Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**ЛАБОРАТОРНАЯ РАБОТА №8**

**Дисциплина: Обработка больших данных**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ К. А. Корнилов

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_В. И. Шиян

**Тема:** Корреляционный и регрессионный анализ данных.

**Цель**: Ознакомиться с понятием корреляционный и регрессионный анализ данных, некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научиться оценивать связь между переменными и оценивать степень этой связи.

**Задание:**

1. Определить с помощью уже известной вам команды количество переменных в наборе данных longley;
2. Определить с помощью уже известной вам команды объем выборки в наборе данных longley;
3. Подвести простую статистику (дескриптивный анализ).
4. С помощью функции cor(longley) получить корреляционную матрицу и пояснить ее.
5. Подкрепить свои пояснения графиками.
6. Проверить на нормальность распределения переменных из набора данных longley;
7. Обосновать полученные результаты графически.
8. Вам необходимо выбрать данные по одной стране, согласно вашему варианту, построить кривую прироста ВВП, проанализировать взаимное влияние атрибутов, отобразить корреляцию:
   1. Роста ВВП и прироста населения
   2. Прироста населения на динамику безработицы
   3. Изменения расходов на медицину и увеличения продолжительности жизни и смертность.
   4. Прирост людей с высшим образованием на рост экспорта товаров и на прирост высокотехнологичного производства.
   5. Расходов на образование на – кумулятивный прирост бакалавров среди женщин
   6. Прирост людей с высшим образованием на развитие высоких технологий (прирост статей в научных журналах)
   7. С помощью регрессионного анализа найдите зависимые переменные и поясните влияние на них независимых переменных.
   8. С помощью функции predict() (см. лекции и help()) постройте прогноз по любому атрибуту

**Ход работы**:

1. Было определено количество независимых переменных и количество строк в таблице Longley

Код:

#Задание 8.2.1

data = longley

data

#1 Определяем количество переменных

variable\_number = ncol(data)

variable\_number

#2 Определяем количество строк

row\_number = nrow(data)

row\_number

Результат представлен на рисунке 1.

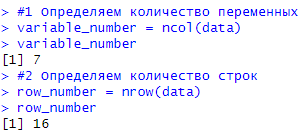


Рисунок 1 – Количество независимых переменных и количество строк в наборе данных

1. Был проведен дискриптивный анализ набора данных.

Код:

#Находим квартили

apply(data,2,summary)

#Находим моду

getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

modas = apply(data,2,getmode)

modas

#Находим размахы

apply(data,2,max) - apply(data,2,min)

#Находим межквартильный размах

quantile\_25 = apply(data,2,function(x){quantile(x,probs=0.25)})

quantile\_75 = apply(data,2,function(x){quantile(x,probs=0.75)})

quantile\_75 - quantile\_25

#Находим квадратичное отклонение и дисперсию

apply(data,2,var)

apply(data, 2, sd)

#Строим боксплоты

par(mfrow=c(3,3))

for(i in 1:variable\_number){

boxplot(data[,i],main=colnames(data)[i],ylab="value")

}

#Строим распределения

#Смотрим распределение

par(mfrow=c(3,3))

for(i in 1:variable\_number){

hist(data[,i],xlab='values',main=colnames(data)[i],col=rainbow(length(unique(data[,i]))))

}

Результаты нахождения мер центральных тенденций, тенденций распределения, построения боксплотов и диаграмм распределения представлен на рисунка 2,3,4,5.

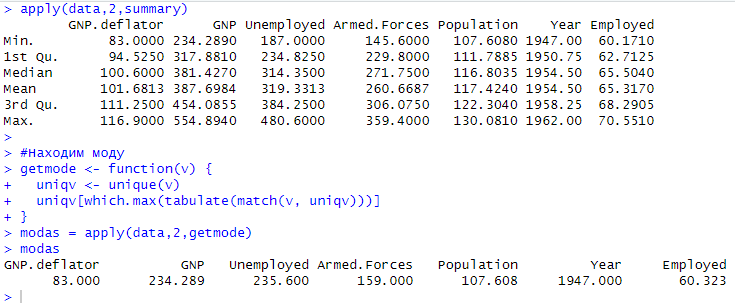


Рисунок 2 – Результат нахождения мер центральных тенденций

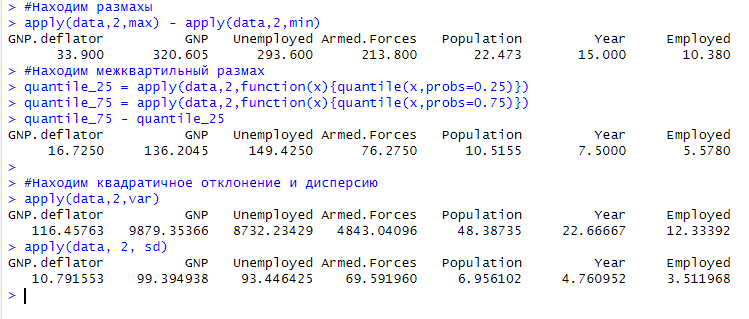


Рисунок 3 – Результат нахождения мер размаха

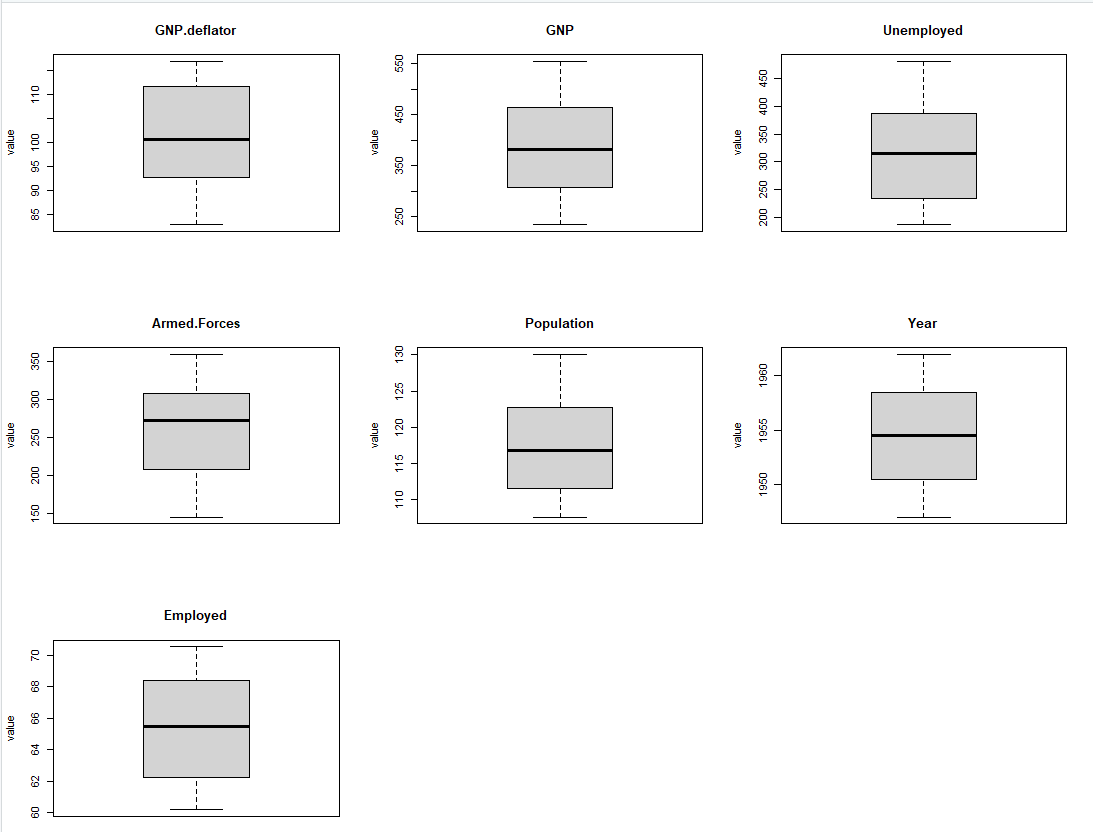


Рисунок 4 – Результат построения боксплотов

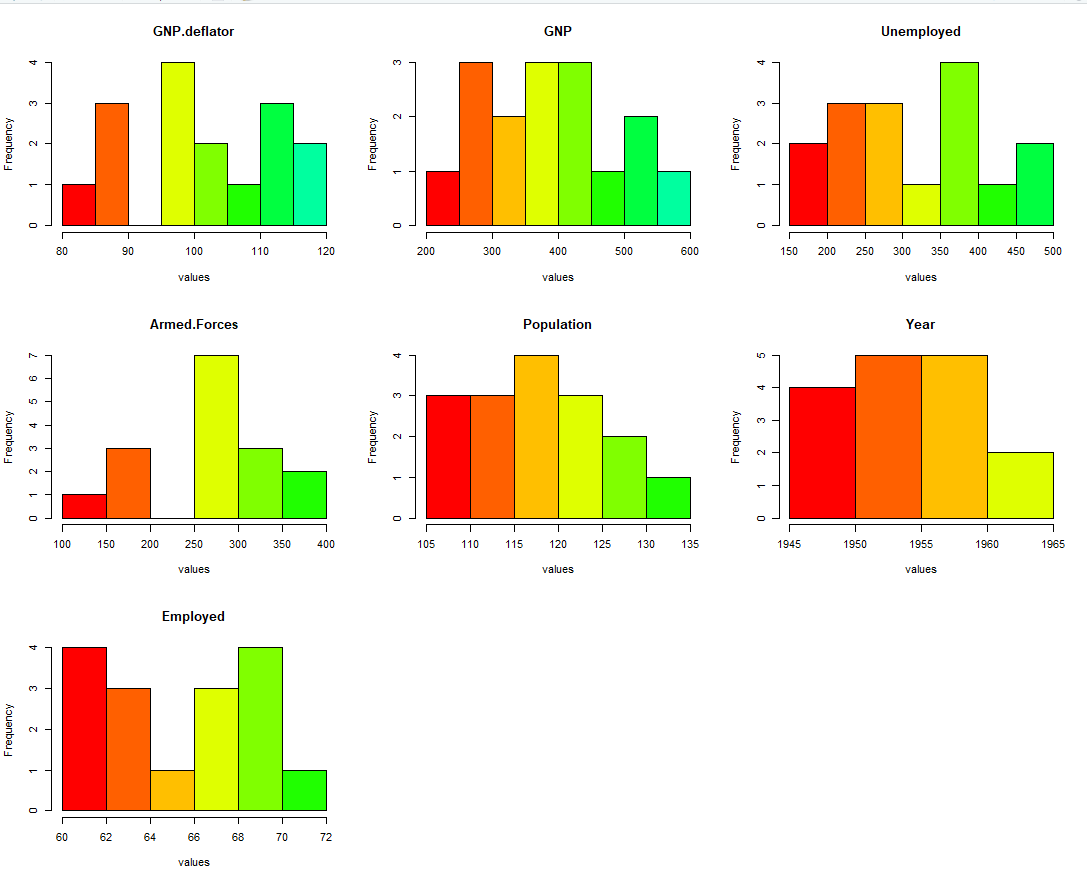


Рисунок 5 – Результат построения гистограмм распределения величин

Согласно полученным диаграммам, можно сделать вывод об отсутствии выбросов в данных. Также можно сделать вывод, что величина Armed.Forces имеет распределения близкое к нормальному. Остальные распределения либо имеют сильное смешение относительно центра, либо являются многомодальными.

