Министерство науки и высшего образования РФ

Севастопольский государственный университет

Кафедра информационных систем

ОТЧЕТ

по лабораторной работе №2

Корреляционный и регрессионный анализ данных

по дисциплине «Интеллектуальный анализ данных»

Выполнил:

Студент группы ИС/б 17-2-о

Черняев Н.Г.

Проверил:

Сырых О.А.

г. 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

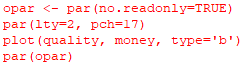
1. По экспериментальным данным проведем построение диаграммы и используя изученные функции и команды проведем модификацию построенных графиков.





Рисунок 2.1 – График с регрессионной прямой

При помощи функции par() изменим характеристики диаграммы.



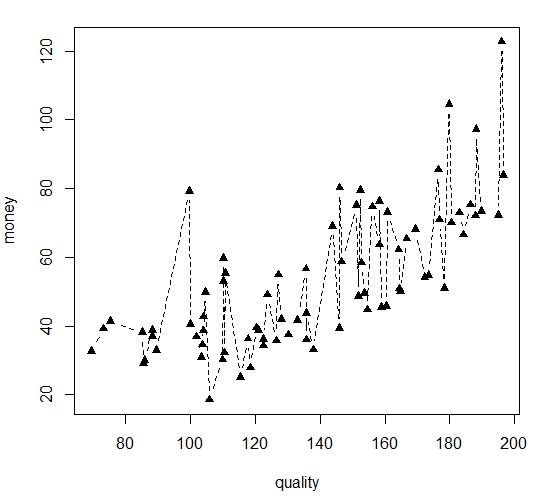
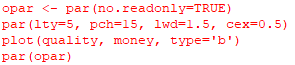


Рисунок 2.2 – График после изменения типа линии и символа

Применим параметры для изменения размера символа и толщины линии.



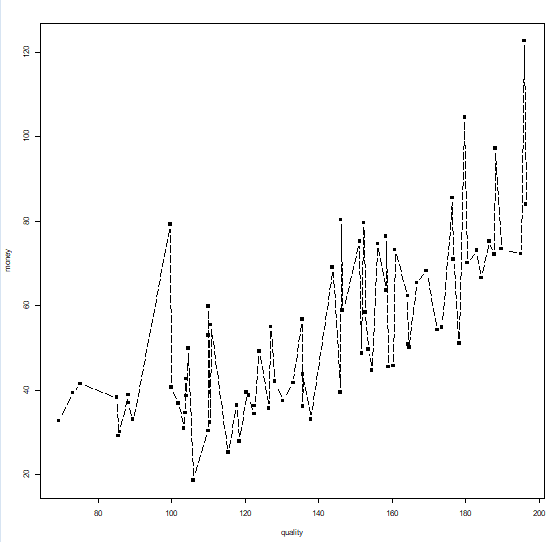


Рисунок 2.3 – График после изменения толщины линии и размера символов

Изменим тип графика и воспользуемся параметром для изменения цвета.



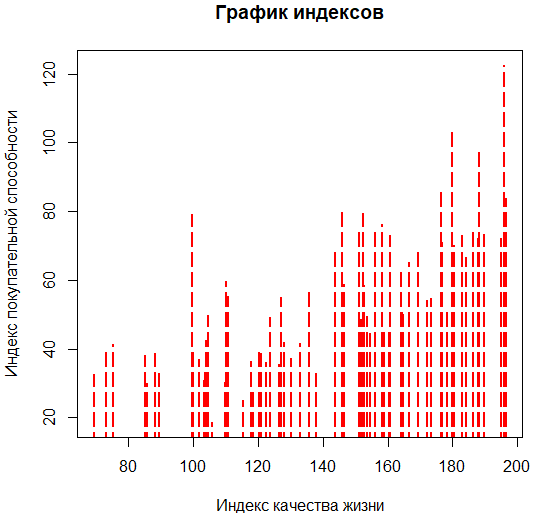


Рисунок 2.4 – График после изменения параметра цвета и добавления подписей

1. Воспользуемся корреляционным анализом:

