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: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Курбатский В.А.

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Направленность (профиль): Компьютерные технологии и прикладная математика.

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Приходько Т. А.

Краснодар

2023

Вариант 7 - Сербия

**Постановка задачи**

1. С указанного ниже ресурса (Всемирный Банк - World Development Indicators database) для студентов сформирована CSV-таблица, содержащая информацию по 18 странам мира и по 22 характеристикам по каждой стране за годы (1989-2018). **За набором данных обращаться к преподавателю.**

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

1. 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 понравившемуся Вам атрибуту.

Данные NA не учитывать. Рекомендую ваши результаты сопоставить с таблицей 8.6. или на сайте <https://data.worldbank.org/country/united-kingdom?view=chart>

1. Форма вашего датасета: Таблица 8.4.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Годы | GDP (ВВП) | Прирост ВВП (GDP %) | Рост рождаемости, обусловленный квалификацией персонала | Динамика роста рождаемости на 1000 человек | Заемщиков в банках на 1000 человек населения | . SL.UEM.ADVN.ZS | SL.UEM.BASC.ZS | IP.JRN.ARTC.SC |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 22 |
| 1989 |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |
| 2017 |  |  |  |  |  |  |  |  |

**Ход работы:**

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

Изображение выглядит как текст, снимок экрана, число, Параллельный

Автоматически созданное описание

Рисунок 1 – Исходная таблица с данными

В данном наборе данных представлены данные о двадцати двух характеристиках восемнадцати стран мира с 1989 года по 2018 год. Удалим все страны, кроме страны согласно варианту, а также повторяющиеся строчки. На рисунке 2 изображена таблица получившихся данных.

Изображение выглядит как текст, снимок экрана, число, линия

Автоматически созданное описаниеРисунок 2 – Измененная таблица с данными

Согласно первому заданию, построим кривую прироста ВВП Франции в процентах и USD. Соответствующие графики изображены на рисунках 3,4.

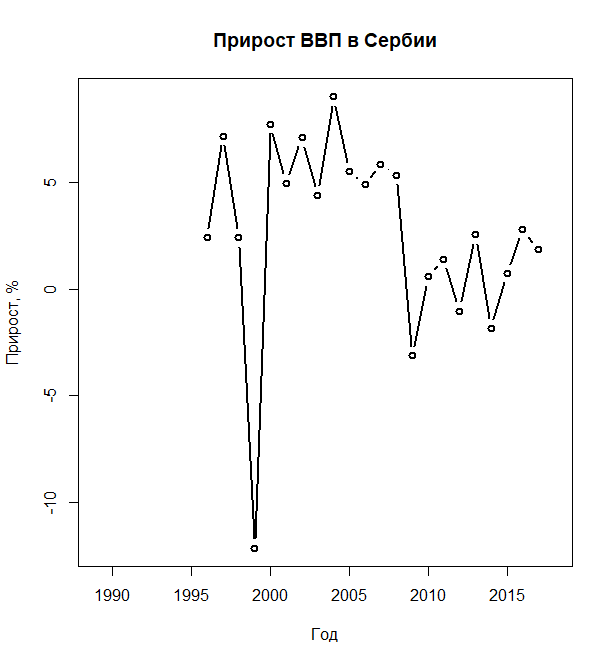


Рисунок 3 – Рост ВВП Франции на протяжении 23 лет в процентах

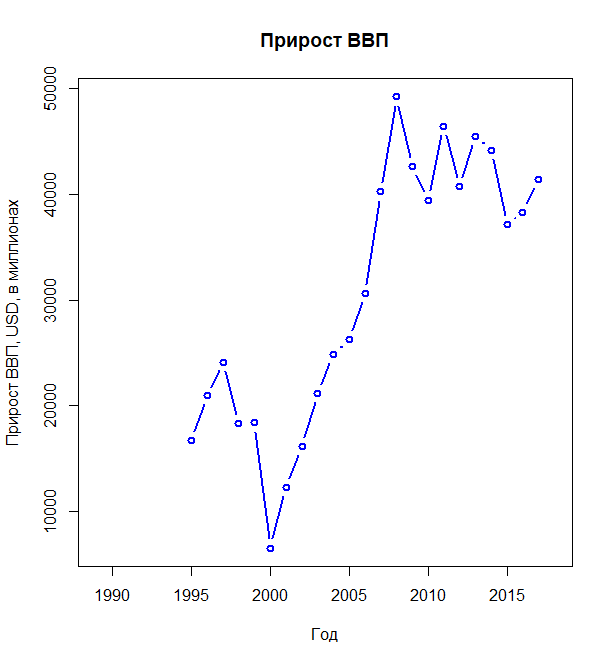


Рисунок 4 – Рост ВВП Сербии на протяжении 23 лет в USD

Далее приступим к поиску корреляции заданных характеристик.

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

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рисунок 5 – Корреляция роста ВВП и прироста населения Сербии

По полученным значениям видно, что 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.

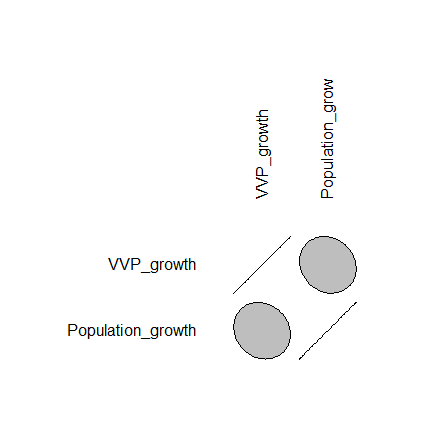


Рисунок 6 – Визуальное представление результатов исследования корреляции

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

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рисунок 7 – Корреляция между изменением расходов на медицину и увеличением продолжительности жизни

Корреляция между характеристиками высокая положительная.

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

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

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.

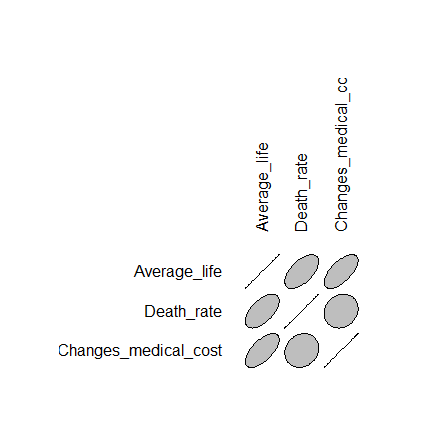


Рисунок 9 – Визуальное представление результатов исследования корреляции

Провести аналогичные действия

* для прироста людей с высшим образованием на рост экспорта товаров и на прирост высокотехнологичного производстваИзображение выглядит как текст, снимок экрана, Шрифт, линия

  Автоматически созданное описание
* для расходов на образование на кумулятивный прирост бакалавров среди женщин

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

* для прироста людей с высшем образованием на развитие высоких технологий.

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

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

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

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.

Изображение выглядит как текст, Шрифт, линия, снимок экрана

Автоматически созданное описание

Рисунок 11 – Прогноз по атрибуту значение смертности

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

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

Рисунок 12 – Графики реальных (слева) и прогнозируемых (справа) значений смертности

**Вывод:** В данной работе были ознакомлены с понятием корреляционного и регрессионного анализа данных, некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Помимо этого, научились оценивать связь между переменными и оценивать степень этой связи.