Лабораторная работа №3. Определение тесноты связи между двумя признаками. Корреляционный и регрессионный анализ

Студент:

DUBOVSKIJ JAN

Вариант

19

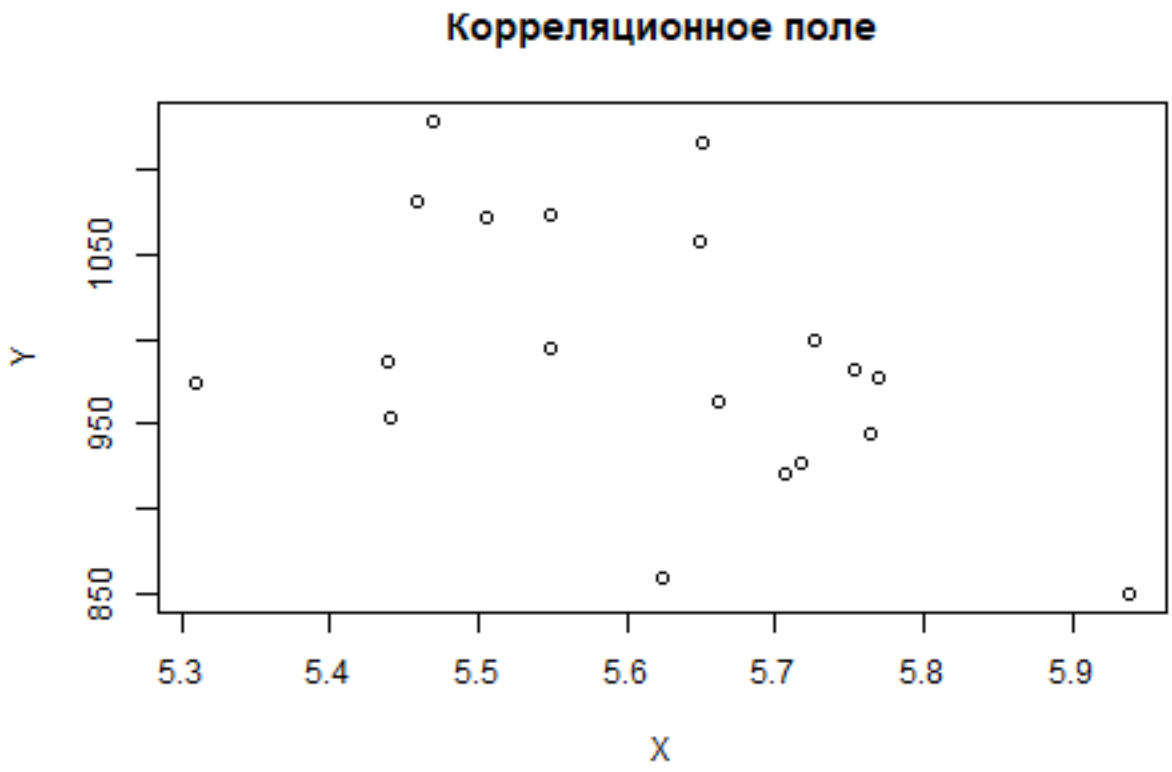
Чтение данных из файла

```
dat <- read.table(file=params$filename, dec=",")
dat
```

```
##        V1     V2
## 1  5.6242  859.0
## 2  5.9389  850.4
## 3  5.7177  926.8
## 4  5.7070  921.0
## 5  5.7637  943.9
## 6  5.4404  953.7
## 7  5.3085  973.8
## 8  5.4386  987.5
## 9  5.6613  963.6
## 10 5.7266  999.7
## 11 5.5486  994.0
## 12 5.6507 1115.3
## 13 5.7525  981.6
## 14 5.7696  977.6
## 15 5.6494 1056.9
## 16 5.5483 1073.7
## 17 5.5054 1071.6
## 18 5.4576 1080.5
## 19 5.4701 1128.5
```

Корреляционное поле

```
plot(dat,type="p",main="Корреляционное поле",xlab="X", ylab="Y")
```



