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**

**Тема: «Корреляционный и регрессионный анализ данных»**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Придава А.А.

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Шиян В.И.

**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.

**Ход работы:**

Из исходного csv-файла были импортированы данные в RStudio. Таблица представлена на рисунке 1.

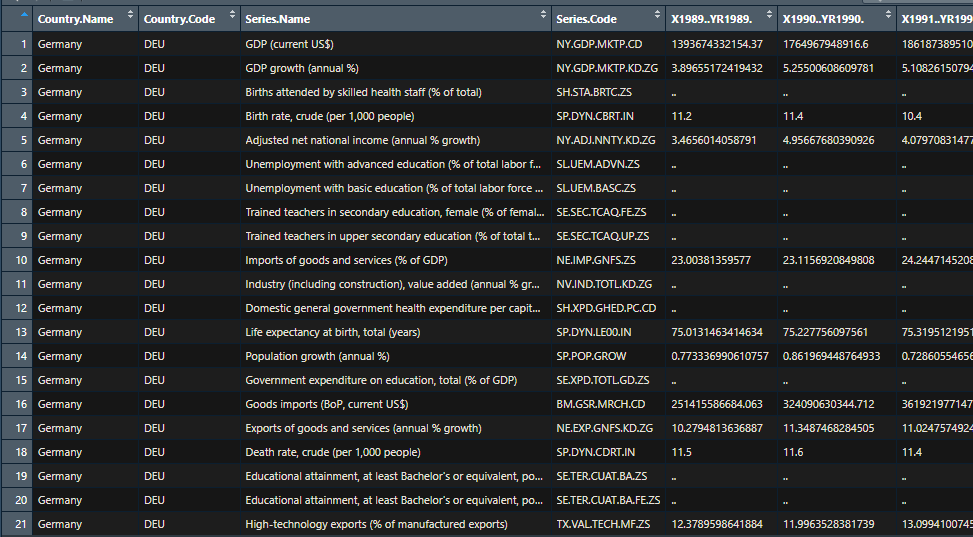


Рисунок 1 – Исходная таблица с данными

В этом датасете представлены данные о 22-ух характеристиках 18-ти стран мира с 1989 по 2018 годы.

Удалим все страны, кроме Франции, а также повторяющиеся строчки.

На рисунке 2 изображена таблица получившихся данных.

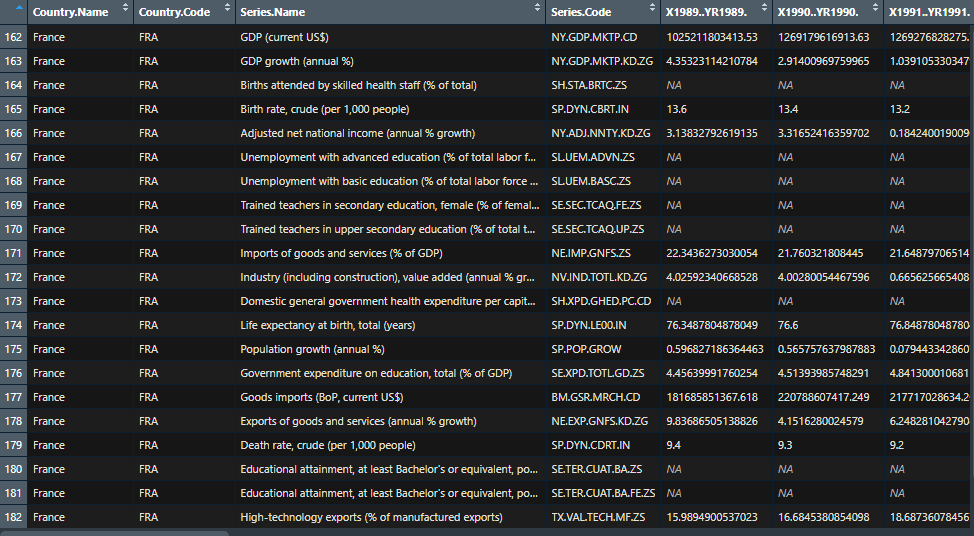


Рисунок 2 – Измененная таблица с данными

Построим кривую прироста ВВП Франции в процентах и USD. Соответствующие графики изображены на рисунках 3,4.

