Лабораторная работа № 1

Никишов Дмитрий

27 12 2020

# Раздел I.

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

**Необходимо построить:**

* графики на одном полотне, отразив фактические и теоретические плотности распределения;
* таблицу со статистикой теста Шапиро-Уилка на нормальность;
* точечные графики взаимного разброса с полупрозрачными маркерами;
* графическое представление корреляционной матрицы.

### Обозначения количественных показателей:

* *GRP.2013* – ВРП.
* *IFA.2012* – Инвестиции в основной капитал.
* *ERF.2012* – Расходы консолидированных бюджетов субъектов Российской Федерации: на национальную экономику.
* *UPC.2012* – Использование информационных и коммуникационных технологий в организациях: персональные компьютеры.
* *IRDC.2012* – Внутренние затраты на научные исследования и разработки.

## Импорт данных:

# импорт данных из .csv  
file.path <- 'Nikishov.csv'  
DF <- read.csv2(file.path, stringsAsFactors = F)

Файл с данными содержит 92 строк и 8 столбцов.

# делаем из столбца "FO" фактор  
DF$FO <- factor(DF$FO)  
   
# оставляем только регионы и выбрасываем столбец меток,   
# чтобы удобнее было считать  
reg.df <- DF[DF$Regcode < 1000, c(-1, -2)]  
# выбрасываем пропущенные  
reg.df <- na.omit(reg.df)

## Описательная статистика

# ручной расчёт  
# средние арифметические  
mns <- round(apply(reg.df[, -1], 2, mean), 1)  
# стандартные отклонения  
sds <- round(apply(reg.df[, -1], 2, sd), 1)  
# коэффициенты вариации  
coef.vars <- round(sds / mns \* 100, 1)  
# делаем свою таблицу только с нужными статистиками   
# по количественным показателям: среднее, СКО, коэффициент вариации  
smm <- rbind(mns, sds, coef.vars)  
# названия статистик -- заголовки строк  
row.names(smm) <- c('Среднее', 'Стандартное отклонение',  
 'Коэффициент вариации, %')  
kable(smm)

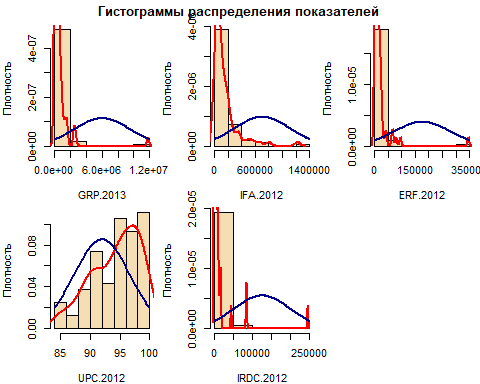
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | GRP.2013 | IFA.2012 | ERF.2012 | UPC.2012 | IRDC.2012 |
| Среднее | 653467.5 | 151577.1 | 19420.9 | 94.3 | 8555.7 |
| Стандартное отклонение | 1373311.9 | 193221.0 | 40452.3 | 4.1 | 30176.4 |
| Коэффициент вариации, % | 210.2 | 127.5 | 208.3 | 4.3 | 352.7 |

**Вывод:** показатели неоднородны.

## Анализ распределения данных

Построим гистограммы с их фактическими плотностями распределений и наложим сверху теоретическую нормальную плотность распределения.

# строим гистограммы на одном полотне по изначальным данным  
par(mfrow = c(2, 3)) # разбить полотно на 6 частей, 2x3  
par(oma = c(0, 0, 1.5, 0)) # внешние поля общего полотна  
par(mar = c(4, 4, 0.5, 0.5)) # внутренние поля каждого графика  
# цикл по номерам столбцов с количественными переменными  
for (i in 2:6) {  
 # данные -- i-ый столбец фрейма  
 x <- reg.df[, i]  
   
 # гистограмма  
 hist(x,  
 freq = F, # по вертикали – плотность (доля)  
 col = 'wheat', # цвет заливки  
 xlab = colnames(reg.df)[i], # название оси X – название столбца   
   
 #ylim = c(0, 0.00001),  
 ylab = 'Плотность', # название оси Y  
 main = '') # без заголовка  
   
 #фактическая плотность  
 lines(density(x),col = "red", lwd = 2)  
   
