength, fWidth.

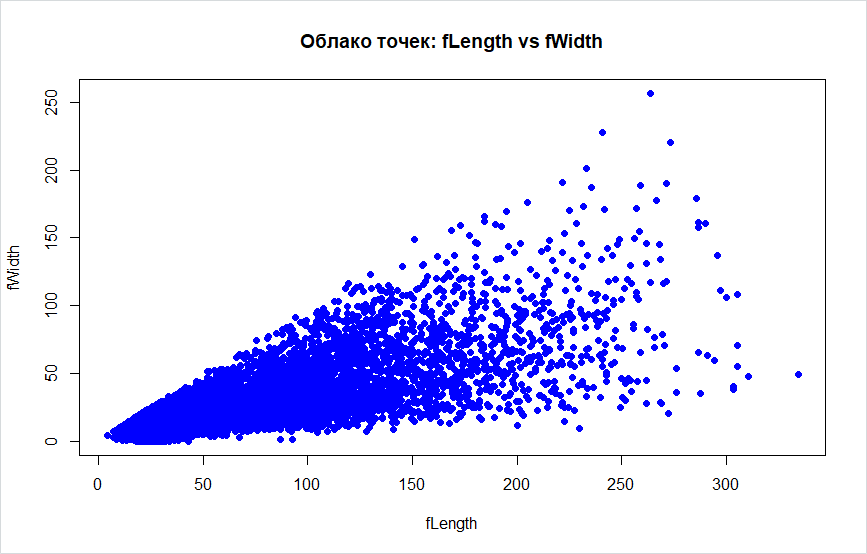


Рисунок 4 – Облако точек для fLength и fWidth

Получаем уравнение линейной регрессии в виде fWidth = 0.3336694\*fLength + 4.4130174.

Добавим уравнение линейной регрессии на наш график облака точек.

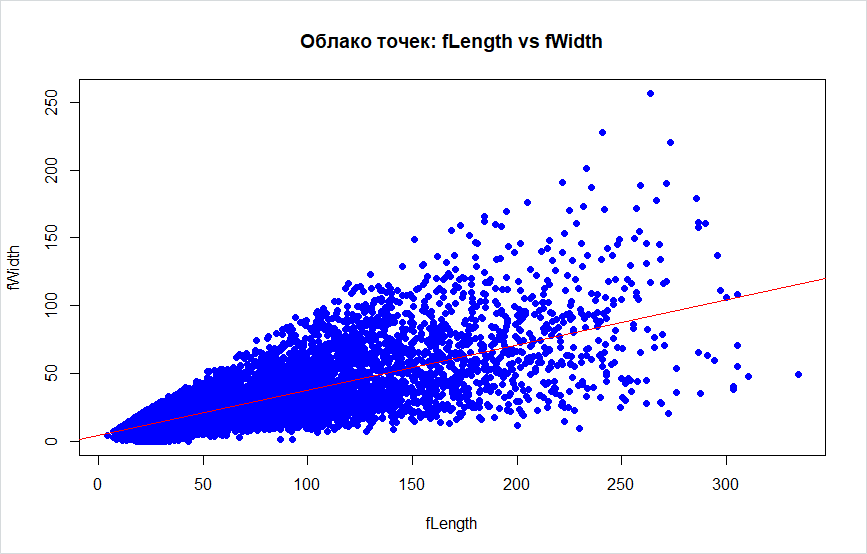


Рисунок 5 – Добавление линии регрессии

Коэффициент корреляции равен 0.770512, что является признаком хорошей корреляции. Критерии Стьюдента для каждого коэффициента:

Для Intercept значение равно 32.4 это показывает, что свободный член значимо отличается от нуля и для fLength = 0, значение fWidth будет около 4.4130174.

Для наклона значение равно 166.7 это означает, что значение сильно отличается от нуля, что значит, что есть сильная линейная зависимость и значение fWidth будет увеличиваться куда больше, чем значение fLength.

Просчитаны значения fWidth для заданных fLength и отображены на графике.

fLength[50,175, 250]

На графике мы можем увидеть тенденцию, к тому что чем больше значение fLength, тем больше разброс в значении fWidth.