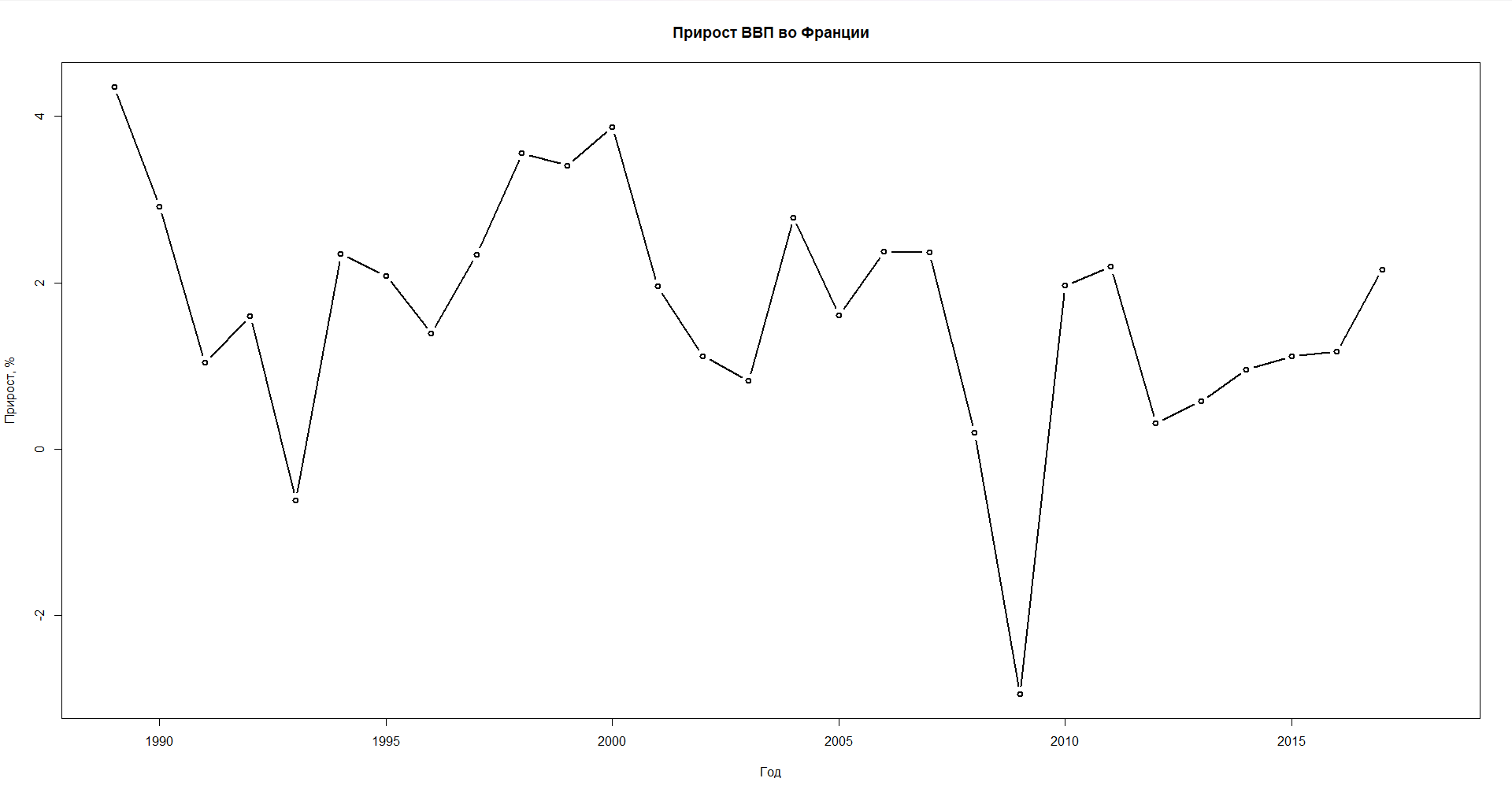


Рисунок 3 – Рост ВВП Франции на протяжении 30 лет в процентах

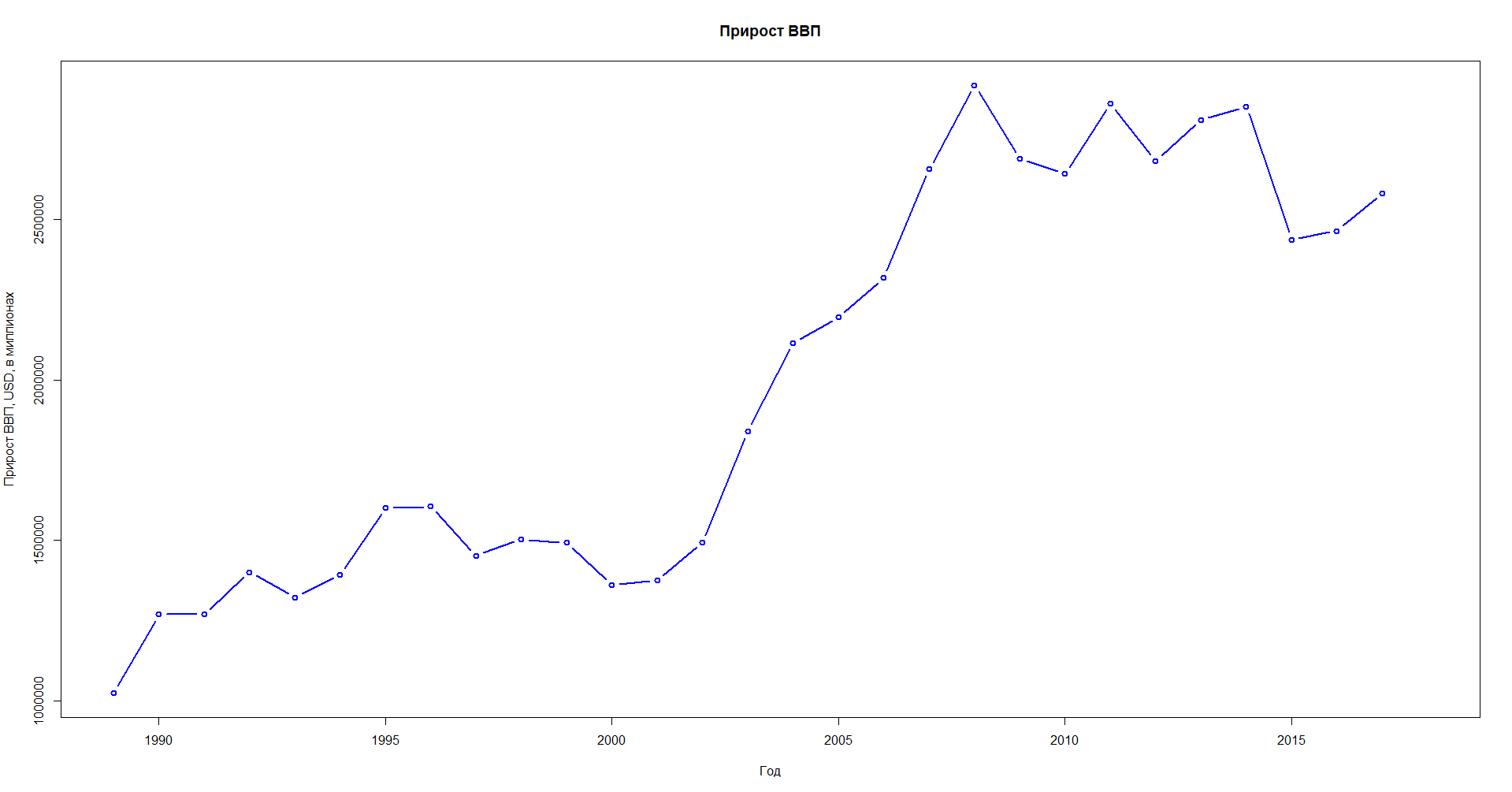


Рисунок 4 – Рост ВВП Франции на протяжении 30 лет в USD

Проанализируем рост ВВП и прирост населения. Выполним корреляцию двумя методами: Спирмена и Пирсона. Результаты работы методов отображены на рисунке 5.

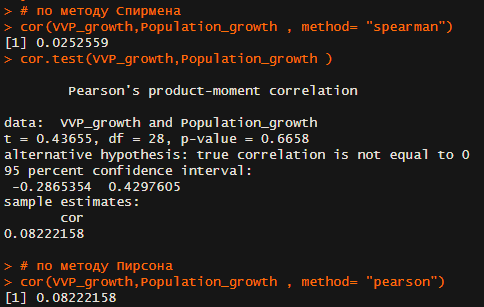


Рисунок 5 – Корреляция роста ВВП и прироста населения Франции

Корреляция между ростом ВВП и приростом населения слабая. Изобразим визуально результаты исследования корреляции между характеристиками. Значения коэффициентов корреляции рисуются в виде эллипсов, отражающих форму плотности двумерного нормального распределения с данным значением корреляции между компонентами. Чем ближе значение коэффициента корреляции к (+1) или (-1), тем более вытянутым становится эллипс. Наклон эллипса отражает знак. На рисунке 6 представлено представление в виде эллипсов для данных характеристик.

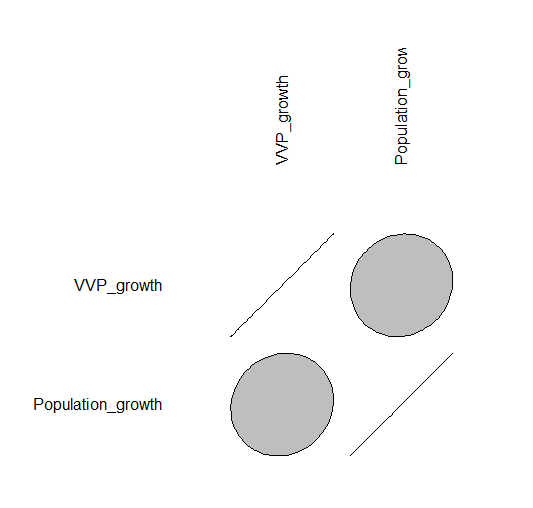


Рисунок 6 – Визуальное представление результатов исследования корреляции

Далее рассмотрим корреляцию между изменением расходов на медицину и увеличением смертности. На рисунке 7 отражена корреляция по аналогичным методам.

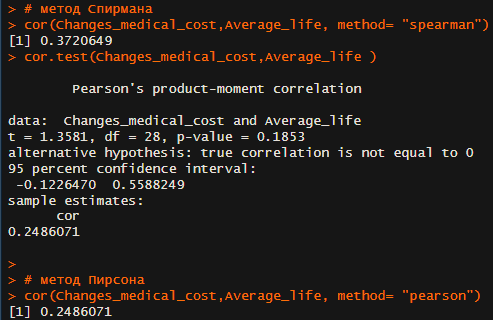


Рисунок 7 – Корреляция между изменением расходов на медицину и увеличением смертности

Как и в предыдущем опыте получаем, что корреляция между характеристиками слабая.

Рассмотрим между изменением расходов на медицину и смертностью. На рисунке 8 представлена корреляция для этих характеристик.

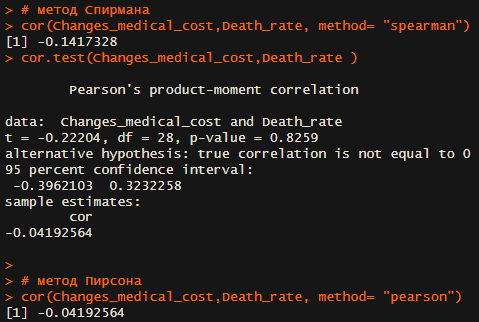


Рисунок 8 – Корреляция между изменением расходов на медицину и смертностью

Коэффициент корреляции между изменением расходов на медицину и смертностью считается средней корреляцией. Вновь изобразим результаты исследования корреляции визуально. В этот раз между тремя характеристиками. Рисунок 9 отображает эту корреляцию.

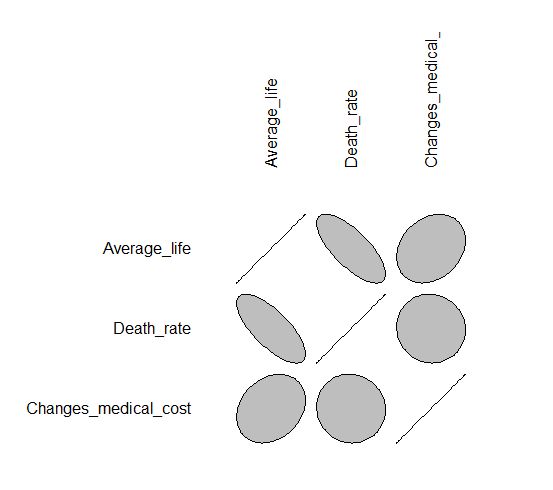


Рисунок 9 – Визуальное представление результатов исследования корреляции

В следующем задании необходимо было с помощью регрессионного анализа выявить зависимые и независимые переменные. Результат показан на рисунке 10.

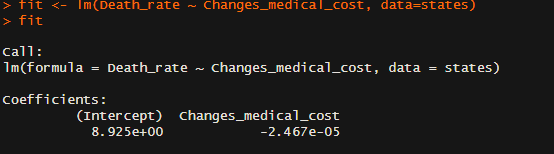


Рисунок 10 – Результат регрессионного анализа

В качестве независимой переменной были взяты расходы на медицину, а в качестве зависимой – значение смертности.

С помощью функции predict() было сделано предсказание по атрибуту значение смертности. Результат представлен на рисунке 11.

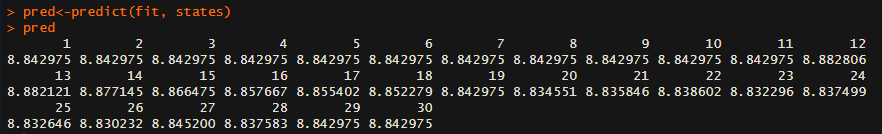


Рисунок 11 – Прогноз по атрибуту значение смертности

Для наглядности был построен график. Он представлен на рисунке 12.

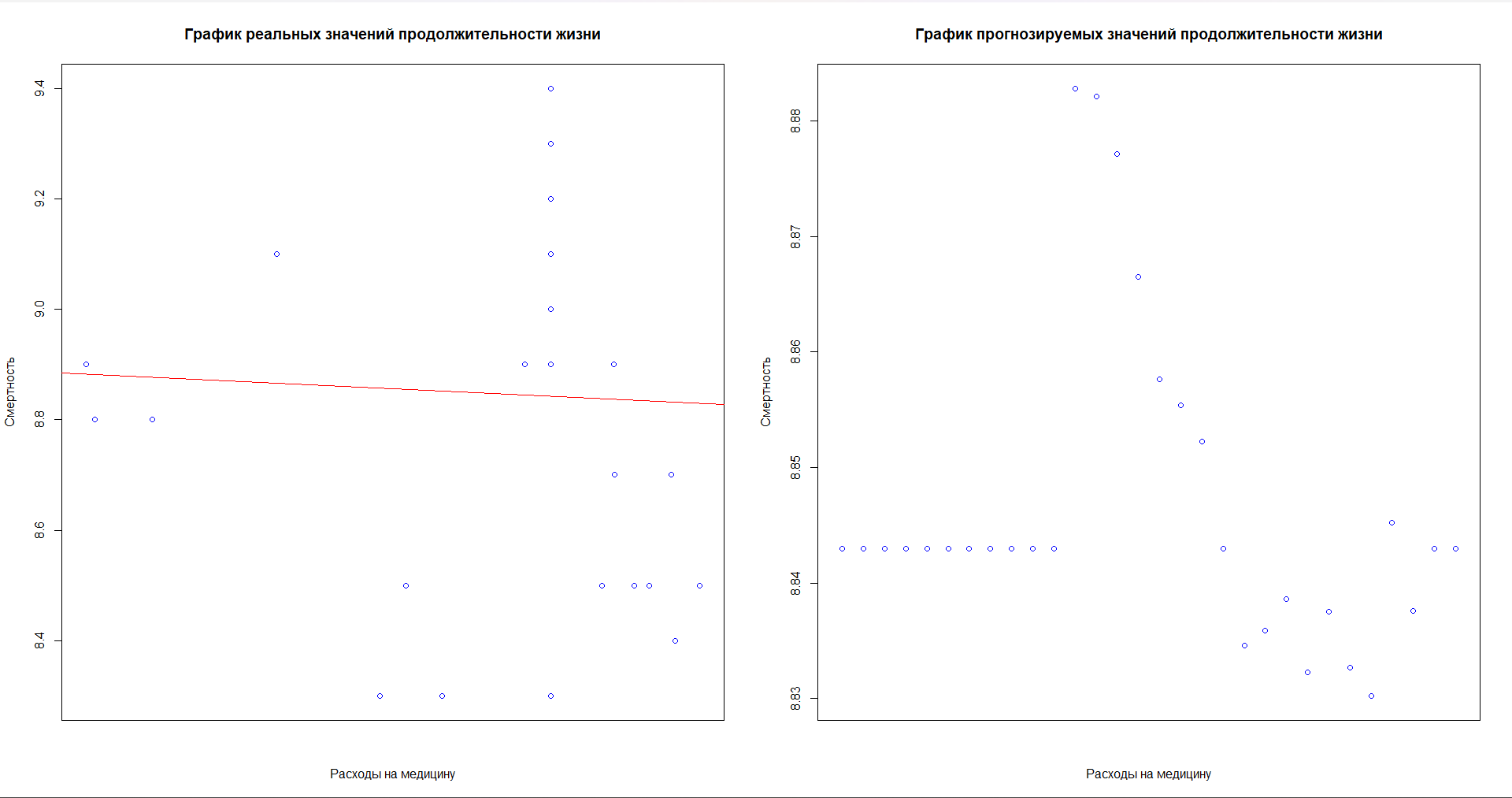


Рисунок 12 – Графики реальных и прогнозируемых значений смертности

**Вывод**: ознакомился с понятием корреляционного и регрессионного анализа данных, а также некоторыми функциями языка R, осуществляющими этот вид анализа и принципами их работы. Кроме того, научился оценивать связь между переменными и степень этой связи.

**Листинг программы**

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

install.packages("car")

library(car)

library(ellipse)

# импорт данных

df1 <- read.csv("C:/Users/vladc/OneDrive/Рабочий стол/Учёба/BigData/ЛР8/LR8\_dataset.csv", sep = "," , header = TRUE)

years=1989:2018

# выборка данных по данной стране

data\_set <- df1[df1$Country.Name == "France", ]

# замена отсутствующих данных на NA

for (i in 1:23){

for (j in 5:34){

if (data\_set[i,j]!=".."){

data\_set[i,j]= as.numeric(data\_set[i,j])

}

else{

data\_set[i,j]= NA

}

}

}

# прирост ВВП

VVP<-rep(0,30)

for (i in 5:34) {

VVP[i-4]=data\_set[2,i]

}

# построение графика прироста ВВП

plot(years, VVP, xlab='Год', ylab='Прирост, %', main='Прирост ВВП во Франции',type='b', lty=1, pch=1, lwd=2)

# прирост населения

Population\_growth<-rep(0,30)

for (i in 5:34) {

Population\_growth[i-4]=as.numeric(data\_set[14,i])

}

# прирост ВВП

VVP\_growth<-rep(0,30)

for (i in 5:34) {

VVP\_growth[i-4]=as.numeric(data\_set[1,i])

}

VVP\_growth <- VVP\_growth/1000000

# построение графика прироста ВВП

plot(

years,

VVP\_growth,

xlab='Год',

ylab='Прирост ВВП, USD, в миллионах',

main='Прирост ВВП',

col='blue',

type='b',

lty=1,

pch=1,

lwd=2

)

# поиск среднего и медианы для каждой выборки

mean(VVP\_growth,na.rm=TRUE)

median(VVP\_growth,na.rm=TRUE)

mean(Population\_growth,na.rm=TRUE)

median(Population\_growth,na.rm=TRUE)

# проверка на нормальность распределения

shapiro.test(Population\_growth)

shapiro.test(VVP\_growth)

# замена NA медианными значениями

for (i in 1:30){

if (is.na(VVP\_growth[i])){

#VVP\_growth[i]=0

VVP\_growth[i]= median(VVP\_growth,na.rm=TRUE)

}

if (is.na(Population\_growth[i])){

#Population\_growth[i]=0

Population\_growth[i]= median(Population\_growth,na.rm=TRUE)

}

}

# корреляция по методу Спирмена

cor(VVP\_growth,Population\_growth , method= "spearman")

cor.test(VVP\_growth,Population\_growth )

# корреляция по методу Пирсона

cor(VVP\_growth,Population\_growth , method= "pearson")

# корреляционная матрицы

df1=data.frame(VVP\_growth,Population\_growth)

# представление в виде эллипсов

plotcorr(cor(df1))

# расходы на медицину

Changes\_medical\_cost<-rep(0,30)

for (i in 5:34) {

Changes\_medical\_cost[i-4]=as.numeric(data\_set[12,i])

}

# продолжительность жизни

Average\_life<-rep(0,30)

for (i in 5:34) {

Average\_life[i-4]=as.numeric(data\_set[13,i])

}

# смертность

Death\_rate<-rep(0,30)

for (i in 5:34) {

Death\_rate[i-4]=as.numeric(data\_set[18,i])

}

# поиск среднего и медианы для каждой выборки

mean(Changes\_medical\_cost,na.rm=TRUE)

median(Changes\_medical\_cost,na.rm=TRUE)

mean(Average\_life,na.rm=TRUE)

median(Average\_life,na.rm=TRUE)

mean(Death\_rate,na.rm=TRUE)

median(Death\_rate,na.rm=TRUE)

# проверка на нормальность распределения

shapiro.test(Changes\_medical\_cost)

shapiro.test(Death\_rate)

shapiro.test(Average\_life)

# замена NA медианными значениями

for (i in 1:30){

if (is.na(Changes\_medical\_cost[i])){

#Changes\_medical\_cost[i]=0

Changes\_medical\_cost[i]= median(Changes\_medical\_cost,na.rm=TRUE)

}

if (is.na(Average\_life[i])){

#Average\_life[i]=0

Average\_life[i]= median(Average\_life,na.rm=TRUE)

}

if (is.na(Death\_rate[i])){

#Death\_rate[i]=0

Death\_rate[i]= median(Death\_rate,na.rm=TRUE)

}

}

# корреляция методом Спирмана средней продолжительности жизни

cor(Changes\_medical\_cost,Average\_life, method= "spearman")

cor.test(Changes\_medical\_cost,Average\_life )

# корреляция методом Пирсона средней продолжительности жизни

cor(Changes\_medical\_cost,Average\_life, method= "pearson")

# корреляция методом Спирмана рейтинга смертности

cor(Changes\_medical\_cost,Death\_rate, method= "spearman")

cor.test(Changes\_medical\_cost,Death\_rate )

# корреляция методом Пирсона рейтинга смертности

cor(Changes\_medical\_cost,Death\_rate, method= "pearson")

# корреляционная матрица

df3=data.frame(Average\_life,Death\_rate,Changes\_medical\_cost)

# представление в виде эллипсов

plotcorr(cor(df3))

# прирост людей с высшим образованием на рост экспорта товаров и на прирост высокотехнологичного производства

# люди в высшим образованием

Higher\_education\_growth<-rep(0,30)

for (i in 5:34) {

Higher\_education\_growth[i-4]=as.numeric(data\_set[19,i])

}

# экспорт товаров

Growth\_exports\_goods<-rep(0,30)

for (i in 5:34) {

Growth\_exports\_goods[i-4]=as.numeric(data\_set[17,i])

}

# высокотехнологическое производство

Growth\_technical\_production<-rep(0,30)

for (i in 5:34) {

Growth\_technical\_production[i-4]=as.numeric(data\_set[21,i])

}

# поиск среднего и медианы для каждой выборки

mean(Higher\_education\_growth,na.rm=TRUE)

median(Higher\_education\_growth,na.rm=TRUE)

mean(Growth\_exports\_goods,na.rm=TRUE)

median(Growth\_exports\_goods,na.rm=TRUE)

mean(Growth\_technical\_production,na.rm=TRUE)

median(Growth\_technical\_production,na.rm=TRUE)

# проверка на нормальность распределения

shapiro.test(Higher\_education\_growth)

shapiro.test(Growth\_technical\_production)

shapiro.test(Growth\_exports\_goods)

# замена NA медианными значениями

for (i in 1:30){

if (is.na(Higher\_education\_growth[i])){

#Higher\_education\_growth[i]=0

Higher\_education\_growth[i]= median(Higher\_education\_growth,na.rm=TRUE)

}

if (is.na(Growth\_exports\_goods[i])){

#Growth\_exports\_goods[i]=0

Growth\_exports\_goods[i]= median(Growth\_exports\_goods,na.rm=TRUE)

}

if (is.na(Growth\_technical\_production[i])){

#Growth\_technical\_production[i]=0

Growth\_technical\_production[i]= median(Growth\_technical\_production,na.rm=TRUE)

}

}

# корреляция методом Спирмана влияния высшего образования на прирост экспорта товаров

cor(Higher\_education\_growth,Growth\_exports\_goods, method= "spearman")

cor.test(Higher\_education\_growth,Growth\_exports\_goods )

# корреляция методом Пирсона влияния высшего образования на прирост экспорта товаров

cor(Higher\_education\_growth,Growth\_exports\_goods, method= "pearson")

# корреляция методом Спирмана влияния высшего образования на прирост высокотехнологичных товаров

cor(Higher\_education\_growth,Growth\_technical\_production, method= "spearman")

cor.test(Higher\_education\_growth,Growth\_technical\_production )

# корреляция методом Пирсона влияния высшего образования на прирост высокотехнологичных товаров

cor(Higher\_education\_growth,Growth\_technical\_production, method= "pearson")

