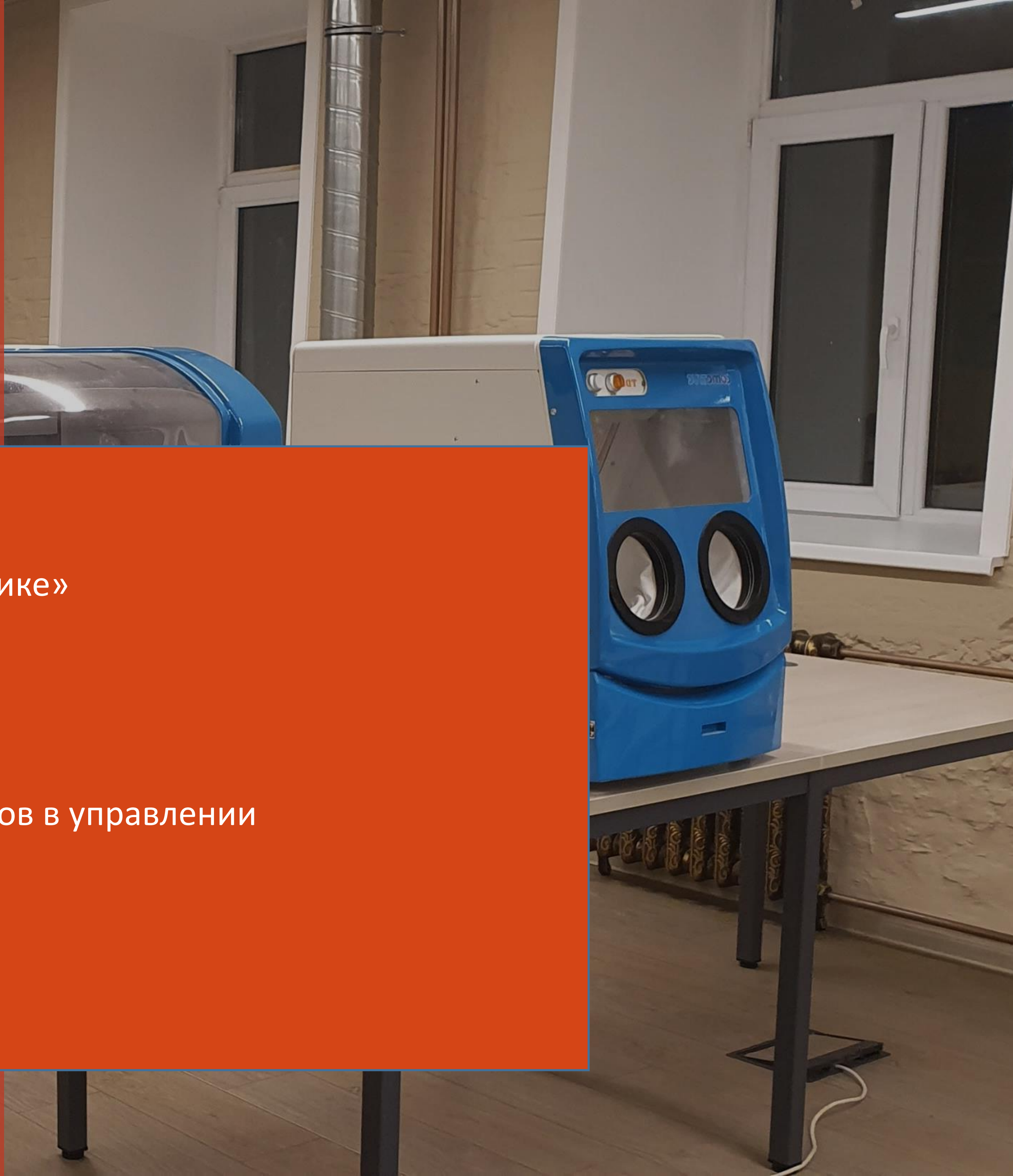
МИРЭА
Российский технологический университет

Институт технологий управления



Программные средства анализа данных

01.03.05 Статистика

Профиль «Анализ данных в бизнесе и экономике»

Квалификация «бакалавр»

Бурцева Татьяна Александровна

Профессор кафедры статистики и математических методов в управлении

Burceva_t@mirea.ru

Москва, 2022

МИРЭА
Российский технологический университет

Тема 3. Корреляционный анализ в R

План лекции

1. Коэффициенты корреляции.
2. Проверка гипотез и обоснование статистической значимости.
3. Визуализации парной и множественной взаимосвязи переменных.

МИРЭА
Российский
технологический
университет

1. Коэффициенты корреляции

Статистические показатели, позволяющие определить, тесноту связи (в одном случае она сильная, устойчивая, в другом - слабая) и форму связи (прямая, обратная, линейная, нелинейная) между признаками.

Факторные связи *между признаками* характеризуются тем, что они проявляются в согласованной вариации изучаемых показателей.
При этом одни показатели выступают как факторные, а другие - как результативные.
Факторные связи могут рассматриваться как функциональные и корреляционные.

Функциональная зависимость (functional dependency)

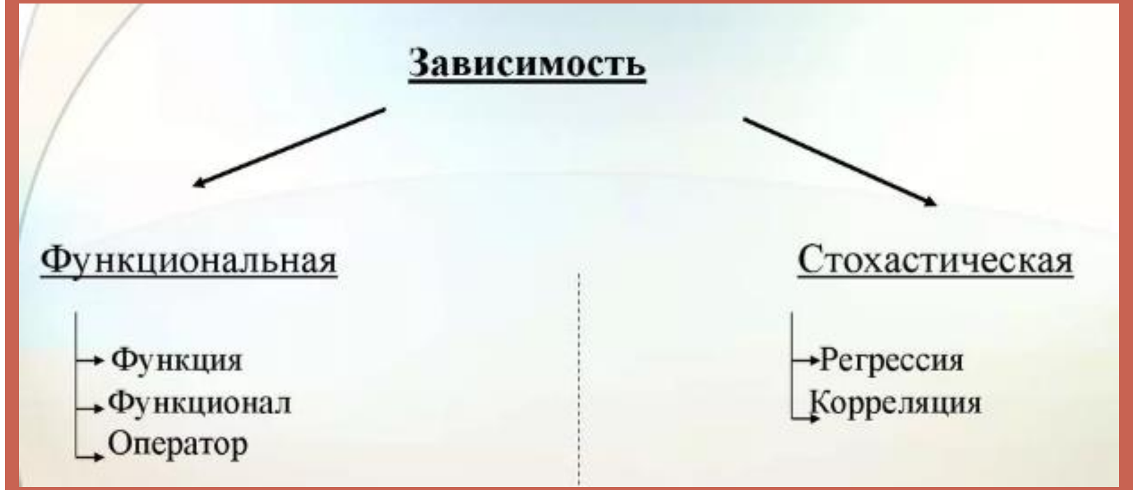
Российский технологический университет

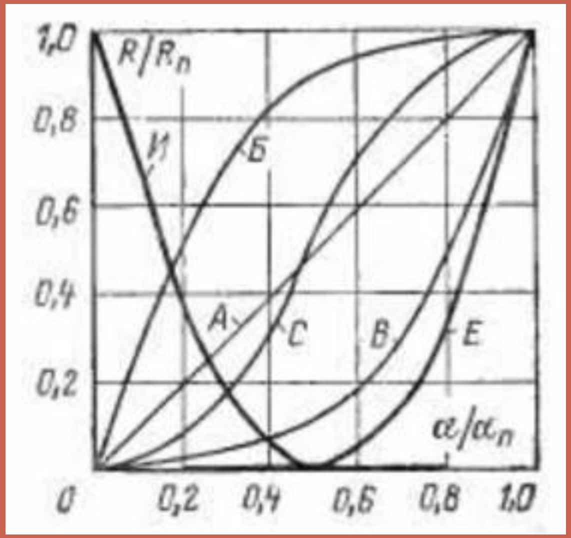
Определение. Функциональной зависимостью между двумя случайными величинами называется зависимость, при которой каждому значению одной переменной соответствует вполне определенное значение другой переменной.

Примеры: 1) $Y = X^2$; 2) $Y = aX + b$;
3) скорость падения от времени; 4) стоимость проданных изделий от их числа.

При функциональной связи изменение результативного признака (y) всецело зависит от изменения факторного признака (x): $y = f(x)$.

Коэффициент корреляции равен 1 или -1.





Является ли зависимость функциональной?
- Температура от времени суток – **да**
- Каждому ученику школы поставлено в соответствие 4-значное число, соответствующее году рождения - **да**
- Каждому дню в году поставлен в соответствие ученик школы, родившийся в этот день – **нет** ;

А – линейная, Б – логарифмическая, В – обратнологарифмическая, С – S-образная, тип Е, тип И

Корреляционная связь

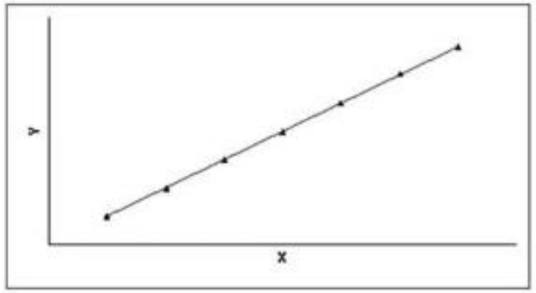
<u>Стохастическая или вероятностная</u> зависимость – такая форма связи, когда при фиксированном значении одной величины другая величина может принимать различные значения.

При корреляционной (статистической) связи изменение результативного признака (*y*) не всецело зависит от факторного признака (*x*), а лишь в среднем, так как возможно влияние прочих факторов (ε): $y=\varphi(x)+\varepsilon$.

Характерной особенностью корреляционных связей является то, что они проявляются не в единичных случаях, а в массе, т.е. в среднем.

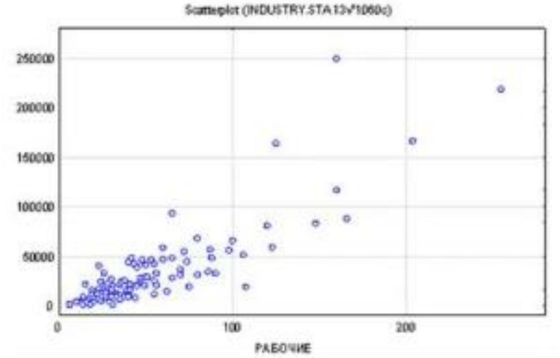
Функциональная зависимость:

X→Y



Статистическая зависимость:

X →Y_1, Y_2, … ,Y_n



Показатели тесноты связи между признаками называются

КОЭФФИЦИЕНТАМИ КОРРЕЛЯЦИИ

(их выбор зависит от вида представления исследуемых признаков (шкал))

- Коэффициент корреляции
  - Количественные шкалы (признаки количественные)
    - Линейный коэффициент корреляции rxy
    - Корреляционное отношение η
  - Порядковые шкалы (признаки представлены рангами)
    - Коэффициент Спирмена p
    - Коэффициент конкордации w
  - Номинальные шкалы (признаки качественные)
    - Коэффициенты сопряжённости Пирсона и Чупрова
    - Коэффициенты ассоциации и Контингенции ka, kk

Область значений коэффициентов корреляции

Коэффициенты корреляции изменяются от -1 до 1.

