





Программирование в среде R

Шевцов Василий Викторович, директор ДИТ РУДН, shevtsov_vv@rudn.university

Корреляции





Коэффициенты корреляции

- Коэффициенты корреляции используются для описания связей между количественными переменными.
- Знак коэффициента (+ или –) свидетельствует о направлении связи (положительная или отрицательная)
- Величина коэффициента показывает силу связи (варьирует от 0 нет связи до 1 – абсолютно предсказуемая взаимосвязь).





Коэффициент корреляции

Линейный коэффициент корреляции принимает значения от —1 до +1. Связи между признаками могут быть слабыми и сильными (тесными). Их критерии оцениваются по шкале Чеддока:

```
0.1 < rxy < 0.3: слабая;
```

0.3 < rxy < 0.5: умеренная;

0.5 < rxy < 0.7: заметная;

0.7 < rxy < 0.9: высокая;

0.9 < rxy < 1: весьма высокая;





cor()

- cor(x, use= , method=)
- Линейный коэффициент корреляции Пирсона (Pearson product moment correlation) отражает степень линейной связи между двумя количественными переменными.
- Коэффициент ранговой корреляции Спирмана (Spearman's Rank Order correlation) – мера связи между двумя ранжированными переменными.
- Коэффициент Тау Кэнделла (Kendall's Tau) также непараметрический показатель ранговой корреляции.





cor(). Параметры

- х Матрица или таблица данных
- use. Упрощает работу с пропущенными данными.
 - all.obs (предполагается, что пропущенные значения отсутствуют; их наличие вызовет сообщение об ошибке),
 - everything (любая корреляция, включающая строку с пропущенным значением, не будет вычисляться – обозначается как missing),
 - complete.obs (учитываются только строки без пропущенных значений)
 - pairwise.complete.obs (учитываются все полные наблюдения для каждой пары переменных в отдельности)
- method. Определяет тип коэффициента корреляции. Возможные значения
 - pearson (по умолчанию)
 - spearman
 - kendall





cor(mtcars)

> mtcars mpg cyl disp hp drat wt qsec vs am gear carb Mazda RX4 21.0 6 160.0 110 3.90 2.620 16.46 Mazda RX4 Wag 21.0 6 160.0 110 3.90 2.875 17.02 Datsun 710 22.8 93 3.85 2.320 18.61 Hornet 4 Drive 21.4 6 258.0 110 3.08 3.215 19.44 18.7 8 360.0 175 3.15 3.440 17.02 Hornet Sportabout ∨aliant. 18.1 6 225.0 105 2.76 3.460 20.22 Duster 360 14.3 8 360.0 245 3.21 3.570 15.84 Merc 240D 24.4 3.190 20.00 Merc 230 22.8 Merc 280 19.2 6 167.6 123 3.92 3.440 18.30 Merc 280C 17.8 6 167.6 123 3.92 3.440 18.90 16.4 8 275.8 180 3.07 4.070 17.40 Merc 450SE

```
> t1 <- mtcars[,c(1,2,9)]

> cor(t1)

mpg cyl am

mpg 1.0000000 -0.852162 0.5998324

cyl -0.8521620 1.000000 -0.5226070

am 0.5998324 -0.522607 1.0000000
```





cor(mtcars, method=)

```
> t1 <- mtcars[,c(1,2,9)]
> cor(t1,method="pearson")
          mpg cyl
                                am
mpg 1.0000000 -0.852162 0.5998324
cyl -0.8521620 1.000000 -0.5226070
    0.5998324 -0.522607 1.0000000
> cor(t1,method="spearman")
          mpq cyl
                                 am
mpq 1.0000000 -0.9108013 0.5620057
cyl -0.9108013 1.0000000 -0.5220712
    0.5620057 -0.5220712 1.0000000
> cor(t1,method="kendall")
                                 am
mpg 1.0000000 -0.7953134 0.4690128
   -0.7953134 1.0000000 -0.4946212
    0.4690128 -0.4946212 1.0000000
```





Ковариация

```
> cov(t1)

mpg cyl am

mpg 36.324103 -9.1723790 1.8039315

cyl -9.172379 3.1895161 -0.4657258

am 1.803931 -0.4657258 0.2489919
```





Результат – прямоугольная матрица





Частные корреляции

- Частная корреляция это корреляция между двумя количественными переменными, зависящими, в свою очередь, от одной или более других количественных переменных.
- Для вычисления коэффициентов частной корреляции можно использовать функцию pcor() из пакета ggm.