- корреляция Пирсона

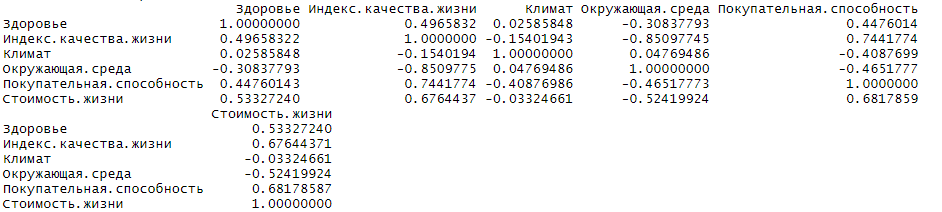


Рисунок 2.5 – Матрица корреляций Пирсона

- корреляция Спирмена

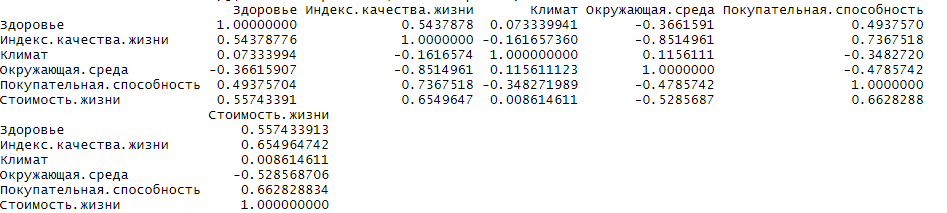


Рисунок 2.6 – Матрица корреляций Спирмена

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

1) индекс качества жизни – окружающая среда (-0.8514)

2) индекс качества жизни – покупательная способность (0.7367)

3) стоимость жизни – покупательная способность (0.6628)

Самые слабые связи:

1) стоимость жизни – климат (0.0086)

2) здоровье – климат (0.0733)

3) окружающая среда – климат (0.1156)

1. Найдем уровень значимости коэффициентов корреляции для нескольких пар.

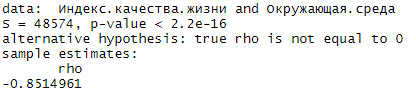


Рисунок 2.7 – Корреляционный тест для пары индекс качества жизни

и окружающая среда

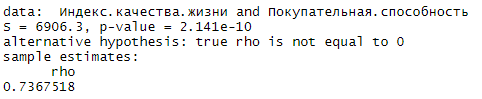


Рисунок 2.8 – Корреляционный тест для пары индекс качества жизни

и покупательная способность

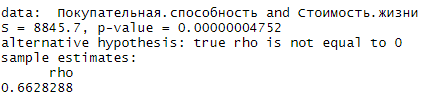


Рисунок 2.9 – Корреляционный тест для пары покупательная способность

и стоимость жизни

1. Построим матрицу точечных графиков для этих же переменных.

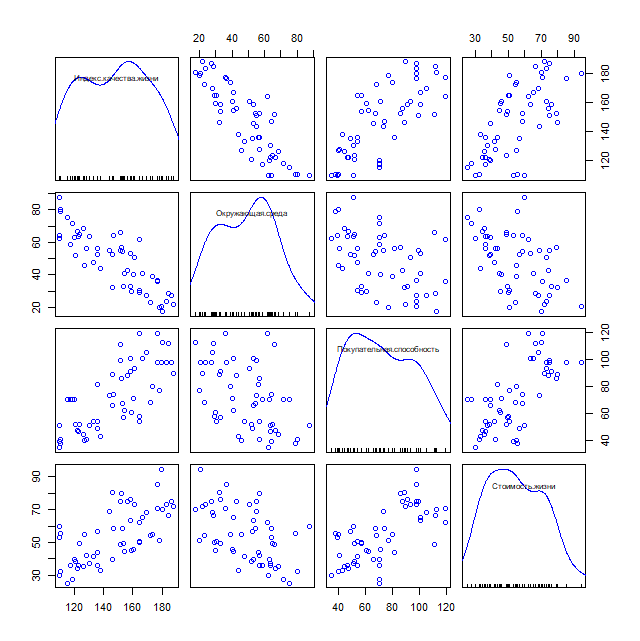


Рисунок 2.10 – Матрица точечных графиков

1. Построим уравнение зависимости индекса качества жизни от окружающей среды:

