Лабораторная работа 1

Вдовин Иван

03 02 2021

# раздел 1. проверка статистических данных.

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

**требуется построить:**

* Гистограммы распределенния показателей с наложением графиков функций распределения.
* Тест Андерсена на нормальность: таблица со статистикой теста 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 овалами.

## 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 94.4 94.7 99.1 92.6 90.2 92.7 97.3 96.1 99.3 93.2 ...  
## $ x1 : num 69254134 24135019 730562 285848 392052 ...  
## $ x2 : num 28306 9531830 298670 219879 200149 ...  
## $ x3 : num 1192 NA 7104 7682 8239 ...  
## $ x4 : int 4764483 1748045 37351 22072 34662 60358 34209 27930 17256 23467 ...

## '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 94.4 94.7 99.1 92.6 90.2 92.7 97.3 96.1 99.3 93.2 ...  
## $ x1 : num 69254134 24135019 730562 285848 392052 ...  
## $ x2 : num 28306 9531830 298670 219879 200149 ...  
## $ x3 : num 1192 NA 7104 7682 8239 ...  
## $ x4 : int 4764483 1748045 37351 22072 34662 60358 34209 27930 17256 23467 ...

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 94.4 94.7 99.1 92.6 90.2 92.7 97.3 96.1 99.3 93.2 ...  
## $ x1 : num 69254134 24135019 730562 285848 392052 ...  
## $ x2 : num 28306 9531830 298670 219879 200149 ...  
## $ x3 : num 1192 NA 7104 7682 8239 ...  
## $ x4 : int 4764483 1748045 37351 22072 34662 60358 34209 27930 17256 23467 ...

2:

## 'data.frame': 77 obs. of 6 variables:  
## $ FO: Factor w/ 9 levels "","ДФО","ПФО",..: 8 8 8 8 8 8 8 8 8 8 ...  
## $ y : num 99.1 92.6 90.2 92.7 97.3 96.1 99.3 93.2 96.5 99.9 ...  
## $ x1: num 730562 285848 392052 841376 179633 ...  
## $ x2: num 298670 219879 200149 487054 147373 ...  
## $ x3: num 7104 7682 8239 11588 6416 ...  
## $ x4: int 37351 22072 34662 60358 34209 27930 17256 23467 22458 240423 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:8] 37 43 44 58 64 76 84 85  
## ..- attr(\*, "names")= chr [1:8] "42" "49" "50" "65" ...

## подсчет описательной статистики

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | y | x1 | x2 | x3 | x4 |
| Среднее | 94.6 | 842124.4 | 336103.2 | 14590.0 | 57970.5 |
| Стандартное отклонение | 5.5 | 1723790.8 | 554122.3 | 27320.6 | 123057.0 |
| Коэффициент вариации, % | 5.8 | 204.7 | 164.9 | 187.3 | 212.3 |

**вывод:** показатели неоднородные кроме Y

## проверка показателей на нормальное распределение с помощью теста Андерсена

# тест Андерсена на нормальность

##   
## Anderson-Darling normality test  
##   
## data: reg.df$y  
## A = 4.0855, p-value = 2.934e-10

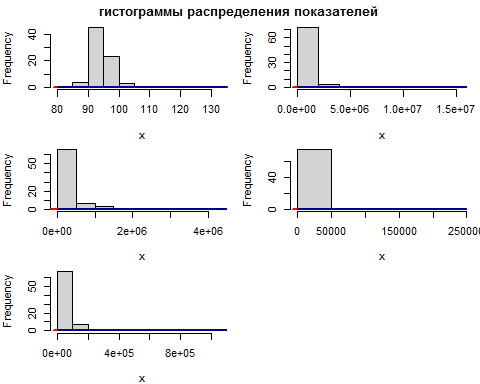
## y x1   
## statistic 4.085474 13.39444   
## p.value 2.933854e-10 3.7e-24   
## method "Anderson-Darling normality test" "Anderson-Darling normality test"  
## data.name "X[[i]]" "X[[i]]"   
## x2 x3   
## statistic 11.68081 14.11769   
## p.value 3.7e-24 3.7e-24   
## method "Anderson-Darling normality test" "Anderson-Darling normality test"  
## data.name "X[[i]]" "X[[i]]"   
## x4   
## statistic 14.63099   
## p.value 3.7e-24   
## method "Anderson-Darling normality test"  
## data.name "X[[i]]"

|  |  |  |
| --- | --- | --- |
|  | st | p.value |
| y.A | 4.0855 | 0e+00 |
| x1.A | 13.3944 | 1e-04 |
| x2.A | 11.6808 | 0e+00 |
| x3.A | 14.1177 | 0e+00 |
| x4.A | 14.6310 | 1e-04 |

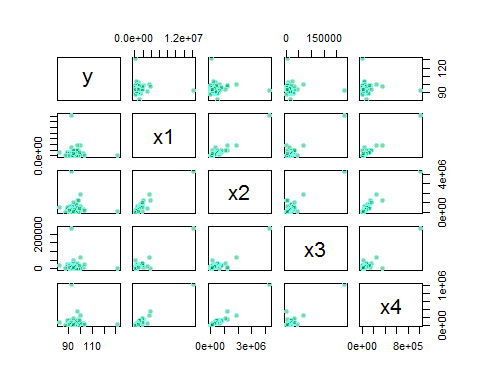
Можно сделать вывод о том, что ни один показатель не имеет нормальное распределение, так как p-значение меньше уровня значимости 0,05.