1. Была получена корреляционная матрица для набора данных. Результаты построения представлены на рисунках 6,7.

Код:

#4 Получаем корреляционную матрицу

cor(data)

#install.packages(pkgs=c("ellipse"))

library(ellipse)

plotcorr(cor(data))

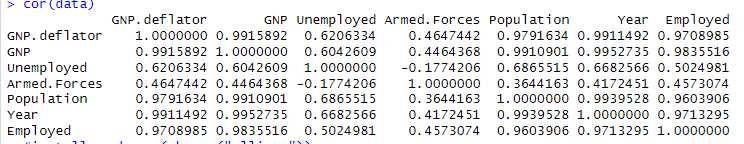


Рисунок 6 – Результат построения матрицы

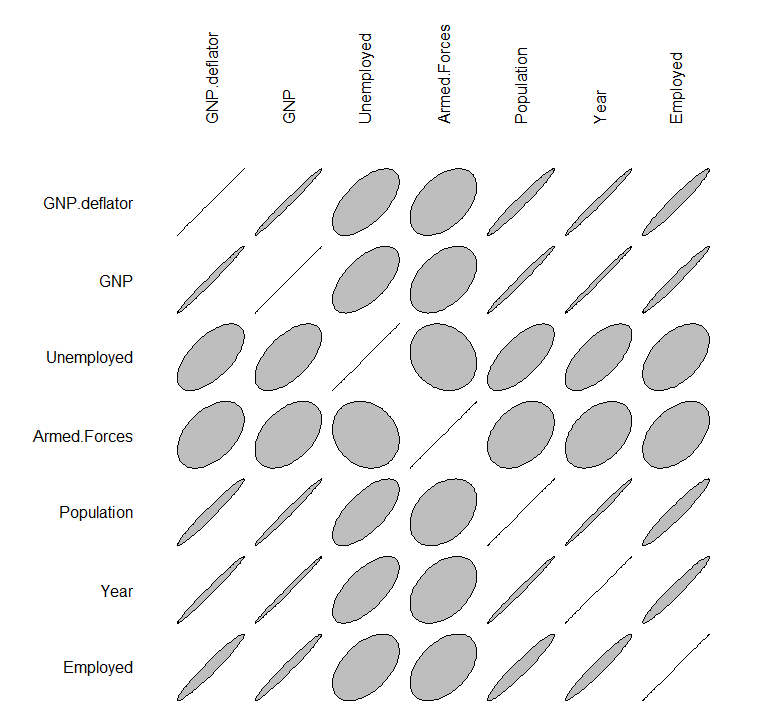


Рисунок 7 – Графическое представление построенной матрицы

Согласно данной матрицы, видно, что GND.deflator имеет следующие лин. зависимости: средняя положительная с GNP, низкая с Population, сильная с Year, низкая с Employed.

Согласно данной матрицы, видно, что GND имеет следующие лин. зависимости: средняя положительная с GNP.deflator, средняя с Population, сильная с Year, средняя с Employed.

Согласно данной матрицы, видно, что Unemployed имеет либо слабые положительные и отрицательные зависимости, либо не имеет их вообще.

Согласно данной матрицы, видно, что Armed Forces имеет либо слабые положительные и отрицательные зависимости, либо не имеет их вообще.

Согласно данной матрицы, видно, что Population имеет следующие лин. зависимости: сильная с Year, средняя с Employed.

Согласно данной матрицы, видно, что Employed имеет следующие лин. зависимости: слабая с Year.

Для подкрепления полученных результатов была построена матрица зависимостей, представленная на рисунке 8. На данной матрице видно, что описанные матрицей корреляции результаты являются в целом правдивыми.

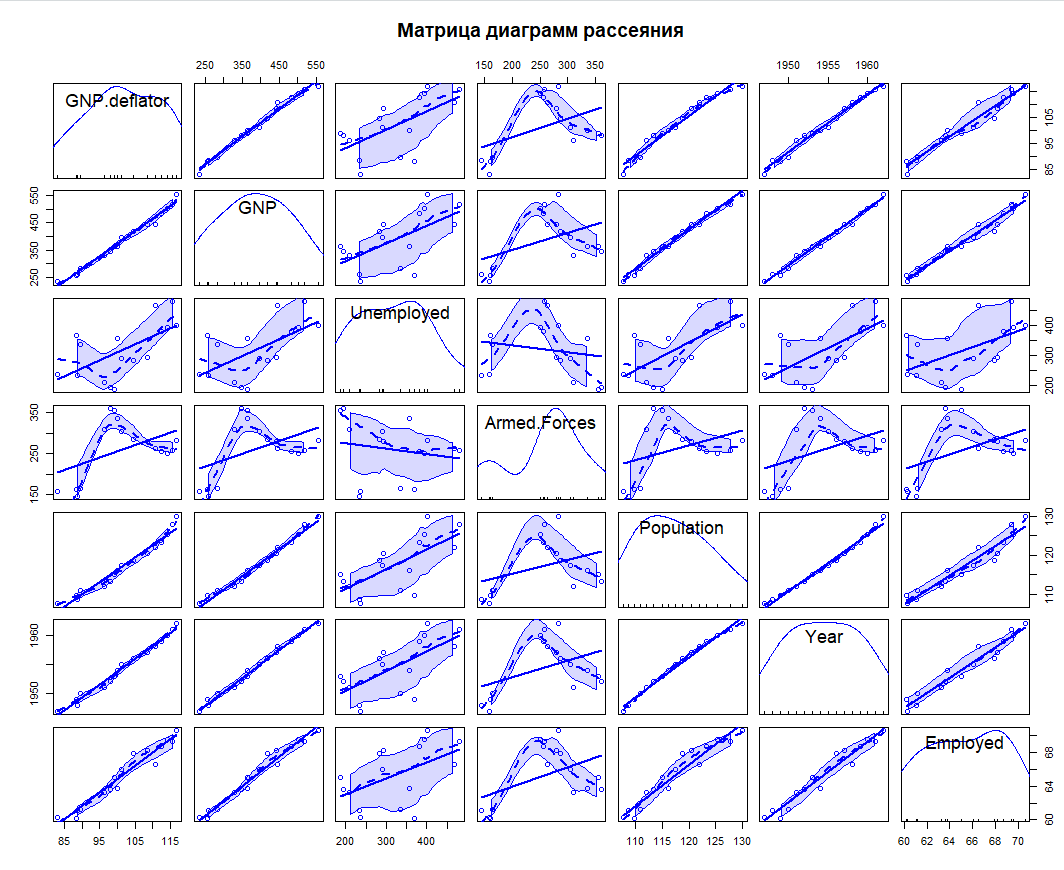


Рисунок 8 – Матрица зависимостей

1. Были проведены теста на нормальность распределения данных. Результаты представлены на рисунке 9.

Код:

#6 Проверяем на нормальность распределения

apply(data,2,shapiro.test)

#Показываем графически, что величины не норм

par(mfrow=c(3,3))

for(i in 1:variable\_number){

hist(data[,i],xlab='values',main=colnames(data)[i],col=rainbow(length(unique(data[,i]))))

}

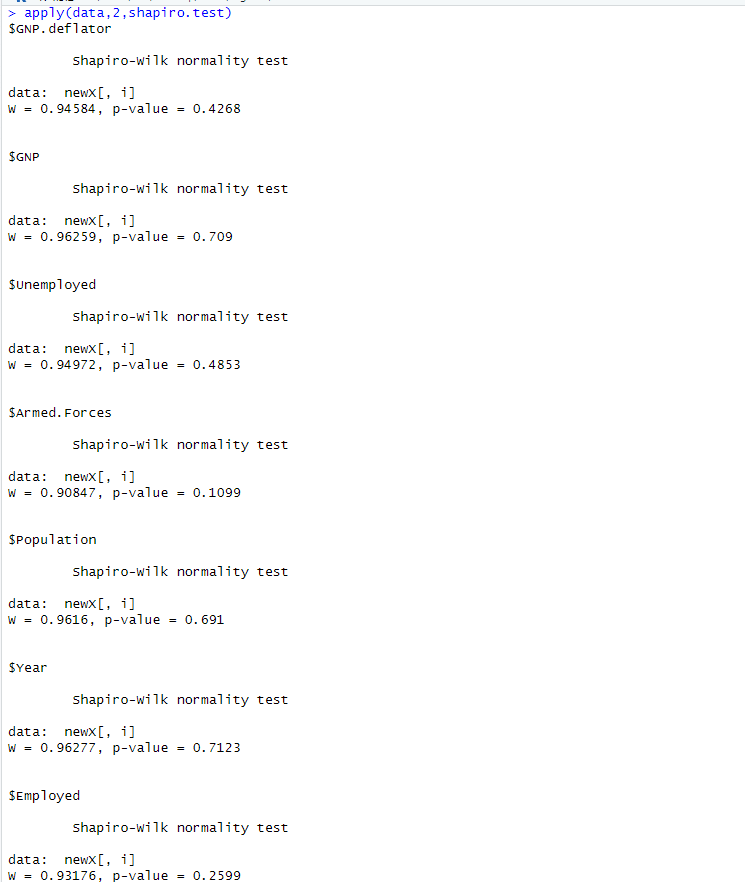


Рисунок 9 – Результаты проведения теста на нормальность распределения данных

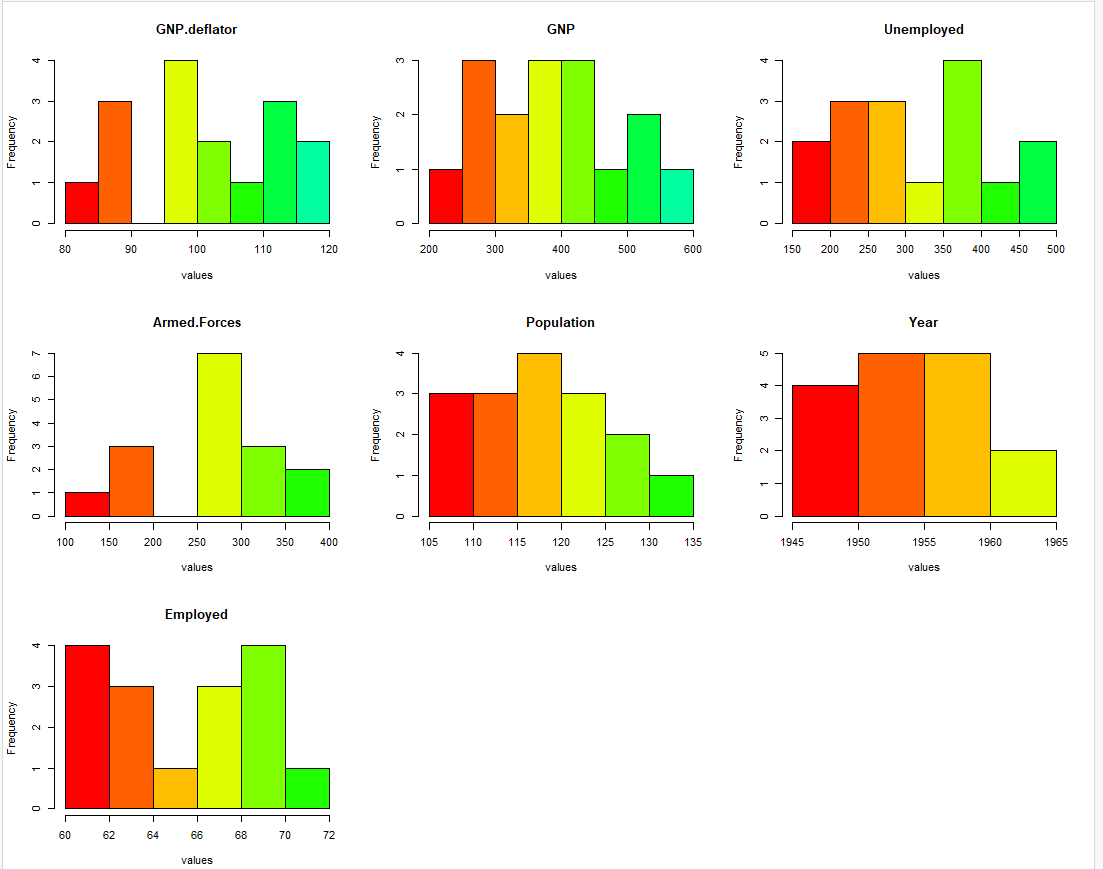


Рисунок 10 – Диаграмма гистограмм для величин

Согласно результатам теста, нормальное распределение имеет только величина Armed.Forces. Данное утверждение подтверждается гистограммами величин на рисунке 10.

1. Был собран набор данных об экономических показателях Франции в период с 1989 по 2018 год.
2. Были построены кривая изменения ВВП и кривая прироста ВВП Франции за каждый из годов. Результаты представлены на рисунка 11 – 12.

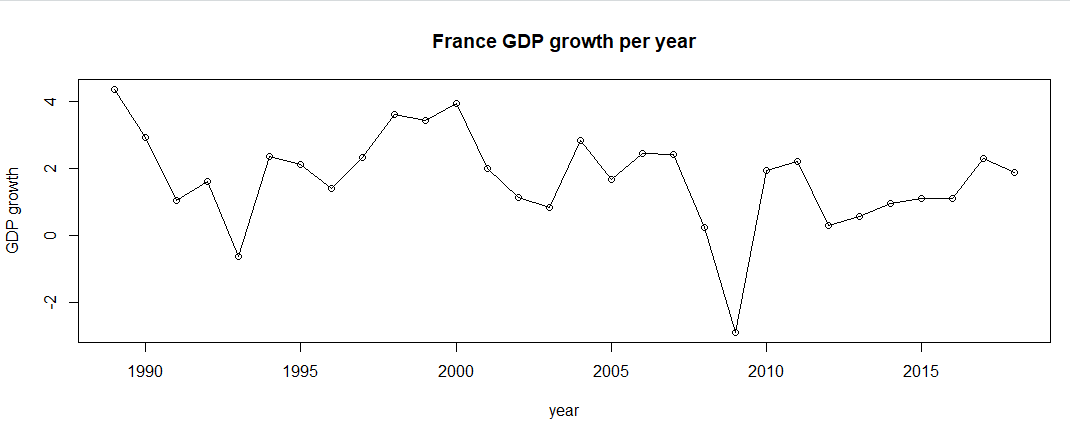


Рисунок 11 – График прироста ВВП Франции по годам

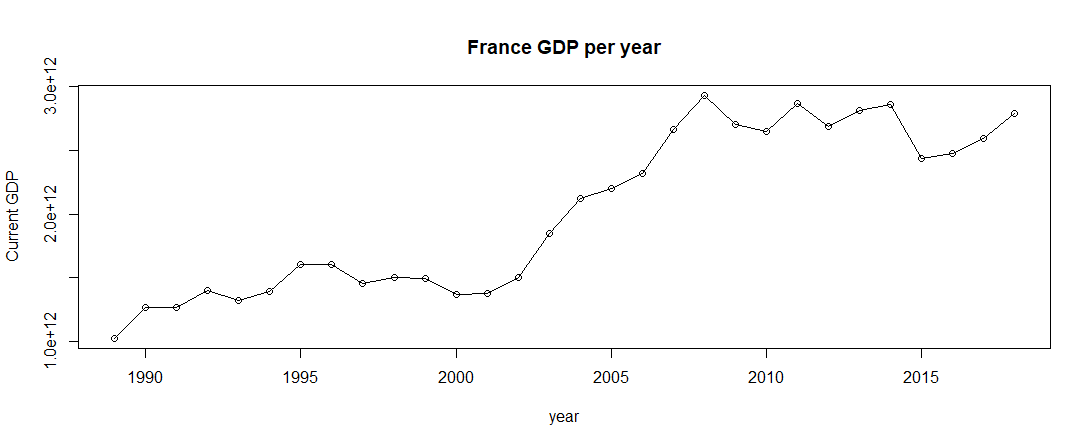


Рисунок 12 – График изменения ВВП Франции по годам

1. Была изучена корреляция между ростом ВВП и приростом населения.

Код:

cor(data[,2],data[,14])

cor(data[,2],data[,14],method="spearman")

cor(data[,2],data[,14],method="kendall")

cor.test(data[,2],data[,14])

cor.test(data[,2],data[,14],method="spearman")

cor.test(data[,2],data[,14],method="kendall")

data\_t = data[!is.na(data[,14]) & !is.na(data[,2]),]

data\_t = data\_t[order(data\_t[,2]),]

par(mfrow=c(2,1))

plot(sort(data\_t[,2]),data\_t[,14],type="o",xlab="GDP growth",ylab="Population growth")

data\_t = data\_t[order(data\_t[,14]),]

plot(sort(data\_t[,14]),data\_t[,2],type="o",ylab="GDP growth",xlab="Population growth")

Результаты изучения корреляции представлены на рисунках 13,14,15,16. Согласно полученным результатам, отсутствует какая – либо корреляция между величинами.

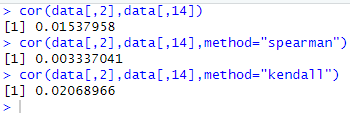


Рисунок 13 – Значение коэффициентов корреляции

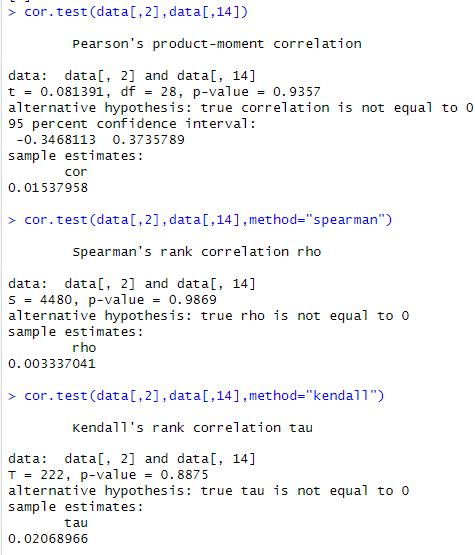
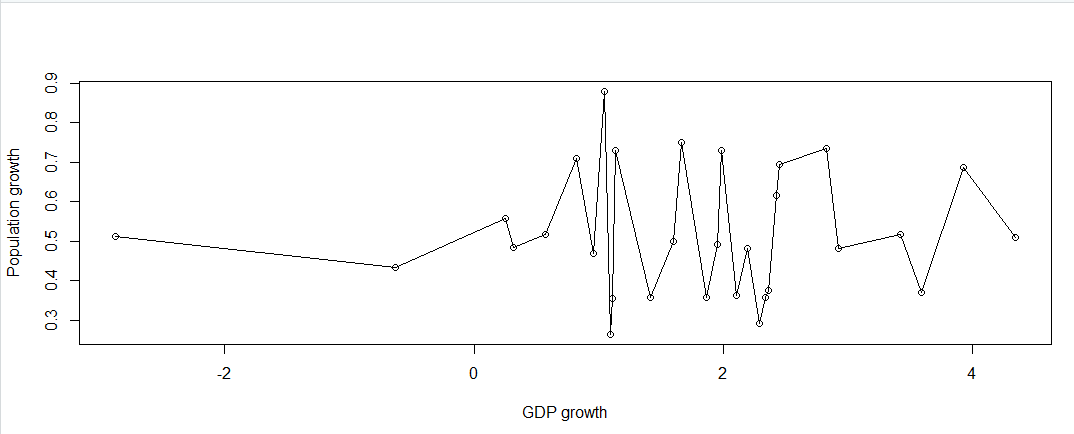


Рисунок 14 – Результаты проверки гипотез



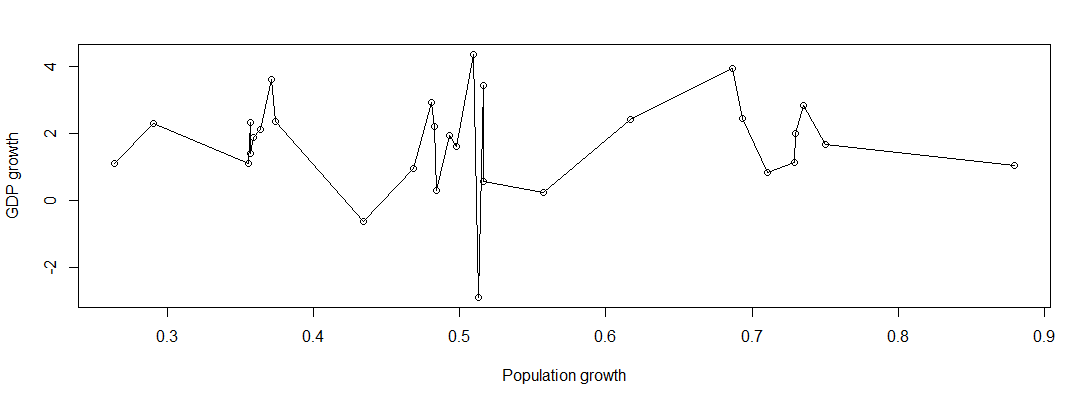


Рисунок 15-16 – Графики зависимости величин друг от друга

1. Была изучена корреляция между приростом населения и динамику безработицы.

Код:

# 2 Прирост населения на динамику безработицы

cor(data[,14],data[,17])

cor(data[,14],data[,17],method="spearman")

cor(data[,14],data[,17],method="kendall")

cor.test(data[,14],data[,17])

cor.test(data[,14],data[,17],method="spearman")

cor.test(data[,14],data[,17],method="kendall")

data\_t = data[!is.na(data[,14]) & !is.na(data[,17]),]

data\_t = data\_t[order(data\_t[,14]),]

par(mfrow=c(2,1))

plot(sort(data[,14]),data[,17],type="o",ylab="Unemployment",xlab="Population growth")

data\_t = data\_t[order(data\_t[,17]),]

plot(sort(data[,17]),data[,14],type="o",xlab="Unemployment",ylab="Population growth")

Результаты изучения корреляции представлены на рисунках 17,18,19,20. Согласно полученным результатам, слабая отрицательная корреляция между величинами.

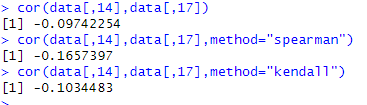


Рисунок 17 – Значение коэффициентов корреляции

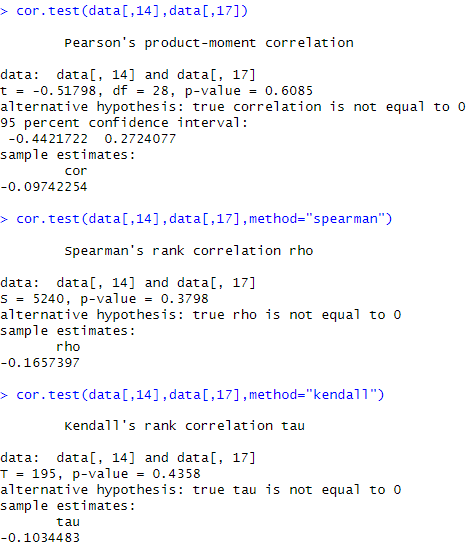


Рисунок 18 – Результаты проверки гипотез

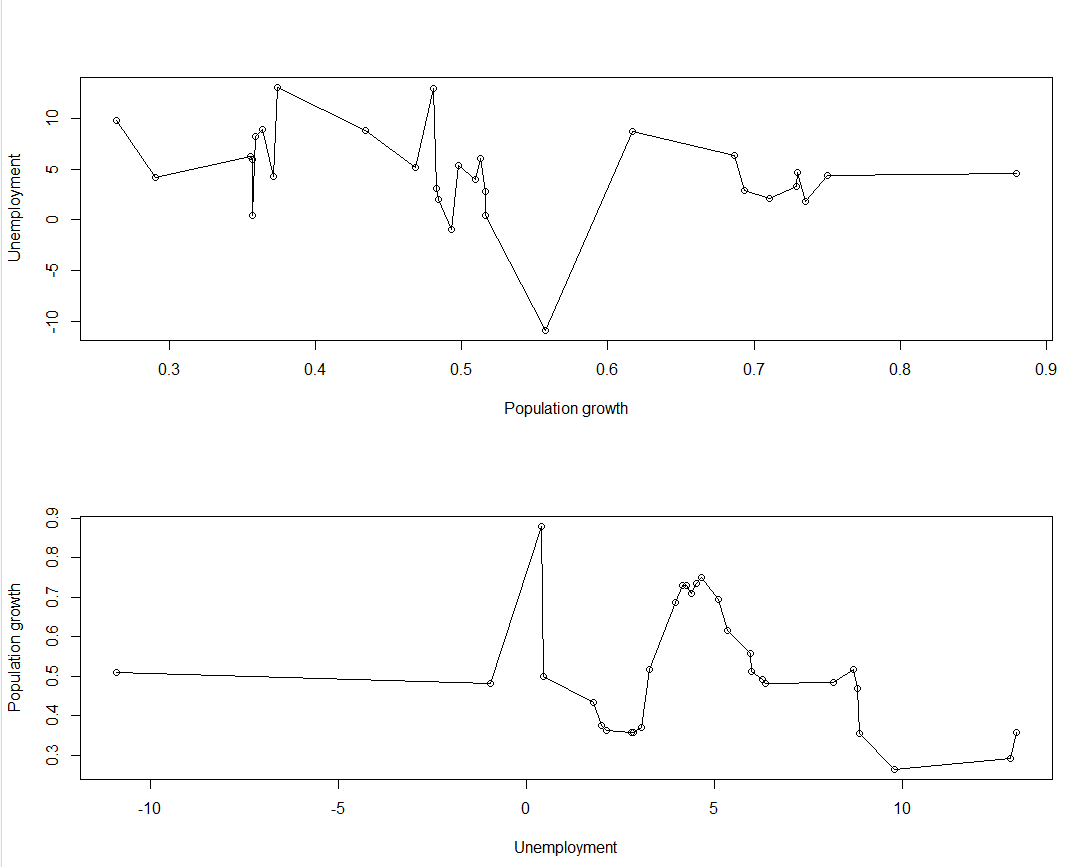


Рисунок 19-20 – Графики зависимости величин друг от друга

1. Была изучена корреляция между изменением расходов на медицину и средней продолжительностью жизни.

Код:

cor(data[,12],data[,13],use="pairwise.complete.obs")

cor(data[,12],data[,13],method="spearman",use="pairwise.complete.obs")

cor(data[,12],data[,13],method="kendall",use="pairwise.complete.obs")

cor.test(data[,12],data[,13],use="pairwise.complete.obs")

cor.test(data[,12],data[,13],method="spearman",use="pairwise.complete.obs")

cor.test(data[,12],data[,13],method="kendall",use="pairwise.complete.obs")

data\_t = data[!is.na(data[,12]) & !is.na(data[,13]),]

par(mfrow=c(2,1))

data\_t = data\_t[order(data\_t[,12]),]

plot(sort(data\_t[,12]),data\_t[,13],type="o",ylab="Life expectancy",xlab="Health expenditure")

data\_t = data\_t[order(data\_t[,13]),]

plot(sort(data\_t[,13]),data\_t[,12],type="o",xlab="Life expectancy",ylab="Health expenditure")

plotcorr(cor(data[,c(12,13)],use="pairwise.complete.obs"))

Результаты изучения корреляции представлены на рисунках 21,22,23,24. Согласно полученным результатам, имеется сильная положительная корреляция между величинами.

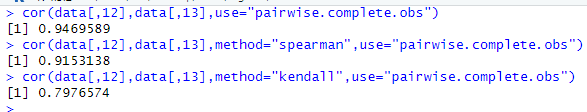


Рисунок 21 – Значение коэффициентов корреляции

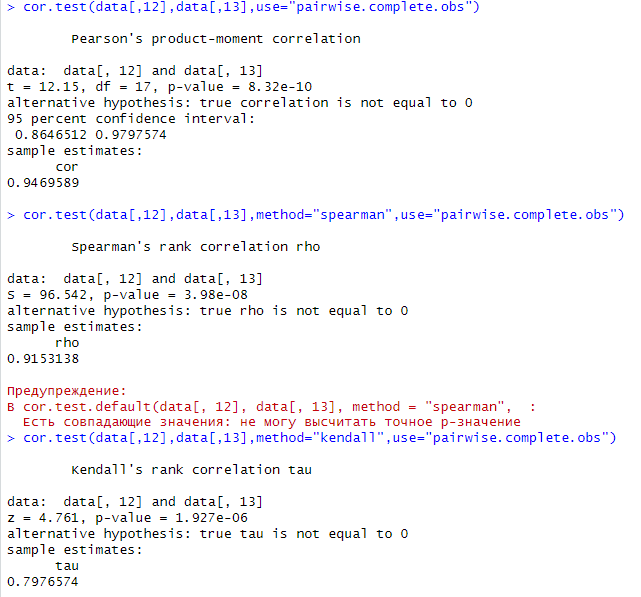


Рисунок 22 – Результаты проверки гипотез

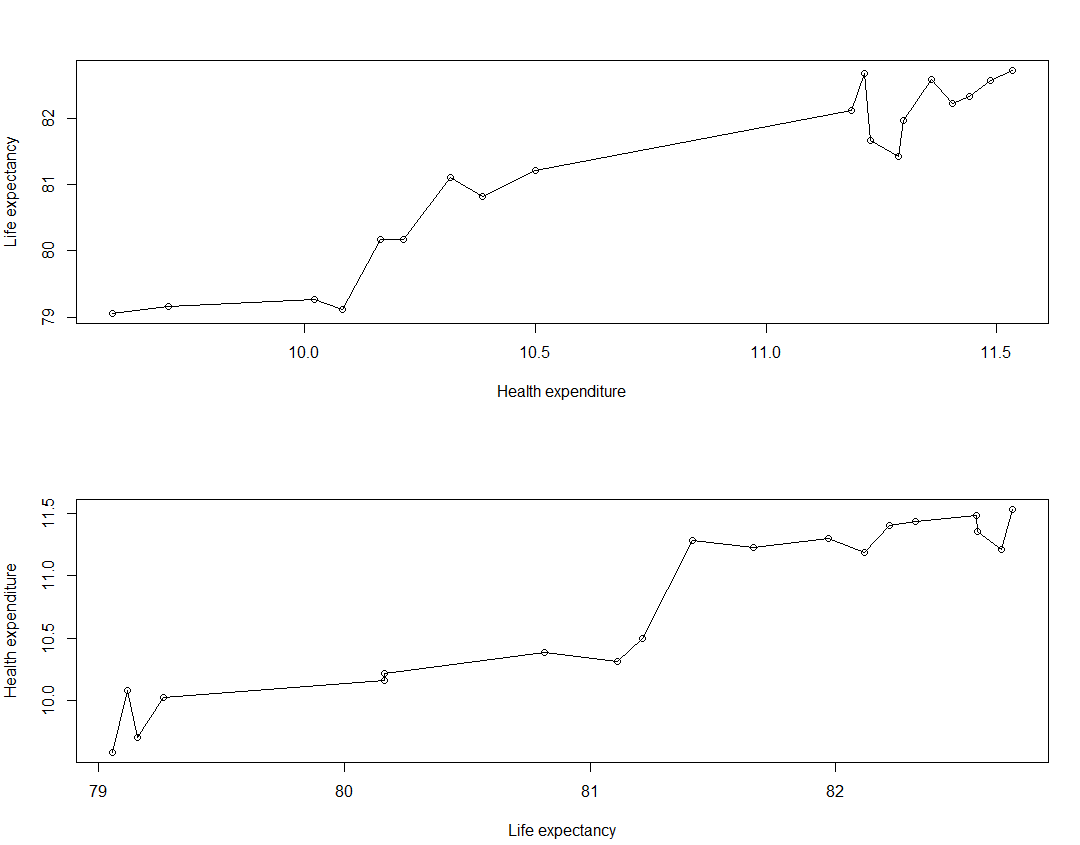


Рисунок 23-24 – Графики зависимости величин друг от друга

1. Была изучена корреляция между изменением расходов на медицину и смертностью.

