**Задание 4**

*Моделирование данных с заданным законом распределения*

*и визуализация данных*

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.

# нормальное распределение

# в качесте предполгаемых данных возьмем массу мозга человека

# средняя масса мозга мужчины: 1350 г.

*sample.norm* **<-** rnorm(*1000*, *1350*, *120*)

hist(*sample.norm*,

*breaks* **=** "Sturges",

*col* **=** "#92140C",

*main* **=** "Нормальное распределение")

summary(*sample.norm*)

boxplot(*sample.norm*, *col* **=** "#92140C")

sd(*sample.norm*)

var(*sample.norm*)

# равномерное распределение

*sample.unif* **<-** runif(*1000*, *min* **=** *5*, *max* **=** *17*)

hist(*sample.unif*,

*breaks* **=** "Sturges",

*col* **=** "#FFF8F0",

*main* **=** "Равномерное распределение")

summary(*sample.unif*)

boxplot(*sample.unif*, *col* **=** "#FFF8F0")

var(*sample.unif*)

# экспоненциальное распределение

*sample.exp* **<-** rexp(*1000*, *4*)

hist(*sample.exp*,

*breaks* **=** "Sturges",

*col* **=** "#F6CA83",

*main* **=** "Экспоненциальное распределение")

summary(*sample.exp*)

boxplot(*sample.exp*, *col* **=** "#F6CA83")

var(*sample.exp*)

# распределение Стьюдента

*sample.t* **<-** rt(*1000*, *3*)

hist(*sample.t*,

*breaks* **=** "Sturges",

*col* **=** "#9CAFB7",

*main* **=** "Распределение Стьюдента")

summary(*sample.t*)

boxplot(*sample.t*, *col* **=** "#9CAFB7")

var(*sample.t*)

# распределение Хи-квадрат

*sample.chisq* **<-** rchisq(*1000*, *7*)

hist(*sample.chisq*,

*breaks* **=** "Sturges",

*col* **=** "#FFCF99",

*main* **=** "Распределение Хи-квадрат")

summary(*sample.chisq*)

boxplot(*sample.chisq*, *col* **=** "#FFCF99")

var(*sample.chisq*)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |