**Задание 6**

*Нахождение параметров множественной линейной регрессии  
Прогнозирование*

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

1. Подберите данные для задачи построения множественной линейной регрессии;
2. Постройте с помощью пакета R уравнение линейной регрессии (сделайте это 2мя способами - по явным формулам и с помощью функции lm - и сравните результаты (запишите найденные 2мя способами коэффициенты уравнения регрессии в таблицу);
3. Проанализируйте степень влияния факторов на переменную отклика;
4. Спрогнозируйте значения переменной отклика при заданных значениях исходных данных - сделайте несколько прогнозов: для нескольких новых данных и для тех, для которых значение переменной отклика уже известно;
5. Оцените реалистичность полученных результатов (запишите в таблицу истинные значения переменной отклика и полученные с помощью прогноза, основанного на использовании построенного уравнения регрессии);
6. Опишите полученные результаты в файле формата Word и прикрепите файл здесь.

Были использованы данные о посещении студентами тренажерного зала в университетском кампусе (к сожалению, нет информации, какого именно университета).

Источник данных: <https://www.kaggle.com/nsrose7224/crowdedness-at-the-campus-gym>.

Данные содержат следующие характеристики:

1. Number\_people — число человек в тренажерном зале;
2. Timestamp — число секунд, прошедших с начала суток;
3. Day\_of\_week — день недели (от 0 до 6);
4. Is\_weekend — является ли день недели субботой или воскресеньем (0 или 1);
5. Is\_holiday — является ли день недели праздничным днем (0 или 1);
6. Apparent\_temperature — ощущаемая температура воздуха, в фаренгейтах;
7. Temperature — температура воздуха, в фаренгейтах;
8. Is\_start\_semester — было ли сделано измерение в начале семестра (0 или 1).

Сам автор данных предлагает прогнозировать число людей в тренажерном зале в зависимости от остальных факторов. Выберем в качестве переменной отклика характеристику 1.

Очевидно, что характеристики 6 и 7, 3 и 4 должны сильно коррелировать между собой. Мы исключим характеристики 3 и 7, т.к. они меньше коррелируют с переменной отклика нежели атрибуты 6 и 4 (см. [Матрица корреляции](#_Матрица_корреляции)).

Согласно матрице корреляции, также откажемся от тех характеристик, которые слабо коррелируют с переменной отклика ().

Итого, используемые факторы:

* Timestamp — число секунд, прошедших с начала суток, столбец №2;
* Is\_weekend — является ли день недели субботой или воскресеньем (0 или 1), столбец №4;
* Apparent\_temperature — ощущаемая температура воздуха, в фаренгейтах, столбец №6;
* Is\_start\_semester — было ли сделано измерение в начале семестра (0 или 1), столбец №8.

Переменная отклика:

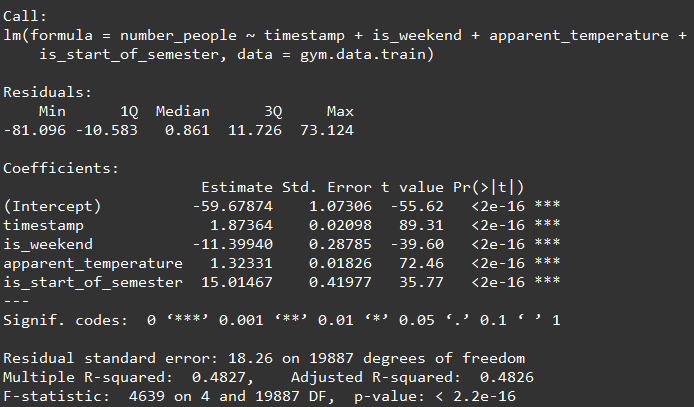
* Number\_people — число человек в тренажерном зале, столбец №1.

Найденные коэффициенты уравнения регрессии:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| Вариант 1 (непосредственное вычисление) |  |  |  |  |  |
| Вариант 2 (функция lm) |  |  |  |  |  |

Нетрудно заметить, что коэффициенты, найденные обоими способами почти идентичны.

