Министерство образования и науки Российской Федерации

Севастопольский государственный университет

Кафедра ИС

Отчет

По дисциплине: “Интеллектуальный анализ данных”

Лабораторная работа №2

“Корреляционный и регрессионный анализ данных”

Выполнил:

ст.гр. ИС/б-32

Долженко И.А.

Проверила:

Сырых О.А.

Севастополь

2020

1 ЦЕЛЬ РАБОТЫ

Исследовать возможности языка R для проведения корреляционного и регрессионного анализа данных. Создание набора данных для проведения корреляционного и регрессионного анализа данных. Исследовать возможности языка R для создания и изменения вида диаграмм. Исследовать возможности языка R для определения тесноты взаимосвязей экспериментальных данных.

2 ХОД РАБОТЫ

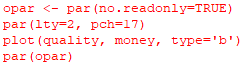
1. Выполним построение графиков для своего набора данных.





Рисунок 1 – График с регрессионной прямой

2. При помощи функции par() изменим характеристики диаграммы.



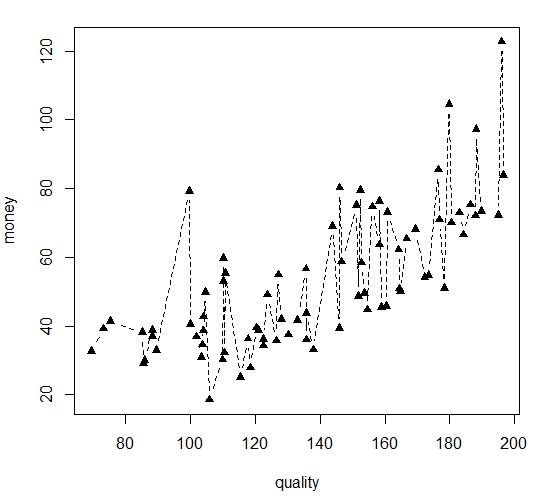
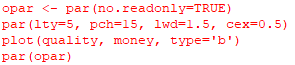


Рисунок 2 – График после изменения типа линии и символа

3. Применим параметры для изменения размера символа и толщины линии.



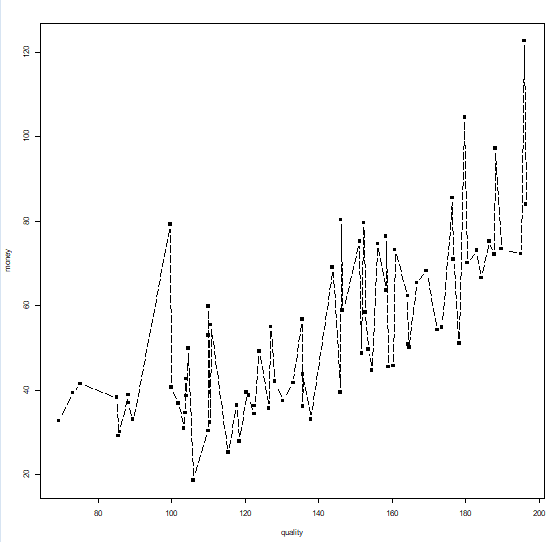


Рисунок 3 – График после изменения толщины линии и размера символов

4. Изменим тип графика и воспользуемся параметром для изменения цвета.



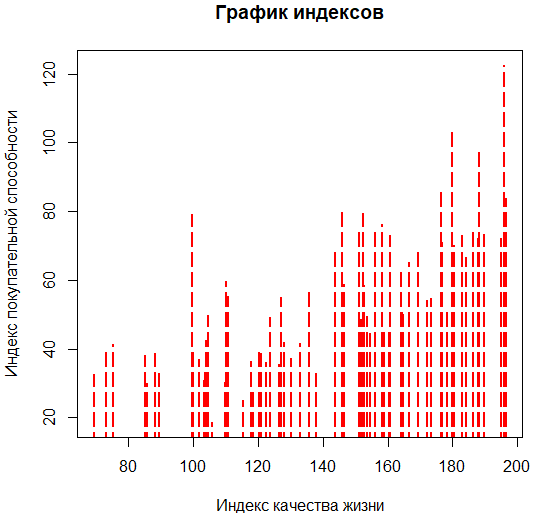


Рисунок 4 – График после изменения параметра цвета и добавления подписей

5. Воспользуемся корреляционным анализом:

- корреляция Пирсона

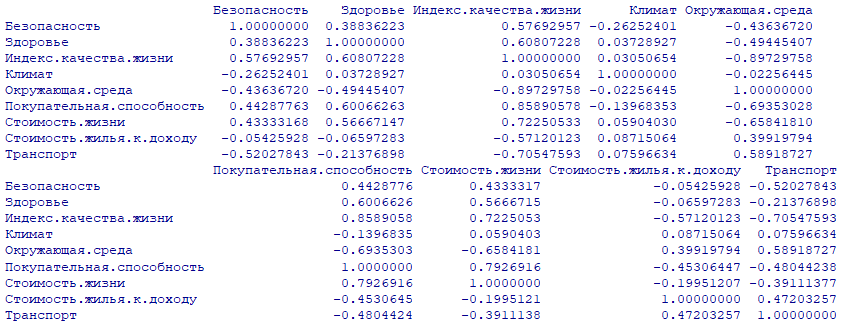


Рисунок 5 – Матрица корреляций Пирсона

- корреляция Спирмена

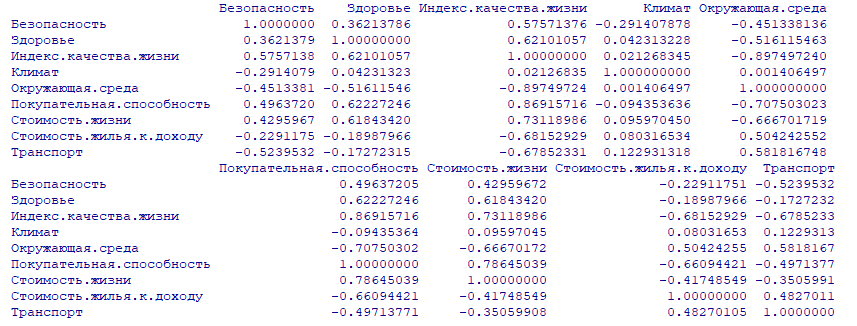


Рисунок 6 – Матрица корреляций Спирмена

Исходя из полученных матриц корреляций можно сделать вывод, что самыми сильными связями являются:

1) индекс качества жизни – окружающая среда (-0.8974)

2) индекс качества жизни – покупательная способность (0.8691)

3) стоимость жизни – покупательная способность (0.7864)

Самые слабые связи:

1) покупательная способность – климат (0.0943)

2) стоимость жизни – климат (0.0959)

3) индекс качества жизни – климат (0.0212)

6. Найдем уровень значимости коэффициентов корреляции для нескольких пар.

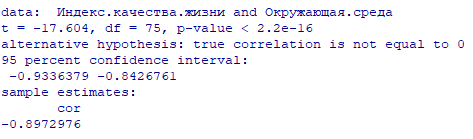


Рисунок 7 – Корреляционный тест для пары Индекс качества жизни

и Окружающая среда

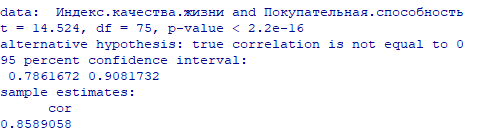


Рисунок 8 – Корреляционный тест для пары Индекс качества жизни

и Покупательная способность

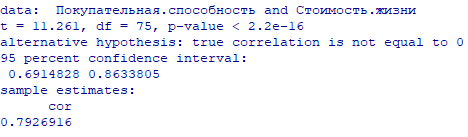


Рисунок 9 – Корреляционный тест для пары Покупательная способность

и Стоимость жизни

7. Построим матрицу точечных графиков для этих же переменных.

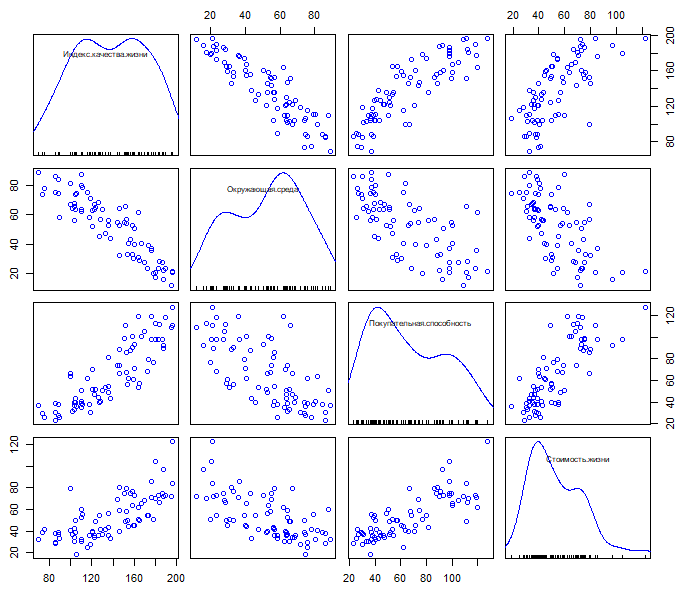


Рисунок 10 – Матрица точечных графиков

8. Построим уравнение зависимости индекса качества жизни от индекса окружающей среды и построим график остатков.