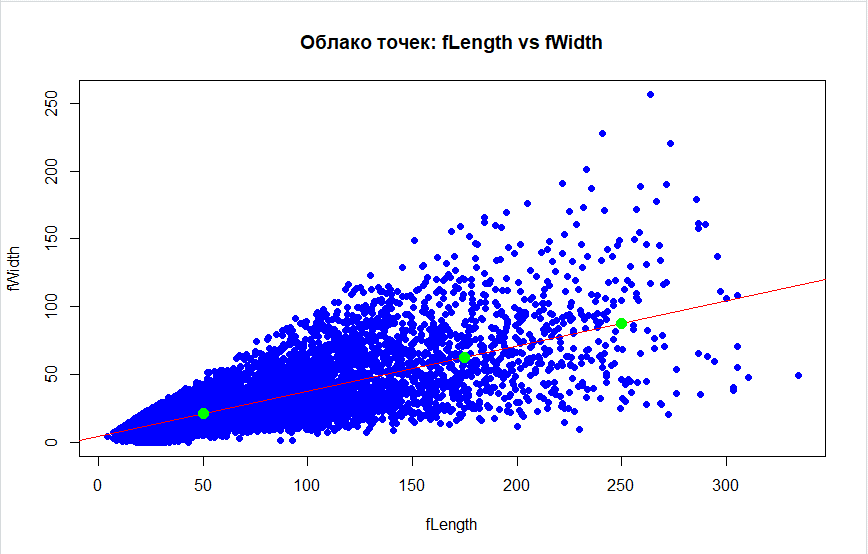


Рисунок 6 – Добавление точек на график

В результате у нас корреляция между параметрами достаточно высокая, критерий Стьюдента показывает, что есть сильная линейная зависимость между fLength и fWidth.

**Задание 3.**

Решите в пакете R задачу построения и анализа уравнения множественной линейной регрессии:

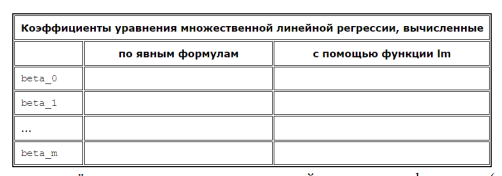
1. Подберите данные для задачи (можно использовать данные из приложения А);

2. Постройте с помощью пакета R уравнение линейной регрессии сделайте это 2мя способами:

- по явным формулам;

- с помощью функции lm.

3. Cравните результаты с помощью таблицы следующего вида:



4. Приведите в отчёте матрицу попарных корреляций и сводную информацию (summary) модели.

5. Проанализируйте степень влияния факторов на переменную отклика согласно каждой из полученных моделей;

6. Сделайте прогноз (с помощью каждой из полученных моделей. если они различны) – на вход модели подайте не менее 10 значений из обучающей выборки и сравните:

- истинное значение переменной отклика;

- прогноз, полученный с помощью модели, коэффициенты которой вычислены явно;

- прогноз, полученный с помощью модели, коэффициенты которой вычислены с помощью функции lm.

7. Проанализируйте и объясните полученные результаты.

В исходном наборе данных класс телескопа был поменян на числовое значение для реализации множественной регрессии. g=1, h=2.

Построим с помощью пакета R уравнение линейной регрессии по явным формулам.

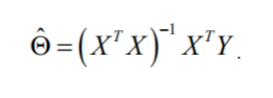


Рисунок 7 – Явная формула

X – матрица независимых переменных с единичным столбцом.

Y – вектор зависимой переменной.

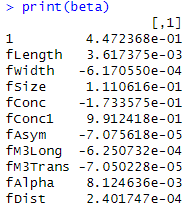


Рисунок 8 – Коэффициенты линейной регрессии полученной по явной формуле

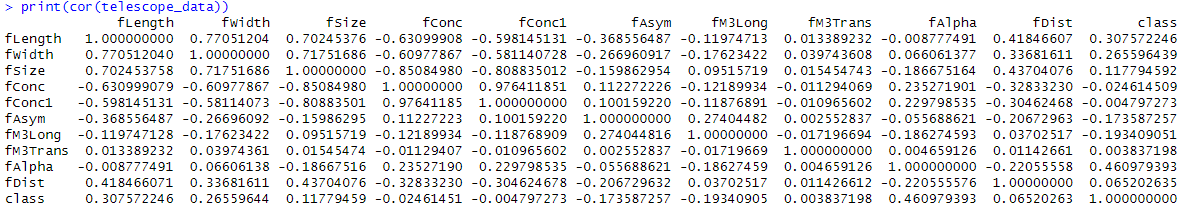


Рисунок 9 – Корреляция между переменными в наборе данных

Заметим, что корреляция между fConc и fConc1 слишком высокая, уберём fConc1 для того, чтобы избежать мультиколлинеарность.

Построим уравнение линейной регрессии без fConc1.

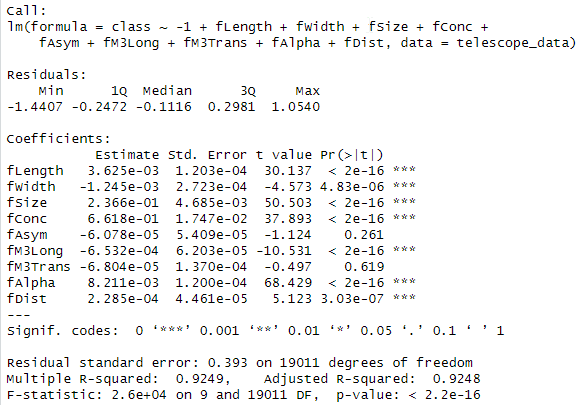


Рисунок 10 – Коэффициенты для class, полученные с помощью lm

Сравним значения коэффициентов полученные с помощью разных способов.