- pcor(*u*, S)
 - и 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.





Частные корреляции

```
> pcor(c(1,2,3,4,5,6,7,8,9,10,11),cov(mtcars))
[1] -0.02326429
> pcor(c(1,2,3,4,5,6,7,8,10,11),cov(mtcars))
[1] -0.0926765
> pcor(c(1,2,3,4,5,6,7,8,10),cov(mtcars))
[1] -0.1127779
```





Проверка статистической значимости корреляций

- Стандартная нулевая гипотеза это отсутствие связи (то есть коэффициент корреляции для генеральной совокупности равен нулю).
- Для проверки значимости отдельных корреляционных коэффициентов Пирсона, Спирмена и Кэнделла можно использовать функцию cor.test().
- cor.test(x, y, alternative = , method =)
 - где x и y − это переменные, корреляция между которыми исследуется,
 - опция alternative определяет тип теста ("two.side", "less" или "greater"),
 - опция method задает тип корреляции ("pearson", "kendall" или "spearman").
 - опция alternative="less" для проверки гипотезы о том, что в генеральной совокупности коэффициент корреляции меньше нуля
 - опция alternative="greater" для проверки того, что он больше нуля. По умолчанию alternative="two.side" (проверяется гипотеза о том, что коэффициент корреляции в генеральной совокупности не равен нулю).





cor.test()

```
> cor.test(mtcars$mpg,mtcars$cyl)

Pearson's product-moment correlation

data: mtcars$mpg and mtcars$cyl

t = -8.9197, df = 30, p-value = 6.113e-10

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.9257694 -0.7163171

sample estimates:

cor

-0.852162
```

Нулевая гипотеза заключается в том, что коэффициент корреляции Пирсона между расходом топлива и количеством цилиндров равен нулю. Если этот коэффициент для генеральной совокупности равен нулю, то его значение для случайной выборки будет равно 0.852 реже, чем в одном случае из 10¹⁰ (p-value = 6.113e – 10).

Учитывая, насколько мала вероятность, мы отвергнем нулевую гипотезу и примем альтернативную – о том, что значение этого коэффициента для генеральной совокупности *не* равно нулю.





corr.test()

- При помощи функции cor.test() одновременно можно проверить значимость только одного коэффициента корреляции.
- В пакете psych есть функция corr.test(), которая позволяет вычислить коэффициенты корреляции Пирсона, Спирмена и Кэнделла между несколькими переменными и оценить их достоверность.
- corr.test(x, use =, method=)
 - use= может принимать значения "pairwise" или "complete" (для попарного или построчного удаления пропущенных значений соответственно).
 - Значения опции method= "pearson" (по умолчанию), "spearman" или "kendall".





corr.test(mtcars, method="pearson")

```
> corr.test(mtcars, method="pearson")
call:corr.test(x = mtcars, method = "pearson")
Correlation matrix
           cyl disp
                      hp drat
                                 wt qsec
      mpg
                                          VS
                                                 am
                                                   gear
     1.00 -0.85 -0.85 -0.78 0.68 -0.87 0.42 0.66 0.60 0.48 -0.55
mpq
cyl -0.85 1.00 0.90 0.83 -0.70 0.78 -0.59 -0.81 -0.52 -0.49 0.53
disp -0.85 0.90 1.00 0.79 -0.71 0.89 -0.43 -0.71 -0.59 -0.56 0.39
    -0.78 0.83 0.79 1.00 -0.45 0.66 -0.71 -0.72 -0.24 -0.13 0.75
drat 0.68 -0.70 -0.71 -0.45 1.00 -0.71 0.09 0.44 0.71 0.70 -0.09
    -0.87 0.78 0.89 0.66 -0.71 1.00 -0.17 -0.55 -0.69 -0.58 0.43
asec 0.42 -0.59 -0.43 -0.71 0.09 -0.17 1.00 0.74 -0.23 -0.21 -0.66
     0.66 -0.81 -0.71 -0.72 0.44 -0.55 0.74 1.00 0.17 0.21 -0.57
     0.60 -0.52 -0.59 -0.24 0.71 -0.69 -0.23 0.17
                                               1.00
                                                   0.79 0.06
gear 0.48 -0.49 -0.56 -0.13 0.70 -0.58 -0.21 0.21
                                               0.79 1.00 0.27
carb -0.55 0.53 0.39 0.75 -0.09 0.43 -0.66 -0.57 0.06 0.27 1.00
Sample Size
[1] 32
Probability values (Entries above the diagonal are adjusted for multiple tests.)
     mpg cyl disp
                  hp drat
                           wt qsec vs
                                        am gear carb
          0 0.00 0.00 0.00 0.00 0.22 0.00 0.01 0.10 0.02
mpg 0.00
cyl 0.00
          0 0.00 0.00 0.00 0.00 0.01 0.00 0.04 0.08 0.04
disp 0.00
          0 0.00 0.00 0.00 0.00 0.20 0.00 0.01 0.02 0.30
    0.00
hp
          drat 0.00
          0.00
          0 0.00 0.00 0.00 0.00 1.00 0.02 0.00 0.01 0.20
wt
gsec 0.02
          0 0.01 0.00 0.62 0.34 0.00 0.00 1.00 1.00 0.00
          VS
    0.00
    0.00
          0 0.00 0.18 0.00 0.00 0.21 0.36 0.00 0.00 1.00
gear 0.01
          0 0.00 0.49 0.00 0.00 0.24 0.26 0.00 0.00 1.00
carb 0.00
          0 0.03 0.00 0.62 0.01 0.00 0.00 0.75 0.13 0.00
```





corr.test(mtcars, method="spearman")

```
> corr.test(mtcars, method="spearman")
call:corr.test(x = mtcars, method = "spearman")
Correlation matrix
         cyl disp
                      hp drat
                              wt qsec
      mpg
                                         VS
                                                am gear
     1.00 -0.91 -0.91 -0.89 0.65 -0.89 0.47 0.71 0.56 0.54 -0.66
mpg
cvl
    -0.91 1.00 0.93 0.90 -0.68 0.86 -0.57 -0.81 -0.52 -0.56 0.58
disp -0.91 0.93 1.00 0.85 -0.68 0.90 -0.46 -0.72 -0.62 -0.59 0.54
    -0.89
         0.90 0.85 1.00 -0.52 0.77 -0.67 -0.75 -0.36 -0.33 0.73
drat 0.65 -0.68 -0.68 -0.52 1.00 -0.75 0.09 0.45 0.69 0.74 -0.13
    -0.89 0.86 0.90 0.77 -0.75 1.00 -0.23 -0.59 -0.74 -0.68 0.50
qsec 0.47 -0.57 -0.46 -0.67 0.09 -0.23 1.00 0.79 -0.20 -0.15 -0.66
     0.71 -0.81 -0.72 -0.75 0.45 -0.59 0.79 1.00 0.17 0.28 -0.63
VS
     0.56 -0.52 -0.62 -0.36 0.69 -0.74 -0.20 0.17 1.00 0.81 -0.06
gear 0.54 -0.56 -0.59 -0.33 0.74 -0.68 -0.15 0.28 0.81
                                                   1.00 0.11
carb -0.66 0.58 0.54 0.73 -0.13 0.50 -0.66 -0.63 -0.06 0.11 1.00
Sample Size
[1] 32
Probability values (Entries above the diagonal are adjusted for multiple tests.)
     mpg cyl disp hp drat wt gsec vs
                                        am gear carb
mpg 0.00
          0 0.00 0.00 0.00 0.00 0.10 0.00 0.02 0.03 0.00
cvl 0.00
          0 0.00 0.00 0.00 0.00 0.01 0.00 0.04 0.02 0.01
disp 0.00
          hp
    0.00
          0 0.00 0.00 0.04 0.00 0.00 0.00 0.46 0.64 0.00
drat 0.00
          0.00
          0 0.00 0.00 0.00 0.00 1.00 0.01 0.00 0.00 0.05
wt
qsec 0.01
          0 0.01 0.00 0.62 0.21 0.00 0.00 1.00 1.00 0.00
    0.00
VS
          0.00
          0 0.00 0.04 0.00 0.00 0.26 0.36 0.00 0.00 1.00
gear 0.00
          0 0.00 0.06 0.00 0.00 0.42 0.12 0.00 0.00 1.00
carb 0.00
          0 0.00 0.00 0.49 0.00 0.00 0.00 0.73 0.53 0.00
```