Проанализируем степень влияния факторов на переменную отклика:

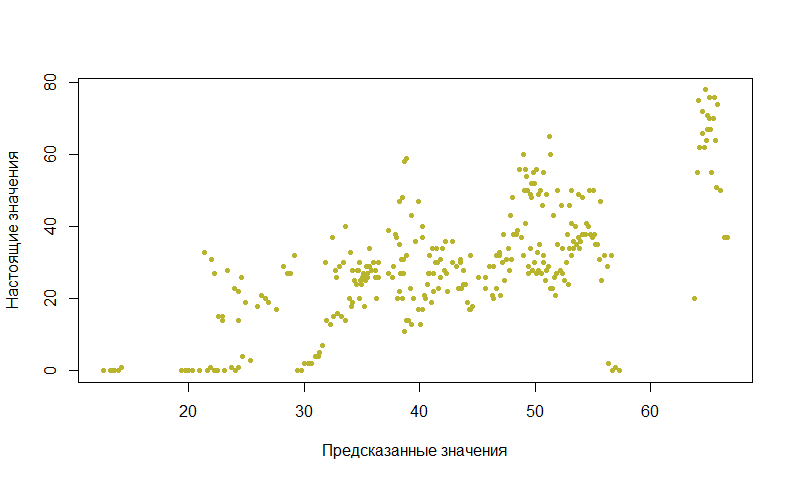


Вероятность превысить значение значения критерия Стьюдента крайне мало, все факторы достаточно хорошо влияют на переменную отклика. Однако, коэффициент детерминации также невелик. Не следует ожидать от модели точного прогноза значений.

Согласно матрице корреляции, наибольшее влияние на количество человек в тренажерном зале влияет время дня, т.е. количество секунд, прошедшее с начала суток.

Для построения модели и последующего прогнозирования данных исходная выборка была разделена на две части: обучающую и тестовую. Объем обучающей выборки составляет *19892* наблюдений, а тестовой — *6174*.

Ввиду большого объема тестовой выборки, приведем следующий график — «облако прогноза»:



Из всей тестовой выборки были взяты только 300 значений для «читаемости» графика.

По оси абсцисс отложим предсказанные значения переменной отклика с помощью функции *predict* по модели, построенной вторым вариантом, а по оси ординат — ее настоящие значения.

Очевидно, что при использовании идеальной модели, точки на графике лежали бы на биссектрисе первого координатного угла. Однако, наша модель не точна (), соответственно прогнозируемые значения не всегда достоверны.

Коэффициенты регрессии, полученные путем их непосредственного вычисления, практически не отличаются от тех, что получены средствами пакета. Следовательно, прогнозируемые значения также будут практически не отличаться друг от друга. Графики их будут также совпадать.

Скрипт программы на R:

options(scipen=5)

gym.data <- read.table("data\_gym.csv",

header = TRUE,

sep = ",")

gym.data[, 2] <- mapply(function(timestamp) timestamp / 3600, gym.data[, 2])

# требуемые нам атрибуты:

# timestamp (2) — кол-во секунд, прошедших с момента начала суток

# is\_weekend (4) — день недели

# apparent\_temperature (6) — ощущаемая температура воздуха, в фаренгейтах

# is\_start\_of\_semester (8) — 1 - начало семестра, 0 - нет

# number\_people (1) — количество людей в зале

write.csv(cor(gym.data), file = "gym\_data\_corr.csv")

cor(gym.data[, 1], gym.data[, c(2, 4, 6, 8)])

indices.train <- c(2375:7643, 9840:18466, 20071:26066)

indices.test <- c(1:2374, 7644:9839, 18467:20070)

length(indices.test) + length(indices.train)

gym.data.train <- gym.data[indices.train,]

gym.data.test <- gym.data[indices.test,]