Код:

cor(data[,12],data[,18],use="pairwise.complete.obs")

cor(data[,12],data[,18],method="spearman",use="pairwise.complete.obs")

cor(data[,12],data[,18],method="kendall",use="pairwise.complete.obs")

cor.test(data[,12],data[,18],use="pairwise.complete.obs")

cor.test(data[,12],data[,18],method="spearman",use="pairwise.complete.obs")

cor.test(data[,12],data[,18],method="kendall",use="pairwise.complete.obs")

data\_t = data[!is.na(data[,12]) & !is.na(data[,18]),]

par(mfrow=c(2,1))

data\_t = data\_t[order(data\_t[,12]),]

plot(sort(data\_t[,12]),data\_t[,18],type="o",ylab="Death rate",xlab="Health expenditure")

data\_t = data\_t[order(data\_t[,18]),]

plot(sort(data\_t[,18]),data\_t[,12],type="o",xlab="Death rate",ylab="Health expenditure")

plotcorr(cor(data[,c(12,18)],use="pairwise.complete.obs"))

Результаты изучения корреляции представлены на рисунках 25, 26, 27, 28. Согласно полученным результатам, имеется слабая положительная корреляция между величинами.

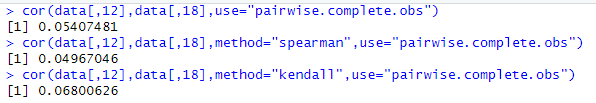


Рисунок 25 – Значение коэффициентов корреляции

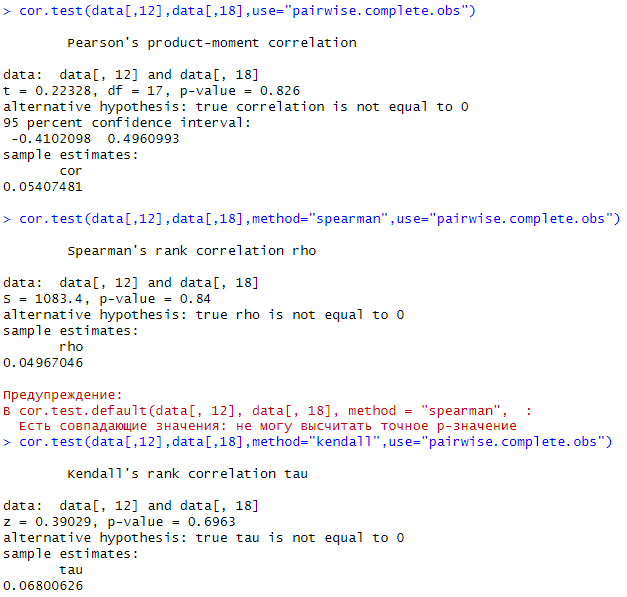


Рисунок 26 – Результаты проверки гипотез

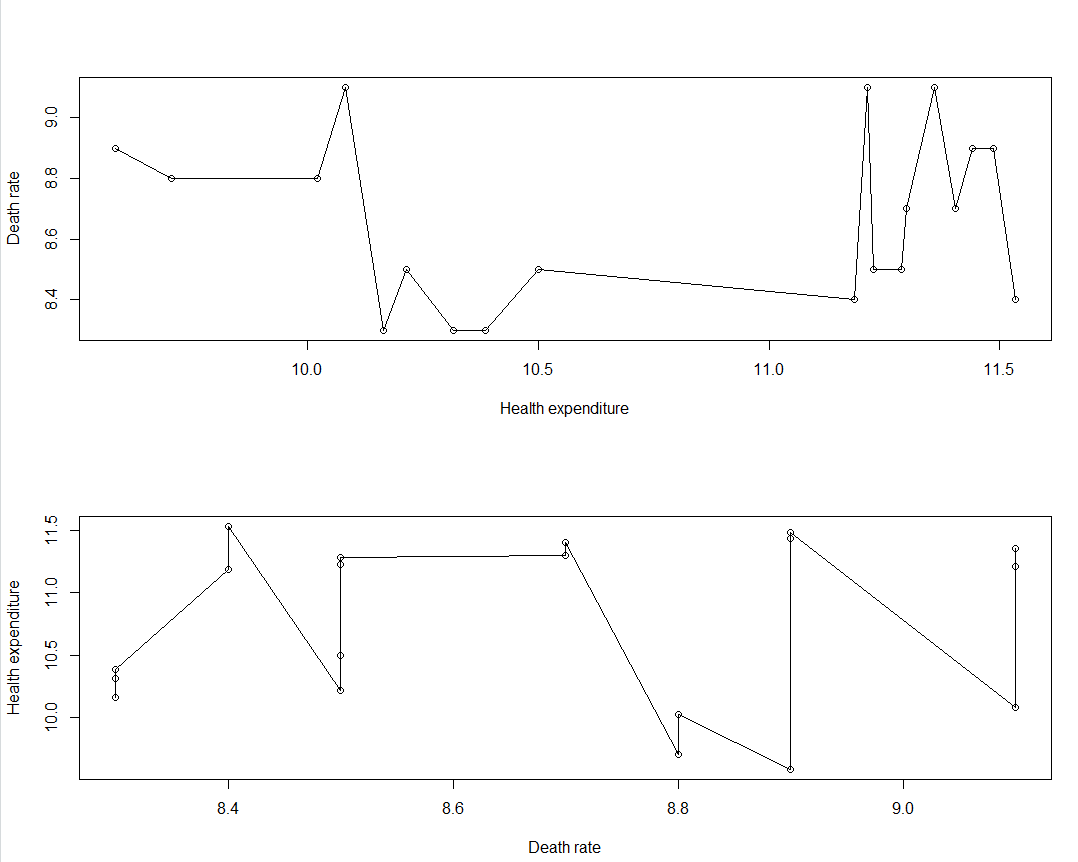


Рисунок 27-28 – Графики зависимости величин друг от друга

1. Была изучена корреляция между изменением расходов на медицину и смертностью.

Код:

cor(data[,19],data[,17],use="pairwise.complete.obs")

cor(data[,19],data[,17],method="spearman",use="pairwise.complete.obs")

cor(data[,19],data[,17],method="kendall",use="pairwise.complete.obs")

cor.test(data[,19],data[,17],use="pairwise.complete.obs")

cor.test(data[,19],data[,17],method="spearman",use="pairwise.complete.obs")

cor.test(data[,19],data[,17],method="kendall",use="pairwise.complete.obs")

data\_t = data[!is.na(data[,19]) & !is.na(data[,17]),]

data\_t = data\_t[order(data\_t[,19]),]

par(mfrow=c(2,1))

plot(sort(data\_t[,19]),data\_t[,17],type="o",ylab="Export",xlab="Бакалавры")

data\_t = data\_t[order(data\_t[,17]),]

plot(sort(data\_t[,17]),data\_t[,19],type="o",xlab="Export",ylab="Бакалавры")

plotcorr(cor(data[,c(19,17)],use="pairwise.complete.obs"))

Результаты изучения корреляции представлены на рисунках 29, 30, 31, 32. Согласно полученным результатам, имеется слабая линейная положительная корреляция между величинами, а также сильная ранговая корреляция

