Labroratory work №1. Research of data by regions of Russian Federation. Variant №8

Andrey Vladimirov

дек 04 20

# SECTION 0. Uploading data

Full list of data you can find at the file Data\_Vladimirov.csv, located in the project directory.

Below are the first five values from the table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Federal\_district | Y | X1 | X2 | X3 | X4 |
| ЦФО | 192504 | 150 | 97.9 | 83670.0 | 99.1 |
| ЦФО | 179767 | 107 | 96.4 | 52847.1 | 92.6 |
| ЦФО | 143643 | 141 | 94.5 | 57072.8 | 90.2 |
| ЦФО | 208638 | 158 | 95.1 | 105084.7 | 92.7 |
| ЦФО | 143568 | 204 | 97.6 | 40014.8 | 97.3 |
| ЦФО | 175286 | 144 | 95.2 | 63379.7 | 96.1 |

File contains 85 observations and 6 variables.

### Designations:

* Regions - full list of the Russian Federation region;
* Federal district - federal district of the region;
* Y - retail trade turnover per capita in 2016;
* X1 - the number of small businesses per 10000 people in 2016;
* X2 - the use of information and communication technologies in organizations: personal computers in 2016;
* X3 - expenditures of the consolidated budgets of the constituent entities of the Russian Federation: total in 2015;
* X4 - real cash income of the population in 2016.

# SECTION 1. Working with variables

### Description:

In this section need to do the following actions:

1. find actual distribution densities of variables with superimposed theoretical normal law densities, and place graphs on one canvas;
2. perform the Shapiro-Wilk test for normality: a table with test statistics, P-value and output for each variable;
3. construct cross scatter point plots, make markers semi-transparent;
4. Study the relationship of indicators: graphical representation of the correlation matrix, hide insignificant correlation coefficients, show the rest with squares.

### Implementation:

#### First step

Calculate factors descriptive statistics, standard deviations and coefficients of variation.

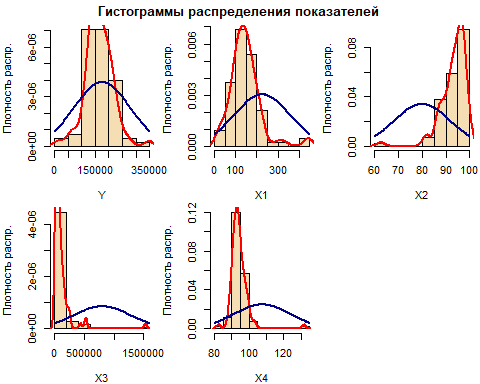
General descriptive statistics (function ‘summary’):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | X3 | X4 |
|  | Min. : 12227 | Min. : 20.0 | Min. : 62.60 | Min. : 11766 | Min. : 82.30 |
|  | 1st Qu.:136870 | 1st Qu.:106.0 | 1st Qu.: 91.40 | 1st Qu.: 40015 | 1st Qu.: 91.80 |
|  | Median :162400 | Median :141.0 | Median : 94.60 | Median : 68027 | Median : 93.90 |
|  | Mean :164267 | Mean :146.8 | Mean : 93.32 | Mean : 115783 | Mean : 94.39 |
|  | 3rd Qu.:195514 | 3rd Qu.:179.0 | 3rd Qu.: 97.00 | 3rd Qu.: 122364 | 3rd Qu.: 96.10 |
|  | Max. :346602 | Max. :445.0 | Max. :100.00 | Max. :1529127 | Max. :131.60 |

Table with mean, standard deviations and coefficients of variation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | X3 | X4 |
| Mean | 164266.66 | 146.85 | 93.32 | 115782.92 | 94.39 |
| Standard deviations | 49948.84 | 72.38 | 5.46 | 183242.89 | 5.31 |
| Coefficient of variation | 30.41 | 49.29 | 5.85 | 158.26 | 5.63 |

**Conclusion**: the coefficients of variation show that the data is heterogeneous, as some percentages are greater than 20 percent.



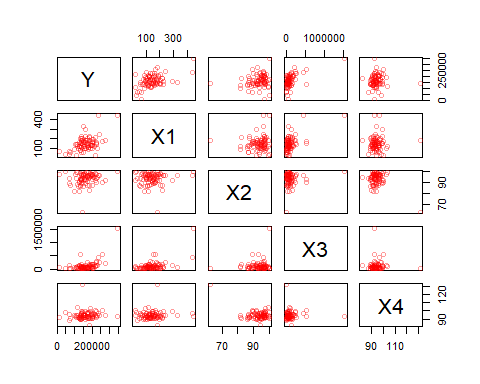
**Conclusion**: distributions of all indicators are asymmetric compared to the normal law.

#### Second step

|  |  |  |
| --- | --- | --- |
|  | p\_value | test\_stat |
| Y | 0.05 | 0.97 |
| X1 | 0.00 | 0.88 |
| X2 | 0.00 | 0.81 |
| X3 | 0.00 | 0.44 |
| X4 | 0.00 | 0.68 |

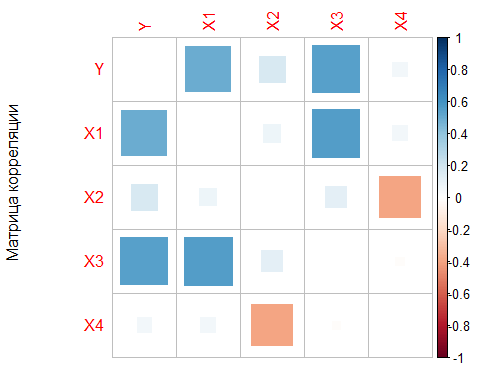
**Conclusion**: the hypothesis of normal distribution is rejected, since all factors have P-values less than the significance level (0.05 or 5%).