## Построение гистограмм, графиков разброса и матрици корреляции

# гистограммы

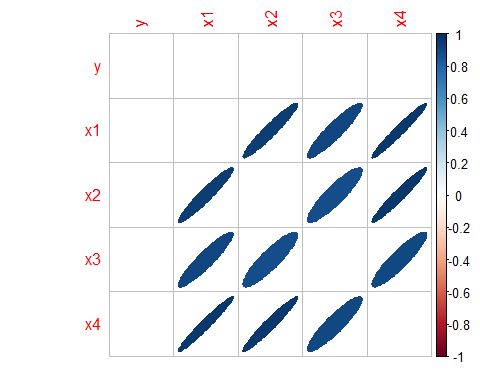


# график разброса



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

# матрица корреляции



# вывод

Наблюдается мультиколлинеарность, так как наблюдается связь с x1 признаков x2, x3, x4.

# раздел 2. работа с логарифмированными данными

## описательная статистика

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | y | x1 | x2 | x3 | x4 |
| Среднее | 4.5480 | 12.9630 | 12.1422 | 9.1029 | 10.2917 |
| Стандартное отклонение | 0.0533 | 1.0844 | 1.0612 | 0.8847 | 1.0731 |
| Коэффициент вариации, % | 1.1719 | 8.3653 | 8.7398 | 9.7189 | 10.4268 |

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

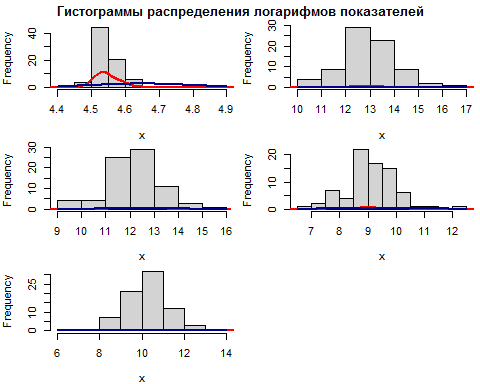
# статистические тесты Андерсена на нормальность

## y x1   
## statistic 2.98857 0.2964952   
## p.value 1.414672e-07 0.5840047   
## method "Anderson-Darling normality test" "Anderson-Darling normality test"  
## data.name "X[[i]]" "X[[i]]"   
## x2 x3   
## statistic 0.9747544 0.5610929   
## p.value 0.01344086 0.1421589   
## method "Anderson-Darling normality test" "Anderson-Darling normality test"  
## data.name "X[[i]]" "X[[i]]"   
## x4   
## statistic 0.5538003   
## p.value 0.1484005   
## method "Anderson-Darling normality test"  
## data.name "X[[i]]"

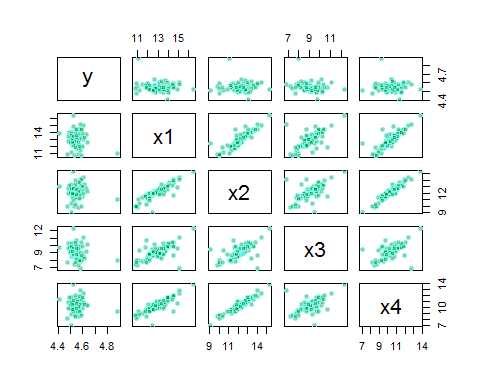
|  |  |  |
| --- | --- | --- |
|  | st1 | p.value1 |
| y.A | 2.9886 | 0 |
| x1.A | 0.2965 | 0 |
| x2.A | 0.9748 | 0 |
| x3.A | 0.5611 | 0 |
| x4.A | 0.5538 | 0 |

## Построение гистограмм, графиков разброса и матрици корреляции

# гистограммы

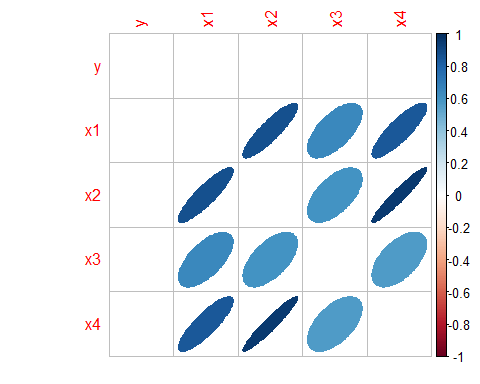


# график разброса



Между зависимой и объясняющими переменными наблюдается положительная свзь и между зависимыми переменными наблюдается мультиколлинеарность.

# Корреляционная матрица



# вывод

Наблюдается влияние x1 x2 на x3 и x3 на x4.