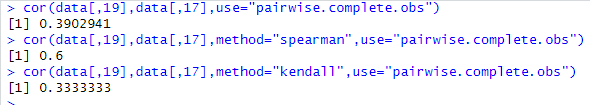


Рисунок 29 – Значение коэффициентов корреляции

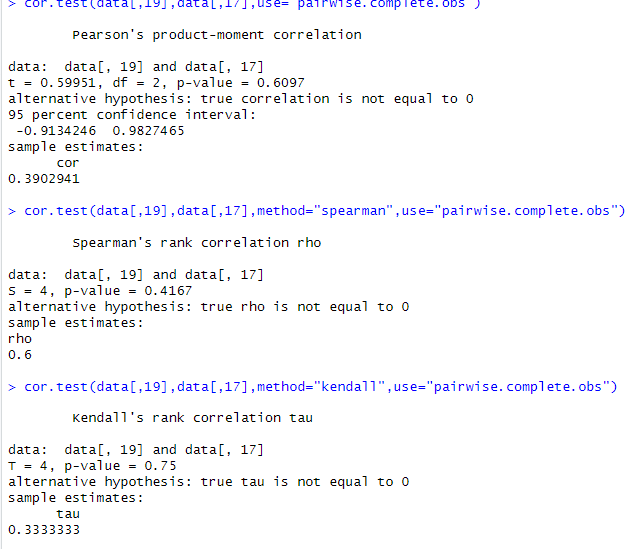


Рисунок 30 – Результаты проверки гипотез

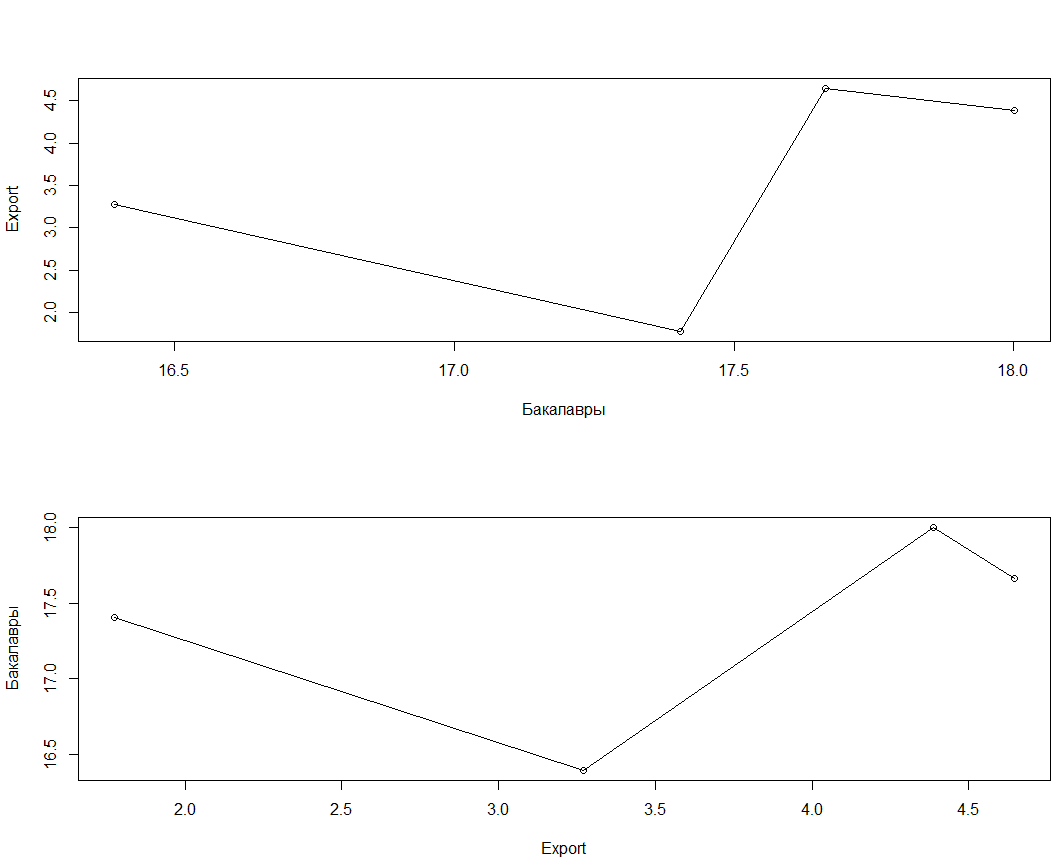


Рисунок 31-32 – Графики зависимости величин друг от друга

1. Была изучена корреляция между приростом населения с высшим образованием и экспортом товаров.

Код:

cor(data[,19],data[,17],use="pairwise.complete.obs")

cor(data[,19],data[,17],method="spearman",use="pairwise.complete.obs")

cor(data[,19],data[,17],method="kendall",use="pairwise.complete.obs")

cor.test(data[,19],data[,17],use="pairwise.complete.obs")

cor.test(data[,19],data[,17],method="spearman",use="pairwise.complete.obs")

cor.test(data[,19],data[,17],method="kendall",use="pairwise.complete.obs")

data\_t = data[!is.na(data[,19]) & !is.na(data[,17]),]

data\_t = data\_t[order(data\_t[,19]),]

par(mfrow=c(2,1))

plot(sort(data\_t[,19]),data\_t[,17],type="o",ylab="Export",xlab="Бакалавры")

data\_t = data\_t[order(data\_t[,17]),]

plot(sort(data\_t[,17]),data\_t[,19],type="o",xlab="Export",ylab="Бакалавры")

plotcorr(cor(data[,c(19,17)],use="pairwise.complete.obs"))

Результаты изучения корреляции представлены на рисунках 33, 34, 35, 36. Согласно полученным результатам, имеется слабая линейная положительная корреляция между величинами, а также сильная ранговая корреляция

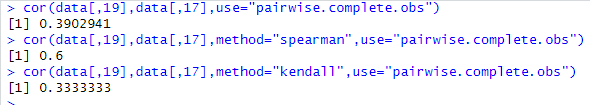


Рисунок 33 – Значение коэффициентов корреляции

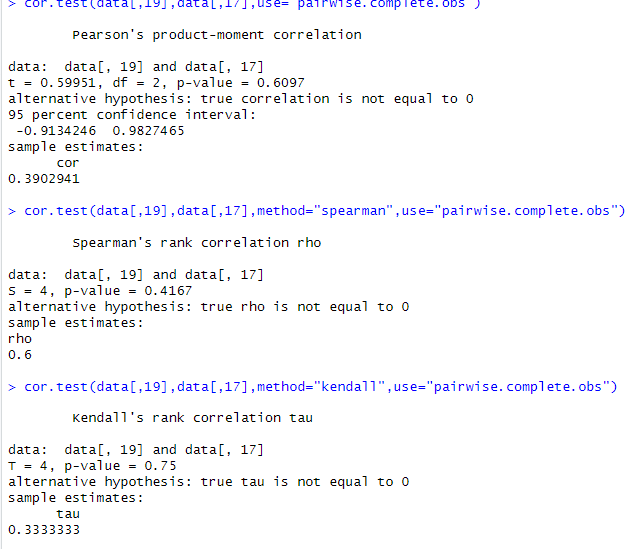


Рисунок 34 – Результаты проверки гипотез

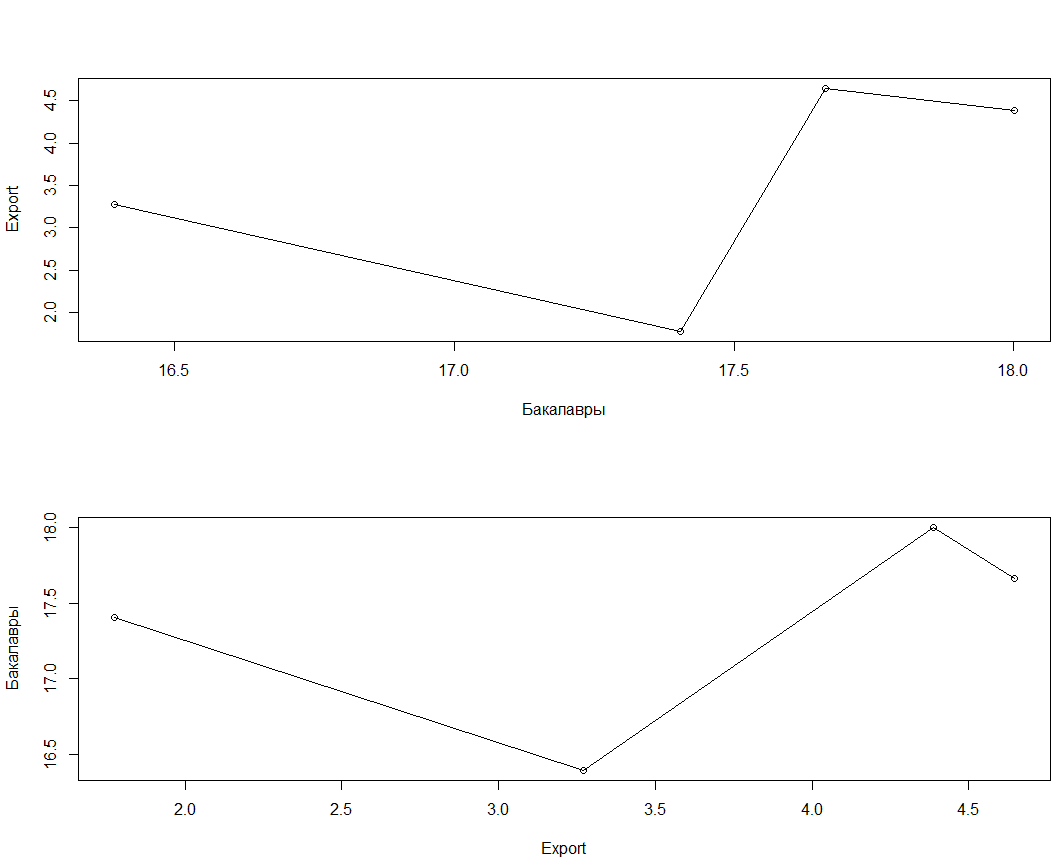


Рисунок 35-36 – Графики зависимости величин друг от друга

1. Была изучена корреляция между приростом населения с высшим образованием и приростом высокотехнологического производства.

Код:

cor(data[,19],data[,22],use="pairwise.complete.obs")

cor(data[,19],data[,22],method="spearman",use="pairwise.complete.obs")

cor(data[,19],data[,22],method="kendall",use="pairwise.complete.obs")

cor.test(data[,19],data[,22],use="pairwise.complete.obs")

cor.test(data[,19],data[,22],method="spearman",use="pairwise.complete.obs")

cor.test(data[,19],data[,22],method="kendall",use="pairwise.complete.obs")

data\_t = data[!is.na(data[,19]) & !is.na(data[,22]),]

data\_t = data\_t[order(data\_t[,19]),]

par(mfrow=c(2,1))

plot(data\_t[,19],data\_t[,22],type="o",ylab="Hightech production",xlab="Бакалавры")

data\_t = data\_t[order(data\_t[,22]),]

plot(data\_t[,22],data\_t[,19],type="o",xlab="Hightech production",ylab="Бакалавры")

plotcorr(cor(data[,c(19,22)],use="pairwise.complete.obs"))

Результаты изучения корреляции представлены на рисунках 37, 38, 39, 40. Согласно полученным результатам, имеется сильная линейная положительная корреляция между величинами, а также средняя ранговая корреляция