Знак коэффициента корреляции характеризует направление взаимосвязи, если он положительный, то связь между признаками прямая, и наоборот, если знак отрицательный, то связь обратная.

Абсолютная величина коэффициента характеризует степень тесноты рассматриваемой связи. Если она равна 1, то связь функциональная, если 0, то связи нет.

Если коэффициент корреляции возвести в квадрат, то получится коэффициент детерминации (изменяется от 0 до 1 и характеризует долю влияния фактора на результат).

ЛИНЕЙНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

В качестве оценки генерального коэффициента корреляции используют коэффициент корреляции r Пирсона:

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

где x_i - значения, принимаемые в выборке X,

y_i - значения, принимаемые в выборке Y.

$$\boxed{r_b = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\delta_x \cdot \delta_y}}$$

где $\overline{XY}$ – среднее значение произведений X на Y,

$\overline{X}, \overline{Y}$ – средние значения соответствующих признаков,

δ_x, δ_y – средние квадратические отклонения, найденные для признака X и для признака Y.

$$x_1, \ldots, x_n \quad y_1, \ldots, y_n$$

r_{xy} измеряет тесноту линейной связи между двумя и более признаками.

где n объем выборки,

x и y- значения признаков.

$$\operatorname{cov}(x, y) = \overline{x \cdot y} - \bar{x} \cdot \bar{y}$$ выборочная ковариация

$$\operatorname{cor}(x, y) = \frac{\operatorname{cov}(x, y)}{\sqrt{s_x^2 \cdot s_y^2}}$$ выборочный коэффициент корреляции

Свойства выборочный коэффициента корреляции

1. $r_{xy} = r_{yx}$.
2. $-1 \le r^*_{xy} \le 1$.
3. Если $\left|r^*_{xy}\right| = 1$ тогда и только тогда, когда между значениями X, Y имеется линейная зависимость.
4. Если r^*_{xy} =0, то между X, Y отсутствует линейная корреляционная связь, но возможно наличие между ними другого типа связи.
5. Если r^*_{xy}>0, то увеличение признака X в среднем приводит к увеличению признака Y. Если r^*_{xy}<0, то с увеличением X в среднем признак Y уменьшается.

Стандартную ошибку коэффициента корреляции находят по формуле

$$\boxed{\delta_r = \sqrt{\frac{1 - r^2}{n - 2}}},$$

где n - объем выборки.

С увеличением n уменьшается δ_r и возрастает точность определения r.

При небольших объемах выборки часто используют более предпочтительные оценки коэффициентов корреляции и детерминации, чем выборочные коэффициенты:

- **более предпочтительная оценка коэффициента корреляции** –

$$\tilde{r}^2 = r(1 + \frac{1 - r^2}{2 \cdot (n-4)}),$$

- **более предпочтительная оценка коэффициента детерминации**

$$\tilde{r}^2 = \frac{(n-1) \cdot r^2 - 1}{n-2}),$$

МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Если необходимо проанализировать линейную связь между результативным признаком (y) и двумя факторными признаками (x, z), тогда используется формула расчета множественного линейного коэффициента корреляции R_{yxz}, *изменяется от 0 до 1.*

$$R_{yxz} = \sqrt{\frac{r_{yx}^2 + r_{yz}^2 - 2 \cdot r_{yx} \cdot r_{yz} \cdot r_{xz}}{1 - r_{xz}^2}}$$

МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Если число факторов-признаков более двух, тогда совокупный множественный коэффициент корреляции имеет вид:

$$R_{yxz\ldots k}=\sqrt{1-\frac{\Delta}{\Delta^{*}}},$$

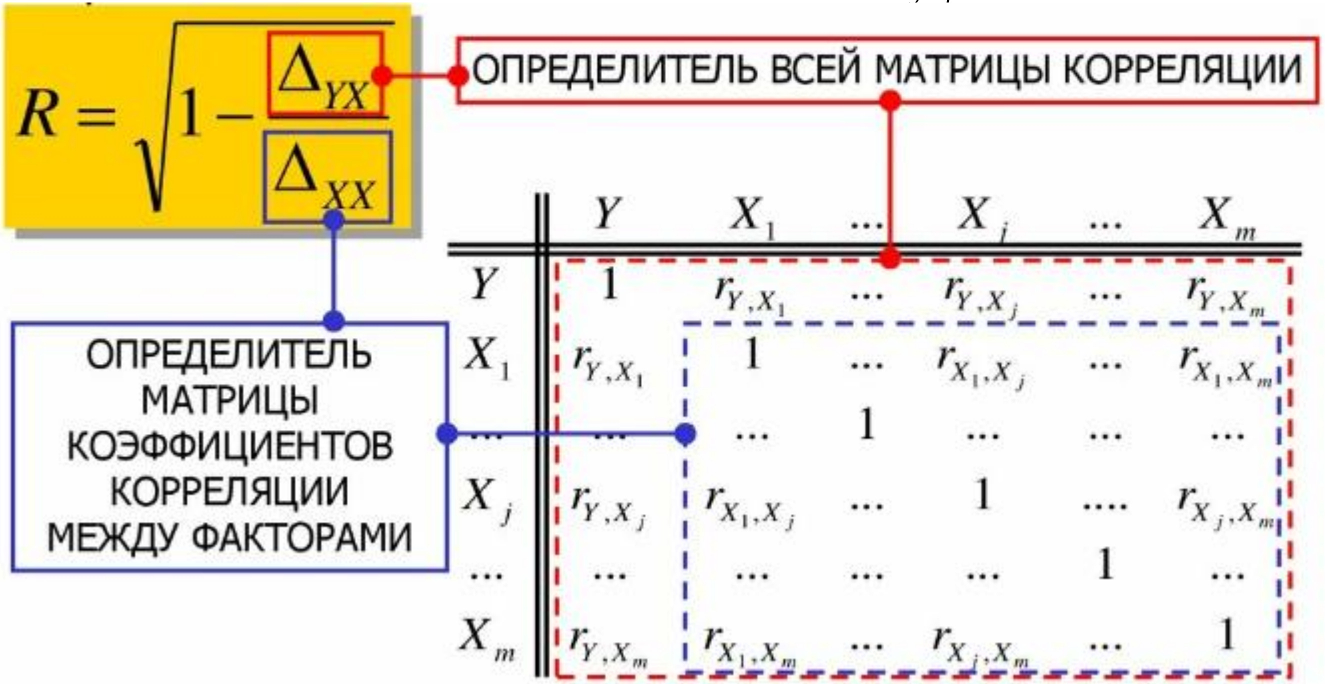
где Δ – *детерминант матрицы парных коэффициентов корреляции,* Δ^{*} – *детерминант этой матрицы без верхней строки и первого столбца, то есть без* r_{yx_i}

R^2 показывает в какой мере вариация результирующего признака обусловлена совместным влиянием признаков-факторов

R – изменяется от 0 до 1,

существенность также проверяется

с помощью критерия Фишера



ЧАСТНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Позволяет установить степень «чистого» влияния факторного признака на результативный признак, при условии, что остальные факторы не влияют, изменяется от 0 до 1, не может быть больше по величине коэффициента множественной корреляции.

$$r_{yx_k}(x_1, x_2 \ldots x_{k-1}) = \sqrt{\frac{R_k^2 - R_{k-1}^2}{1 - R_k^2}}$$

Где R^2k – коэффициент множественной детерминации между y и $x_1 \ldots x_к$;

$R^2{}_{k-1}$ – коэффициент множественной детерминации между y и $x_1 \ldots x_{к-1}$;

ЧАСТНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Если парный коэффициент корреляции между x и y больше частного коэффициента корреляции между x и y, то существует фактор, усиливающий влияние x на y, если наоборот, то существует фактор, ослабляющий это влияние. В R используется функция pcor().

Формат применения этой функции таков:

```
pcor(u, S)
```

где *u* – это числовой вектор, в котором первые два числа – это номера переменных, для которых нужно вычислить коэффициент, а остальные числа – номера «влияющих» переменных (воздействие которых должно быть отделено). *S* – это ковариационная матрица для всех этих переменных. Проиллюстрируем это на примере.

```
> library(ggm)
> # частная корреляция между численностью населения и уровнем
> # преступности, освобожденная от влияния дохода, доли
> # неграмотного населения и долей людей со средним образованием
> pcor(c(1,5,2,3,6), cov(states))
[1] 0.346
```

В данном случае 0.346 – это коэффициент корреляции между численностью населения и уровнем преступности без влияния дохода, доли неграмотного населения и долей людей со средним образованием. Частные корреляции обычно используются в социологии.

Функция hetcor() из пакета polycor позволяет вычислять комбинированную корреляционную матрицу, содержащую коэффициенты корреляции Пирсона для числовых переменных, многорядные корреляции между числовыми и порядковыми переменными, полихорические корреляции между порядковыми переменными и тетрахорические корреляции между двумя дихотомическими переменными. Многорядные, полихорические и тетрахорические корреляции могут быть вычислены для порядковых и дихотомических переменных, которые происходят из нормального распределения. Дополнительную информацию об этих типах корреляций можно получить из справочного материала для данного пакета.

Пример:

Подставим данные в формулу и найдем *r* :

$$\operatorname{cov}(x, y) = \overline{x \cdot y} - \overline{x} \cdot \overline{y}$$

$$s_x^2 = \overline{x^2} - \left(\overline{x}\right)^2$$

$$s_y^2 = \overline{y^2} - \left(\overline{y}\right)^2$$

Ковариация	17,47
Выборочная дисперсия по x	3,14
Выборочная дисперсия по y	114,47
Коэффициент корреляции	0,92

Ответ. Значение коэффициента корреляции равно 0,92. Это означает, что существует сильная положительная связь.

РАНГОВЫЕ КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ

Ранг - это порядковый номер значений признака, расположенных в порядке возрастания или убывания их величин. Если отдельные значения признака имеют одинаковую количественную оценку, то ранг всех этих значений принимается равным средней арифметической от соответствующих им номеров. Данные ранги называются связанными. Ранговые показатели связи используются для ее оценки как между количественными, так и между качественными признаками, если их значения могут быть проранжированы. Наиболее распространены ранговые парный коэффициент Спирмена (ρ) и множественный коэффициент конкордации (w) .

КОЭФФИЦИЕНТ СПИРМЕНА

Когда нет связных рангов рассчитывается по формуле:

$$\rho = 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)}$$

где *di²* - квадраты разности рангов; *n* - число наблюдений.