Однородность и нормальность распределения факторных признаков

```
sr <- mean(dat[,1])
sigma <- sd(dat[,1])
v<-(sigma/sr)*100
v
```

```
## [1] 2.746907
```

```
rows <- paste0("(", round(sr - (1:3) * sigma, 2), ", ", round(sr + (1:3) * sigma, 2), ")")
tab1 <- rep(0, 3*4)
dim(tab1) <- c(3,4)
tab1<- as.data.frame(tab1)
tab1[,1] = rows
tab1[1,4] = 68.3
tab1[2,4] = 95.4
tab1[3,4] = 99.7
for(i in 1:3){
  counter = 0
```

```
  for (variable in dat[,1]) {
    if(sr - i * sigma < variable && sr + i * sigma > variable){
      counter = counter + 1
    }
  }
  tab1[i,2] = counter
  tab1[i,3] = (counter / length(dat[,1])*100)
}
colnames(tab1) = c("Интервалы значений признака фактора",
                   "Число единиц, входящих в интервал",
                   "Удельный вес единиц, входящих в интервал в их общем числе, %",
                   "Удельный вес единиц, входящих в интервал, при нормальном распределении, %"
tab1
```

```
##   Интервалы значений признака фактора Число единиц, входящих в интервал
## 1                        (5.46, 5.77)                                 13
## 2                        (5.31, 5.92)                                 18
## 3                        (5.15, 6.08)                                 19
##   Удельный вес единиц, входящих в интервал в их общем числе, %
## 1                                                     68.42105
## 2                                                     94.73684
## 3                                                    100.00000
##   Удельный вес единиц, входящих в интервал, при нормальном распределении, %
## 1                                                                      68.3
## 2                                                                      95.4
## 3                                                                      99.7
```

Выводы: … Поскольку коэффициент вариации <33%, то выборка считается однородной по исследуемому признаку.При сопоставлении 3 и 4 столбцов получаем наличие нормальности распределения факторных признаков.

Аналитическая группировка

```
k = 1 + round(log(length(dat[,1]), 2), 0)
h = (max(dat[,1]) - min(dat[,1])) / k
tab2 = rep(0, k*4)
dim(tab2) = c(k,4)
tab2 = as.data.frame(tab2)
rows2 = paste("(", min(dat[,1]) + (1:k) * h - h, ", ", min(dat[,1]) + (1:k) * h, ")", sep = "
tab2[,1] = rows2
for(i in 1:(k-1)){
  counter = 0
  sm = 0
  for(j in 1:length(dat[,1])){
    if(min(dat[,1]) + (i - 1) * h <= dat[j,1] && min(dat[,1]) + i * h > dat[j,1]){
      counter = counter + 1
      sm = sm + dat[j,2]
    }
  }
  tab2[i,2] = counter
```

```
  tab2[i,3] = sm
}
counter = 0
sm = 0
for(variable in dat[,1]){
  if(min(dat[,1]) + (k - 1) * h <= variable && min(dat[,1]) + k * h >= variable){
    counter = counter + 1
    sm = sm + dat[j,2]
  }
}
tab2[k,2] = counter
tab2[k,3] = sm
for (i in 1:k) {
  tab2[i,4] = tab2[i,3] / tab2[i,2]
}
colnames(tab2) = c("Интервалы",
                   "Число вариант, попавших в i-ый интервал",
                   "Сумма результирующего фактора i-ого интервала",
                   "Средняя величина результирующего фактора в группе")
tab2
```

```
##            Интервалы Число вариант, попавших в i-ый интервал
## 1  (5.3085, 5.43458)                                        1
## 2 (5.43458, 5.56066)                                        7
## 3 (5.56066, 5.68674)                                        4
## 4 (5.68674, 5.81282)                                        6
## 5  (5.81282, 5.9389)                                        1
##   Сумма результирующего фактора i-ого интервала
## 1                                         973.8
## 2                                        7289.5
## 3                                        3994.8
## 4                                        5750.6
## 5                                        1128.5
##   Средняя величина результирующего фактора в группе
## 1                                          973.8000
## 2                                         1041.3571
## 3                                          998.7000
## 4                                          958.4333
## 5                                         1128.5000
```