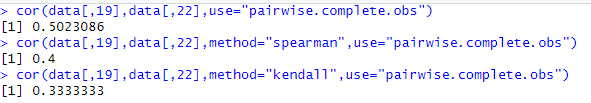


Рисунок 37 – Значение коэффициентов корреляции

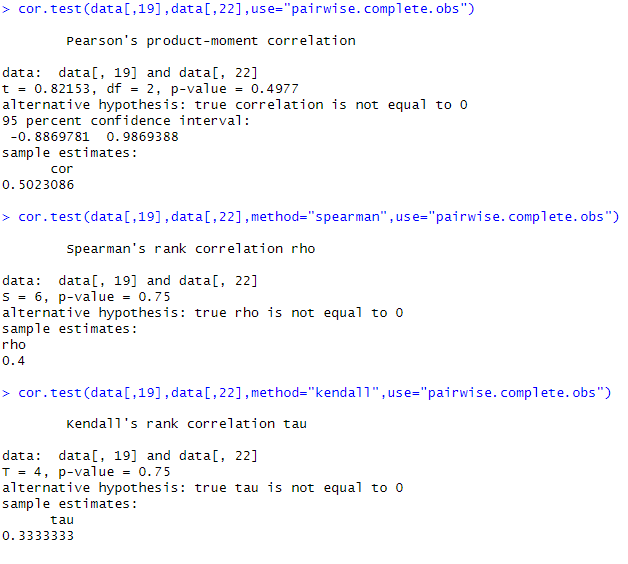


Рисунок 38 – Результаты проверки гипотез

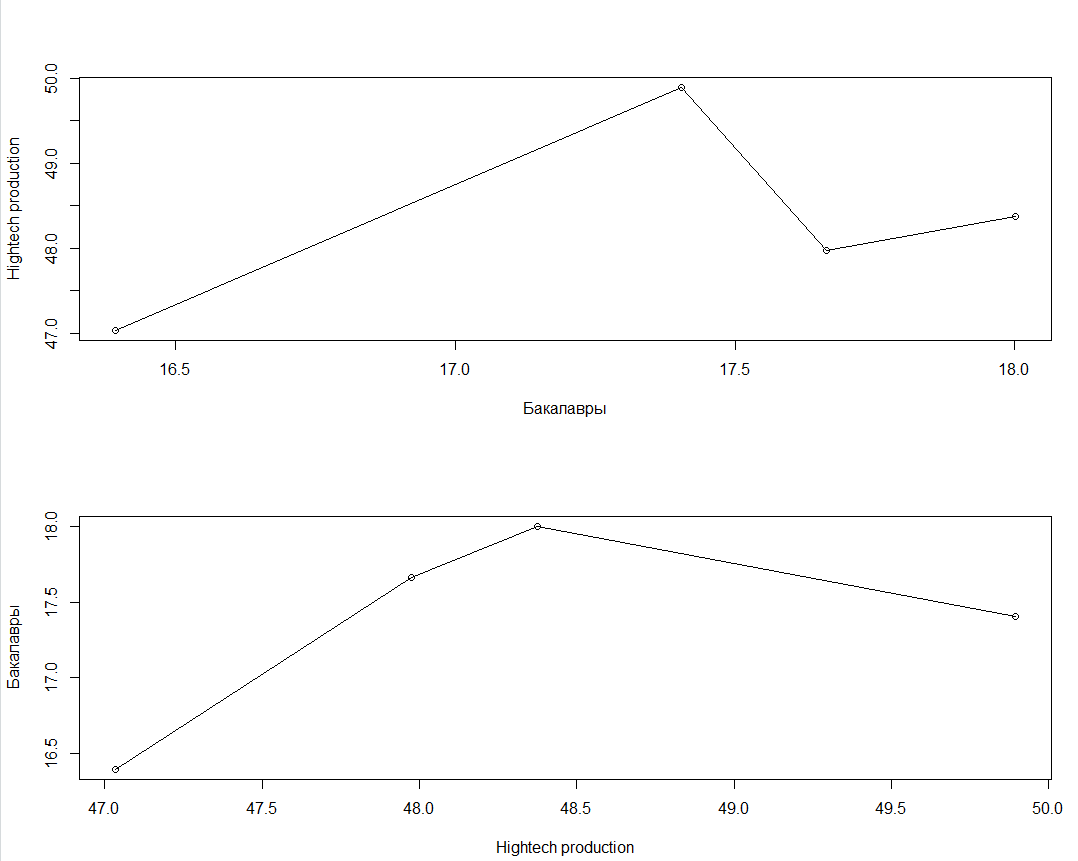


Рисунок 39-40 – Графики зависимости величин друг от друга

1. Была изучена корреляция между изменением расходов на образование и кумулятивным приростом бакалавров среди женщин.

Код:

cor(data[,15],data[,20],use="pairwise.complete.obs")

cor(data[,15],data[,20],method="spearman",use="pairwise.complete.obs")

cor(data[,15],data[,20],method="kendall",use="pairwise.complete.obs")

cor.test(data[,15],data[,20],use="pairwise.complete.obs")

cor.test(data[,15],data[,20],method="spearman",use="pairwise.complete.obs")

cor.test(data[,15],data[,20],method="kendall",use="pairwise.complete.obs")

data\_t = data[!is.na(data[,15]) & !is.na(data[,20]),]

data\_t = data\_t[order(data\_t[,15]),]

par(mfrow=c(2,1))

plot(data\_t[,15],data\_t[,20],type="o",xlab="Education expenditure",ylab="Бакалавры среди женщин")

data\_t = data\_t[order(data\_t[,20]),]

plot(data\_t[,20],data\_t[,15],type="o",ylab="Education expenditure",xlab="Бакалавры среди женщин")

plotcorr(cor(data[,c(15,20)],use="pairwise.complete.obs"))

Результаты изучения корреляции представлены на рисунках 41, 42, 43, 44. Согласно полученным результатам, имеется сильная линейная отрицательная корреляция между величинами, а также средняя ранговая корреляция

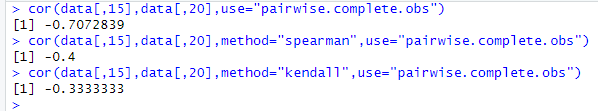


Рисунок 41 – Значение коэффициентов корреляции

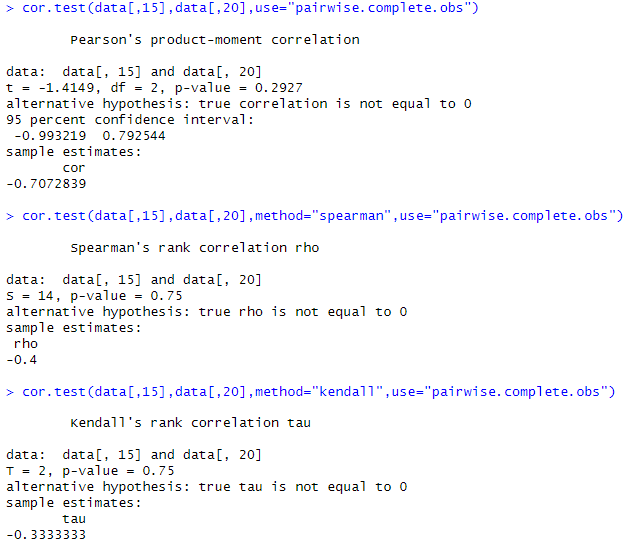


Рисунок 42 – Результаты проверки гипотез

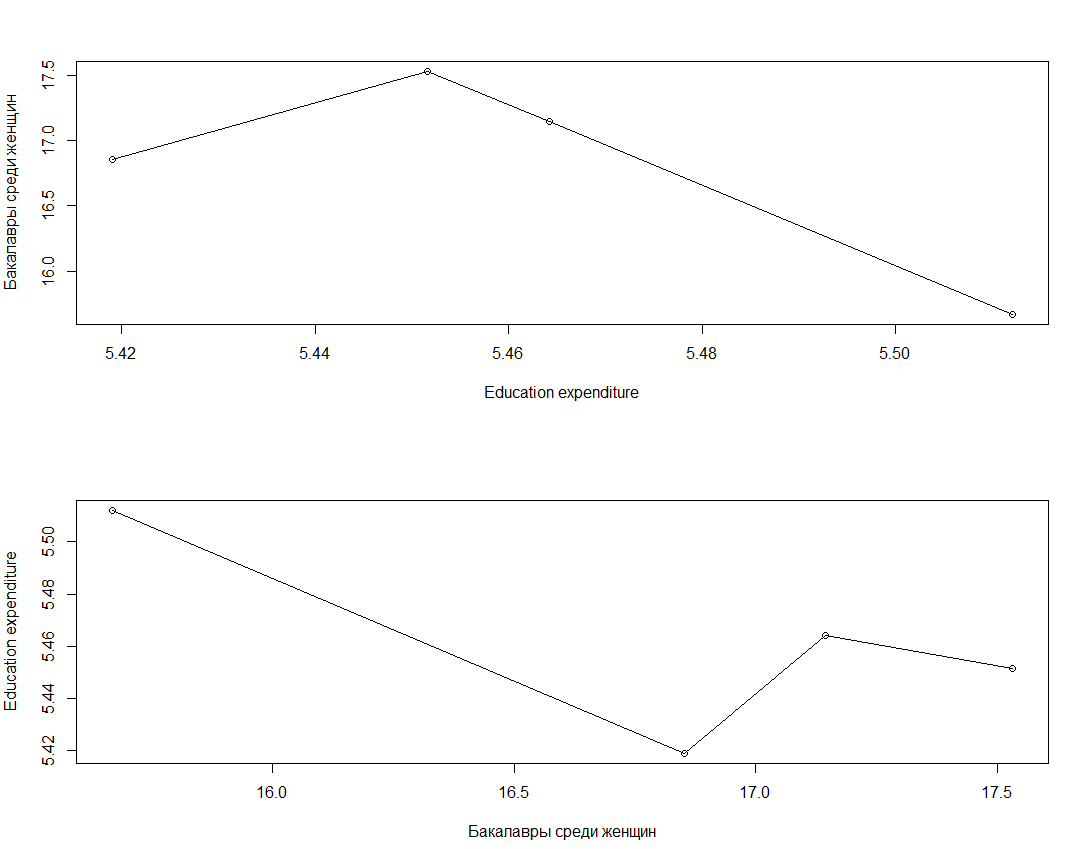


Рисунок 43-44 – Графики зависимости величин друг от друга

1. Была изучена корреляция между приростом бакалавров и количеством научных статей.

Код:

cor(data[,19],data[,22],use="pairwise.complete.obs")

cor(data[,19],data[,22],method="spearman",use="pairwise.complete.obs")

cor(data[,19],data[,22],method="kendall",use="pairwise.complete.obs")

cor.test(data[,19],data[,22],use="pairwise.complete.obs")

cor.test(data[,19],data[,22],method="spearman",use="pairwise.complete.obs")

cor.test(data[,19],data[,22],method="kendall",use="pairwise.complete.obs")

data\_t = data[!is.na(data[,19]) & !is.na(data[,22]),]

data\_t = data\_t[order(data\_t[,19]),]

par(mfrow=c(2,1))

plot(data\_t[,19],data\_t[,22],type="o",xlab="Прирост бакалавров",ylab="Количество научных статей")

data\_t = data\_t[order(data\_t[,22]),]

plot(data\_t[,22],data\_t[,19],type="o",ylab="Прирост бакалавров",xlab="Количество научных статей")

plotcorr(cor(data[,c(19,22)],use="pairwise.complete.obs"))

Результаты изучения корреляции представлены на рисунках 45, 46, 47, 48. Согласно полученным результатам, имеется средняя линейная положительная корреляция между величинами, а также средняя ранговая корреляция

