Лабораторная работа №2

«Корреляционный и регрессионный анализ данных»

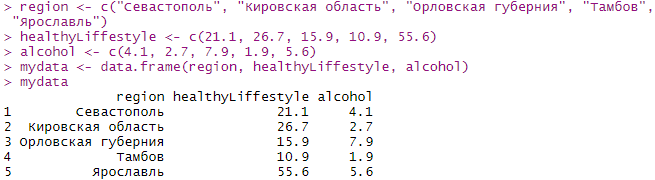
1 ЦЕЛЬ РАБОТЫ

– исследовать возможности языка R для проведения корреляционного и регрессионного анализа данных;

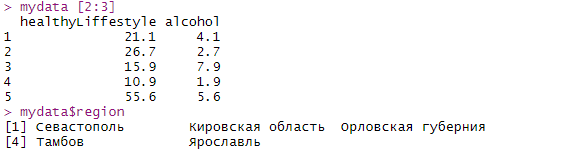
– создание набора данных для проведения корреляционного и регрессионного анализа данных.

2 ХОД РАБОТЫ

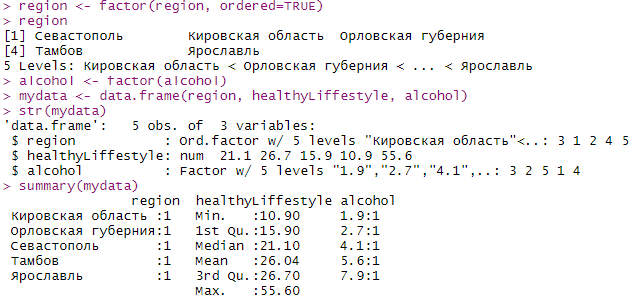
1. Создадим таблицу данных при помощи функции data.frame():



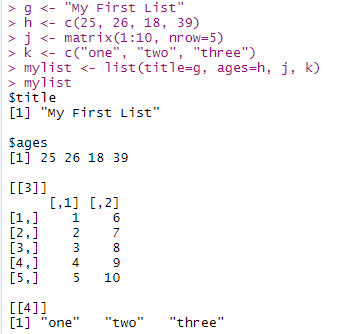
Выведем второй и третий столбцы таблицы данных и с помощью знака $ обозначим переменную в таблице данных:



С помощью функции factor() сохраним категориальные данные в виде вектора из целых чисел. Далее выведем информацию о таблице данных с помощью функции str() и, воспользовавшись функцией summary() обработаем все переменные, то есть вычислим минимум, максимум, среднее значение и квартили (числа, которые делят набор данных на четыре равные части или подсчитывают частоту встречаемости каждого значения).

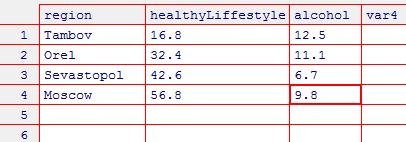


Создадим список, являющийся упорядоченным набором объектов и объединяющий различные объекты под одним именем:

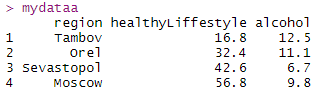


1. Произведём ввод данных с клавиатуры с помощью команды edit():





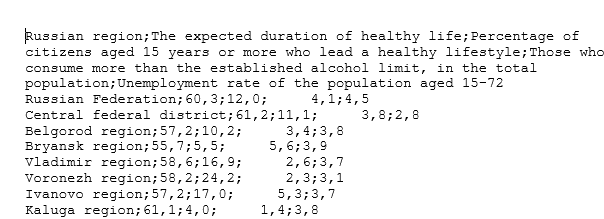
Проверяем, сохранились ли данные:



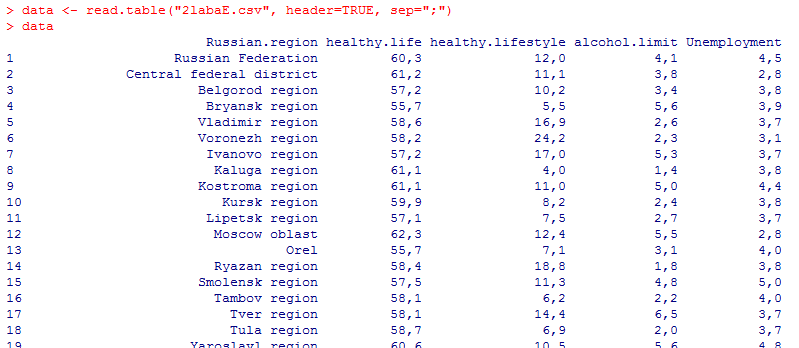
1. Произведём импорт данных из текстового файла с разделителями.

После формирования таблицы Excel с данными, которые будут подвергаться анализу, сохраним их в формате csv, то есть в формате текстового файла с разделителями.

Данные в файле выглядят следующим образом:

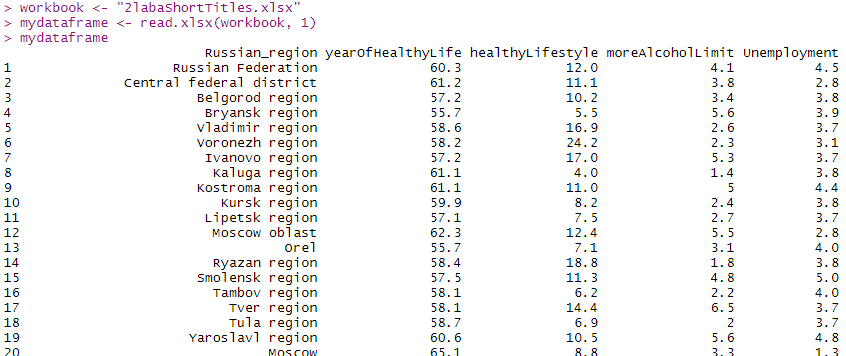


Функция read.table() поможет сохранить данные из текстового файла в виде таблицы:



1. Воспроизведём импорт данных из Excel.

Сохраним в переменную workbook путь к Excel файлу, затем с помощью команды read.xlsx(workbook, 1) импортируем первый лист книги:



1. Ответим на контрольные вопросы.

5.1 Функциональная связь (жестко детерминированная) – это такая связь, при которой заданному значению фактора X соответствует строго определенные значения параметра Y. Это свойственно строго детерминированным процессам, например, связь температуры и объема, давления и объема).

5.2 Статистической связью (стохастически детерминированной) называется такая связь, при которой причинная связь проявляется не в каждом отдельном случае, а в общем, среднем при большом числе наблюдений.

5.3 Корреляционная связь – это связь, при которой заданному значению фактора X может соответствовать множество возможных значений параметра Y. