МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«Кубанский государственный университет»**

**(ФГБОУ ВО «КубГУ»)**

**Кафедра вычислительных технологий**

**ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №8**

**Дисциплина: Обработка больших данных**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_К.С. Фисун

(подпись)

Факультет Компьютерных технологий и прикладной математики

Направление подготовки02.03.02 Фундаментальная информатика и информационные технологии курс 3

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Приходько Т.А.

Краснодар 2022

**Лабораторная работа № 8**

**«Корреляционный и регрессионный анализ данных»**

**Цель работы:**

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:**

Классическим инструментом для измерения *лилейной зависимости* между двумя наборами данных является коэффициент корреляции. Коэффици­ент корреляции — это числовая величина, находящаяся в интервале от -1 до +1. Чем она больше по модулю (т.е. ближе к 1 или к -1), тем выше линейная связь между наборами данных. Знак коэффициента корреляции показывает, в одном ли направлении изменяются наборы данных. Если один из наборов возрастает, а второй убывает, то коэффициент корреляции отрицателен, а если оба набора од­новременно возрастают или убывают, то коэффициент корреляции положителен. Значение коэффициента корреляции по модулю равное 1 соответствует точной линейной зависимости между двумя наборами данных. Линейная зависимость является самой любимой у экспериментаторов всех мастей.

Набор исходных данных – различные экономические показатели стран за период 1989-2017 (ВВП, Процентный рост ВВП, Процент рождения скилованным персоналом, Рождаемость, Национальный доход, Процент безработных и т.п). Обработанный датасет для Индии представлен на рисунке 1.



Рисунок 1 – Фрагмент исходных данных.

Построим график прироста ВВП за весь период (рисунок 2).

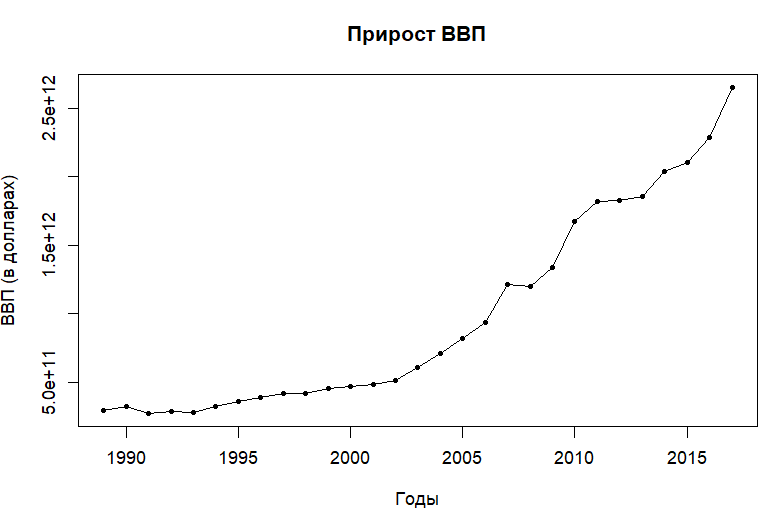
****

Рис. 2 – График роста ВВП.

Индия за последнии 30 лет имеет постоянный рост ВВП, особенно заметен рост начиная с 2005 года.

Вычислим корреляционную матрицу для каждого показателя и построим их визуализацию (рисунок 3).

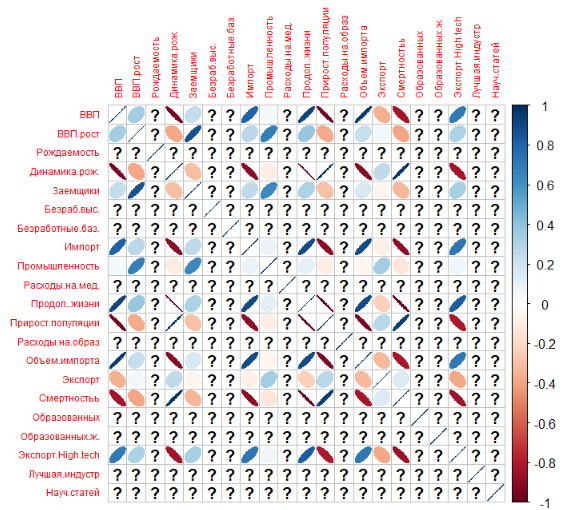
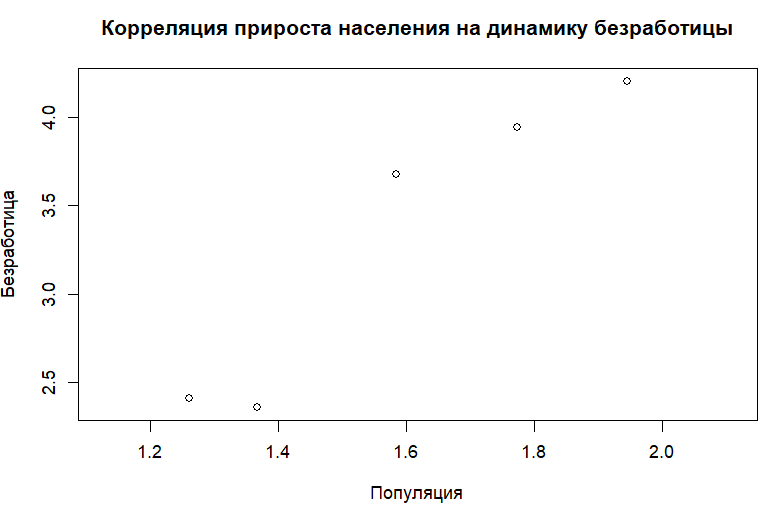
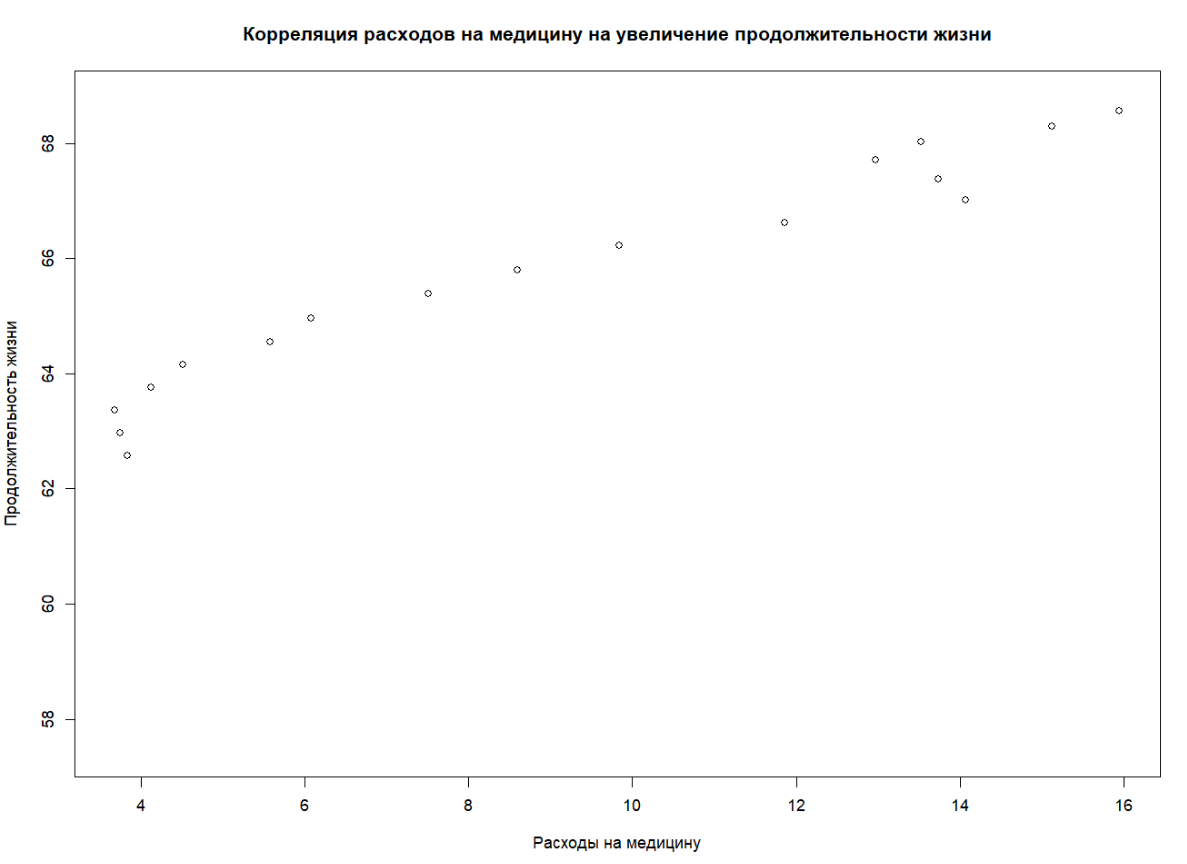
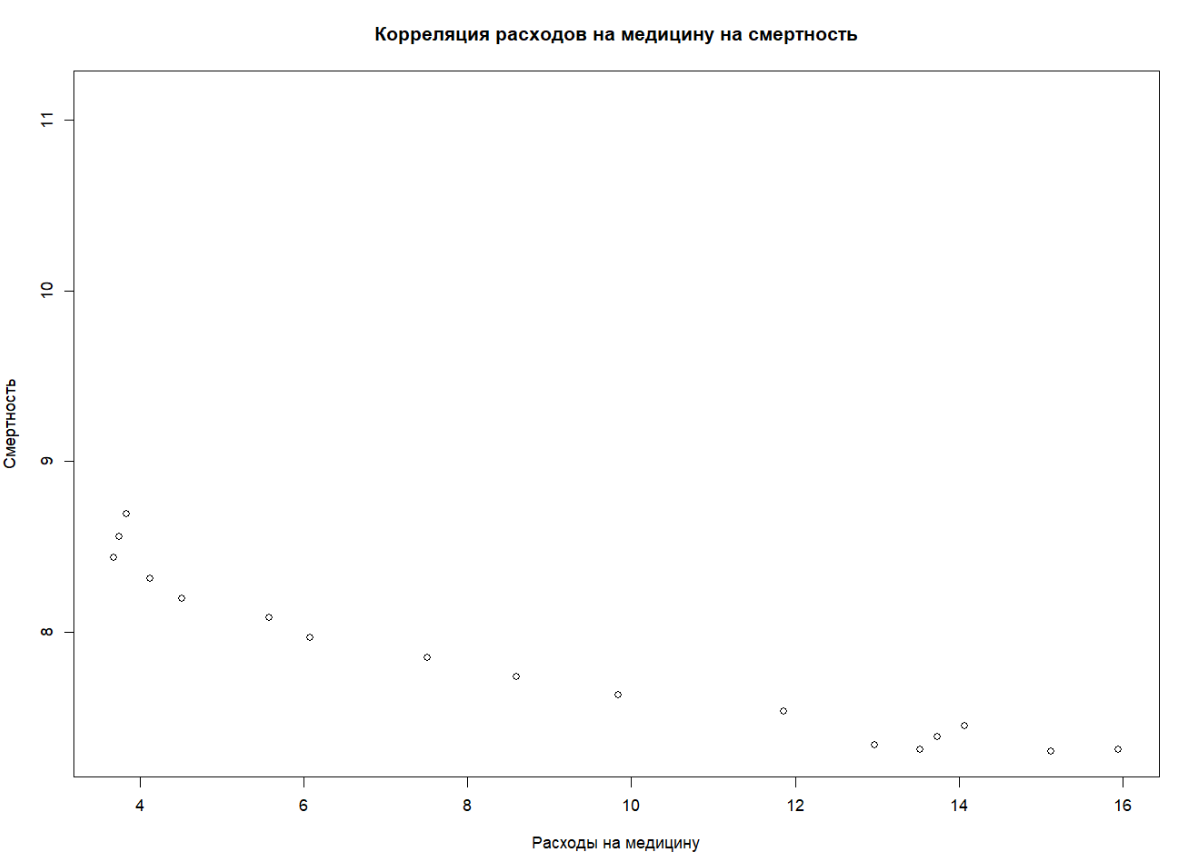
****

Рис. 3 – Корреляционная матрица.

Исходя из данных матрицы можно сделать вывод о том, как некоторые параметры влияют на другие. Далее представлены корреляция на параметрах, которые имеют заметную линейную зависимость.







Регрессионный анализ можно использовать для обнаружения независимых переменных, которые имеют отношение к зависимой, для описания типа взаимосвязи и для составления уравнения, позволяющего предсказать значения зависимой переменной по значениям независимых. Важный первый шаг в множественной регрессии – исследование парных взаимосвязей между переменными. Двухмерные корреляции вычисляются при помощи функции cor(), а диаграммы рассеяния создаются при помощи функции scatterplotMatrix() из пакета car () (рисунок 4).

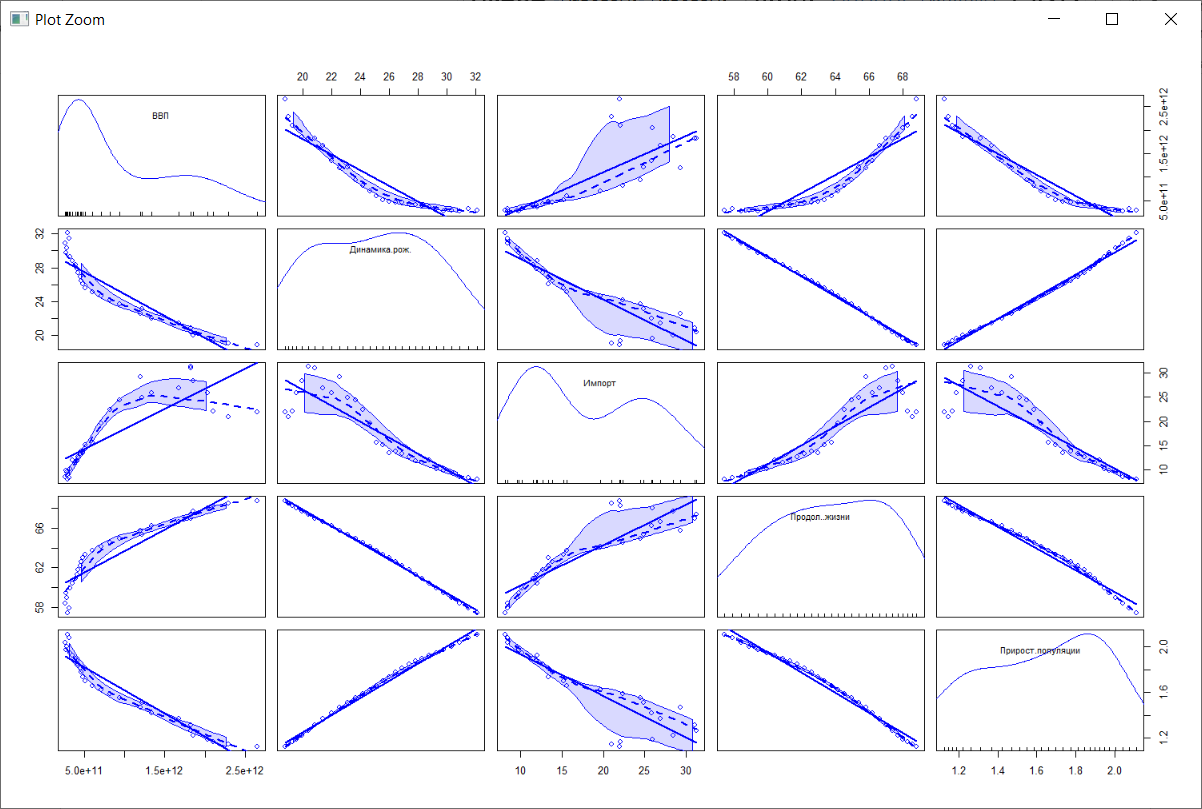


Рис. 4 ­­­– Матрица диаграмм рассеивания.

Построим полиномиальная регрессионную модель зависимости продолжительности жизни от ВВП (рисунок 5).

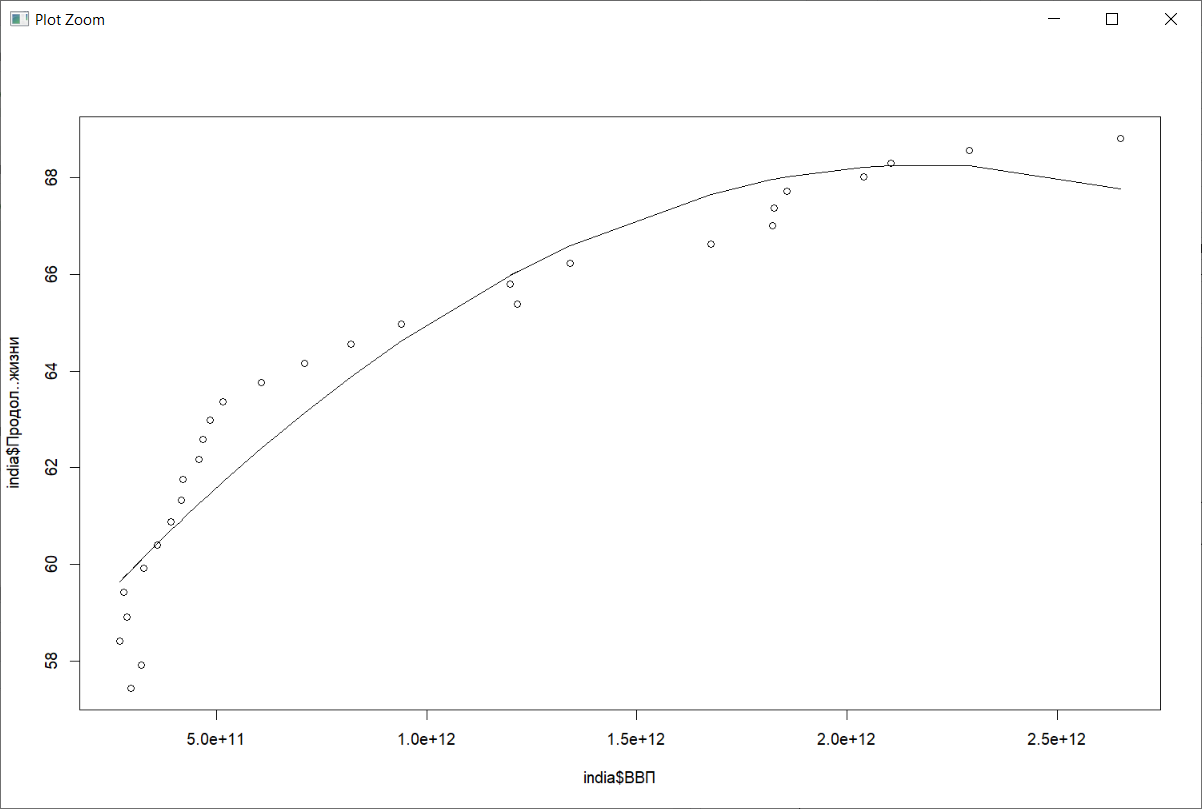
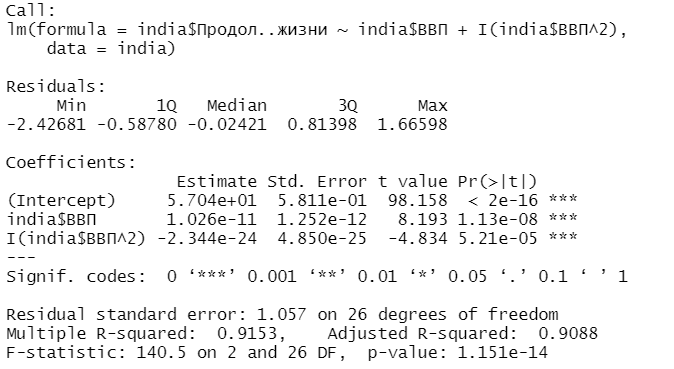


Рис. 4 ­­­– Регрессионная модель.



Если существует больше одной независимой переменной, то регрессионные коэффициенты показывают, на сколько увеличится значение зависимой переменной при изменении данной независимой переменной на единицу при условии, что все остальные независимые переменные останутся неизменными.

Вывод: Ознакомился с некоторыми статистическими тестами, принципами их работы. Научился оценивать нормальность распределения выборки, а также выполнил оценку статистических гипотез.