**Задание 5**

*Нахождение, исследование и использование уравнения однофакторной линейной регрессии*

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.
8. Прикрепите doc-файл c отчётом здесь.

Для исследования были взяты достаточно популярные данные о поездках с использованием велопроката (т.н. Bike Sharing) в г. Сан-Франциско за ~700 дней и данные о погоде (средняя температура воздуха, облачность и т.п.) за соответствующие дни.

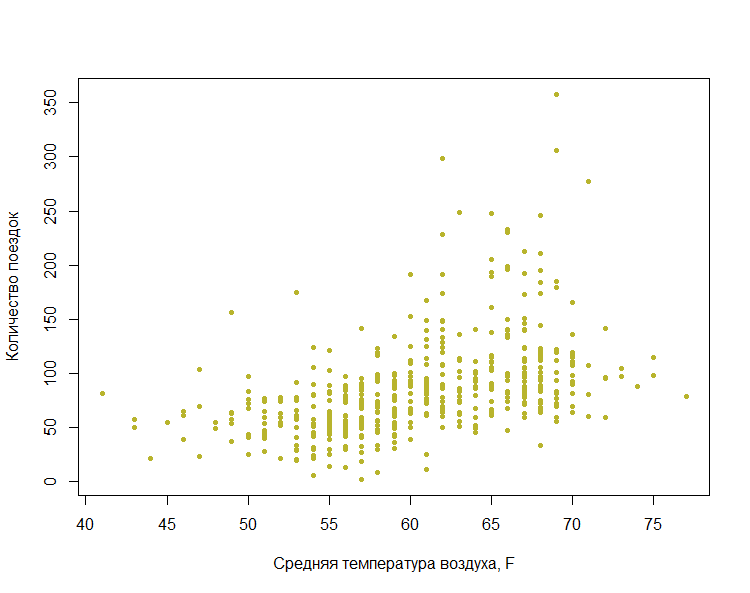
Источник: <https://www.kaggle.com/benhamner/sf-bay-area-bike-share>

Была проведена следующая обработка данных:

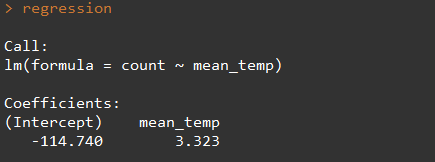
1. Данные о поездках содержат информацию о каждой поездке, сделанной за определенный из 733 дней, в том есть информации о человеке, совершившем поездку. Люди подразделяются на два типа — Customer и Subscriber. Subscriber — подписчик — человек, имеющий годовую подписку на пользование велопрокатом. Customer — покупатель — человек, купивший суточный или трехдневный «абонементы» на использование велопроката. Причем суточный или трехдневный абонементы можно приобрести непосредственно в киоске, находящемся рядом со станцией велопроката, в то время как годовую подписку можно оформить на официальном сайте проекта.  
   Выберем пользователей с типом Customer.
2. Известно, однако не отражено в данных, что за поездки длительностью более 30 минут, взимается дополнительная плата. Отберем те, поездки, чья длительность составляет менее 30 минут (1800 секунд).
3. Сгруппируем поездки по дате для определения количество поездок пользователей типа Customer с длительностью менее получаса.

Предполагается, что есть зависимость между средней температурой воздуха (в Фаренгейтах) и количеством поездок, совершенных за день.

Облако точек:



Найдем коэффициенты уравнения регрессии:

