Выполнение корелляционного и регрессионного анализа для 2-мерной выборки.

Пенкин С.В. ИВТ-21

Вариант 16

Задание:

1) По выборке значений двумерной нормально распределенной СВ найти точечную оценку коэффициента корреляции её компонент. При уровне значимости 0.05 проверить гипотезу о независимости компонент X и Y.

2) С помощью МНК получить эмпирическое уравнение линейной регрессии Y на X. Построить график эмпирической прямой с изображением на нем выборочных точек.

	5	10	15	20	25	30	m
8	2	4	0	0	0	0	6
12	0	3	7	0	0	0	10
16	0	0	5	30	10	0	45
20	0	0	7	10	8	0	25
24	0	0	0	5	6	3	14
n	2	7	19	45	24	3	100

Выборочный коэффициент корелляции.

Найдем $\bar{x}, \bar{y}, \overline{xy}$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i x_i \qquad \bar{y} = \frac{1}{n}\sum_{j=1}^{l} n_j y_j \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij} x_i y_j$$

```
# обьем выборки
n = 100
# переменные для вычислений
x = c() ; xkv = c() ; y = c() ; ykv = c() ; xy = c()
count = 1
# вычисление x, y, xy
for (i in 1:length(cols)) {
    for (k in 1:length(rows)) {
        x[count] = rows[k] * df[k,i]
        xkv[count] = rows[k] ^ 2 * df[k,i]
        y[count] = cols[i] * df[k,i]
        ykv[count] = cols[i] ^ 2 * df[k,i]
        xy[count] = rows[k] * cols[i] * df[k,i]
        count = count + 1
    }
}
x = sum(x) / n ; x
```

```
## [1] 17.24
```

```
y = sum(y) / n ; y
```

```
## [1] 19.55
```

```
xy = sum(xy) / n ; xy
```

```
## [1] 350.8
```

Найдем σ_x и σ_y:

$$\sigma_x = \sqrt{\bar{x^2} - (\bar{x})^2} \qquad \sigma_y = \sqrt{\bar{y^2} - (\bar{y})^2} \qquad \bar{x^2} = \frac{1}{n}\sum_{i=1}^{k} m_i x_i^2 \qquad \bar{y^2} = \frac{1}{n}\sum_{j=1}^{l} n_j y_j^2$$

```
xkv = sum(xkv) / n ; xkv
```

```
## [1] 314.08
```

```
ykv = sum(ykv) / n ; ykv
```

```
## [1] 407.25
```

```
vx = sqrt(xkv - x ^ 2) ; vx
```

```
## [1] 4.106385
```

```
vy = sqrt(ykv - y ^ 2) ; vy
```

```
## [1] 5.004748
```

Найдем выборочный коэффициент корелляции:

$$r = \frac{\overline{xy} - \bar{x} * \bar{y}}{\sigma_x * \sigma_y}$$

```
r = (xy - x * y) / (vx * vy) ; r
```

```
## [1] 0.6694427
```

Проверка гипотезы о независимости компонент X и Y с помощью статистики Стьюдента.

$$H_0 : r = 0$$

Найдем наблюдаемое значение критерия:

$$T_n = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

```
# уровень значимости
a = 0.05
# число степеней свободы
k = n - 2
# критический критерий (по таблице)
Tk = 1.985

Tn = r * (sqrt(k) / sqrt(1 - r ^ 2)) ; Tn
```

```
## [1] 8.921078
```

Так как $T_k < T_n$, то нулевая гипотеза о равенстве нулю генерального коэффициента корелляции отвергается. Следовательно X и Y кореллированы.

Эмпирическое уравнение линейной регрессии Y на X.

$$y = ax + b$$

$$F(a,b) = \sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

$$\frac{dF}{da} = 2\sum_{i=1}^{n}(ax_i^2 + bx_i - x_i y_i) \qquad \frac{dF}{db} = 2\sum_{i=1}^{n}(ax_i + b - y_i)$$

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases} \qquad \begin{cases} a6400 + b80 = 6949.968 \\ a80 + b5 = 86.8746 \end{cases}$$

Методом Крамера получаем:

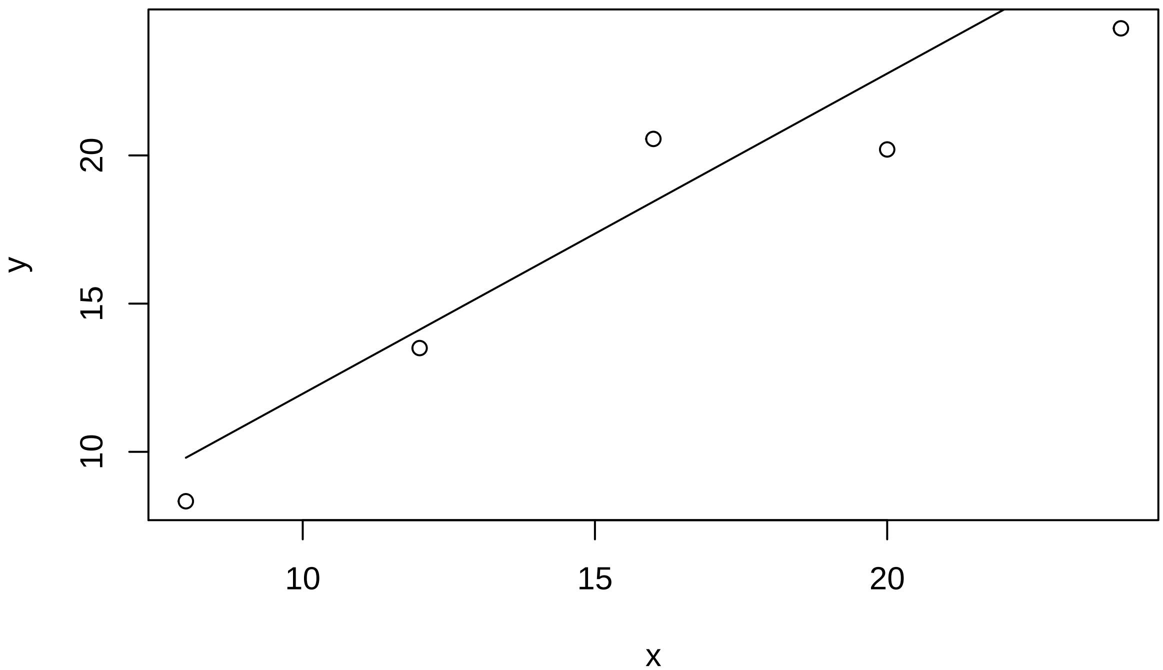
$$a = 1.08, b = 1.16$$

```
# переменные для расчетов
tmp = c(); y = c()
m = c(6,10,45,25,14)
# расчет y для координат графика
for (i in 1:length(rows)) {
    for (k in 1:length(cols)) {
        tmp[k] = cols[k] * df[i,k]
    }
    y[i] = sum(tmp) / m[i]
}
```

```
a = 1.08
b = 1.16

y1 = a * rows + b

plot(xy.coords(rows, y), xlab = "x", ylab = "y")
lines(rows, y1)
```


10
15
20
x
10
15
20
y