corr.test(mtcars, method="kendall")

```
> corr.test(mtcars, method="kendall")
call:corr.test(x = mtcars, method = "kendall")
Correlation matrix
           cvl disp
                        hp drat
                                   wt gsec
                                            VS
                                                    am gear
    1.00 -0.80 -0.77 -0.74 0.46 -0.73 0.32 0.59 0.47 0.43 -0.50
mpg
cyl -0.80 1.00 0.81 0.79 -0.55 0.73 -0.45 -0.77 -0.49 -0.51 0.47
disp -0.77 0.81 1.00 0.67 -0.50 0.74 -0.30 -0.60 -0.52 -0.48 0.41
    -0.74 0.79 0.67 1.00 -0.38 0.61 -0.47 -0.63 -0.30 -0.28 0.60
drat 0.46 -0.55 -0.50 -0.38 1.00 -0.55 0.03 0.38 0.58 0.58 -0.10
    -0.73 0.73 0.74 0.61 -0.55 1.00 -0.14 -0.49 -0.61 -0.54 0.37
gsec 0.32 -0.45 -0.30 -0.47 0.03 -0.14 1.00 0.66 -0.17 -0.09 -0.51
     0.59 -0.77 -0.60 -0.63 0.38 -0.49 0.66 1.00 0.17 0.27 -0.58
     0.47 -0.49 -0.52 -0.30 0.58 -0.61 -0.17 0.17 1.00 0.77 -0.06
dear 0.43 -0.51 -0.48 -0.28 0.58 -0.54 -0.09 0.27 0.77 1.00 0.10
carb -0.50 0.47 0.41 0.60 -0.10 0.37 -0.51 -0.58 -0.06 0.10 1.00
Sample Size
[1] 32
Probability values (Entries above the diagonal are adjusted for multiple tests.)
     mpg cyl disp hp drat
                             wt gsec vs
                                           am gear carb
mpg 0.00 0.00 0.00 0.00 0.15 0.00 1.00 0.01 0.15 0.24 0.09
cyl 0.00 0.00 0.00 0.00 0.04 0.00 0.19 0.00 0.10 0.08 0.15
disp 0.00 0.00 0.00 0.00 0.10 0.00 1.00 0.01 0.07 0.14 0.32
    0.00 0.00 0.00 0.00 0.49 0.01 0.14 0.00 1.00 1.00 0.01
drat 0.01 0.00 0.00 0.03 0.00 0.04 1.00 0.52 0.02 0.02 1.00
    0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.11 0.01 0.04 0.52
gsec 0.08 0.01 0.09 0.01 0.86 0.44 0.00 0.00 1.00 1.00 0.09
    0.01 0.00 0.00 0.09 0.00 0.00 0.36 0.36 0.00 0.00 1.00
gear 0.01 0.00 0.01 0.12 0.00 0.00 0.62 0.14 0.00 0.00 1.00
carb 0.00 0.01 0.02 0.00 0.60 0.04 0.00 0.00 0.75 0.59 0.00
```





Поиск зависимости

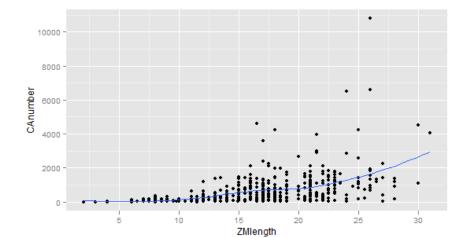
head(dat)					
	Month	Lake	Site	${\it ZMlength}$	CAnumber
1	May	Batorino	S 3	14.9	36
2	May	Batorino	S 3	14.0	30
3	May	Batorino	S 3	13.0	331
4	May	Batorino	S 3	14.0	110
5	May	Batorino	53	12.0	4

May Batorino

Таблица dat содержит данные по уровню зараженности двустворчатого моллюска Dreissena polymorpha инфузорией-комменсалом Conchophthirus acuminatus в трех озерах Беларуси, различающихся по уровню трофности. Нас интересуют, в частности, две переменные: длина раковины моллюска (ZMlength, мм) и число обнаруженных в моллюлюске инфузорий (CAnumber). На приведенном ниже рисунке прослеживается положительная связь между этими двумя переменными (замечание: в рассматриваемом примере анализируются все данные, без разделения по озерам). Вопрос, однако, состоит в том, насколько сильна эта связь. Оценить ее поможет коэффициент корреляции.

14.0

171







Поиск зависимости

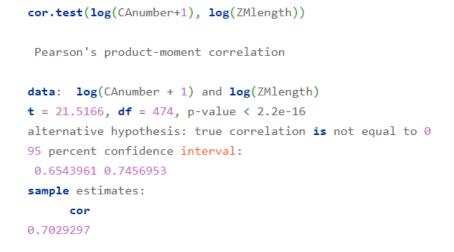
```
attach(dat)

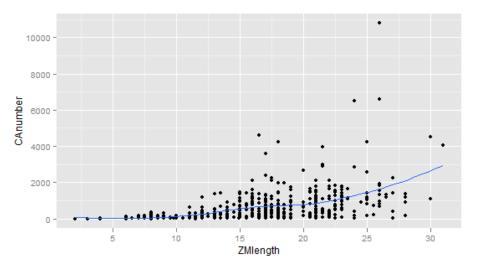
cor.test(CAnumber, ZMlength)

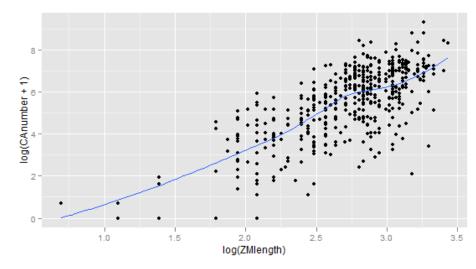
Pearson's product-moment correlation

data: CAnumber and ZMlength
t = 11.4964, df = 474, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
    0.3935877    0.5343949

sample estimates:
    cor
    0.466946</pre>
```











Визуализация





Кореллограммы





Кореллограммы

- Матрицы корреляции это один из основных элементов многомерной статистики. Какие переменные из рассматриваемых сильно коррелируют друг с другом, а какие нет? Существуют ли кластеры переменных, которые связаны между собой определенным способом? С увеличением числа переменных ответить на такие вопросы становится все сложнее.
- Кореллограммы это способ для визуализации корреляционных матриц.
- Пример corr.test(mtcars, method="pearson")

```
Correlation matrix
            cyl disp
                         hp
                            drat
                                        qsec
     1.00 -0.85 -0.85 -0.78 0.68 -0.87 0.42 0.66
    -0.85 1.00 0.90 0.83 -0.70 0.78 -0.59 -0.81 -0.52 -0.49
cyl
disp -0.85 0.90 1.00 0.79 -0.71
                                  0.89 -0.43 -0.71 -0.59 -0.56
                 0.79 1.00 -0.45
                                  0.66 -0.71 -0.72 -0.24 -0.13
drat 0.68 -0.70 -0.71 -0.45 1.00 -0.71
                                        0.09 0.44
                       0.66 - 0.71
                                  1.00 -0.17 -0.55 -0.69 -0.58
gsec 0.42 -0.59 -0.43 -0.71 0.09 -0.17
                                        1.00
                                             0.74 -0.23 -0.21 -0.66
     0.66 -0.81 -0.71 -0.72
                            0.44 - 0.55
                                        0.74
                                              1.00
                                                    0.17
     0.60 -0.52 -0.59 -0.24
                            0.71 -0.69 -0.23
                                              0.17
                                                    1.00
gear 0.48 -0.49 -0.56 -0.13 0.70 -0.58 -0.21
                                              0.21
                                                    0.79
carb -0.55 0.53 0.39 0.75 -0.09 0.43 -0.66 -0.57
```



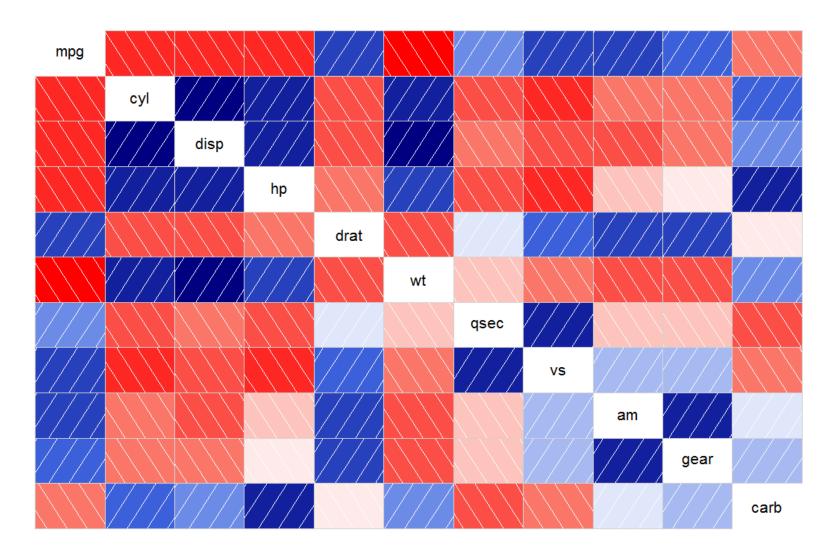
Аннотация к corrgram(). пакет corrgram

- По умолчанию голубой цвет и штриховка из левого нижнего угла к правому верхнему соответствуют положительной корреляции между двумя переменными, на пересечении которых находится данная ячейка.
- Напротив, красный цвет и штриховка из верхнего левого угла к правому нижнему соответствуют отрицательной корреляции. Чем темнее и насыщеннее цвет, тем сильнее корреляция.
- Слабые, близкие к нулю корреляции будут представлены
 "выцветшими" ячейками. На представленной диаграмме порядок
 строк и столбцов был автоматически изменен по результатам
 анализа главных компонент так, чтобы переменные со сходной
 корреляционной структурой формировали кластеры.





corrgram(). пакет corrgram

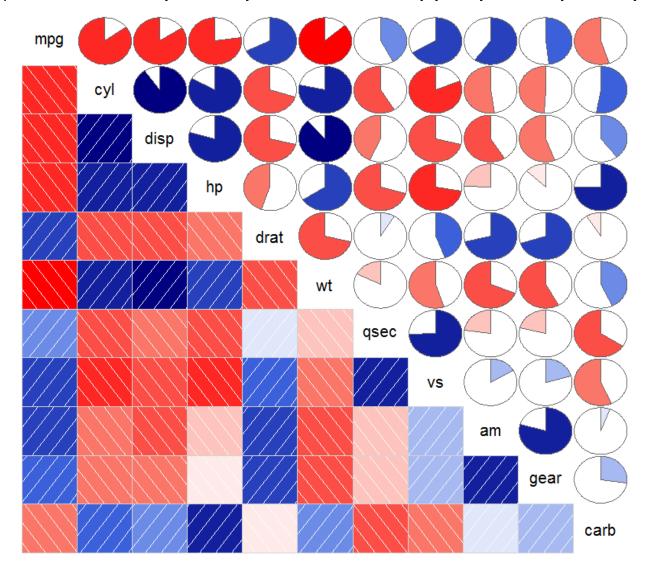






corrgram(). пакет corrgram

corrgram(mtcars,lower.panel=panel.shade, upper.panel=panel.pie)







Аннотация к corrgram()

- На верхнем треугольнике диаграммы та же информация представлена в виде круговых диаграмм.
- Цвета имеют такое же значение, а сила корреляции выражена в размере закрашенного сегмента круговой диаграммы.
- Сегменты, соответствующие положительным корреляциям, начинаются от положения «12 часов» и заполняют круг по часовой стрелке.
- Сегменты, соответствующие отрицательным корреляциям, заполняют круг против часовой стрелки.





Параметры

- corrgram(x, order=, panel=, text.panel=, diag.panel=),
- x это таблица данных с одним наблюдением на строку.
- Если order=TRUE, то порядок переменных изменяется согласно результатам анализа главных компонент корреляционной матрицы
- Параметр panel определяет вид диаграммы (кроме главной диагонали – ее свойства задаются отдельно). Вместо него можно использовать параметры lower.panel и upper.panel, чтобы отдельно определять вид нижней и верхней (по отношению к главной диагонали) половин диаграммы.
- Параметры text.panel и diag.panel относятся к главной диагонали.





Допустимые значения параметра panel

Положение	Значение параметра	Описание
Не на главной диагонали	panel.pie	Закрашенный сегмент круговой диаграммы соответствует силе корреляции
lower.panel=	panel.shade	Интенсивность цвета соответствует силе корреляции
upper.panel=	panel.ellipse	Изображаются доверительный эллипс и сглаженная линия
	panel.pts	Изображается диаграмма рассеяния
diag.panel=	panel.minmax	Приводятся минимальное и максимальное значения переменной
text.panel=	panel.txt	Отображается название переменной

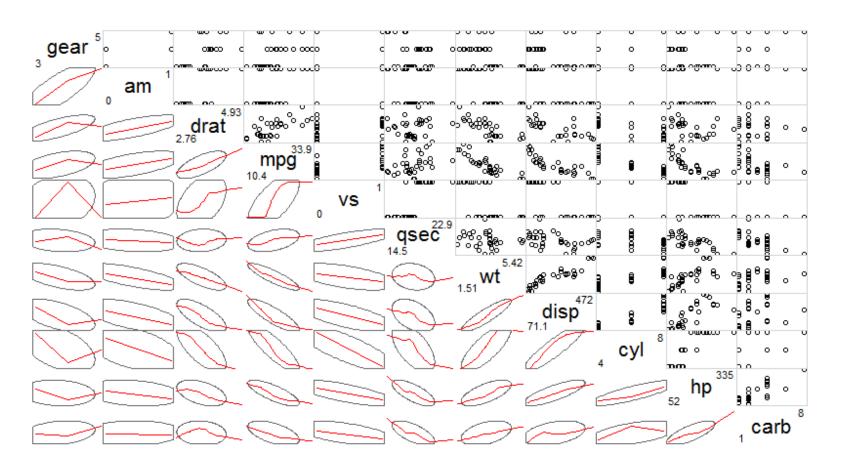




Параметры

corrgram(mtcars, order=TRUE, lower.panel=panel.ellipse,upper.panel=panel.pts, text.panel=panel.txt,diag.panel=panel.minmax,main="Кореллограмма")

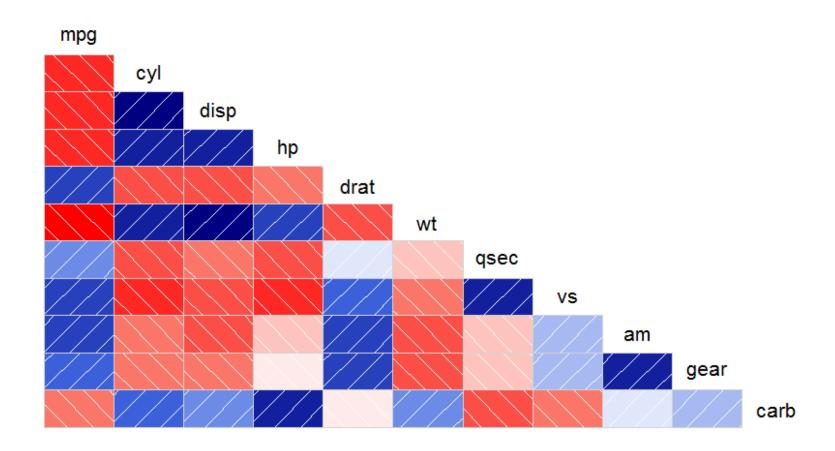
Кореллограмма



Параметры

corrgram(mtcars, lower.panel=panel.shade,upper.panel=NULL, text.panel=panel.txt,main="Pacxoд топлива")

Расход топлива



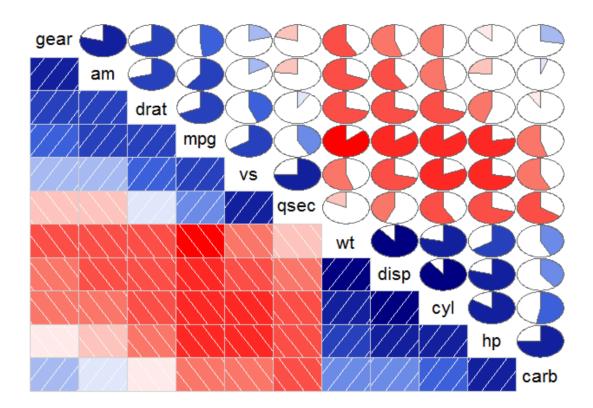


Управление цветом. Определяются 4 цвета

col.corrgram <- function(ncol){colorRampPalette(c("darkgoldenrod4", "burlywood1", "darkkhaki", "darkgreen"))(ncol)}

corrgram(mtcars, order=TRUE, lower.panel=panel.shade, upper.panel=panel.pie, text.panel=panel.txt, main="Другие цвета")

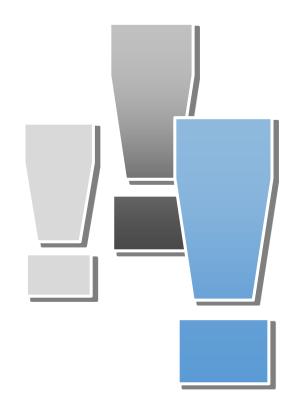
Другие цвета







Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university +7(903)144-53-57