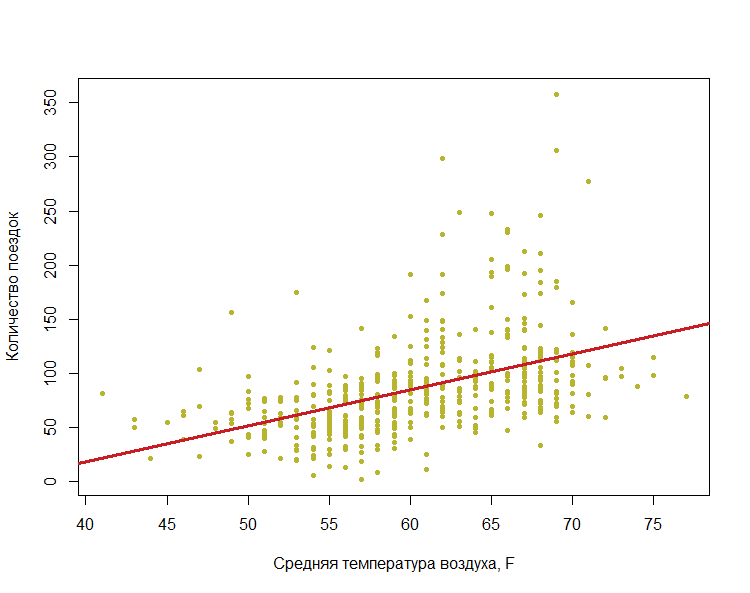


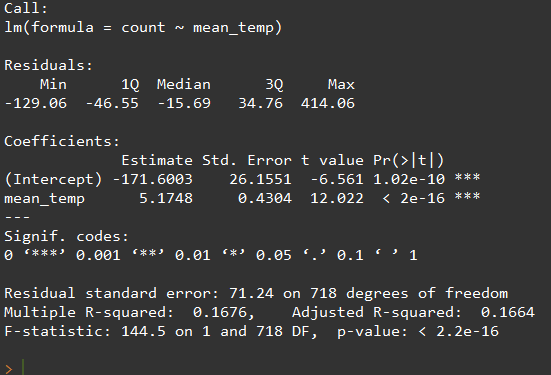
Уравнение регрессии имеет вид:

*,*

где — переменная отклика — отвечает за количество поездок покупателей с длительностью менее получаса, произведенных за день, а — фактор — отражает среднюю температуру воздуха в этот день.

Облако точек, совмещенное с графиком уравнения регрессии:





Значения критериев Стьюдента отличны от нуля. Вероятность превысить значения критериев по модулю крайне малы. Полагаем, что гипотеза о независимости признаков ложна и отвергаем ее.

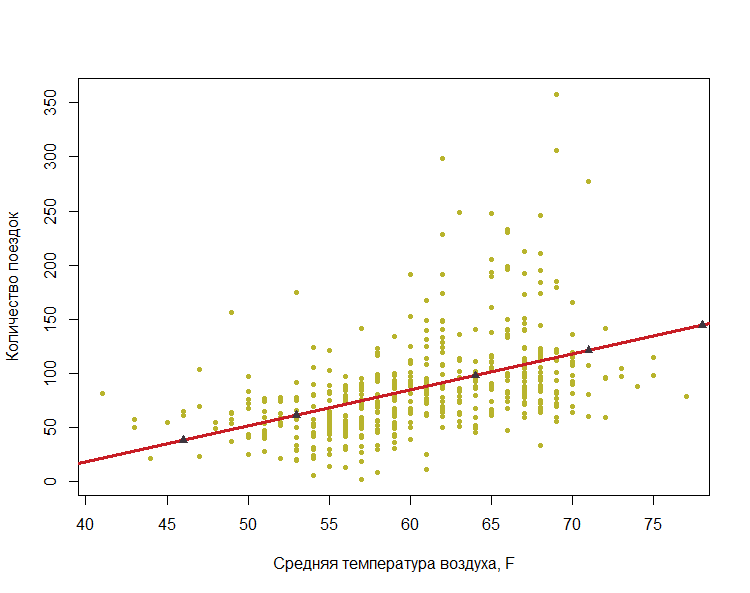
Коэффициент корреляции:

*.*

Нельзя говорить о сильной зависимости, однако она присутствует.

Для прогноза используем следующие данные:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Средняя температура воздуха за день, F** | 53 | 71 | 64 | 78 | 46 |



Темно-синие маркеры на графике — прогнозируемые значения.

Скрипт программы на R:

data <- read.table("cust\_less\_1800\_cnt\_wkds.txt",

header=TRUE,

sep=";")

na.fail(data)

data <- na.omit(data)

options(scipen=5)

cor(data.duration[c(3:19, 21:23)])

count <- data[,22]

mean\_temp <- data[,3]

plot(mean\_temp, count, col="#B8B42D", type="p",

pch=20,

xlab="Средняя температура воздуха, F",

ylab="Количество поездок")

regression <- lm(formula=count ~ mean\_temp)

regression

summary(regression)

abline(regression, col="#C81D25", lwd="3")

prognosis.temps <-c(53, 71, 64, 78, 46)

prognosis.counts <- predict(regression, data.frame(mean\_temp=prognosis.temps), level=0.9, interval="confidence")

prognosis.counts

points(prognosis.temps,

prognosis.counts[,1],

col="#3E363F",

pch=17,

add=TRUE)