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**

**Отчёт лабораторная №7**

**Дисциплина: ОБРАБОТКА БОЛЬШИХ ДАННЫХ**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Козлов Э.Д.

Направление подготовки 02.03.02 Фундаментальная информатика и

информационные технологии

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Яхонтов А.А.

**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.

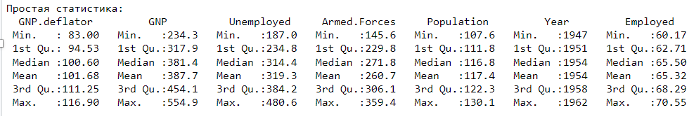
**Ход работы:**

**Задание 7.2.1**

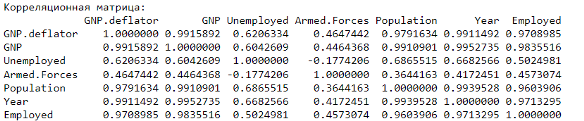
Количество переменных и объем выборки в наборе данных longley:



Дескриптивный анализ:

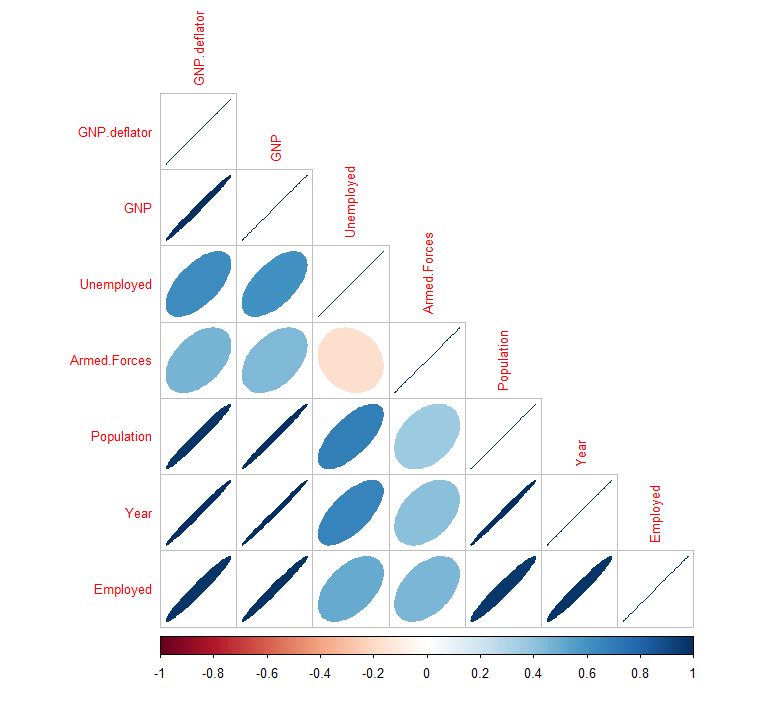


Корреляционная матрица:



Корреляционный анализ показал сильную положительную связь между GNP, Population и Employed (> 0.98), что говорит о росте занятости и населения при увеличении ВНП. Переменная Year также сильно связана с другими, отражая общий рост показателей со временем. Unemployed коррелирует слабо, возможна нелинейная зависимость.

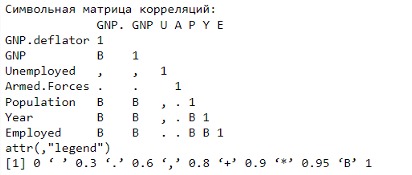
График корелляций:



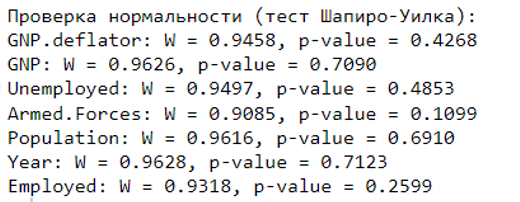
На графике визуально подтверждается сильная положительная корреляция между переменными GNP, Population, Employed и Year – эллипсы вытянуты по диагонали вверх, окрашены в тёмно-синий. Это говорит о тесной линейной зависимости.

Самая слабая и единственная отрицательная корреляция наблюдается между Unemployed и Armed.Forces – эллипс слегка наклонён вниз и окрашен в светло-красный.

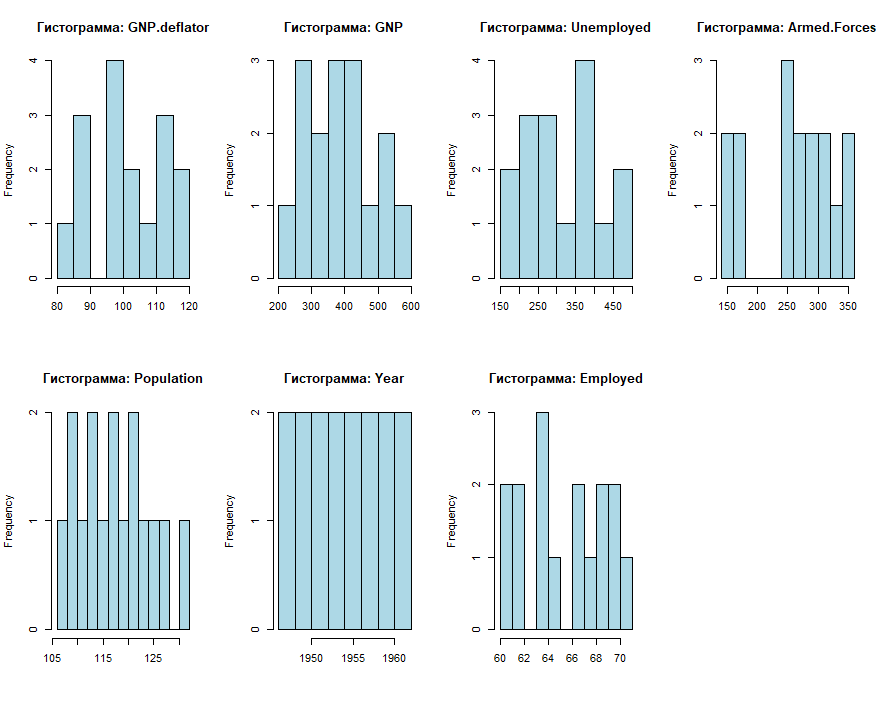
Символьный вариант:

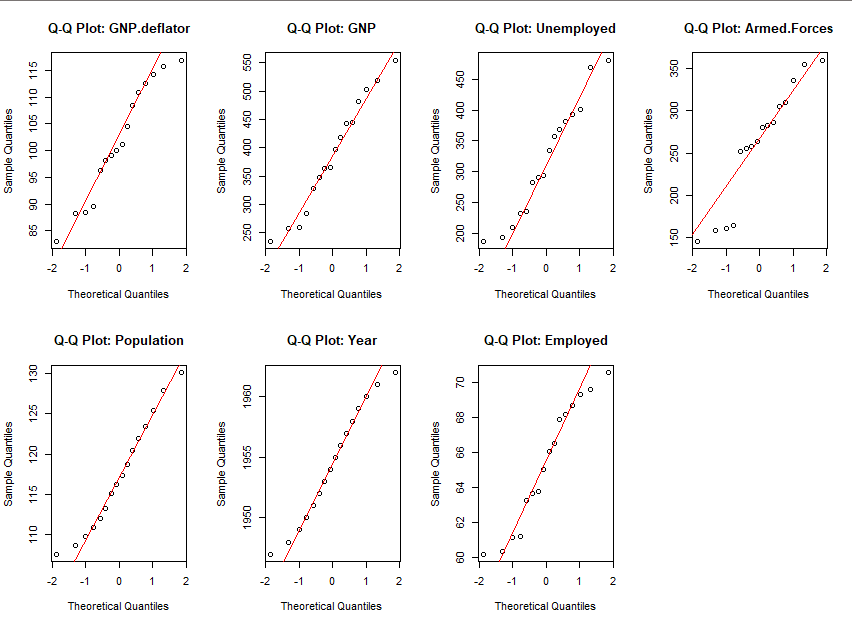


Проверка на нормальность распределения:



Графическое подтверждение нормальности:



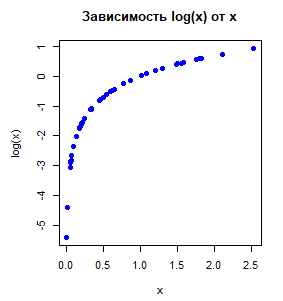


**Задание 7.2.2**

Для переменных x и log(x), где x распределена экспоненциально, был рассчитан коэффициент ранговой корреляции Спирмена. Результат равен 1, что указывает на идеальную монотонную зависимость между переменными.

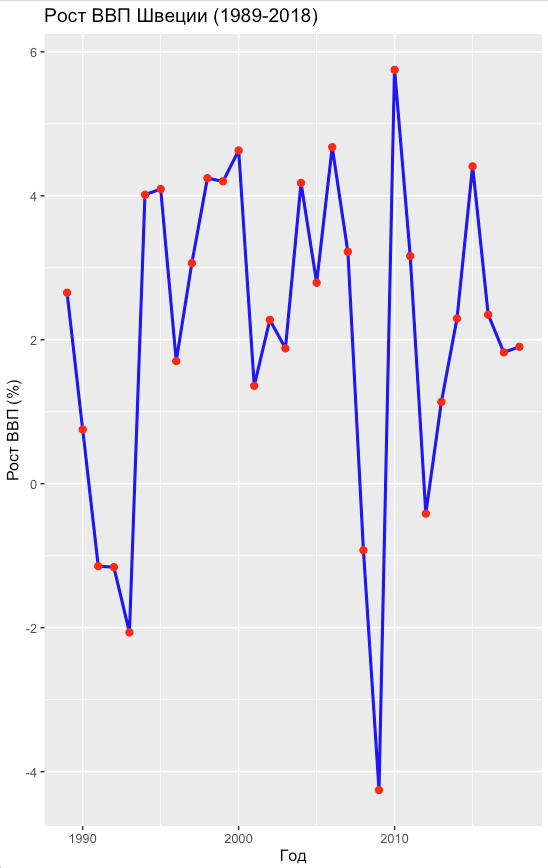
На графике видно, что точки образуют строго возрастающую кривую – логарифмическую. Это подтверждает наличие устойчивой зависимости, несмотря на её нелинейную форму.

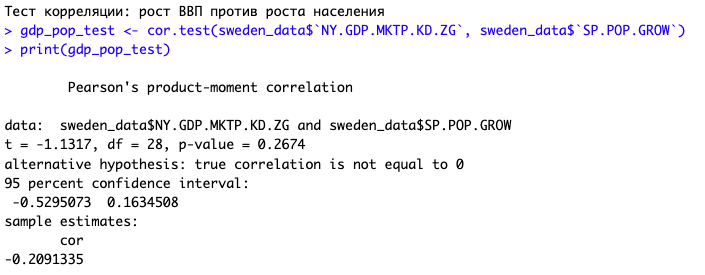




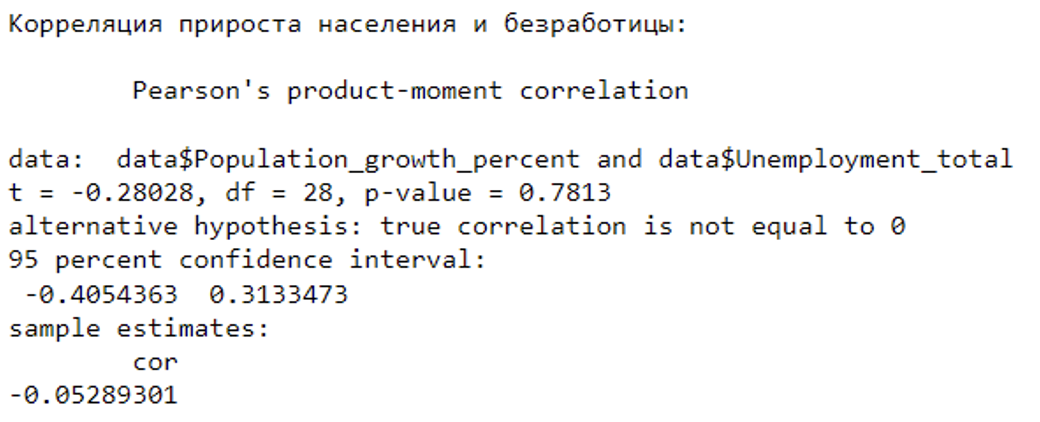
**Задание 7.5**

Вариант 5.

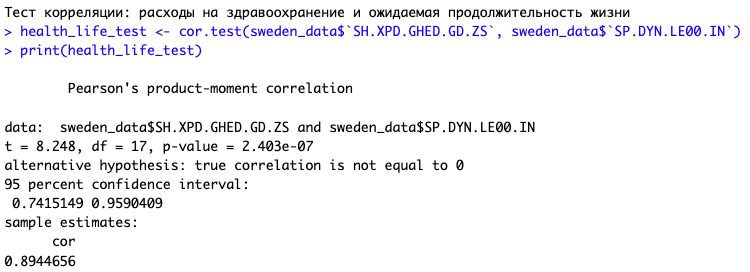




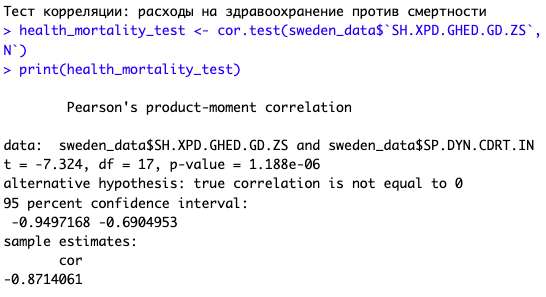
Слабая отрицательная связь (r = -0.21), но статистически не значима (p = 0.267 > 0.05). Это означает, что есть слабая тенденция к снижению темпов роста ВВП при увеличении темпов роста населения, однако эта связь могла возникнуть случайно.



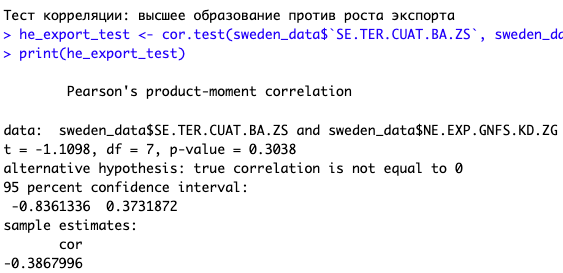
Корреляция близка к нулю (r = -0.05), связь отсутствует (p = 0.78). Изменение безработицы не связано с динамикой населения.



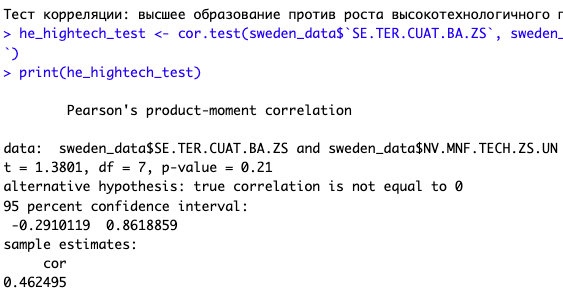
Очень сильная положительная связь (r = 0.89), статистически значима (p < 0.001). Это означает, что увеличение государственных расходов на здравоохранение связано с увеличением ожидаемой продолжительности жизни, и такая связь вряд ли возникла случайно.