 # теоретическая плотность  
 curve(dnorm(x, mean = mean(x), sd = sd(x)), col = 'darkblue',   
 lwd = 2, add = TRUE)  
}  
# общий заголовок для всех графиков  
title(main = 'Гистограммы распределения показателей',   
 outer = TRUE, cex = 1.5)  
# вернуть настройки обратно, 1x1  
par(mfrow = c(1, 1))



**Вывод:** Судя по гистограммам, распределения всех показателей ассиметричны по сравнению с нормальным законом.(Особенно это заметно у *“GRP.2013”*, *“IFA.2012”*, *“ERF.2012”* и *“IRDC.2012”*) Кроме того, показатели имеют более островершинные распределения по сравнению с нормальным законом.

## Тест Шапиро-Уилка на нормальность распределения

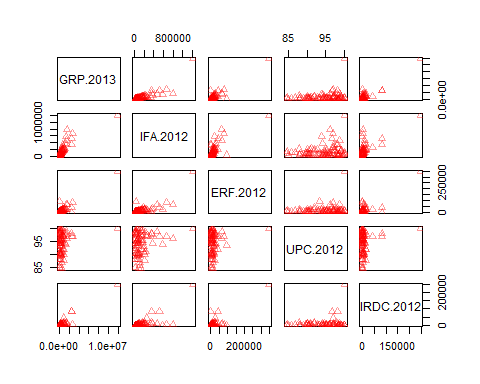
# применяем ко всем столбцам и вытаскиваем только тестовую статистику  
# для изначальных значений  
W <- sapply(reg.df[, 2:6], function(x) {  
 round(shapiro.test(x)$statistic, 4)  
})  
p <- sapply(reg.df[, 2:6], function(x) {  
 round(shapiro.test(x)$p.value, 4)  
})  
table <- data.frame(p, W)  
kable(table)

|  |  |  |
| --- | --- | --- |
|  | p | W |
| GRP.2013 | 0.0000 | 0.3534 |
| IFA.2012 | 0.0000 | 0.6523 |
| ERF.2012 | 0.0000 | 0.3409 |
| UPC.2012 | 0.0011 | 0.9415 |
| IRDC.2012 | 0.0000 | 0.2726 |

**Вывод:** Проверяя гипотезу нормального распределения для переменных на уровне значимости 0.05, мы отвергаем все, так как p-значения < 0.05 .

## Графики разброса

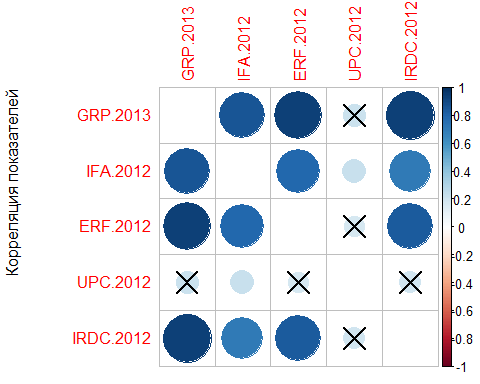
# графики взаимного разброса по изначальным данным  
pairs(reg.df[, -1], # фрейм без первого столбца-фактора  
 pch = 2, # тип символов для точек  
 col = rgb(1, 0, 0, alpha = 0.4), # цвет заливки точек  
 bg = rgb(0, 0, 1, alpha = 0.4), # цвет границы точек  
 cex = 1.1) # масштаб символов для точек



**Вывод:** по первой строке графиков попытаемся определить наличие связи между *GRP.2013* и остальными показателями. Вероятно, что связь отсутствует с показателем *UPC.2012*. Со всеми остальными показателями связь достаточно высокая. Чтобы проверить эти предположения, рассчитаем корреляционную матрицу и проверим коэффициенты на значимость.

## Корреляционная матрица

# коэффициенты Пирсона с P-значениями  
r.corr <- rcorr(as.matrix(reg.df[, -1]))  
# Визуализация корреляционной матрицы   
# сохраняем корреляционную матрицу  
matrix.cor <- r.corr$r  
# сохраняем p-значения  
matrix.p <- r.corr$P  
# изображаем матрицу графически  
corrplot(matrix.cor, # сама корреляционная матрица  
 order = 'original', # порядок отображения показателей   
 # в матрице  
 diag = F, # не отображать значения на главной   
 # диагонали  
 p.mat = matrix.p, # p-значения  
 #insig = 'blank', # метод отображения незначимых  
 sig.level = 0.05) # уровень значимости  
title(ylab = 'Корреляция показателей')



**Вывод:** Взаимосвязь *GRP.2013* значима с показателями *IFA.2012*, *ERF.2012* и *IRDC.2012*.

# Раздел II.

В этом разделе необходимо прологарифмировать данные из первого раздела и провести с новыми данными те же махинации.