Таблица 1 – Сравнение коэффициентов множественной регрессии

|  |  |  |
| --- | --- | --- |
|  | По явным формулам | С помощью функции lm |
| fLength | 0.003617375 | 0.003625 |
| fWidth | -0.0006170550 | -0.001245 |
| fSize | 0.1110616 | 0.2366 |
| fConc | -0.1733575 | 0.6618 |
| fAsym | 0.9912418 | -0.00006078 |
| fM3Long | -0.00007075618 | -0.0006532 |
| fM3Trans | -0.00007050228 | -0.00006804 |
| fAlpha | 0.008124636 | 0.008211 |
| fDist | 0.0002401747 | 0.0002285 |

Сделаем предсказание для 10 значений.

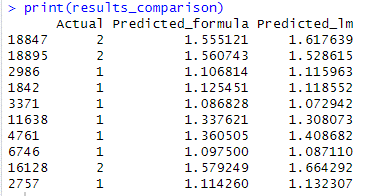


Рисунок 11 – Предсказанные значения

Предсказанные значения не точны, по причине большого количества признаков, создающих шум и не позволяющих задать правильную зависимость. Так же возможны нелинейные зависимости в признаках, что так же понижает точность предсказаний, выбросы в данных так же могут повлиять на точность предсказаний.

Вывод

Были исследованы возможности языка R для построения линейной однофакторной регрессионной модели и множественной линейной регрессий.

Приложение А

Task 1  
X <- scan("C:\\Users\\1\\Desktop\\Univer\\5kurs\\IAD\\LR2\\X5.txt")

Y <- scan("C:\\Users\\1\\Desktop\\Univer\\5kurs\\IAD\\LR2\\Y5.txt")

plot(X, Y, main = "Облако точек", xlab = "X", ylab = "Y", pch = 19, col = "blue")

regr = lm(formula = Y ~ X)

abline(regr, col="red", lwd=3, add=TRUE)

regr

cor(X,Y)  
  
Task 2

telescope\_data <- read.csv("C:\\Users\\1\\Desktop\\Univer\\5kurs\\IAD\\LR2\\magic04.data", header = FALSE)

colnames(telescope\_data) <- c("fLength","fWidth","fSize","fConc", "fConc1", "fAsym", "fM3Long", "fM3Trans", "fAlpha", "fDist", "class")

plot(telescope\_data$fLength, telescope\_data$fWidth, main = "Облако точек: fLength vs fWidth", xlab = "fLength", ylab = "fWidth", pch = 19, col = "blue")

#Проведение исследования линейной регрессии

modellm <- lm(fWidth ~ fLength, data = telescope\_data)

coefficients(modellm)

abline(modellm, col = "red")

cor(telescope\_data$fLength,telescope\_data$fWidth)

summary(modellm)

new\_fLength <- c(50, 175, 250)

predicted\_fWidth <- predict(modellm, newdata = data.frame(fLength = new\_fLength))

points(new\_fLength, predicted\_fWidth, col = "green", pch = 19, cex = 1.5)

Task 3

telescope\_data <- read.csv("C:\\Users\\1\\Desktop\\Univer\\5kurs\\IAD\\LR2\\magic04.data", header = FALSE)

colnames(telescope\_data) <- c("fLength","fWidth","fSize","fConc", "fConc1", "fAsym", "fM3Long", "fM3Trans", "fAlpha", "fDist", "class")

X <- as.matrix(cbind(1, telescope\_data[,1:(ncol(telescope\_data)-1)]))

Y <- as.matrix(telescope\_data[,ncol(telescope\_data)])

beta <- solve(t(X) %\*% X) %\*% t(X) %\*% Y

print(beta)

print(cor(telescope\_data))

model = lm(class ~ -1 + fLength + fWidth + fSize + fConc + fAsym+ fM3Long + fM3Trans + fAlpha + fDist, data=telescope\_data)

summary(model)

set.seed(123)

sample\_indices <- sample(1:nrow(telescope\_data),10)

predicted\_lm <- predict(model, newdata = telescope\_data[sample\_indices,])

predicted\_formula <- as.vector(X[sample\_indices,] %\*% beta)

actual\_values <- telescope\_data[sample\_indices,ncol(telescope\_data)]

results\_comparison <- data.frame(

Actual = actual\_values,

Predicted\_formula = predicted\_formula,

Predicted\_lm = predicted\_lm)

print(results\_comparison)