Если в совокупности есть связные ранги, то

$$\rho = 1 - \frac{6\sum d_i^2 - T_x - T_y}{\sqrt{(n^3 - n - 12T_x)(n^3 - n - 12T_y)}}$$

где $T = \frac{1}{12}\sum^k (t^3 - t)$ по x и по y соответственно, *к* - число связных рангов,

t - число значений признака, имеющих один ранг.

Значимость его проверяется на основе t-критерия Стьюдента:

$$t_p = \rho \times \sqrt{\frac{n-2}{1-\rho^2}}$$

Если расчётное значение критерия больше табличного *t(α ; к = n-2),*

то значение коэффициента корреляции считается значимым.

Функции cor.test(), cov(), cor() и cov2cor() в R

cor.test() позволяет вычислить коэффициент корреляции для заданных выборок

cov() позволяет построить ковариационную матрицу для заданных выборок,

cor() — строит матрицу коэффициентов корреляции.

cov2cor() создаёт корреляционную матрицу на основе заданной ковариационной.

Описание функции

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

Параметры

x	Вектор, матрица или data.frame
y	Второй вектор (или NULL, если первый аргумент – матрица или фрейм данных)
method	Вычисляемый коэфициент корреляции (по умолчанию – pearson)

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> y<-c(2.7,8.9,6.5,8.8,6.4)
> cor(x, y)
0.4668
```

Аргументы функций

В R мы можем использовать функцию cor (). Требуется три аргумента и метод.

```
cor(x, y, method)
```

Аргументы :

- x: первый вектор
- y: второй вектор
- Метод: формула, используемая для вычисления корреляции. Три строковых значения:
  - «Pearson»
  - «Kendall»
  - «Копьеносец»

Копьеносец – ρ Спирмена

- **x** и **y** — числовые векторы, матрицы или таблицы данных, причём аргумент **x** — обязательный.
- **na.rm** — логический аргумент — позволяет исключать из рассмотрения отсутствующие значения — **NA**.
- **use** — дополнительный символьный аргумент, определяющий как вычислять ковариацию или коэффициент корреляции при отсутствующих значениях (**NA**). Его возможные значения (полностью или сокращённо):
  - «**everything**» (по умолчанию) — **NA** остаются в выборке и учитываются при нахождении выборочной ковариации (коэффициента корреляции). Если **NA** есть, то результатом будет также **NA**.
  - «**all.obs**» — **NA** остаются ка элементы выборки, но по вычислении выводится сообщение об ошибке
  - «**complete.obs**» — **NA** не рассматриваются при вычислениях. Но если вся выборка состоит из **NA**, то выводится сообщение об ошибке.
  - «**na.or.complete**» — аналогично предыдущему, но в случае, если вся выборка состоит из **NA**, результатом будет **NA**.
  - «**pairwise.complete.obs**» — при нахождении ковариации или корреляции, если хотя бы одна из пары переменных принимает значение **NA**, то вся пара значений отбрасывается. Используется для **cov()**. если только **method**=«**pearson**».
- **method** — символьный аргумент, определяющий на основе какого метода нужно вычислять коэффициент корреляции. Названия методов (полностью или сокращённо):
  - «**pearson**» (по умолчанию) — вычисление обычной выборочной ковариации или коэффициента корреляции.
  - «**kendall**» и «**spearman**» — ранговые коэффициенты корреляции.
- **V** — симметричная числовая матрица (положительно определённая), рассматриваемая в качестве матрицы ковариаций и преобразуемая в матрицу коэффициентов корреляции.

Если в ваших данных есть одинаковые наблюдения, но вы хотите рассчитать непараметрическую корреляцию, используйте функцию spearman_test из пакета coin

library(coin) spearman_test(~ mpg + disp, mtcars)


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

Параметры

`x, y`	Числовые вектора x и y одинаковой длины .
`alternative`	Выбирает альтернативную гипотезу одну из "two.sided" (по умолчанию)-двустороняя критическая область, "greater" -правостороняя критическая область или "less"-левостороняя критическая область.
`method`	Выбирает какой коэфициент корреляции используется в тесте. Один из"pearson", "kendall", или "spearman".
`conf.level`	Доверительная вероятность

Примечание

Для проверки нулевой гипотезы H0 о равенстве показателя корреляции нулю необходимо в `alternative` выбрать `"two.sided"`.

Критическое значение находят по таблице критических точек распределения Стьюдента с числом степеней свободы $f = n - 2$ (в R используется функция вычисления квантилей распределения Стьюдента `qt(p,df)`).

- ПРИМЕРЫ
- cor.test(x = t$X, y = t$У)
- cor.test(x = t$X, y = t$У, method = "spearman")
- cor.test(x = t$X, y = t$У, method = "kendall")
- cor.test(x = t$X[t$Status=="Регионы-дотационные"], y = t$У[t$Status=="Регионы-дотационные"])

2. Проверка гипотез и обоснование статистической значимости

Приведем список некоторых основных пакетов, содержащих стандартные статистические тесты (многие критерии находятся в пакете stats, который загружается автоматически):

`ctest` - классические тесты (Фишера, "Стьюдента", Пирсона, Бартлетта, Колмогорова- Смирнова...)

`eda` - методы, используемые в "Разведочном анализе данных"

`lqs` - регрессия и оценка ковариации

`modreg` – современные методы построения регрессионных моделей: сглаживание и локальные регрессии

`mva` - многомерный анализ

`nls` – нелинейные модели регрессии

`splines` - сплайны

`stepfun` - эмпирические функции распределения

`ts` - исследования временных рядов

Для загрузки пакета используется фукция: **`library()`** с именем соответствующего пакета:

```
> library(eda)
```

При помощи функции cor.test() одновременно можно проверить значимость только одного коэффициента корреляции. В пакете psych есть функция corr.test(), с ее помощью можно вычислить коэффициенты корреляции Пирсона, Спирмена и Кэнделла между несколькими переменными и оценить их достоверность

Оценка значимости r_{xy}, ρ_{xy}

Выборочный r_{xy} рассчитывается по конечному набору данных, поэтому необходимо проверить гипотезу о не случайности связи, то есть что $r_{xy} \neq 0$.

$$t_{pac} = \sqrt{\frac{r^2}{1-r^2}(n-2)}$$

Для этого используется статистический t-критерий Стьюдента:

Если $t_{pac} > t_{\alpha}(n-2)$, то связь между признаками существенная с вероятностью 1-ą. То есть, гипотеза о том, что $r_{xy}=0$ отвергается и связь между признаками значима с вероятностью 1-ą.

Обозначения:

Выборочный коэффициент корреляции Спирмена r_s

Коэффициент корреляции генеральной совокупности ρ_s

Требуется:

Проверить гипотезу о равенстве нулю коэффициента ранговой корреляции генеральной совокупности на основании значения коэффициента ранговой корреляции выборки:

$$H_0 : \rho_s = 0$$

$$H_1 : \rho_s \neq 0$$

Примеры

```
> cor.test(x = t$X, y = t$Y, method = "spearman")

        Spearman's rank correlation rho

data:  t$X and t$Y
S = 74676, p-value = 0.02534
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2439432
```

```
> cor.test(x = t$X, y = t$Y, method = "kendall")

        Kendall's rank correlation tau

data:  t$X and t$Y
z = 2.3075, p-value = 0.02103
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.1727681
```

```
> cor.test(x = t$X, y = t$Y)

        Pearson's product-moment correlation

data:  t$X and t$Y
t = 4.0447, df = 82, p-value = 0.0001178
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2119683 0.5721941
sample estimates:
      cor
0.4078297
```

```
> library(readxl)
> types = c(rep('numeric', 22))
> t <- as.data.frame(read_excel("C:/Users/компьютер/Documents/reg-18.xlsx", 1,
+                               col_types = types))
> # вычисление корреляционной матрицы по 4 первым факторам
> t1=t[1:4]
> cor(t1)
           x1         x2          x3          x4
x1  1.0000000 -0.3524325 -0.31135084 -0.19479258
x2 -0.3524325  1.0000000  0.29339968  0.35998691
x3 -0.3113508  0.2933997  1.00000000 -0.02852365
x4 -0.1947926  0.3599869 -0.02852365  1.00000000
```

Вывод: гипотеза H1 о том, что коэффициент корреляции отличен от нуля принимается, если p-value <0,05

Создание матрицы коэффициентов корреляции и проверка их значимости при помощи функции corr.test()

МИРЭА
Российский технологический университет

```
> library(psych)
> corr.test(states, use="complete")

Call:corr.test(x = states, use = "complete")
Correlation matrix
           Population Income Illiteracy Life Exp Murder HS Grad
Population       1.00   0.21       0.11    -0.07   0.34   -0.10
Income           0.21   1.00      -0.44     0.34  -0.23    0.62
Illiteracy       0.11  -0.44       1.00    -0.59   0.70   -0.66
Life Exp        -0.07   0.34      -0.59     1.00  -0.78    0.58

Murder           0.34  -0.23       0.70    -0.78   1.00   -0.49
HS Grad         -0.10   0.62      -0.66     0.58  -0.49    1.00
Sample Size
[1] 50
Probability value
           Population Income Illiteracy Life Exp Murder HS Grad
Population       0.00   0.15       0.46     0.64   0.01     0.5
Income           0.15   0.00       0.00     0.02   0.11     0.0
Illiteracy       0.46   0.00       0.00     0.00   0.00     0.0
Life Exp         0.64   0.02       0.00     0.00   0.00     0.0
Murder           0.01   0.11       0.00     0.00   0.00     0.0
HS Grad          0.50   0.00       0.00     0.00   0.00     0.0
```

Опция `use=` может принимать значения `"pairwise"` или `"complete"` (для попарного или построчного удаления пропущенных значений соответственно). Значения опции `method=` бывают следующими: `"pearson"` (по умолчанию), `"spearman"` или `"kendall"`. Из приведенного примера видно, что коэффициент корреляции между численностью населения и долей людей со средним образованием (−0.10) не отличается от нуля ($p = 0.5$).

Функция r.test() из пакета psych позволяет проводить ряд полезных тестов на статистическую значимость.
Эту функцию можно использовать, чтобы проверять значимость:
- коэффициента корреляции;
- различий между двумя независимыми корреляциями;
- различий между двумя зависимыми корреляциями, у которых есть одна общая переменная;
- различий между двумя зависимыми корреляциями между разными парами переменных.

Основные примеры

библиотека	Задача	метод	код
База	двумерная корреляция	Pearson	`cor(dfx2, method = "pearson")`
База	двумерная корреляция	копьеносец	`cor(dfx2, method = "spearman")`
База	Многомерная корреляция	пирсон	`cor(df, method = "pearson")`
База	Многомерная корреляция	копьеносец	`cor(df, method = "spearman")`
Hmisc	Значение P		`rcorr(as.matrix(data[,1:9]))[["P"]]`
Ggally	Тепловая карта		`ggcorr(df)`

Статистические критерии в R

Критерий X^2 Пирсона (Проверка гипотезы о нормальном распределении генеральной совокупности).

МИРЭА Российский технологический университет

Описание

Критерий $\chi2$ используется для анализа таблиц сопряженности признаков и сравнения законов распределения непрерывных случайных величин. Анализируются номинальные или приведенные к номинальной шкале данные, представленные в виде таблицы сопряженности признаков. Для непрерывных случайных величин используется принадлежность значений заданным интервалам, выбираемых таким образом, чтобы в каждом из них было не менее 5-7 значений (интервалы с меньшим числом значений объединяются). Простейшим выбором является равный шаг интервалов, равный $\lambda = \frac{x_{\max} - x_{\min}}{k}$, $k = 1 + 3.32 \cdot \lg(n)$ или $k = 5 \cdot \lg(n)$.

Вычисление критериальной статистики производится по формуле:

$$\chi^2 = \sum_i \frac{(n_i - n_i^{'})^2}{n_i^{'}}$$

где n_i- эмпирические частоты, $n_i^{'}$ - теоретические частоты попадания элементов выборки в группы (заданные интервалы).

Число степеней свободы находят по формуле:

$f = k - 1 - r$

где k-число групп выборки, r-число параметров предполагаемого распределения, которые оценены по данным выборки.

Если предполагаемое распределение – нормальное, то по выборке оценивают два параметра (математическое ожидание и дисперсию), поэтому r=2 и $f = k - 3$.Одной из функций, осуществляющей проверку данного критерия в R является `chisq.test()`.

Описание функции