## Описательная статистика логарифмированных данных:

#фрейм с логарифмированными данными  
DF1 <- reg.df  
DF1$GRP.2013 <- log(reg.df$GRP.2013)  
DF1$IFA.2012 <- log(reg.df$IFA.2012)  
DF1$ERF.2012 <- log(reg.df$ERF.2012)  
DF1$UPC.2012 <- log(reg.df$UPC.2012)  
DF1$IRDC.2012 <- log(reg.df$IRDC.2012)  
# ручной расчёт  
# средние арифметические  
mns1 <- round(apply(DF1[, -1], 2, mean), 3)  
# стандартные отклонения  
sds1 <- round(apply(DF1[, -1], 2, sd), 3)  
# коэффициенты вариации  
coef.vars1 <- round(sds1 / mns1 \* 100, 3)  
# делаем свою таблицу только с нужными статистиками   
# по количественным показателям: среднее, СКО, коэффициент вариации  
smm1 <- rbind(mns1, sds1, coef.vars1)  
# названия статистик -- заголовки строк  
row.names(smm1) <- c('Среднее', 'Стандартное отклонение',  
 'Коэффициент вариации, %')  
kable(smm1)

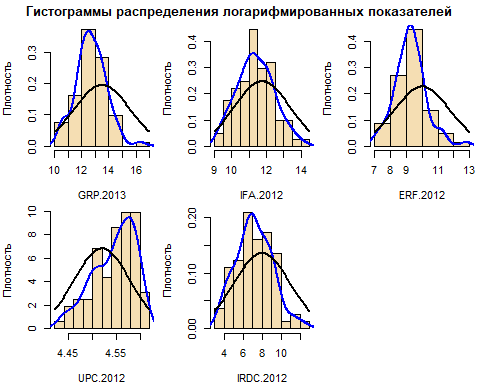
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | GRP.2013 | IFA.2012 | ERF.2012 | UPC.2012 | IRDC.2012 |
| Среднее | 12.708 | 11.385 | 9.273 | 4.546 | 7.188 |
| Стандартное отклонение | 1.096 | 1.043 | 0.978 | 0.044 | 1.871 |
| Коэффициент вариации, % | 8.624 | 9.161 | 10.547 | 0.968 | 26.029 |

**Вывод:** показатели считаются однородными, т.к. коэффициенты вариации не превышают 10%.

## Анализ распределения данных

Построим гистограммы с их фактическими плотностями распределений и наложим сверху теоретическую нормальную плотность распределения.

# строим гистограммы на одном полотне по логарифмированным данным  
par(mfrow = c(2, 3)) # разбить полотно на 6 частей, 2x3  
par(oma = c(0, 0, 1.5, 0)) # внешние поля общего полотна  
par(mar = c(4, 4, 0.5, 0.5)) # внутренние поля каждого графика  
# цикл по номерам столбцов с количественными переменными  
for (i in 2:6) {  
 # данные -- i-ый столбец фрейма  
 x <- DF1[, i]  
   
 # гистограмма  
 hist(x,  
 freq = F, # по вертикали – плотность (доля)  
 col = 'wheat', # цвет заливки  
 xlab = colnames(reg.df)[i], # название оси X – название столбца   
 ylab = 'Плотность', # название оси Y  
 main = '') # без заголовка  
   
 #фактическая плотность  
 lines(density(x),col = "blue", lwd = 2)  
   
 # теоретическая плотность  
 curve(dnorm(x, mean = mean(x), sd = sd(x)), col = 'black',   
 lwd = 2, add = TRUE)  
   
}  
# общий заголовок для всех графиков  
title(main = 'Гистограммы распределения логарифмированных показателей',   
 outer = TRUE, cex = 1.5)  
# вернуть настройки обратно, 1x1  
par(mfrow = c(1, 1))



**Вывод:** Судя по гистограммам, распределения всех показателей близки к нормальному закону, но графики имеют более остроконечные вершины, нежели нормальное.