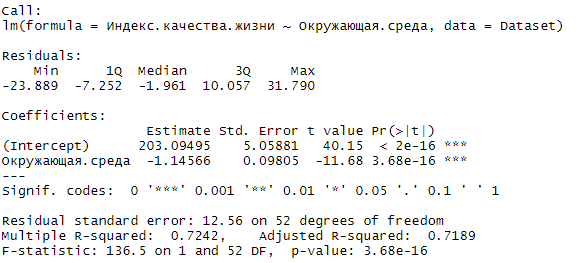


Рисунок 2.11 – Зависимость индекса качества жизни от окружающей среды

Построим график остатков.

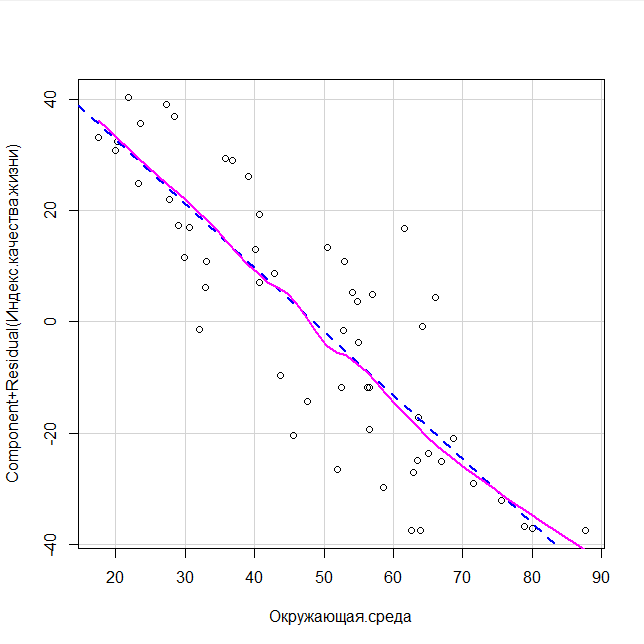


Рисунок 2.12 – График остатков

1. Построим уравнение зависимости индекса качества жизни от покупательной способности:

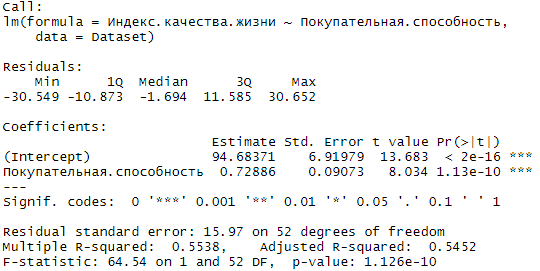


Рисунок 2.13 – Зависимость индекса качества жизни от покупательной способности

Построим график остатков.