#### Third step



**Conclusion**: strict straight linear relationship between Y and X1, Y and X3, X1 and X3, and strict reverse relationship between X2 and X4.

#### Fourth step



**Conclusion**: meaningful connection between Y and X1, Y and X3, X1 and X3, X2 and X4.

# SECTION 2

### Description:

In this section we need to do the same actions as in the first section, but with the logarithm data.

### Implementation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Federal\_district | Y | X1 | X2 | X3 | X4 |
| ЦФО | 12.17 | 5.01 | 4.58 | 11.33 | 4.60 |
| ЦФО | 12.10 | 4.67 | 4.57 | 10.88 | 4.53 |
| ЦФО | 11.88 | 4.95 | 4.55 | 10.95 | 4.50 |
| ЦФО | 12.25 | 5.06 | 4.55 | 11.56 | 4.53 |
| ЦФО | 11.87 | 5.32 | 4.58 | 10.60 | 4.58 |
| ЦФО | 12.07 | 4.97 | 4.56 | 11.06 | 4.57 |

#### First step

Calculate factors descriptive statistics, standard deviations and coefficients of variation.

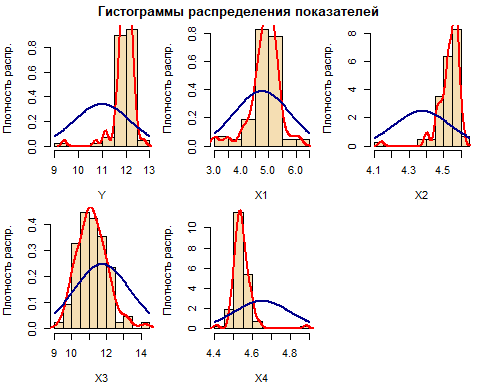
General descriptive statistics (function ‘summary’):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | X3 | X4 |
|  | Min. : 9.41 | Min. :3.000 | Min. :4.140 | Min. : 9.37 | Min. :4.410 |
|  | 1st Qu.:11.83 | 1st Qu.:4.660 | 1st Qu.:4.520 | 1st Qu.:10.60 | 1st Qu.:4.520 |
|  | Median :12.00 | Median :4.950 | Median :4.550 | Median :11.13 | Median :4.540 |
|  | Mean :11.95 | Mean :4.866 | Mean :4.534 | Mean :11.20 | Mean :4.546 |
|  | 3rd Qu.:12.18 | 3rd Qu.:5.190 | 3rd Qu.:4.570 | 3rd Qu.:11.71 | 3rd Qu.:4.570 |
|  | Max. :12.76 | Max. :6.100 | Max. :4.610 | Max. :14.24 | Max. :4.880 |

Table with mean, standard deviations and coefficients of variation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | X3 | X4 |
| Mean | 11.95 | 4.87 | 4.53 | 11.20 | 4.55 |
| Standard deviations | 0.42 | 0.54 | 0.06 | 0.87 | 0.05 |
| Coefficient of variation | 3.51 | 11.09 | 1.32 | 7.77 | 1.10 |

**Conclusion**: the coefficients of variation show that the data is homogeneous, as some percentages are less than 20 percent.



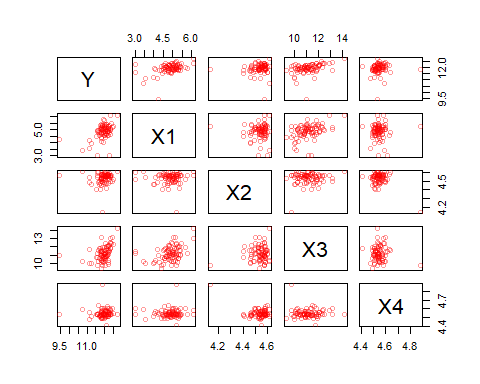
**Conclusion**: distributions of all indicators are asymmetric compared to the normal law.

#### Second step

|  |  |  |
| --- | --- | --- |
|  | p\_value\_log | test\_stat\_log |
| Y | 0.00 | 0.76 |
| X1 | 0.00 | 0.92 |
| X2 | 0.00 | 0.75 |
| X3 | 0.11 | 0.98 |
| X4 | 0.00 | 0.75 |

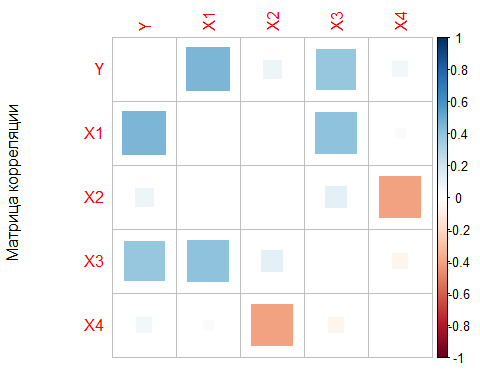
**Conclusion**: the hypothesis of normal distribution is rejected, since Y, X1, X2, X4 factors have P-values less than the significance level (0.05 or 5%). X3 have P-value greater than significance level(0.05 or 5%), therefore we do not reject the hypothesis.

#### Third step



**Conclusion**: strict straight linear relationship between Y and X1, Y and X3, X1 and X3, and strict reverse relationship between X2 and X4.

#### Fourth step



**Conclusion**: meaningful connection between Y and X1, Y and X3, X1 and X3, X2 and X4.