## Тест Шапиро-Уилка на нормальность распределения

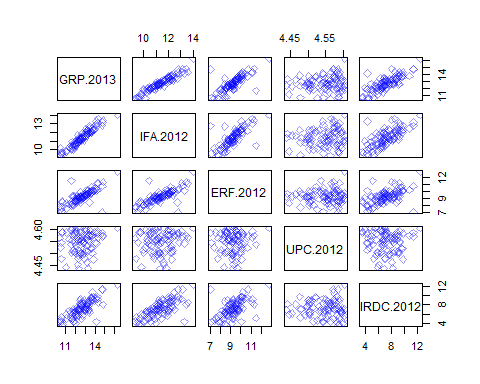
# применяем ко всем столбцам и вытаскиваем только тестовую статистику  
# для изначальных значений  
W <- sapply(DF1[, 2:6], function(x) {  
 round(shapiro.test(x)$statistic, 4)  
})  
p <- sapply(DF1[, 2:6], function(x) {  
 round(shapiro.test(x)$p.value, 4)  
})  
table1 <- data.frame(p, W)  
kable(table1)

|  |  |  |
| --- | --- | --- |
|  | p | W |
| GRP.2013 | 0.4383 | 0.9845 |
| IFA.2012 | 0.8683 | 0.9913 |
| ERF.2012 | 0.1050 | 0.9745 |
| UPC.2012 | 0.0005 | 0.9353 |
| IRDC.2012 | 0.4841 | 0.9853 |

**Вывод:** Проверяя гипотезу нормального распределения для переменных на уровне значимости 0.05, мы отвергаем все, кроме *“UPC.2012”* , так как остальные p-значения < 0.05 .

## Графики разброса

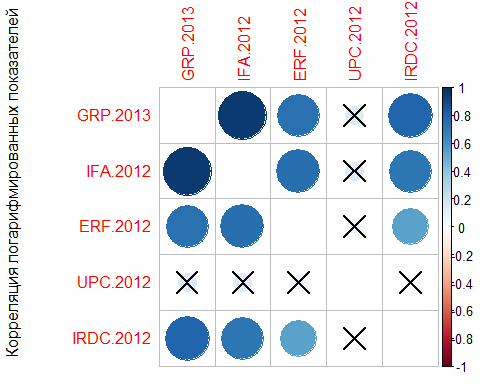
# графики взаимного разброса по логарифмированным данным  
pairs(DF1[, -1], # фрейм без первого столбца-фактора  
 pch = 5, # тип символов для точек  
 col = rgb(0, 0, 1, alpha = 0.4), # цвет заливки точек  
 bg = rgb(0, 0, 1, alpha = 0.4), # цвет границы точек  
 cex = 1.1) # масштаб символов для точек



**Вывод:** по первой строке графиков попытаемся определить наличие связи между *“GRP.2013”* и остальными показателями. Можно предположить, что c *IFA.2012*, *ERF.2012* и *IRDC.2012* показателями существует CBKMYFZ положительная связь . Чтобы проверить эти предположения, рассчитаем корреляционную матрицу и проверим коэффициенты на значимость.

## Корреляционная матрица

# коэффициенты Пирсона с P-значениями  
r.corr <- rcorr(as.matrix(DF1[, -1]))  
# Визуализация корреляционной матрицы =========================================  
# сохраняем корреляционную матрицу  
matrix.cor <- r.corr$r  
# сохраняем p-значения  
matrix.p <- r.corr$P  
# изображаем матрицу графически  
corrplot(matrix.cor, # сама корреляционная матрица  
 order = 'original', # порядок отображения показателей   
 # в матрице  
 diag = F, # не отображать значения на главной   
 # диагонали  
 p.mat = matrix.p, # p-значения  
 #insig = 'blank', # метод отображения незначимых  
 sig.level = 0.05) # уровень значимости  
title(ylab = 'Корреляция логарифмированных показателей')



**Вывод:** Взаимосвязь *GRP.2013* значима с показателями *IFA.2012*, *ERF.2012* и *IRDC.2012*, и при этом она стала сильнее.

# 5. Сохранение рабочего пространства ------------------------------------------  
# сохраняем нужные объекты в файл  
save(list = c('DF', 'reg.df', 'DF1'), file = 'test\_lab1\_Nikishov.RData')  
save.image('test\_lab1\_Nikishov.RData')  
rm()