```
chisq.test (x, y = NULL, p = rep(1/length (x), length (x)))
```

Параметры

x	вектор или матрица.
y	вектор; игнорируемый, если x - матрица.
p	вектор теоретических вероятностей той же длины, что x.

Примечание

Если *x* – матрица с одной строкой или столбцом, или если *x* – вектор, и *y* не дан, *x* – одномерная таблица сопряженности признаков. В этом случае, проверенная гипотеза - равняются ли вероятности совокупности тем, что в *p*, или все равны, если *p* не дается. Если *x* - матрица с двумя строками (или столбцами), содержащими неотрицательные целые числа, то она рассматривается как таблица сопряженности признаков. Если *x* и *y* – два вектора, содержащих факторы (номинальные или ординальные значения), то по ним строится таблица сопряженности.

Критические значения (квантили) находятся с использованием функции `qchisq(p,df)` или по таблице $\chi2$-распределения

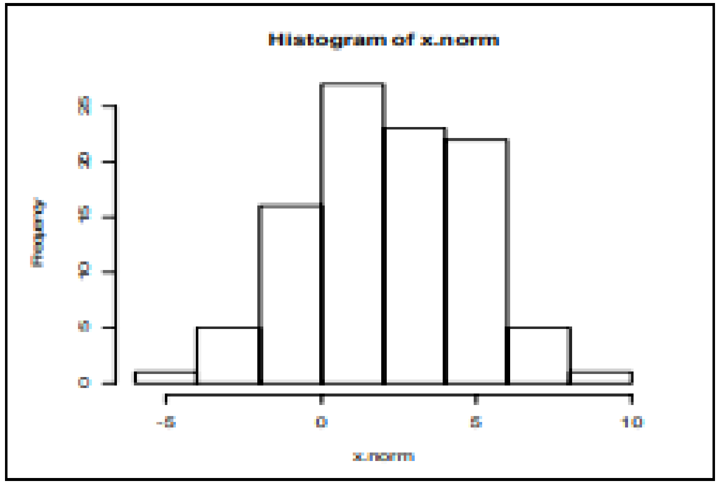
Статистические критерии в R

Критерий X^2 Пирсона (Проверка гипотезы о нормальном распределении генеральной совокупности).

Пример

Сгенерируем случайную выборку из нормального распределения, и проверим ее нормальность.

```
N<-100  # объем выборки
x.norm<-rnorm(N,mean=2,sd=2.5) # задаем среднее и СКО
```



Вычисляем квантили выборки с шагом 10% (по 10 элементов в интервале)

```
> x.norm.q <- quantile(x.norm,probs=seq(0,1,0.1))
> round(x.norm.q,2)
   0%   10%   20%   30%   40%   50%   60%   70%   80%
90%  100%
 -4.12 -1.51 -0.14  0.59  1.52  2.15  2.70  3.89  4.51
5.22  8.15
> summary(x.norm)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.4490  0.6675  2.0330  2.0670  3.1050  6.5830
```

Выбираем интервалы:

```
> k<-6    # число интервалов
> x.q <- c(-10, -1.0, 0.5, 2.0, 3.5, 5.0,12.0)
```

Вычисляем фактические частоты

```
> x.norm.hist<-hist(x.norm,breaks=x.q,plot=FALSE)
> x.norm.hist$counts
12 15 22 18 18 15
```

Вычисляем (по выборке) теоретические вероятности для каждого интервала

```
> x.q[1]<-(-Inf);x.q[k+1]<-(+Inf)#«раздвигаем» границы
по бесконечности

> x.norm.p.theor<-
pnorm(x.q,mean=mean(x.norm),sd=sd(x.norm))
> x.norm.p.theor<-(x.norm.p.theor[2:(k+1)]-
x.norm.p.theor[1:k])
> round(x.norm.p.theor,2)
0.12 0.15 0.21 0.22 0.16 0.14
```

Сравниваем фактические и теоретические частоты

```
> chisq.test(x.norm.hist$counts,p=x.norm.p.theor)
        Chi-squared test for given probabilities
data:  x.norm.hist$counts
X-squared = 0.9691, df = 5, p-value = 0.965
```

Поскольку для проверки нулевой гипотезы H_0 о нормальности распределения генеральной совокупности в нашем случае используется правосторонний критерий, а уровень значимости (*p*-value) равен 0.965 (96.5%), то нужно допустить разрешить вероятность ошибки, равную 96.5%, чтобы считать выборку не принадлежащей нормальному распределению. Следовательно, гипотеза о нормальности принимается.

Статистические критерии в R

Критерий Стьюдента (Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых неизвестны и одинаковы).

Описание

Если предположить, что неизвестные генеральные дисперсии равны между собой, то для решения этой задачи можно применить критерий Стьюдента. Т.е. нужно, пользуясь критерием Фишера, предварительно проверить гипотезу о равенстве генеральных дисперсий. В случае независимых выборок в качестве критериальной статистики для проверки гипотезы принимают случайную величину:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)\cdot S_x^2 + (m-1)\cdot S_y^2}} \cdot \sqrt{\frac{n\cdot m\cdot(n+m-2)}{n+m}},$$

где $\bar{X}, \bar{Y}$ - выборочные средние, S_x^2, S_y^2 – выборочные дисперсии, n, m – объемы выборки и $f = n + m - 2$ - число степеней свободы для распределения критериальной статистики (если дисперсии не равны, то критерий остается применимым, но требует коррекции приводенной формулы и числа степеней свободы – необходимость такой коррекции указывается при вызове функции). Если выборки зависимые (парная выборка), то проверяется гипотеза о равенстве матожидания нулю для новой случайной величины $z_i = x_i - y_i$, также

имеющей нормальное распределение. В этом случае используется критериальная статистика Стьюдента

$$t = \frac{\bar{Z}}{\sqrt{S_z^2}} \cdot \sqrt{n}$$

Одной из функций, осуществляющей проверку данного критерия в R является `t.test()`

Описание функции

```
t.test (x, y = NULL, alternative = c ("two.sided",
"less", "greater"), var.equal = FALSE, conf.level = 0.95,
paired = FALSE,...)
```

Параметры

x	Числовой вектор значений.
y	Числовой вектор значений (используется для парного теста, см. ниже).
paired	Признак парного теста: проверяется гипотеза для x-y, поэтому вектор y должен присутствовать и соответствовать по длине вектору x.
alternative	Символьная строка, определяющая альтернативную гипотезу, должна быть одна из **"two.sided"** (по умолчанию)-двустороняя критическая область, **"greater"** -правостороняя критическая область или **"less"-левостороняя** критическая область.
var.equal	Логическая переменная, указывающая на равенство дисперсий
conf.level	Доверительная вероятность.

Примечание

По умолчанию var.equal=FALSE (дисперсии предполагаются неравными), в этом случае для вычислений используется оценка Велча (Welch).

МИРЭА
Российский технологический университет

Статистические критерии в R. Пример

```
> x<-c(3.5, 3.6, 7.8, 9.6, 5.7, 8.9, 6.3)
> y<-c(1.0, 2.7, 8.9, 6.5, 8.9, 6.5,12.5,10.2, 1.2)
>
t.test(x,y,alternative=c("two.sided"),var.equal=TRUE,conf
.level=0.95)

        Two Sample t-test

 data:  x and y
 t = -0.0018, df = 14, p-value = 0.9986
 alternative hypothesis: true difference in means is not
equal to 0
 95 percent confidence interval:

  -3.760075  3.753726
 sample estimates:
 mean of x mean of y
  6.485714  6.488889
```

Значения

t = -0.0018 (значение критериальной статистики), число степеней свободы равно 14.

p-value = 0.9986, т.е. чтобы отвергнуть гипотезу, нужно допустить 99.86% ошибки.

95% доверительный интервал (-3.760075, 3.753726). Поскольку наше значение в него попадает, то нулевая гипотеза принимается на 5% уровне значимости.

Критерии Бартлетта и Кохрана (Сравнение нескольких дисперсий нормальных генеральных совокупностей по выборкам).

Если равенство дисперсий не проверялось, или гипотеза о равенстве не принимается, то вызов критерия выглядит так:

```
 >
t.test(x,y,alternative=c("two.sided"),var.equal=FALSE,
conf.level=0.95)

        Welch Two Sample t-test

 data:  x and y
 t = -0.0019, df = 13.242, p-value = 0.9985
 alternative hypothesis: true difference in means is not
equal to 0
 95 percent confidence interval:
  -3.545004  3.538655
           tes:
           n of y
  6.485714  6.488889
```

Число степеней свободы теперь 13.242 вместо 14, и границы доверительного интервала несколько изменились.

Статистические критерии в R

Критерий Фишера (Сравнение дисперсий двух нормальных генеральных совокупностей).

МИРЭА Российский технологический университет

Описание

В качестве критериальной статистики для проверки гипотезы о равенстве генеральных дисперсий нормально распределенных генеральных совокупностей используют отношение выборочных дисперсий, т. е. случайную величину:

$$F = \frac{S_x^2}{S_y^2}$$

где S_x^2, S_y^2 – исправленные выборочные дисперсии для выборок объемом n_1 и n_2 соответственно. В качестве нулевой гипотезы H_0 формулируется гипотеза о равенстве генеральных дисперсий.

Величина F при условии справедливости нулевой гипотезы имеет распределение Фишера-Снедекора со степенями свободы $f_1 = n_1 - 1, f_2 = n_2 - 1$. Одной из функций, осуществляющей проверку данного критерия в R является **`var.test()`**

Описание функции