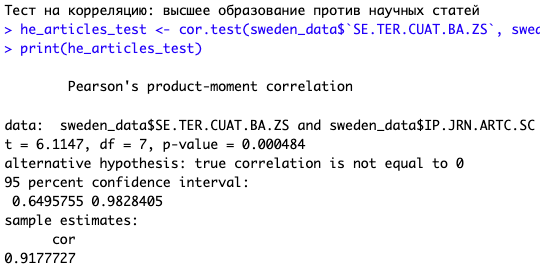
Очень сильная отрицательная связь (r = -0.87), статистически значима (p < 0.001). Это означает, что с увеличением государственных расходов на здравоохранение уровень смертности снижается, и такая связь крайне маловероятно могла возникнуть случайно.



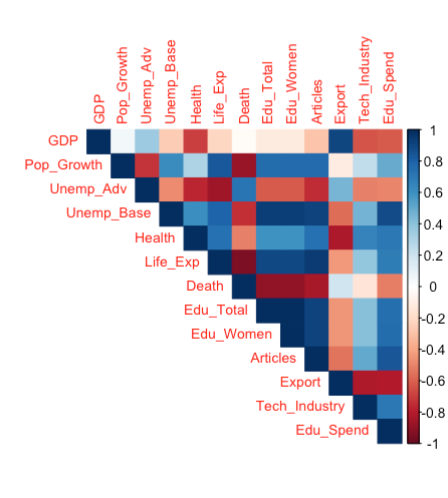
Умеренная отрицательная связь (r = -0.39), но статистически не значима (p = 0.304 > 0.05). Это означает, что наблюдается некоторая тенденция к снижению темпов роста экспорта при увеличении доли населения с высшим образованием, однако эта связь могла возникнуть случайно.

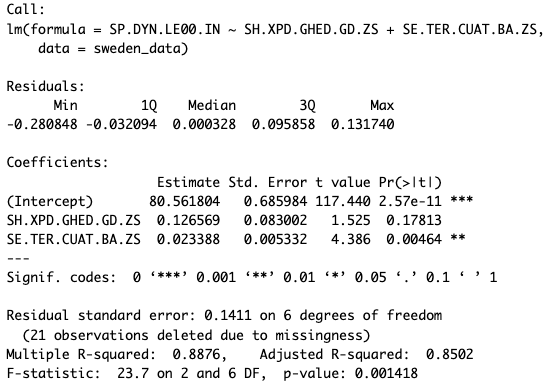


Умеренная положительная связь (r = 0.46), но статистически не значима (p = 0.210 > 0.05). Это означает, что есть тенденция к росту доли высокотехнологичного производства при увеличении доли населения с высшим образованием, однако эта связь могла возникнуть случайно.



Очень сильная положительная связь (r = 0.92), статистически значима (p < 0.001). Это означает, что увеличение доли населения с высшим образованием связано с ростом числа опубликованных научных статей, и такая связь вряд ли возникла случайно.





**Модель:**  
Продолжительность жизни = 80.56 + 0.13 × Расходы на здравоохранение + 0.023 × Высшее образование

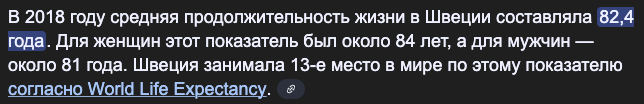
**Качество модели:**  
R² = 0.888: модель объясняет 88.8% вариации продолжительности жизни.  
Adjusted R² = 0.850: с учётом числа предикторов модель всё ещё хорошо объясняет данные.  
F-statistic = 23.7 (p = 0.001): модель в целом статистически значима (p < 0.05).

**Коэффициенты:**  
Высшее образование: p = 0.0046 — статистически значимо, положительное влияние.  
Расходы на здравоохранение: p = 0.178 — не значимо (p > 0.05), влияние может быть случайным.

Прогноз:



По данной модели, если условия (например, расходы на медицину и уровень высшего образования) сохранятся такими же, то ожидается, что в 2018 году продолжительность жизни Швеции возрастет до примерно 82.46 лет.



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