# корреляционная матрица

df4=data.frame(Growth\_exports\_goods,Growth\_technical\_production,Higher\_education\_growth)

# представление в виде эллипсов

plotcorr(cor(df4))

# расходы на образование

Edication\_cost<-rep(0,30)

for (i in 5:34) {

Edication\_cost[i-4]=as.numeric(data\_set[15,i])

}

# бакалавры женщины

Female\_bachelors<-rep(0,30)

for (i in 5:34) {

Female\_bachelors[i-4]=as.numeric(data\_set[20,i])

}

# поиск среднего и медианы для каждой выборки

mean(Edication\_cost,na.rm=TRUE)

median(Edication\_cost,na.rm=TRUE)

mean(Female\_bachelors,na.rm=TRUE)

median(Female\_bachelors,na.rm=TRUE)

# проверка на нормальность распределения

shapiro.test(Edication\_cost)

shapiro.test(Female\_bachelors)

# замена NA медианными значениями

#Female\_bachelors - 1 значение, нет смысла

for (i in 1:30){

if (is.na(Edication\_cost[i])){

#Edication\_cost[i]=0

Edication\_cost[i]= median(Edication\_cost,na.rm=TRUE)

}

if (is.na(Female\_bachelors[i])){

#Female\_bachelors[i]=0

Female\_bachelors[i]= median(Female\_bachelors,na.rm=TRUE)

}

}

# метод Спирмана

cor(Edication\_cost,Female\_bachelors, method= "spearman")

cor.test(Edication\_cost,Female\_bachelors )

# метод Пирсона

cor(Edication\_cost,Female\_bachelors, method= "pearson")

# корреляционная матрица

df5=data.frame(Edication\_cost,Female\_bachelors)

# представление в виде эллипосв

plotcorr(cor(df5))

# научные статьи

Scientific\_Articles<-rep(0,30)

for (i in 5:34) {

Scientific\_Articles[i-4]=as.numeric(data\_set[23,i])

}

# поиск среднего и медианы для каждой выборки

mean(Higher\_education\_growth,na.rm=TRUE)

median(Higher\_education\_growth,na.rm=TRUE)

mean(Scientific\_Articles,na.rm=TRUE)

median(Scientific\_Articles,na.rm=TRUE)

# проверка на нормальность распределения

shapiro.test(Higher\_education\_growth)

shapiro.test(Scientific\_Articles)

# замена NA медианными значениями

#Higher\_education\_growth - 1 значение в таблице, нет смысла

for (i in 1:30){

if (is.na(Scientific\_Articles[i])){

#Article[i]=0

Scientific\_Articles[i]= median(Scientific\_Articles,na.rm=TRUE)

}

}

# метод Спирмана

cor(Higher\_education\_growth,Scientific\_Articles, method= "spearman")

cor.test(Higher\_education\_growth,Scientific\_Articles)

# метод Пирсона

cor(Higher\_education\_growth,Scientific\_Articles, method= "pearson")

# корреляционная матрица

df6=data.frame(Higher\_education\_growth,Scientific\_Articles)

# представление в виде эллипсов

plotcorr(cor(df6))

# все выборки

df2=data.frame(VVP\_growth,Population\_growth,Growth\_exports\_goods,Scientific\_Articles,Growth\_technical\_production,Female\_bachelors,Changes\_medical\_cost,Death\_rate,Average\_life)

# представление в виде эллипсов

plotcorr(cor(df2))

# хранение всех выборок в виде матрицы

states <- as.data.frame(df2[,c('VVP\_growth','Population\_growth',

'Growth\_exports\_goods','Scientific\_Articles',

'Growth\_technical\_production',

'Changes\_medical\_cost',

'Death\_rate','Average\_life')])

cor(states)

# независимая переменная - расходы на медицину,зависимая – рейтинг смертности.

fit <- lm(Death\_rate ~ Changes\_medical\_cost, data=states)

fit

pred<-predict(fit, states)

pred

par(mfrow=c(1, 2))

plot(states$`Changes\_medical\_cost`, states$`Death\_rate`, xaxt="n", xlab="Расходы на медицину", ylab="Смертность", col="blue", main="График реальных значений продолжительности жизни")

abline(fit, col="red")

plot(pred, xaxt="n", xlab="Расходы на медицину", ylab="Смертность", col="blue", main="График прогнозируемых значений продолжительности жизни")