```
var.test(x,y,alternative = c("two.sided",
"less","greater"), conf.level = 0.95, ...)
```

Параметры

x,y	вектор или объекты линейной модели (например, класса lm).
conf.level	доверительная вероятность
Alternative	альтернативная гипотеза. Может быть одна из **"two.sided"** (по умолчанию)-двустороняя критическая область, **"greater"** -правостороняя критическая область или **"less"-левостороняя** критическая область.

Пример

```
> x<-c(3.5, 3.6, 7.8, 9.6, 5.7, 8.9, 6.3)
> y<-c(1.0, 2.7, 8.9, 6.5, 8.9, 6.5,12.5,10.2, 1.2)
```

n_1=7, S_x^2=5.86; n_2=9, S_y^2=16.75

```
> var.test(x, y, alternative = c("two.sided"),
conf.level = 0.95)

        F test to compare two variances
 data:  x and y
 F = 0.3498, num df = 6, denom df = 8, p-value = 0.2174
 alternative hypothesis: true ratio of variances is not
equal to 1
 95 percent confidence interval:
  0.07519108 1.95855792
```

Результат теста:

F = 0.3498 (значение F статистики), число степеней свободы для x - 6, для y - 8

p-value = 0.2174, т.е. уровень ошибки, при котором можно отвергнуть гипотезу о равенстве дисперсий, равен 21.74%

95% доверительный интервал: (0.075, 1.959) - полученное нами значение *F*-статистики в него попадает, следовательно гипотеза о равенстве дисперсий принимается на 5% уровне значимости.

Критерии Бартлетта и Кохрана (Сравнение нескольких дисперсий нормальных генеральных совокупностей по выборкам).

Описание

Критерий Барлетта используется для проверки гипотезу об однородности (равенстве) нескольких дисперсий, полученных по выборкам разного объема. Для этого рассчитывают среднюю арифметическую исправленных дисперсий, взвешенную по числам степеней свободы:

$$\overline{S}^2 = \frac{1}{f}\sum_{i=1}^{k} f_i \cdot S_i^2 ,$$

где $f_i = n_i - 1$ число степеней свободы для i-й выборки объема n_i, S_i^2 - выборочная дисперсия дисперсия i-й выборки, $f = \sum_i f_i$ - общее число степеней свободы, и k - число выборок.

В качестве критериальной статистики для проверки гипотезы об однородности дисперсий используют критерий Бартлетта:

$$B = {}^{V}\!/_{C},$$

имеющая распределение χ^2, где

$$V = 2.303 \cdot [f \cdot \lg \overline{S}^2 - \sum_{i=1}^{l} f_i \cdot \lg S_i^2],$$

$$C = 1 + \frac{1}{3(k-1)} \cdot \left[\left(\sum_{i=1}^{k} \frac{1}{f_i}\right) - \frac{1}{f}\right]$$

Одной из функций, осуществляющей проверку данного критерия в R является `bartlett.test()`

Описание функции

`bartlett.test (x, g...)`

Параметры

x числовой вектор значений, или список числовых значений векторов, или объекты линейной модели (класса "lm").

g вектор или фактор, дающий группу для соответствующих элементов x. Игнорируемый, если x - список.

Примечание

Если x - список, его элементы будут взяты как выборки и g игнорируется, и можно просто использовать `bartlett.test(x)`. Если выборки еще не содержатся в списке, используют `bartlett.test(list (x...))`.

Критические значения (правосторонний критерий) находятся по таблице распределения χ^2 с $k-1$ степенями свободы [2,стр.329] или используют функцию вычисления квантилей распределения Хи-квадрат `qchisq(p,df)`.

Пример

```
> x1<-c(3.5, 3.6, 7.8, 9.6, 5.7, 8.9, 6.3)
> x2<-c(1.0, 2.7, 8.9, 6.5, 8.9, 6.5,12.5,10.2, 1.2)
> x3<-c(3.6,7.8,9.6,5.7,8.9)
> x4<-c(2.7,8.9,6.5,8.9)
```

Дисперсии выборок равны соотвественно 5.86, 16.75, 6.05 и 8.57, нулевая гипотеза H_0 –дисперсии всех генеральных совокупностей равны между собой, уровень значимости – 5%.

```
> bartlett.test(list(x1,x2,x3,x4))
        Bartlett test of homogeneity of variances

data:  list(x1, x2, x3, x4)
Bartlett's K-squared = 2.2368, df = 3, p-value = 0.5247
```

Значения

Bartlett's K-squared = 2.2368 (значение критериальной статистики теста Бартлетта), число степеней свободы 3,

p-value = 0.5247, т.е. отвергнуть гипотезу H_0 можно только при допустимой ошибке в 52.47%. Следовательно, гипотеза об однородности дисперсий принимается на 5% уровне значимости.

Критерии Бартлетта и Кохрана (Сравнение нескольких дисперсий нормальных генеральных совокупностей по выборкам).

Если объем выборок (примерно) одинаковый, то может использоваться тест экстремальных значений Кохрана (Cochran) из пакета **outliers**, реализуемый функцией `cochran.test()`.

Описание функции

`cochran.test(object,data)`

Параметры

`object` числовой вектор, содержащий значения дисперсий для каждой выборки S_i^2

`data` числовой вектор, содержащий объем каждой выборки

В в качестве критериальной статистики используется

$$C = \frac{\max_i \{S_i^2\}}{\sum_i S_i^2}$$

а для вычисления критических значений – функция вычисления квантилей распределения Кохрана qcochran(p, n, k) из того же пакета, где *p*-доверительная вероятность, *n* – объем одной выборки (если объемы различаются, то берется среднее значение), *k* – число выборок.

Пример

Используем в примере те же выборки, что и в предыдущем случае, объем выборок 7, 9, 5 и 4 элементов соответственно. Нулевая гипотеза H_0 – дисперсии всех генеральных совокупностей равны между собой, уровень значимости – 5%.

```
> cochran.test(object=
c(var(x1),var(x2),var(x3),var(x4)), data=c(7,9,5,4))
          Cochran test for outlying variance

 data:  c(var(x1), var(x2), var(x3), var(x4))
 C = 0.4499, df = 6.25, k = 4.00, p-value = 0.3083
 alternative hypothesis: Group 2 has outlying variance
```

Значения

Cochran C = 0.4499 (значение критериальной статистики теста Кохрана), число степеней свободы (средний объем выборки) 6.25, число групп 4, *p*-value 0.3083. Альтернативная гипотеза – дисперсия второй выборки значительно больше остальных (является «выбросом»). Поскольку *p*-value = 0.3083, то отвергнуть гипотезу H_0 можно только при допустимой ошибке в 30.83%. Следовательно, гипотеза об однородности дисперсий принимается на 5% уровне значимости.

Дисперсионный анализ

Описание

Данный метод основан на разложении общей дисперсии численного признака на составляющие ее компоненты (отсюда и название метода ANalysis Of VAriance или ANOVA), сравнивая которые с друг другом посредством F-критерия Фишера можно определить, какую долю (по отношению к совокупности случайных причин) общей вариации признака обуславливает действие на него известных величин (факторов).

Метод основан на сравнении межгрупповой и внутригрупповой изменчивости признака. Каждую группу образуют значения признака при фиксированных значениях (уровнях) известных факторов, поэтому единственным источником дисперсии (изменчивости) внутри каждой группы является суммарное воздействие совокупности случайных причин. Общая модель дисперсионного анализа (на примере двух факторов) выглядит следующим образом:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

где μ - среднее значение признака, α_i - влияние первого фактора на i-м уровне (при i-м значении), β_j - влияние второго фактора на j-м уровне (при j-м значении), $(\alpha\beta)_{ij}$ - влияние взаимодействия факторов на указанных уровнях (если факторы не независимы), и ε_{ijk} - суммарное влияние на признак случайных факторов, имеющее нормальное распределение с нулевым матожиданием и дисперсией $\sigma^2_{ош}$. Предполагается, что ε_{ijk} не зависит от уровней факторов, поэтому общая дисперсия признака (точнее, общая сумма квадратов $SS_{общ} = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$, где точки в индексе среднего показывают, по каким из них проводилось осреднение, может быть разложена на компоненты (частные суммы), соотвествущие вкладу в общую дисперсию каждой составляющей.

В простейшем случае, если имеется всего один фактор, такое разложение представляется в виде таблицы дисперсионного анализа:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Источник дисперсии	SS сумма квадратов	Степеней свободы	Средний квадрат	F статистика
Фактор (межгрупповая)	$SS_{факт} = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$k-1$	$S^2_{факт} = \frac{SS_{факт}}{k-1}$	$F = \frac{S^2_{факт}}{S^2_{ош}}$
Случайная составляющая	$SS_{ош} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	$\sum_{i=1}^{k}(n_i - 1)$	$S^2_{ош} = \frac{SS_{ош}}{\sum_{i=1}^{k}(n_i - 1)}$	
Общая	$SS_{общ} = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$	$N-1$	$S^2_{общ} = \frac{SS_{общ}}{N-1}$	

где k - число групп, n_i - число наблюдений в i-ой группе, $N = \sum_i n_i$ - общее число наблюдений.

Для проведения однофакторного дисперсионного анализа в R используется линейная модель, в которой единственной независимой переменной выступает этот фактор.

Описание функции

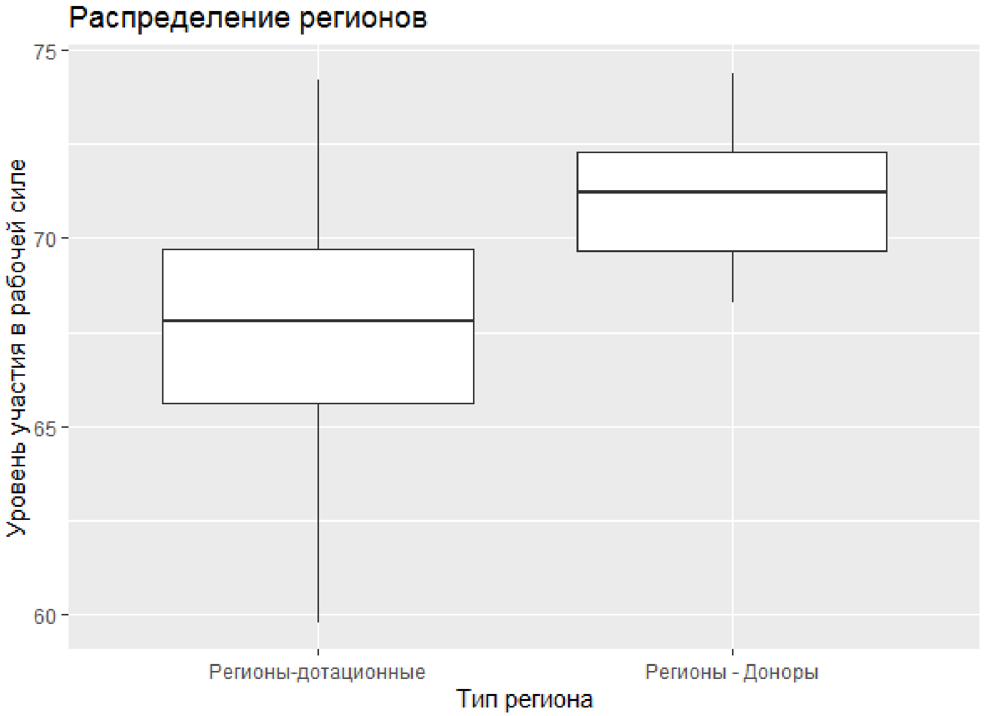
`anova(object)`

Параметры

`object` Объект класса lm, glm .

Используется для исследования зависимости переменной от фактора

Условия применения однофакторного дисперсионного анализа
#1) соответствие распределения анализируемых групп генеральным совокупностям,
#имеющим нормальный закон распределения или близкий к нему;
#2) независимость распределения наблюдений в группах;
#3) наличие частоты (повторяемости) наблюдений.




