#Лабораторный практикум c R

**Преподаватель**

Дорошко Ольга Валерьевна

**Студент**

Клюев Владислав Дмитриевич, 1 курс, Аналитическая логистика

**Вариант № 10**

# Задание 1. Управление данными в R

Скопировать файл “QC10.txt” (10 - номер варианта) из папки “DATA” в рабочую директорию (узнать путь к рабочей директории можно вызовом getwd() в консоли).  
Ввести данные из текстового файла согласно варианту (функция “read.table”). Сохранить в переменную qc (таблица данных).

qc=read.table("QC10.txt")

Вывести размерность таблицы данных(dim).

dim(qc)

## [1] 100 9

Вывести первые 3 записи таблицы (head).

head(qc, 3)

## V1 V2 V3 V4 V5 V6 V7 V8 V9  
## 1 61.39 56.08 4 2 1 4 5.290 47.685 5  
## 2 63.80 57.00 1 3 1 2 4.500 46.800 5  
## 3 61.59 56.93 5 1 2 4 5.285 51.330 5

Изменить имена переменных в таблице (names) согласно: “б”-“kit”, “п”-“nbug”,“№изг”-“maker”,“№пост”-“vendor”,“№вд”-“bug” “н” - последовательно проименовать как “с1”, “c2”,“c3”,“c4” (порядок переменных - согласно варианту).  
Вывести часть таблицы функцией head.

colnames(qc) = c('c1', 'c2', 'kit','nbug', 'maker', 'vendor','c3', 'c4', 'bug')  
head(qc)

## c1 c2 kit nbug maker vendor c3 c4 bug  
## 1 61.39 56.08 4 2 1 4 5.290 47.685 5  
## 2 63.80 57.00 1 3 1 2 4.500 46.800 5  
## 3 61.59 56.93 5 1 2 4 5.285 51.330 5  
## 4 58.89 53.70 3 0 1 2 4.455 49.605 3  
## 5 59.17 53.83 4 1 1 1 4.845 53.235 4  
## 6 62.15 54.80 0 1 1 4 5.845 48.510 9

Вывести общую статистику по переменной vendor (summary из Раздела 2). Обращаться к переменной через “$”.

summary(qc$vendor)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 1.75 2.00 2.49 4.00 4.00

Преобразовать vendor в переменную-фактор (c помощью as.factor, для доступа к переменной обратиться к столбцу с помощью оператора []).  
Вывести общую статистику для преобразованной переменной vendor (summary).

qc[,6] = as.factor(qc[,6])  
summary(qc[,6])

## 1 2 3 4   
## 25 28 20 27

Подсчитать количество изделий, у которых число дефектов больше 3 (функция sum и оператор условия).

sum(qc$nbug > 3)

## [1] 11

Вывести наблюдение, для которого количество некретических дефектов равно максимальному (оператор [] к переменной qc, функция max).

qc[qc$nbug == max(qc$nbug),]

## c1 c2 kit nbug maker vendor c3 c4 bug  
## 15 60.02 53.93 3 6 2 1 5.080 47.535 2  
## 72 63.10 57.19 7 6 1 1 4.925 45.405 3

# Задание 2. Описательная статистика и графический анализ

Присоединить таблицу данных qc к списку текущих переменных (attach). Далее обращаться к именам переменных напрямую.  
Загрузить дополнительную библиотеку (для вычисления коэффициентов асимметрии и эксцесса).

attach(qc)  
library("e1071")

В данном пункте статистики и графики выводятся только для переменной c1.

Вычислить среднее (mean) и медиану (median).

mean(c1)

## [1] 60.9134

median(c1)

## [1] 60.9

Вычислить выборочные квартили Q1,Q2,Q3 (quantile), а также интерквартильный размах с помощью указанной функции.

quantile(c1)

## 0% 25% 50% 75% 100%   
## 55.0600 59.1775 60.9000 62.2925 66.5400

quantile(c1)[4] - quantile(c1)[2]

## 75%   
## 3.115

Вывести комплексную статистику по переменной (summary).

summary(c1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 55.06 59.18 60.90 60.91 62.29 66.54

Вычислить дисперсию (var) для переменных с1 и c1\*2+100 в указанном порядке и сравнить результаты.

var(c1)

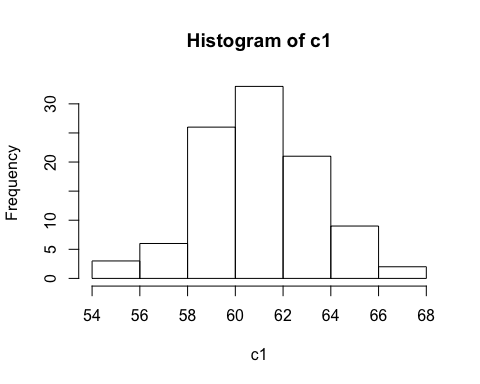
## [1] 5.880716

var(c1\*2+100)

## [1] 23.52286

**Вторая дисперсия больше в четыре раза.** Построить гистограмму (hist).

hist(c1)



Вывести асимметрию(skewness) и эксцесс (kurtosis) в указанном порядке.

skewness(c1)

## [1] 0.06109714

kurtosis(c1)

## [1] -0.3512945

Проверить нормальность остатков с помощью статистического теста Лиллиефорса (lillie.test - Раздел 2).

library(nortest)  
lillie.test(c1)

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: c1  
## D = 0.054421, p-value = 0.6607

# Задание 3. Анализ статистических зависимостей

Перекодировать переменную kit так, чтобы она принимала только два значения: 1 и 2 (1 при kit<3, 2 - иначе).

qc$kit <- ifelse(qc$kit<3,1,2)

Вычислить ранговый коэффициент корреляции Спирмена (cor,method=“spearman”) для переменных kit и maker.

cor(qc$kit,qc$maker)

## [1] 0.08173077

Вывести таблицу сопряженности (table) по переменным kit и maker.

table(qc$kit, qc$maker)

##   
## 1 2  
## 1 27 21  
## 2 25 27

Для переменных с1,с3 вычислить коэффициент корреляции с проверкой статистической значимости (cor.test).

cor.test(qc$c1, qc$c3, method="pearson")

##   
## Pearson's product-moment correlation  
##   
## data: qc$c1 and qc$c3  
## t = 5.8081, df = 98, p-value = 7.836e-08  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3437945 0.6389502  
## sample estimates:  
## cor   
## 0.5060409

Вывести корреляционную матрицу по всем числовым характеристикам с1-c4 (cor). Для выбора нужных столбцов таблицы воспользоваться оператором “[]” (выбирается сразу несколько столбцов, последовательность индексов задается с помощью функции c - Раздел 1, таблицы данных).  
Между какими переменными есть значимая корреляция? **Значимая корреляция наблюдается между меременными c1 и c2.**

cor(qc[c('c1','c2','c3','c4')])

## c1 c2 c3 c4  
## c1 1.0000000 0.9045418 0.50604092 -0.19607400  
## c2 0.9045418 1.0000000 0.38674144 -0.20389934  
## c3 0.5060409 0.3867414 1.00000000 -0.09152175  
## c4 -0.1960740 -0.2038993 -0.09152175 1.00000000

Представить матрицу корреляции в более удобном виде(symnum, в качестве аргумента данной функции использовать результат функции cor).

symnum(cor(qc[c('c1','c2','c3','c4')]),diag = FALSE)

## c1 c2 c3 c4  
## c1   
## c2 \*   
## c3 . .   
## c4   
## attr(,"legend")  
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '\*' 0.95 'B' 1

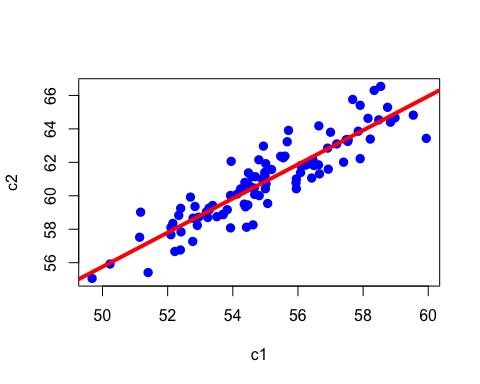
Построить модель линейной регрессии для пары переменных, для которых коэффициент корреляции имеет наибольшее значение. Использовать функцию lm, зависимая переменная - с меньшим номером. Полученную модель сохранить в переменную linmod.  
Вывести общую статистику по модели (summary).

linmod <- lm(formula = c1~c2)  
summary(linmod)

##   
## Call:  
## lm(formula = c1 ~ c2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.46433 -0.64486 -0.04164 0.58828 2.32865   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.77846 2.67499 1.786 0.0771 .   
## c2 1.01962 0.04855 21.001 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.039 on 98 degrees of freedom  
## Multiple R-squared: 0.8182, Adjusted R-squared: 0.8163   
## F-statistic: 441 on 1 and 98 DF, p-value: < 2.2e-16

Построить диаграмму рассеяния (plot) с линией регрессии (abline) для переменных, которые участвуют в оцененной модели.

plot(qc$c2, qc$c1, pch = 16, cex = 1.3, col = "blue", xlab = "c1", ylab = "c2")  
abline(linmod,col="red",lwd=4)



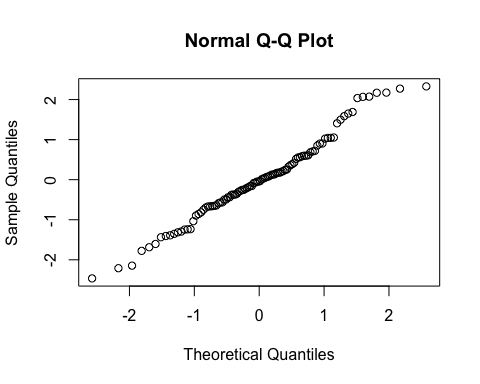
Вывести отдельно коэффициенты модели (coefficients).

linmod$coefficients

## (Intercept) c2   
## 4.778455 1.019618

Сохранить значения остатков (residuals) модели linmod в переменную res. Построить график “Квантиль-квантиль” (qqnorm) для остатков.

res = linmod$residuals  
qqnorm(res)



Проверить нормальность остатков с помощью статистического теста Колмогорова-Смирнова (ks.test - Раздел 2).

m=mean(res); s=sd(res)  
ks.test(res,pnorm,mean=m,sd=s)

##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: res  
## D = 0.073138, p-value = 0.6586  
## alternative hypothesis: two-sided

# Задание 4. Анализ неоднородных данных

Подсчитать частоты значений для переменной maker (table).

table(qc$maker)

##   
## 1 2   
## 52 48

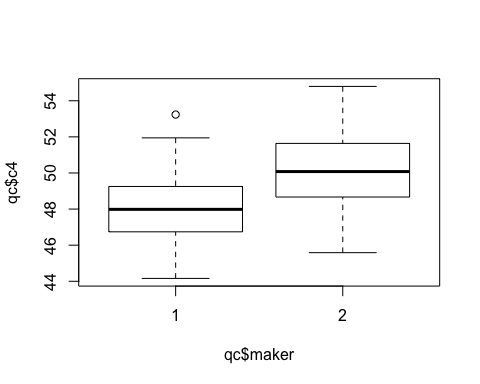
Последовательно применить двухвыборочный t-критерий (t.test) к переменным c1,c2,c3,с4. Ввборки значений каждой переменной разделяются на две подвыборки по значениям переменной maker (1 или 2).  
Вывести результаты (статистику по тесту) для той переменной (одной из c1,c2,c3,с4), для которой наблюдается значимое (на уровне 0.05) различие средних в двух подвыборках (значение p-value близко к нулю).  
Далее анализировать только выбранную переменную.

#t.test(qc$c1[qc$maker==1],qc$c1[qc$maker==2])  
#t.test(qc$c2[qc$maker==1],qc$c2[qc$maker==2])  
#t.test(qc$c3[qc$maker==1],qc$c3[qc$maker==2])  
t.test(qc$c4[qc$maker==1],qc$c4[qc$maker==2])

##   
## Welch Two Sample t-test  
##   
## data: qc$c4[qc$maker == 1] and qc$c4[qc$maker == 2]  
## t = -5.036, df = 95.109, p-value = 2.253e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2.913325 -1.265857  
## sample estimates:  
## mean of x mean of y   
## 48.03635 50.12594

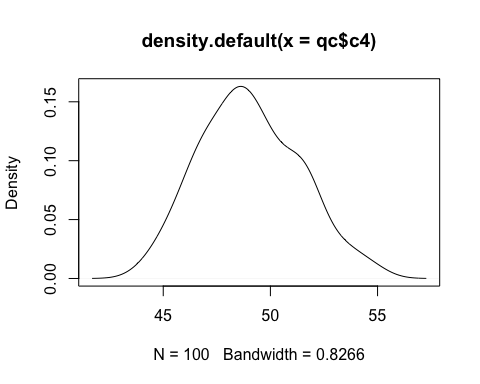
Построить “ящики с усами” для анализируемой переменной в разрезе по номеру производителя (maker).

boxplot(qc$c4~qc$maker)



Построить график ядерной оценки плотности распределения (density) для анализируемой переменной.

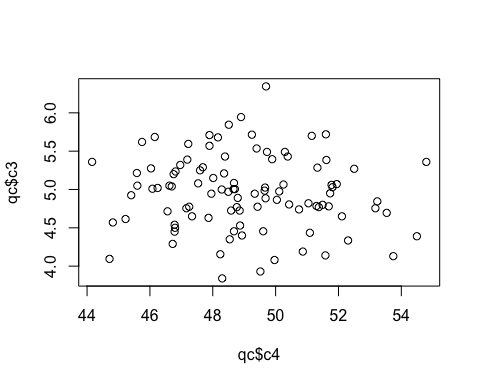
plot(density(qc$c4))



# Задание 5. Классификация неоднородных данных

Построить диаграмму рассеяния (plot) для двух переменных, которые имеют наименьшую корреляцию (см. матрицу корреляции).

plot(qc$c3~qc$c4)



Объединить выбранные переменные в одну матрицу (cbind) и сохранить ее в новую переменную. Вывести первые три строки полученной матрицы (head).

cnew = cbind(qc$c3,qc$c4)  
head(cnew,3)

## [,1] [,2]  
## [1,] 5.290 47.685  
## [2,] 4.500 46.800  
## [3,] 5.285 51.330

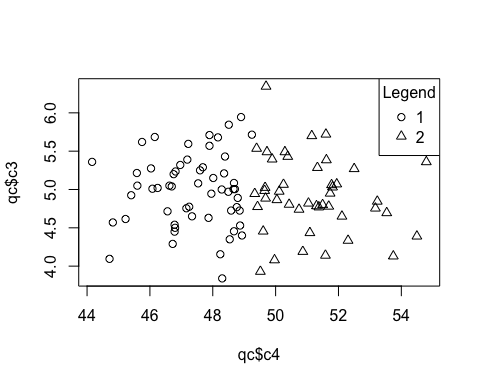
Выполнить кластерный анализ (kmeans) с разбиением на два класса в пространстве выбранных переменных (kmeans). В качестве первого аргумента функции kmeans необходимо передавать матрицу, а второго - количество классов (2), на которое производится разбиение.  
Сохранить результаты в переменную kmres и вывести частоты наблюдений в каждом оцененном классе (table).

kmres = kmeans(cnew,2)  
table(kmres$cluster)

##   
## 1 2   
## 56 44

Вывести график (plot) с легендой (legend), на котором должна быть обозначена классовая принадлежность каждого наблюдения различными символами (параметр pch).

plot(qc$c3~qc$c4, pch=kmres$cluster)  
legend("topright", c("1","2"), pch=c(1,2), title= "Legend")



Отсоединить таблицу данных qc от списка текущих переменных (работа с данными закончена).

detach(qc)