Лабораторная работа 1

Бочкарев Никита

10 12 2020

# раздел 1. проверка статистических данных.

В этом разделе необходимо собрать данные по социально-экономическим показателям в регионах РФ и изучить их взаимосвязи.

**требуется построить:**

* коробчатые диаграммы распределения показателей.
* Тест Шапира-Уилка на нормальность: таблица со статистикой теста, критическим значением и выводом по каждой переменной.
* Точечные графики взаимного разброса (?pairs), сделать маркеры полупрозрачными.
* Взаимосвязь показателей: графическое представление корреляционной матрицы с помощью функции corrplot(). Незначимые коэффициенты корреляции скрыть, остальные показать кругами.

## импорт данных:

## [1] 94 8

## 'data.frame': 94 obs. of 8 variables:  
## $ Reg.code: int 1000 1001 31 32 33 36 37 40 44 46 ...  
## $ Region : chr "Российская федерация" "Центральный федеральный округ" "Белгородская область" "Брянская область" ...  
## $ FO : chr "" "ЦФО" "ЦФО" "ЦФО" ...  
## $ y : num 1.39e+08 2.41e+07 7.31e+05 2.86e+05 3.92e+05 ...  
## $ x1 : int 27794378 3578186 147214 62255 71513 264660 25713 92707 27345 73745 ...  
## $ x2 : num 1865774 NA 24973 12972 10305 ...  
## $ x3 : num 8653.6 95 97.5 96.1 96.1 ...  
## $ x4 : num 1828493 482661 1921 548 3767 ...

## 'data.frame': 94 obs. of 8 variables:  
## $ Reg.code: int 1000 1001 31 32 33 36 37 40 44 46 ...  
## $ Region : chr "Российская федерация" "Центральный федеральный округ" "Белгородская область" "Брянская область" ...  
## $ FO : Factor w/ 9 levels "","ДФО","ПФО",..: 1 8 8 8 8 8 8 8 8 8 ...  
## $ y : num 1.39e+08 2.41e+07 7.31e+05 2.86e+05 3.92e+05 ...  
## $ x1 : int 27794378 3578186 147214 62255 71513 264660 25713 92707 27345 73745 ...  
## $ x2 : num 1865774 NA 24973 12972 10305 ...  
## $ x3 : num 8653.6 95 97.5 96.1 96.1 ...  
## $ x4 : num 1828493 482661 1921 548 3767 ...

были созданы 2 фрэйма DF и reg.df в которых находятся данные вместе с названиями округов и с показателями для дальнейших подсчетов соответственно.

Эти 2 фрейма содержут: DF 94 строки и 8 столбцов reg.df - 76 строк и 6 столбцов.

Структура этих фреймов:

1:

## 'data.frame': 94 obs. of 8 variables:  
## $ Reg.code: int 1000 1001 31 32 33 36 37 40 44 46 ...  
## $ Region : chr "Российская федерация" "Центральный федеральный округ" "Белгородская область" "Брянская область" ...  
## $ FO : Factor w/ 9 levels "","ДФО","ПФО",..: 1 8 8 8 8 8 8 8 8 8 ...  
## $ y : num 1.39e+08 2.41e+07 7.31e+05 2.86e+05 3.92e+05 ...  
## $ x1 : int 27794378 3578186 147214 62255 71513 264660 25713 92707 27345 73745 ...  
## $ x2 : num 1865774 NA 24973 12972 10305 ...  
## $ x3 : num 8653.6 95 97.5 96.1 96.1 ...  
## $ x4 : num 1828493 482661 1921 548 3767 ...

2:

## 'data.frame': 76 obs. of 6 variables:  
## $ FO: Factor w/ 9 levels "","ДФО","ПФО",..: 8 8 8 8 8 8 8 8 8 8 ...  
## $ y : num 730562 285848 392052 841376 179633 ...  
## $ x1: int 147214 62255 71513 264660 25713 92707 27345 73745 116134 623918 ...  
## $ x2: num 24973 12972 10305 20994 5909 ...  
## $ x3: num 97.5 96.1 96.1 95.6 96.3 97.4 87.9 94.6 96.7 95.6 ...  
## $ x4: num 1921 548 3767 6380 713 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:9] 37 44 58 64 76 82 83 84 85  
## ..- attr(\*, "names")= chr [1:9] "42" "50" "65" "72" ...

## подсчет описательной статистики

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | y | x1 | x2 | x3 | x4 |
| Среднее | 849746.6 | 167742.2 | 21683.6 | 93.3 | 11553.0 |
| Стандартное отклонение | 1733487.8 | 233613.5 | 48111.8 | 5.4 | 40954.5 |
| Коэффициент вариации, % | 204.0 | 139.3 | 221.9 | 5.8 | 354.5 |

**вывод:** показатели неоднородные кроме x3

## проверка показателей на нормальное распределение с помощью теста Шапиро-Уилка

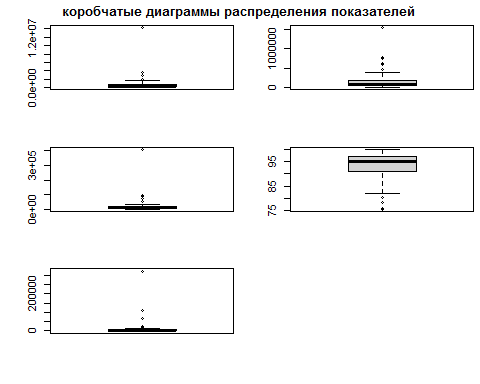
# тест Шапиро Уилка на нормальность

|  |  |  |  |
| --- | --- | --- | --- |
|  | st | krit.zn | p |
| y.W | 0.3740 | 4.2734 | 0e+00 |
| x1.W | 0.6088 | 6.2597 | 0e+00 |
| x2.W | 0.3050 | 3.9290 | 0e+00 |
| x3.W | 0.8593 | 149.3074 | 6e-07 |
| x4.W | 0.2751 | 2.4592 | 0e+00 |

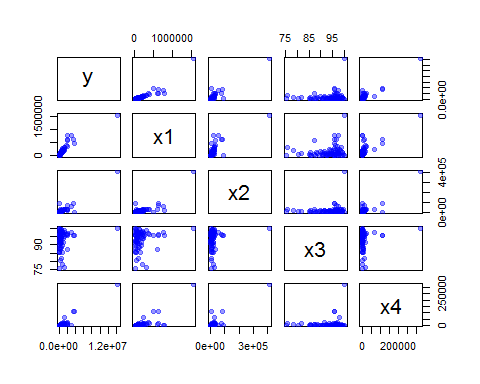
Можно сделать вывод о том, что только x3 имеет нормальное распределение, так как p-значение больше уровня значимости 0,05.

## Построение диаграмм, графиков разброса и матрици корреляции

# коробчатые диаграммы

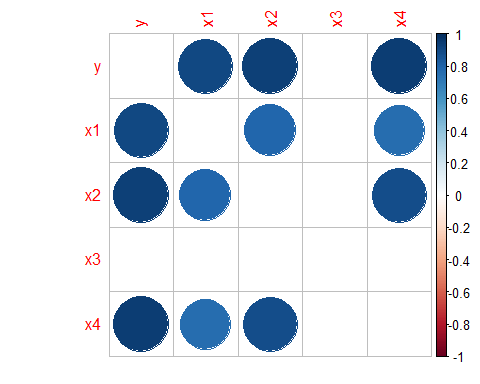


# график разброса



Можно сделать вывод о том, что связь между главным и зависимыми переменными непрослеживается.

# матрица корреляции



# вывод

с y связаны переменные x1, x2, x4, x3

# раздел 2. работа с логарифмированными данными

## описательная статистика

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | y | x1 | x2 | x3 | x4 |
| Среднее | 12.9749 | 11.4274 | 9.4011 | 4.5341 | 7.5015 |
| Стандартное отклонение | 1.0836 | 1.0955 | 0.9220 | 0.0612 | 1.7898 |
| Коэффициент вариации, % | 8.3515 | 9.5866 | 9.8074 | 1.3498 | 23.8592 |

После логарифмирования данные стали однородными, так как коэффициенты вариации не привышают 30 %.

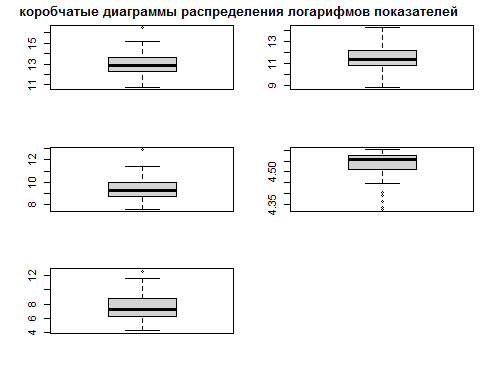
# статистические тесты (Шапиро Уилка на нормальность)

|  |  |  |  |
| --- | --- | --- | --- |
|  | st1 | krit.zn1 | p1 |
| y.W | 0.9825 | 104.3889 | 0.3778478 |
| x1.W | 0.9934 | 90.9380 | 0.9644872 |
| x2.W | 0.9615 | 88.8918 | 0.0209815 |
| x3.W | 0.8358 | 645.4875 | 0.0000001 |
| x4.W | 0.9807 | 36.5380 | 0.3006009 |

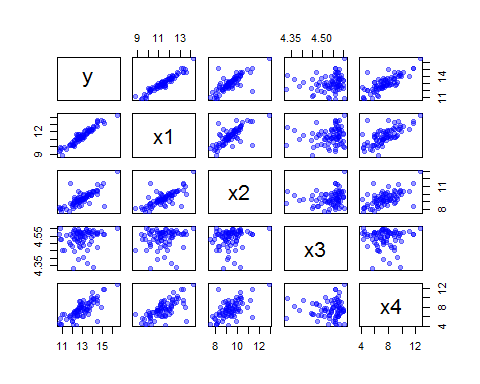
Большинство переменных больше уровня значимости.

## Построение диаграмм, графиков разброса и матрици корреляции

# коробчатые диаграммы

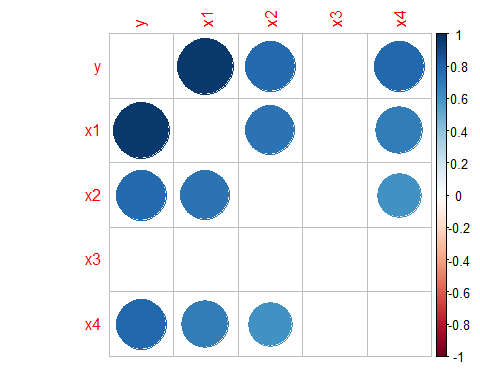


# график разброса



Между зависимой и объясняющими переменными наблюдается положительная свзь.

# Корреляционная матрица



# вывод

с Y наибольшую связь имеет переменная x1 и x4 ,x2