```
# One-way ANOVA
#переменная fit
fit <- aov(tmod$Y ~ tmod$Status, data=tmod)
#результаты анализа ANOVA
summary(fit)
#Ответ
             Df Sum Sq Mean Sq F value  Pr(>F)
tmod$Status   1   72.6   72.60   9.214 0.00322 **<0,05
Residuals    82  646.1    7.88
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Вывод: гипотеза принимается на уровне значимости 0,01, то есть с 99% вероятностью, то есть тип региона определяет уровень участия в рабочей силе

3. Визуализации парной и множественной взаимосвязи переменных

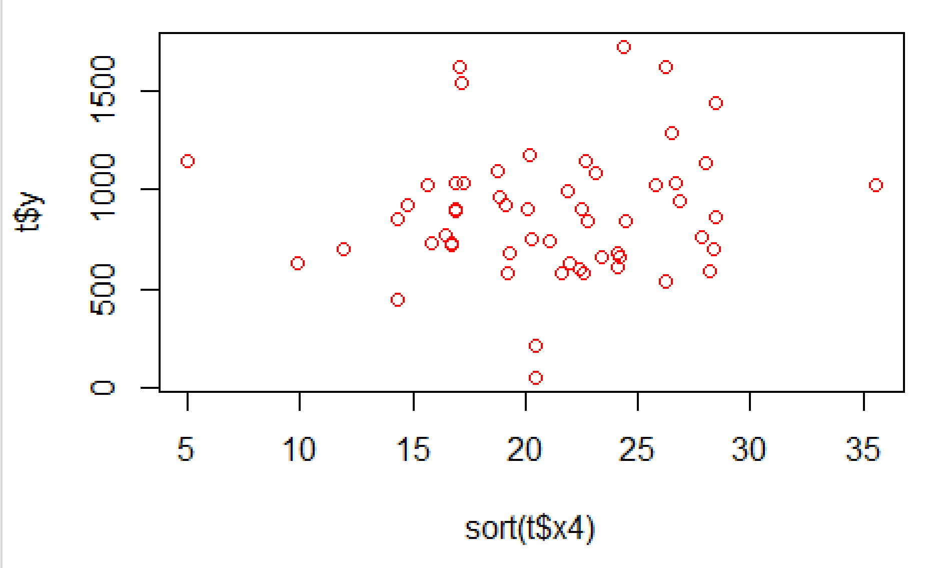
МИРЭА
Российский технологический университет

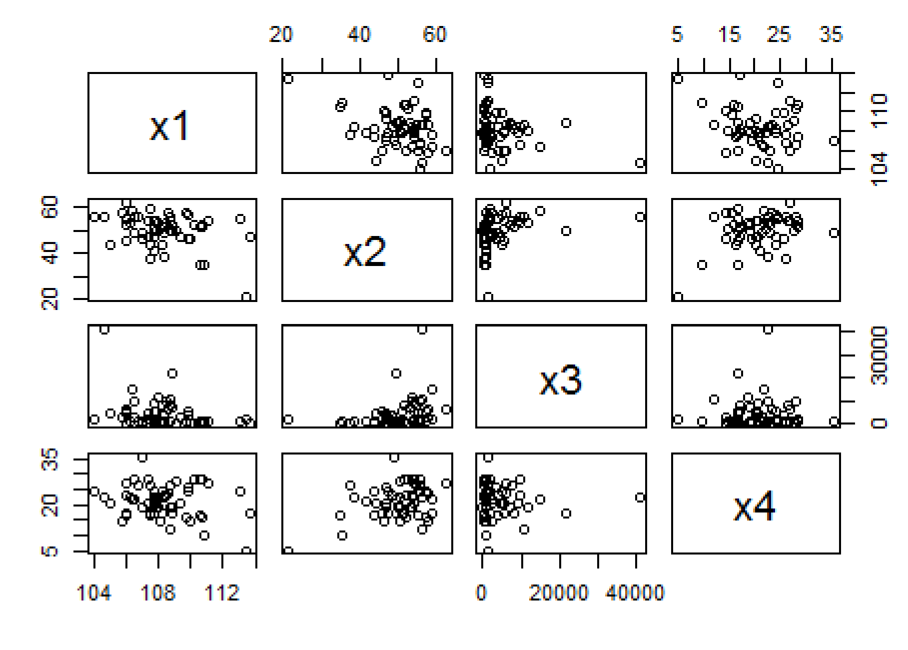
Связи между парами переменных можно визуализировать при помощи диаграмм рассеяния и составленных из них матриц. Коррелограммы – это непревзойденный действенный метод сравнения большого числа коэффициентов корреляции в легко интерпретируемой форме.

Диаграмма рассеивания

Парная: plot(sort(t$x4), t$y, type='p', col='red')

Множественная: plot(t1), где t1 – таблица из 4 переменных

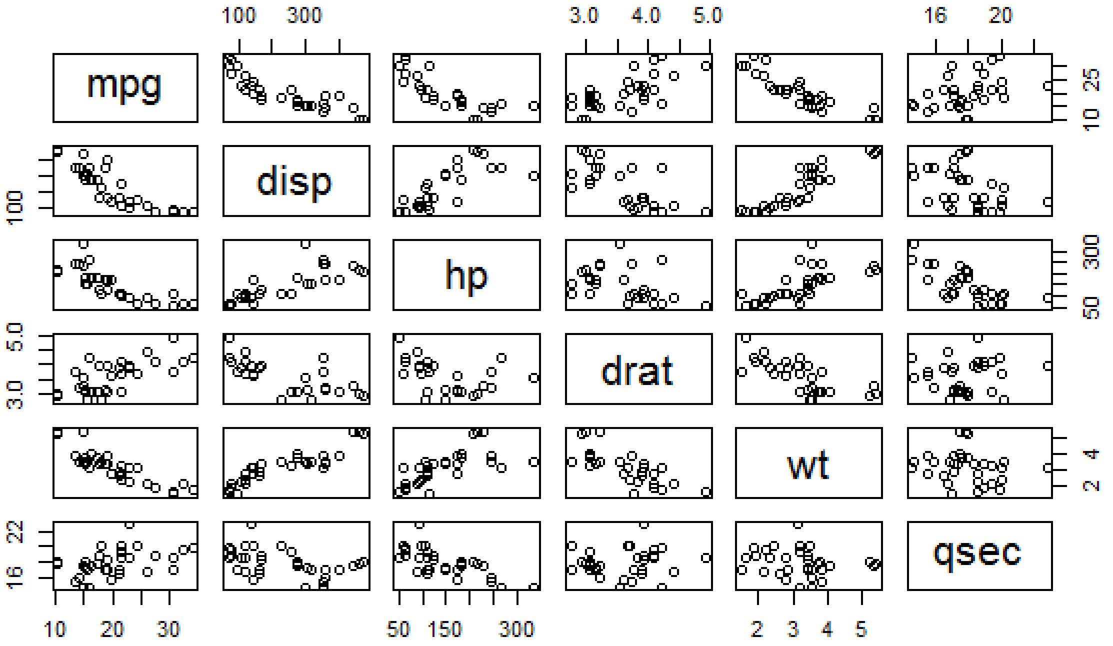




Построение и визуализация корреляционной матрицы

```
> df   <- mtcars
> df_numeric  <- df[, c(1,3:7)]
>
> pairs(df_numeric)
>
> cor(df_numeric)
            mpg       disp         hp        drat         wt        qsec
mpg   1.0000000 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403
disp -0.8475514  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788
hp   -0.7761684  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339
drat  0.6811719 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476
wt   -0.8676594  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588
qsec  0.4186840 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000
```

- cor(df_numeric)



Матрица корреляций

coeffs <- abs(cor(t[18:20]))

Диаграмма рассеивания по фактору

Российский технологический университет

```
library(readxl)
types = c("text", rep("numeric", 2), "text")
t <-
as.data.frame(read_excel("C:/Users/компьютер/Documents/tab3.xlsx", 1,
                         col_types = types))
#Диаграммы рассеяния по фактору
library(ggplot2)
ggplot(data = t, mapping = aes(x = t$X, y = t$У, color = as.factor(Status))) + geom_point()

ggplot(data = t, mapping = aes(x = t$X, y = t$У, size = as.factor(Status), color = as.factor(Status))) + geom_point()
```