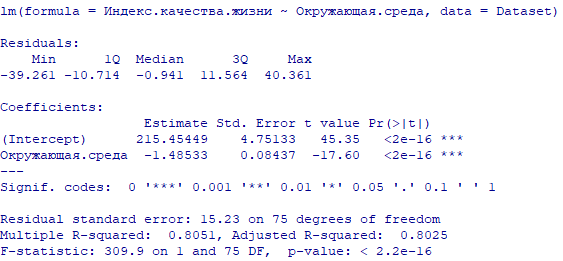


Рисунок 11 – Зависимость индекса качества жизни

от индекса окружающей среды

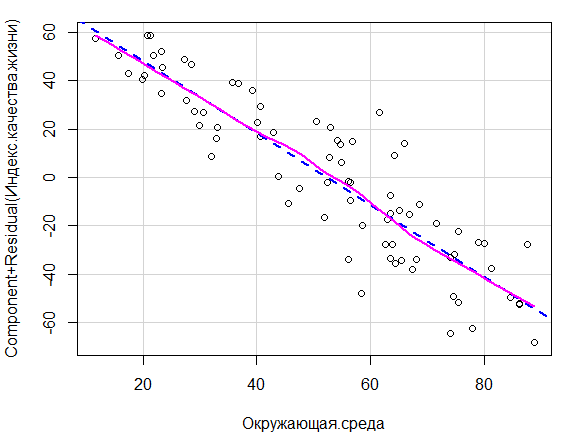


Рисунок 12 – График остатков

9. Проведем подгонку множественной регрессионной модель при помощи функции lm() и построим графики остатков.

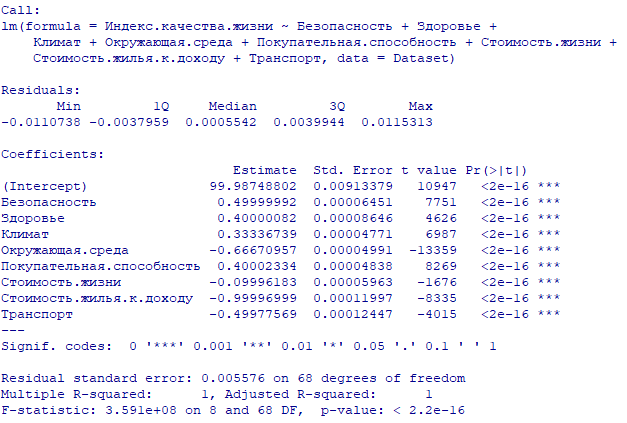


Рисунок 13 – Результат подгонки множественной регрессионной модели

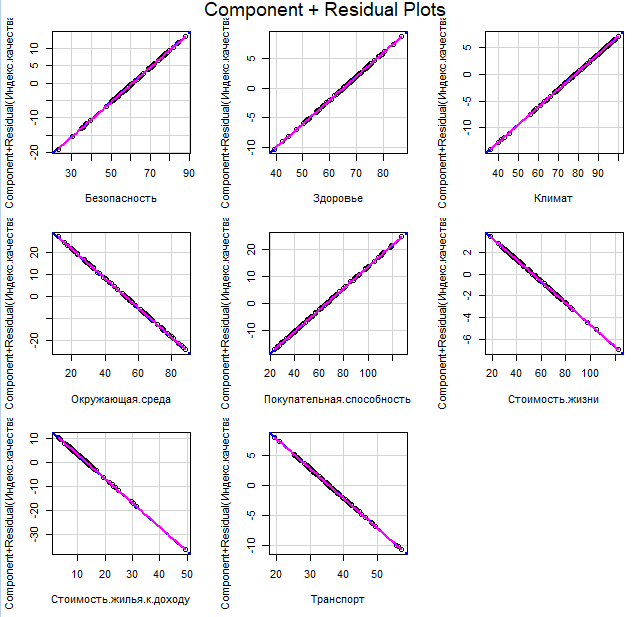


Рисунок 14 – Графики остатков

10. Выполним пошаговое построение регрессии по направлению вперед и назад.

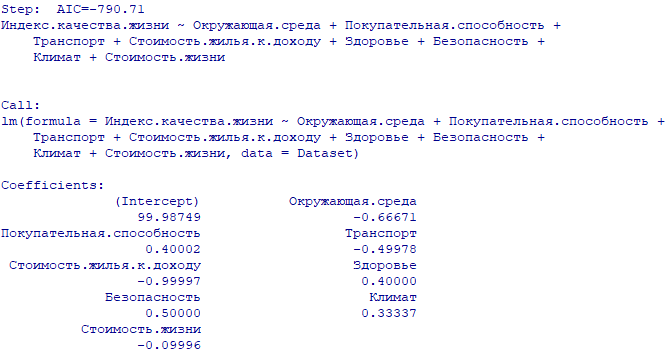


Рисунок 15 – Пошаговое построение регрессии по направлению вперед

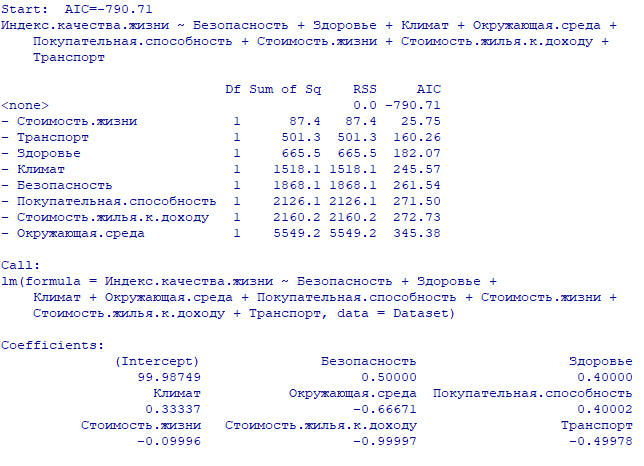


Рисунок 16 – Пошаговое построение регрессии по направлению назад

11. Проверим коэффициент VIF.

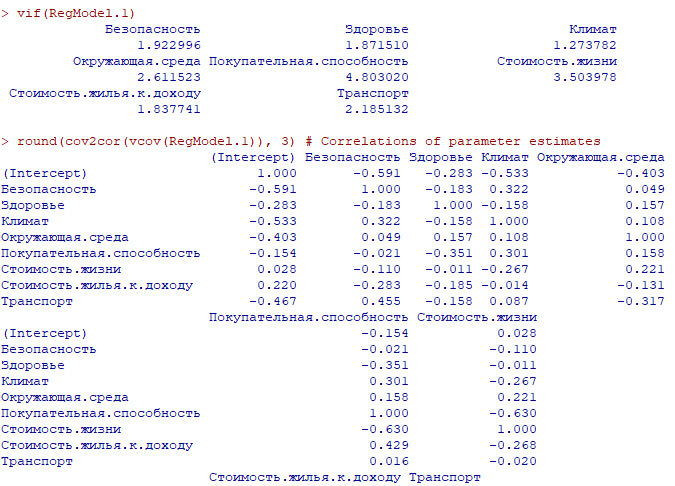


Рисунок 17 – Коэффициент VIF

В результате проверки коэффициента VIF, можно выявить отсутствие мультиколлинеарности, из-за наличия низких (<10) значений коэффициента VIF.

ВЫВОД

В ходе выполнения лабораторной работы были построены графики по своему набору данных, изменены их параметры и типы. Была построена матрица корреляций Пирсона и Спирмена и по их результатам были найдены самые сильные связи (индекс качества жизни – окружающая среда (-0.8974), индекс качества жизни – покупательная способность (0.8691), стоимость жизни – покупательная способность (0.7864)). Была построена матрица точечных графиков для этих переменных. Проведена подгонка множественной регрессионной модель при помощи функции lm() и построены графики остатков. Выполнено пошаговое построение регрессии по направлению вперед и назад, в результате чего была получена наилучшая модель, состоящая из всех переменных модели и найденный коэффициент AIC равен значению -790.71. В результате проверки коэффициента VIF, было выявлено отсутствие мультиколлинеарности, из-за наличия низких (<10) значений коэффициента VIF.