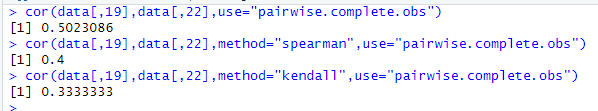


Рисунок 45 – Значение коэффициентов корреляции

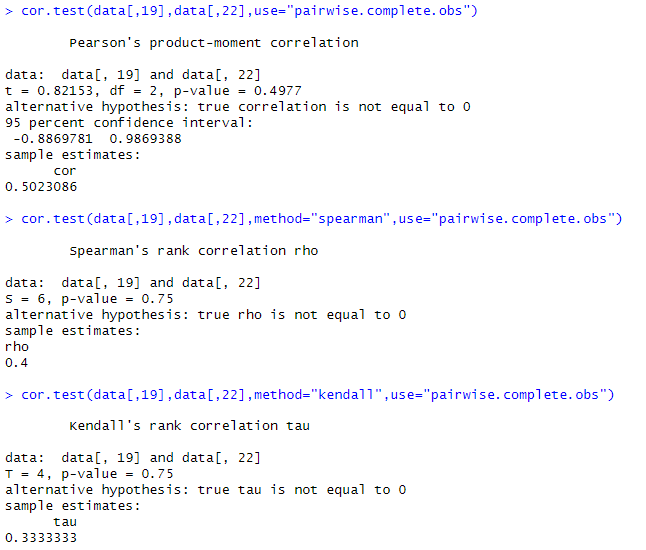


Рисунок 46 – Результаты проверки гипотез

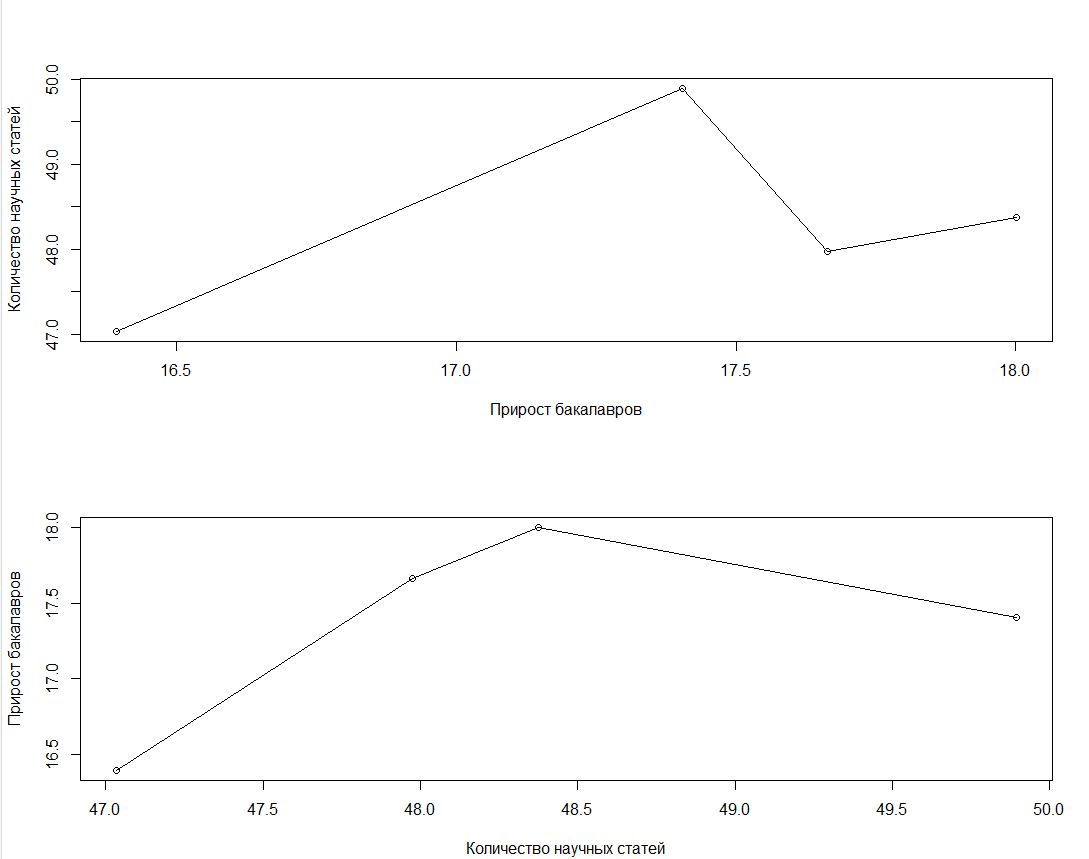


Рисунок 47-48 – Графики зависимости величин друг от друга

1. Были обнаружены зависимые переменные с помощью матрицы корреляции, представленной на рисунке 49.

Код:

library(car)

library(ellipse)

data\_t = data[, colSums(is.na(data)) < nrow(data)]

data\_t

colnames(data\_t)

colnames(data\_t)=1:length(data\_t\_col)

scatterplotMatrix(data\_t,main="Матрица диаграмм рассеяния")

plotcorr(cor(data\_t,use="pairwise.complete.obs"))

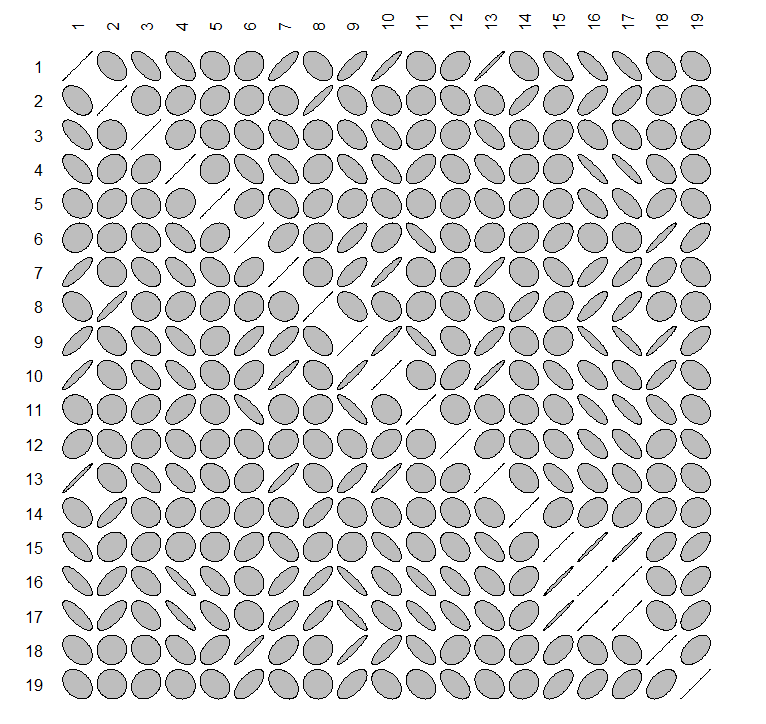


Рисунок 49 – Матрица корреляции всех переменных

На основании одной из переменных была решена задача регрессии. В качестве зависимой переменной выступает 1(GDP (current US$)). В качестве независимых выступают 7(Imports of goods and services (% of GDP)), 13(Goods imports (BoP, current US$)), 10(Life expectancy at birth, total (years)). Результаты предсказаний модели представлены на рисунках 50 и 51.

Код:

data\_t = data[, colSums(is.na(data)) < nrow(data)]

data\_fit = as.data.frame(data\_t[,c(1,7,10,13)])

colnames(data\_fit)

colnames(data\_fit)[1] = "gdp"

colnames(data\_fit)[2] = "iofas"

colnames(data\_fit)[3] = "lexp"

colnames(data\_fit)[4] = "gim"

data\_fit = na.omit(data\_fit)

fit <- lm(formula=gdp~iofas+lexp+gim,data=as.data.frame(data\_fit[1:28,]))

fit

summary(fit)

residuals(fit)

new\_results = predict(fit,newdata=data\_fit[29:30,])

new\_results

par(mfrow=c(2,1))

style <- c(rep(1,28), rep(2,4))

plot(1989:2018,c(data\_fit[1:28,1], new\_results),

ylab="GDP",

xlab="Year", pch=style, col=style)

grid(nx = NULL, ny = NULL,

lty = 2, # Grid line type

col = "gray", # Grid line color

lwd = 2) # Grid line width

plot(1989:2018,data\_fit$gdp,

ylab="GDP",

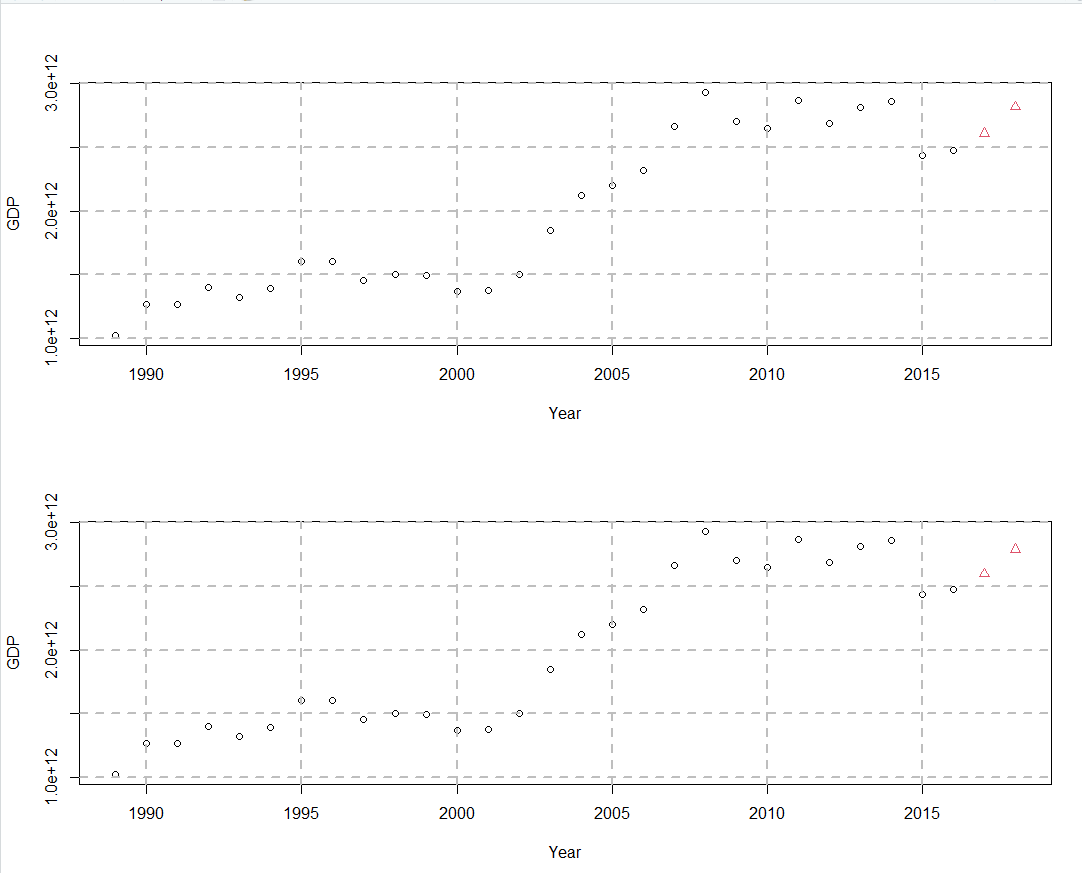
xlab="Year", pch=style, col=style)

grid(nx = NULL, ny = NULL,

lty = 2, # Grid line type

col = "gray", # Grid line color

lwd = 2) # Grid line width



Рисунки 50 – 51 – Предсказанные и истинные результаты

**Вывод:** Было проведено ознакомление с понятием корреляционный и регрессионный анализ данных, некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Получены навыки оценивания связь между переменными и оценивать степень этой связи.