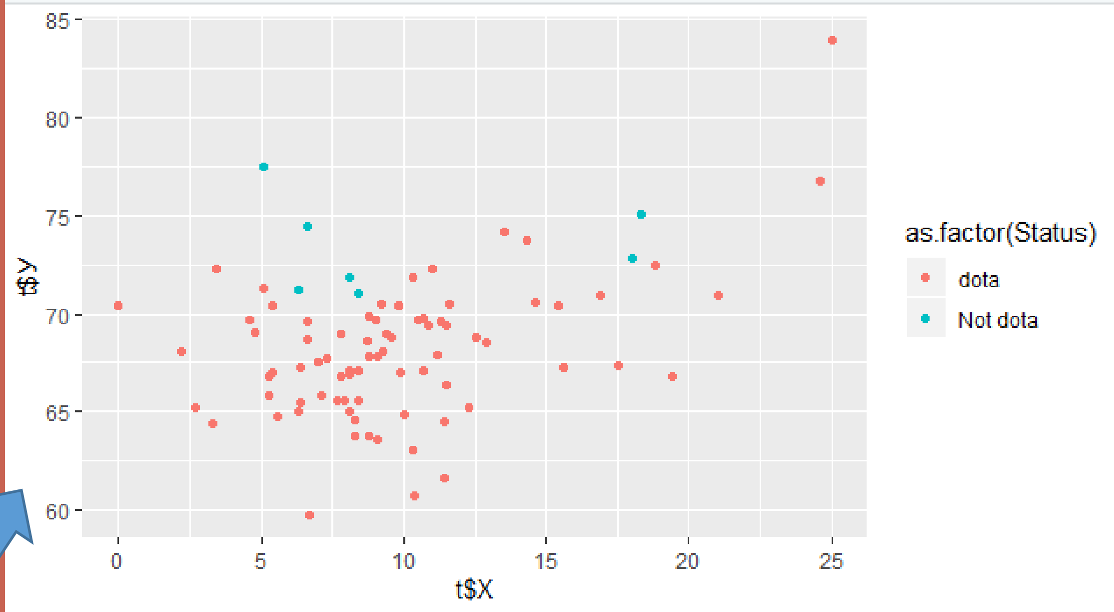
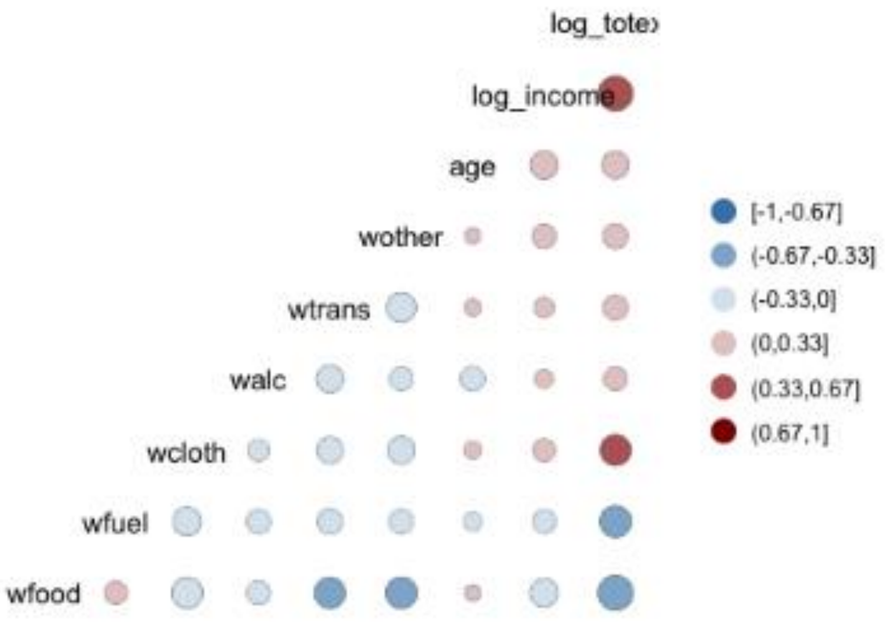
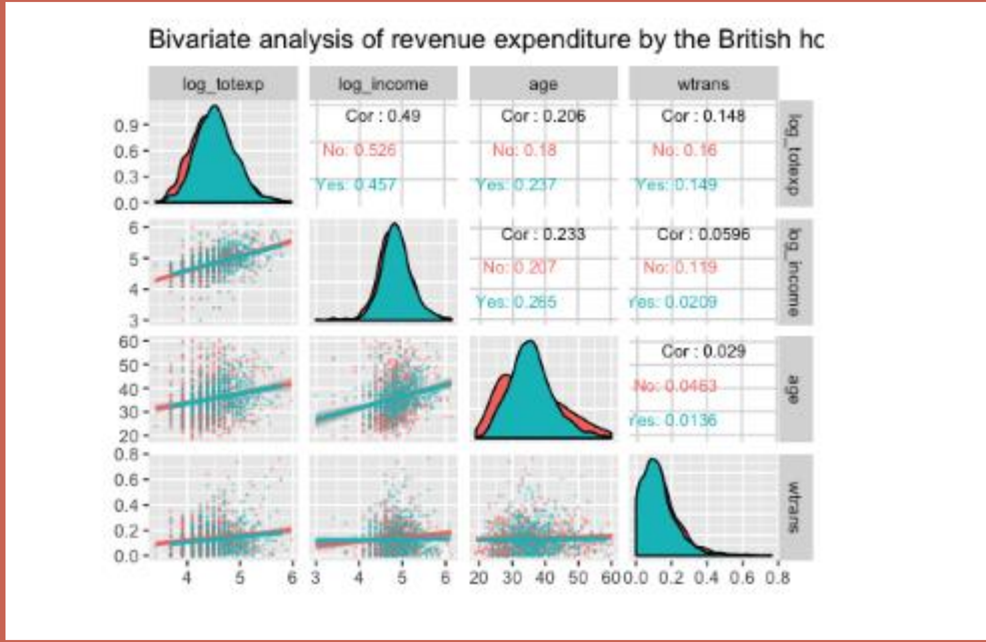


Диаграмма рассеивания по фактору

```
p=ggplot(data = t, aes(x = t$X, y = t$У, shape = factor(Status)))
p+geom_point(aes(colour  =  factor(Status)), size  =  4) +
geom_point(colour  =  "green", size  =  1.5)
```

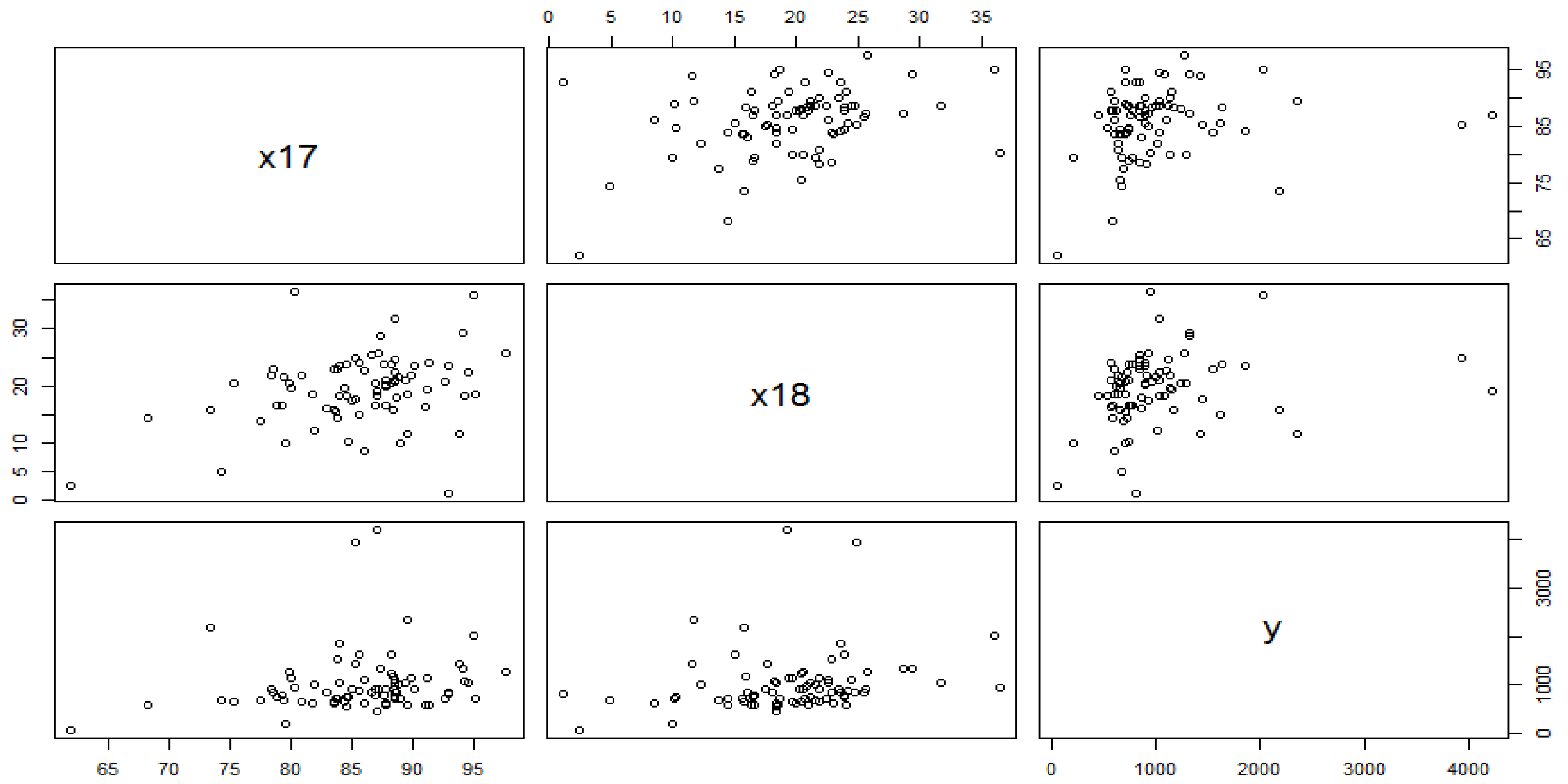
Тепловая карта — это способ показать корреляционную матрицу.
Легенда графика показывает цвет градиента от — 1 до 1, причём горячий цвет указывает на сильную положительную корреляцию, а холодный цвет — на отрицательную корреляцию.
Библиотека GGally является расширением ggplot2.



