Лабораторная работа №6
Корреляция и регрессия

Коэффициенты корреляции

Теория

Если нам дана двумерная выборка (X_i, Y_i) $(i = \overline{1,n})$, то можно оценить степень линейной зависимости между случайными величинами X_i и Y_i. Для этого есть коэффициент линейной корреляции Пирсона:

$$r = \sum_{i=1}^{n} \frac{x_i - m_x}{s_x} \cdot \frac{y_i - m_y}{s_y},$$

где $m_x = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$, $m_y = \frac{1}{n}\sum\limits_{i=1}^{n} y_i$, $s_x^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i - m_x)^2$, $s_y^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(y_i - m_y)^2$.

Если же данные измерены в порядковой шкале или если от исходных данных, измеренных в шкалах разностей или отношений, был произведен переход к порядковой шкале с помощью ранжирования, то вычисляются коэффициенты корреляции Спирмена

$$\rho = 1 - \frac{6\sum\limits_{i=1}^{n}(a_i - b_i)^2}{n^3 - n}$$

и Кендалла

$$\tau = \frac{4R}{n^2 - n} - 1$$

(здесь a_i и b_i — ранги элементов x_i и y_i соответственно, полученные при ранжировании значений выборок x_i и y_i каждой в отдельности, $i \in \overline{1,n}$, $R = R_1 + R_2 + \cdots + R_{n-1}$; чтобы получить R_i, необходимо выборку значений (x_i, y_i) упорядочить по возрастанию рангов элементов x_i, выписать этот порядок в вертикальную таблицу, и для элемента y_i подсчитать количество R_i рангов, больших ранга элемента y_i, элементов, находящихся ниже него по этой таблице).

Для вычисления указанных коэффициентов (приведем синтаксис команды, рассчитывая, что в выборках `x` и `y` или датафрейме `x` нет значений `NA`) используется

```
cor(
    x,
    y = NULL,
    method = c("pearson", "kendall", "spearman")
   )
```

Для проверки гипотезы H_0 : Коэфф. корр. $= 0$ (где Коэфф. корр. — какой-либо из коэффициентов корреляции), используется команда

```
cor.test(
         x,
         y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         exact = NULL,
         conf.level = 0.95,
         continuity = FALSE,
     ...)
```

Для исследования линейной регрессии (вектора `y` на вектор `x`) используются команды:

- `lm(x~y)` генерирует модель линейной регрессии и выводит коэффициенты линии регрессии (если только результат не присвоен в некоторую переменную)
- `summary.lm(linmod)` если модель линейной регрессии присвоена в переменную `linmod`, то указанная функция выводит суммарную информацию по этой модели
- `linmod[i]` если модель линейной регрессии присвоена в переменную `linmod`, то указанная функция выводит `i`-й объект модели линейной регрессии

Для визуализации модели линейной регрессии (вектора `y` на вектор `x`) используется последовательность команд:

```
linmod<-lm(x~y);
plot(x,y);
abline(linmod);
```

Задания

Задание 1. Загрузите данные из файла `связи.csv` и найдите коэффициенты корреляции между выборками, значения которых находятся в указанном файле. Проверьте значимость коэффициентов корреляции.

Задание 2. По данным из файла `связи.csv` постройте линейную модель регрессии вектора `y` на вектор `x`. Получите суммарную информацию по модели. Определите, сколько объектов находится «внутри» этой модели (каково последнее `i` для `linmod[i]`) и попытайтесь понять, что они значат. Постройте визуализацию линейной модели.