Корреляционный анализ

```
dat[,1] -> x
dat[,2] -> y
cor(x,y)
```

```
## [1] -0.457757
```

```
cor.test(x,y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = -2.1229, df = 17, p-value = 0.04875
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.754990578 -0.004479029
## sample estimates:
##       cor
## -0.457757
```

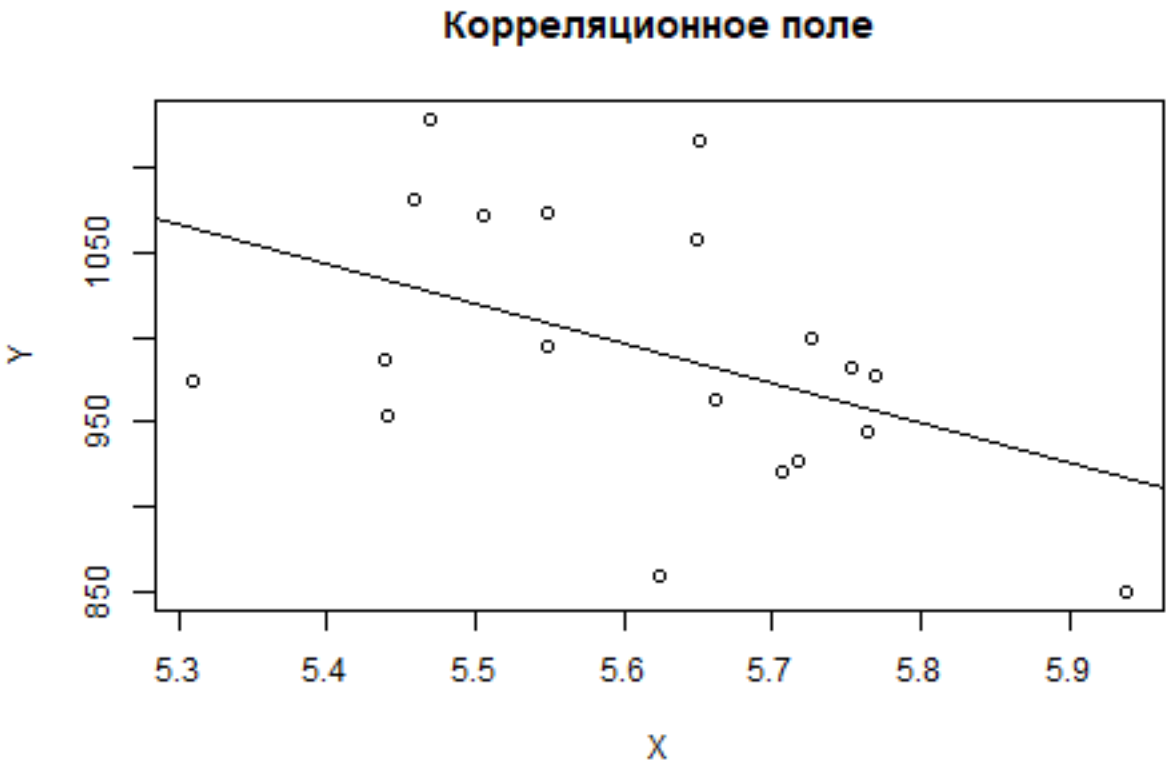
Выводы: …p-уровень значимости равен 0.04875 > 0,01, а значит нельзя отклонить гипотезу о равенстве коэффициента корреляции нулю. следовательно корреляция не является значимой

Регрессионный анализ

```
x = dat[,1]
y = dat[,2]
lm(y ~ x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      2302.6       -233.3
```

```
plot(dat,type="p",main="Корреляционное поле",xlab="X", ylab="Y")
abline(lm(y ~ x))
```


Корреляционное поле
Y
1050
950
850
5.3
5.4
5.5
5.6
5.7
5.8
5.9
X