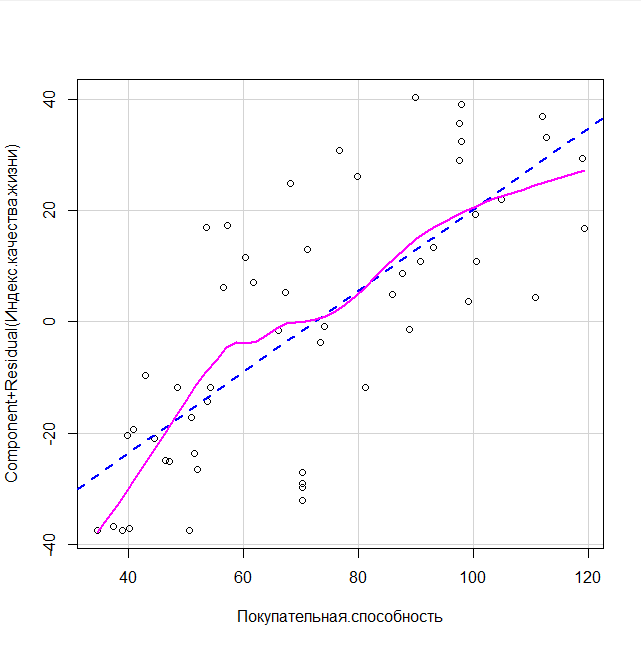


Рисунок 2.14 – График остатков

1. Построим уравнение зависимости стоимости жизни от покупательной способности:

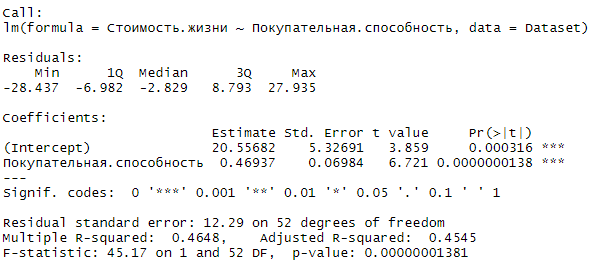


Рисунок 15 – Зависимость стоимости жизни от покупательной способности

Построим график остатков.

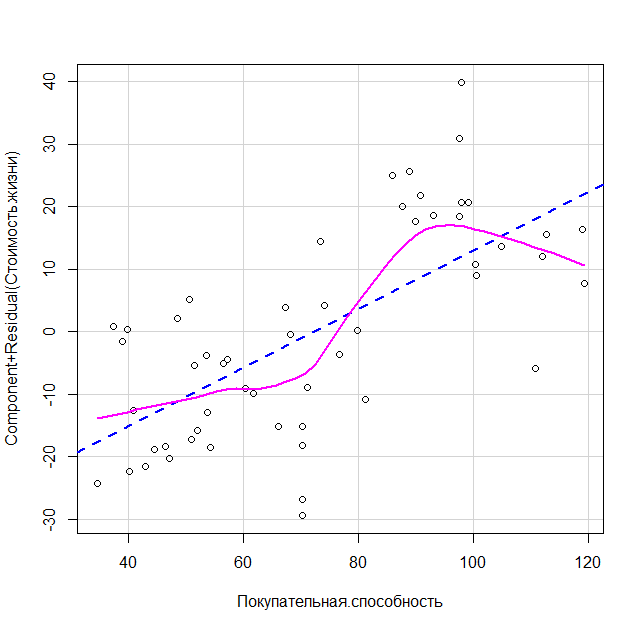


Рисунок 16– График остатков

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

Получим уравнение:

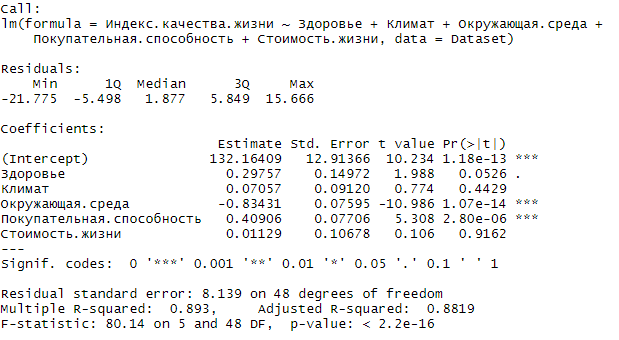


Рисунок 2.17 – Результат подгонки множественной регрессионной модели

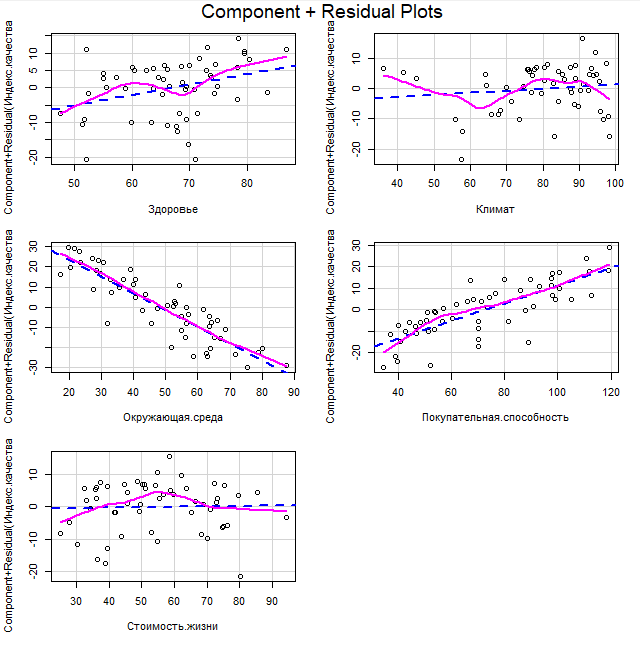


Рисунок 2.18 – Графики остатков

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

Получим уравнение для регрессии по направлению вперед (рисунок 19):

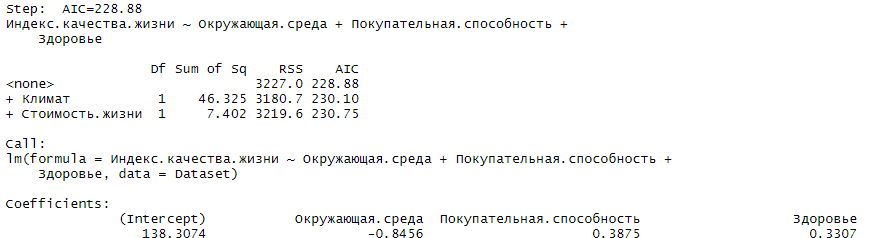


Рисунок 2.19 – Пошаговое построение регрессии по направлению вперед

Уравнение регрессии по направлению назад аналогично уравнению регрессии по направлению вперед (рисунок 20).

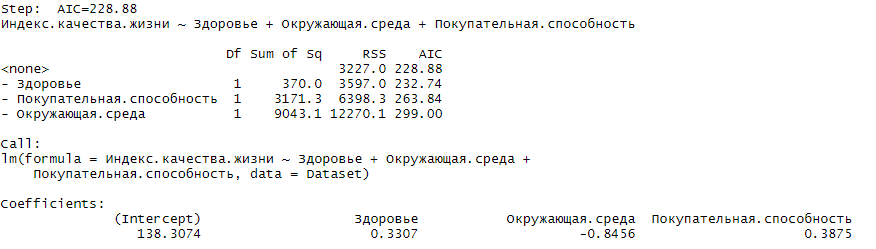


Рисунок 2.20 – Пошаговое построение регрессии по направлению назад

1. Проверим коэффициент VIF.

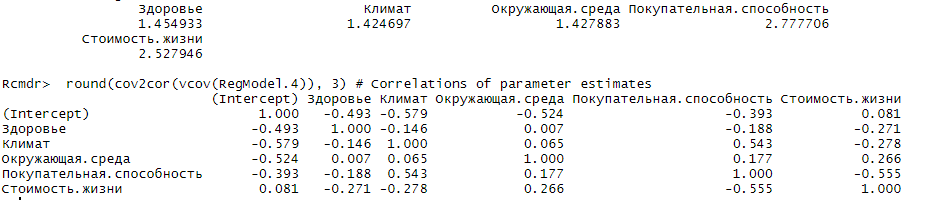


Рисунок 2.21 – Коэффициент VIF

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

3. 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 (-0.8514), индекс качества жизни – покупательная способность (0.7367), стоимость жизни – покупательная способность (0.6628)).

Была построена матрица точечных графиков для этих переменных, проведена подгонка множественной регрессионной модель при помощи функции lm() и построены графики остатков.

Было 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 228.88.

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