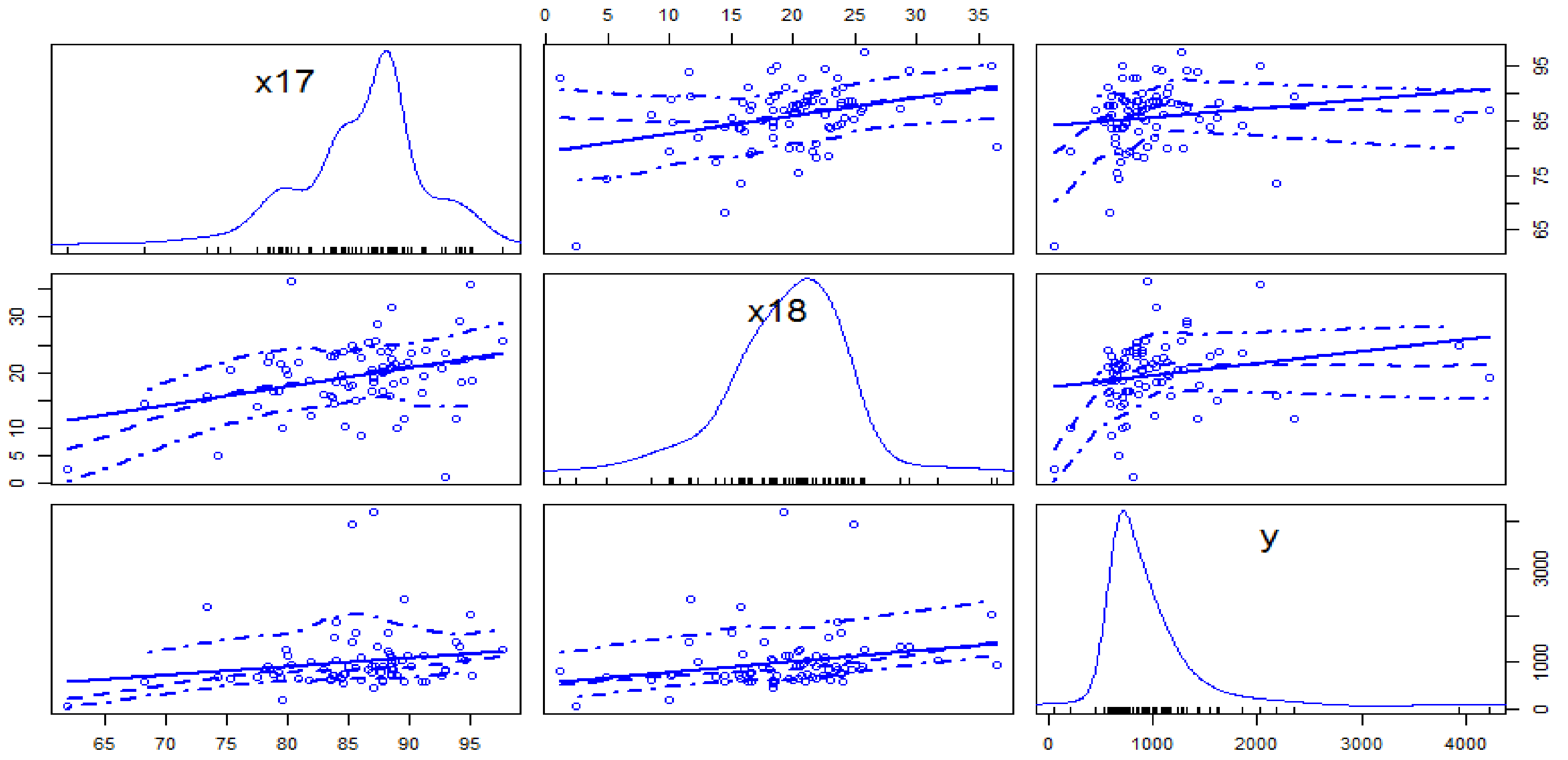


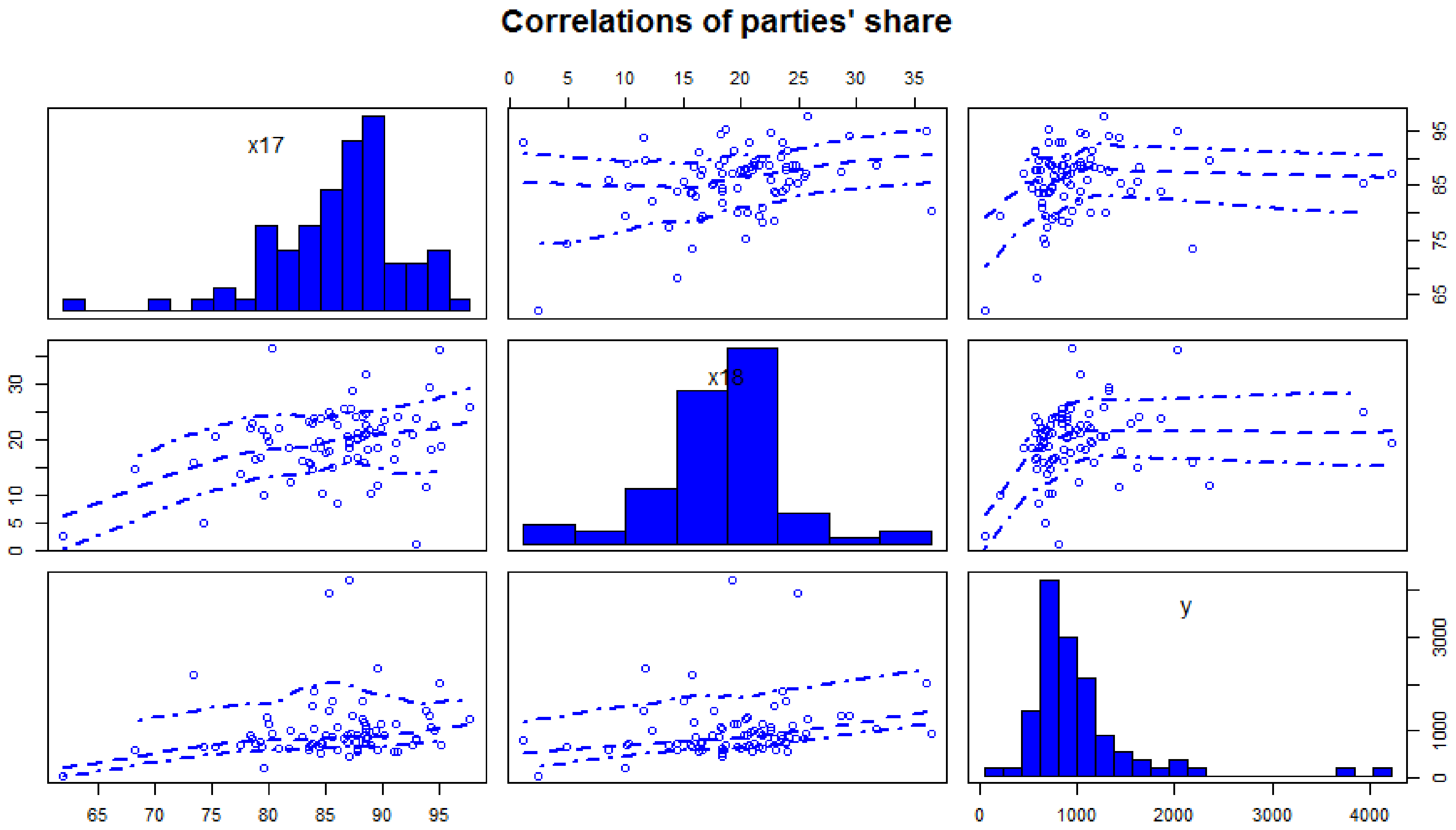
Матрица диаграмм рассеяния (scatterplot matrix)

- Скрипт:
- install.packages("car")
- library(car)
- pairs(t[18:20])



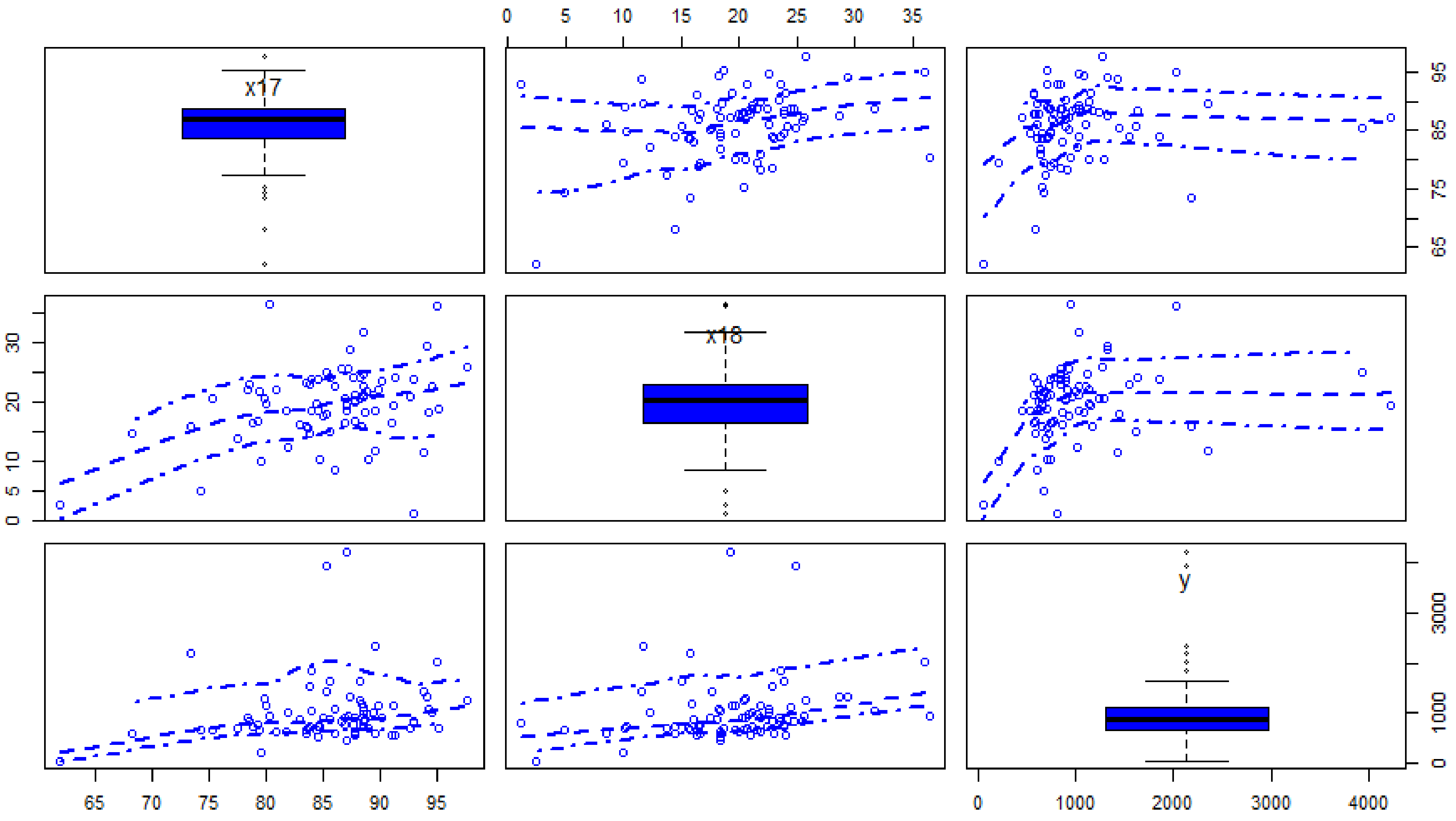
Парные регрессии: scatterplotMatrix(t[18:20])






```
scatterplotMatrix(t[18:20], regLine = FALSE, smooth = TRUE,
        var.labels = labs,
        cex.labels = 1.3,
        main = "Correlations of parties' share",
        diagonal = list(method = "histogram"))
```

Correlations of parties' share



Вопросы по теме

Какие виды взаимосвязей между признаками бывают и какая взаимосвязь анализируется с помощью линейного коэффициента корреляции?
Какой критерий применяется для проверки значимости линейного коэффициента корреляции?
Как рассчитать коэффициенты корреляции в R?
Какие возможности визуализации взаимосвязи переменных есть в R?