

Множественная линейная регрессия. Реализация в R. Расчёты по явным формулам.

Рассмотрим нахождение коэффициентов линейной регрессионной модели (согласно ранее введённым обозначениям, это – вектор $\hat{\Theta}$) и (на их основе) прогнозирование значения переменной отклика по ранее найденной формуле: $\hat{y}_i = X_i \hat{\Theta}$, $i = n+1, \dots$

Файл с данными имеет вид:

TotalSquare (m2)	LivingSquare (m2)	DistCenter (km)	DistMetro (km)	Price
80	53	17	2,1	14 612 000,00 ₽
76	51	1	0,7	16 931 128,00 ₽
96	72	16	1,3	18 905 472,00 ₽
56	37	16	2,2	14 829 304,00 ₽
75	56	6	2,8	19 214 025,00 ₽
75	56	11	1,1	19 582 950,00 ₽
97	65	12	1,4	19 123 259,00 ₽
30	24	14	2,3	6 035 280,00 ₽
84	63	7	1,9	20 058 696,00 ₽
50	33	11	2,4	13 807 800,00 ₽
55	44	6	1,7	13 087 745,00 ₽
94	71	10	1,7	17 337 266,00 ₽
91	68	5	1,4	17 189 900,00 ₽
32	26	7	1,8	6 405 792,00 ₽
86	65	4	0,3	19 267 698,00 ₽
55	41	2	1,6	13 827 495,00 ₽
65	49	18	1,7	11 242 920,00 ₽
45	34	9	2,2	12 004 470,00 ₽
47	38	4	1,3	10 586 844,00 ₽

Скопируем данные (без заголовков) в текстовый файл, предварительно преобразовав данные из денежного формата. В текстовом файле заменим десятичные запятые точками. Сохраним текстовый файл как `flats.txt`. Выполним расчёты двумя способами:

- 1) явным, по полученным выше формулам;
- 2) с помощью библиотечной функции `lm`.

Ниже приведён код R-программы для вычисления коэффициентов линейной регрессионной модели и для получения прогноза, т.е. оценки стоимости квартиры по заданным значениям факторов.

R-код

```
#
# МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ
#
# Назначим текущий директорий (в котором находится наш файл с
# данными)
setwd("D://Olga/Multiple_Regression")
#
N=50 # Число наблюдений (квартир)
M=4  # Число факторов

# Создадим "шаблоны" матриц
Data = matrix(1 : (N*(M+1)), ncol=(M+1)) # Данные из файла
X      = matrix(1 : (N*M),      ncol=M)   # Матрица значений
# факторов, расширенная единичным столбцом
Y      = matrix(1 : N,          ncol=1)    # Столбец
# значений переменной отклика (цен на квартиры)
T      = matrix(1 : (M+1)*(M+1), ncol = (M+1) )
Teta = matrix(1 : (M+1),        ncol = 1) # Вектор
# коэффициентов ЛРМ

# Прочтём файл с характеристиками квартир и ценами и запишем
# данные в заготовленную матрицу
Data = read.table("flats.txt", sep = '\t')
#edit(Data)

# Разделим значения факторов и переменной отклика
X = Data[, 1:M]
Y = Data[, (M+1)]

# Создадим вектор из единиц (как набор из N единиц)
odin=vector(length=N,mode='numeric')
odin = rep(1,N)

# Заготовим матрицу нужного размера
X1 = matrix(1 : (N*(M+1)),ncol=(M+1))
# Заполним первый столбец матрицы X1 единицами
X1[,1] = odin
#edit(X1)

# Остальные столбцы возьмём из матрицы X
for(i in 1:M) X1[,i+1] = X[,i]

# Найдём вектор параметров множественной линейной регрессии Teta
# Вычислим сначала матрицу t(X1)
T1 = t(X1)

# Вычислим матрицу T=t(X1)%*%X1
T=T1*%*%X1
```

```
# Подключим библиотеку для работы с матрицами
library(MASS)
# Найдём обратную матрицу для T
obr = ginv(T)

# Вычислим вектор Teta - вектор коэффициентов линейной модели
Teta=obr%*%T1%*%Y
Teta=Teta[,1]
print(Teta)

#
=====
# Введём данные нашей квартиры (на первое место сразу поставим
единицу)
myx = c(1, 65.7, 46.2, 9, 1.5)
myx
myX=matrix(myx,ncol=M+1,byrow=TRUE)
tmyX=t(myX)

# Оценим стоимость нашей квартиры
myprice=Teta%*%tmyX
print(myprice)
# =====
```

Получили результат: 14539.25.

Таким образом, согласно построенной модели, квартира со следующими характеристиками: общая площадь – 65.7 м², жилая – 46,2 м², находящаяся на расстоянии 9 км от центра города и 1,5 км от метро, оценивается в 14 млн. 539 тыс. 250 рублей.