# СПОСОБ 2

crowd.mul.regression <- lm(number\_people ~ timestamp +

is\_weekend +

apparent\_temperature +

is\_start\_of\_semester,

gym.data.train)

summary(crowd.mul.regression)

crowd.predict <- predict(crowd.mul.regression, gym.data.test)

crowd.true <- gym.data.test[,1]

plot(crowd.predict[70:370], crowd.true[70:370], col="#B8B42D", type="p",

pch=20,

xlab="Предсказанные значения",

ylab="Настоящие значения")

# СПОСОБ 1

f <- 4

matrix.gym <- **matrix**(nrow = 0, ncol = f + 1)

temp <- read.table("data\_gym.csv",

header = TRUE,

sep = ",")

temp[, 2] <- mapply(function(timestamp) timestamp / 3600, temp[, 2])

matrix.gym <- temp[, c(1, 2, 4, 6, 8)]

n <- length(matrix.gym[,1])

# первый столбец отвечает за переменную отклика, возьмем только факторные переменные

X.gym <- matrix.gym[, 2:(f + 1)]

Y.gym <- matrix.gym[, 1]

X1.gym <- **matrix**(nrow = n, ncol = f + 1)

for (k in 1:(f + 1)) {

print(k)

if (k == 1) {

X1.gym[, k] <- runif(n, 1:1)

}

else {

X1.gym[, k] <- X.gym[, k - 1]

}

}

X1.gym.train <- X1.gym[indices.train,]

T <- t(X1.gym.train) %\*% X1.gym.train

library(MASS)

T.inverse <- ginv(T)

th <- (T.inverse%\*%t(X1.gym.train)%\*%Y.gym[indices.train])[, 1]

prognosis <- th%\*%t(X1.gym[indices.test,])

plot(prognosis[70:370], Y.gym[indices.test][70:370], col="#B8B42D", type="p",

pch=20,

xlab="Предсказанные значения",

ylab="Настоящие значения")

Приложение

## Матрица корреляции

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **number\_people** | **timestamp** | **day\_of\_week** | **is\_weekend** | **is\_holiday** | **apparent\_temperature** | **temperature** | **is\_start\_of\_semester** |
| **number\_people** | 1 | 0.516720990417555 | -0.198662160014743 | -0.212955300512139 | -0.023753774524768 | 0.371589002309815 | 0.365246974879143 | 0.188742879419299 |
| **timestamp** | 0.516720990417555 | 1 | 0.00218839321020537 | 0.00577365351315505 | 0.0000150144115856291 | 0.135662608734084 | 0.134208955558041 | -0.00802885395824057 |
| **day\_of\_week** | -0.198662160014743 | 0.00218839321020537 | 1 | **0.795181803054361** | -0.0905684630744177 | -0.0115532170034277 | -0.0110302480349575 | 0.00580063302204169 |
| **is\_weekend** | -0.212955300512139 | 0.00577365351315505 | 0.795181803054361 | 1 | -0.0692562742323101 | 0.00309527353042509 | 0.00307858343691695 | 0.00343411028382523 |
| **is\_holiday** | -0.023753774524768 | 0.0000150144115856291 | -0.0905684630744177 | -0.0692562742323101 | 1 | -0.0680552354854267 | -0.0592059777826788 | 0.0838017038629485 |
| **apparent\_temperature** | 0.371589002309815 | 0.135662608734084 | -0.0115532170034277 | 0.00309527353042509 | -0.0680552354854267 | 1 | 0.995838690710553 | 0.0926224573445956 |
| **temperature** | 0.365246974879143 | 0.134208955558041 | -0.0110302480349575 | 0.00307858343691695 | -0.0592059777826788 | **0.995838690710553** | 1 | 0.0922549344307029 |
| **is\_start\_of\_semester** | 0.188742879419299 | -0.00802885395824057 | 0.00580063302204169 | 0.00343411028382523 | 0.0838017038629485 | 0.0926224573445956 | 0.0922549344307029 | 1 |

Выделены атрибуты, коррелирующие между собой.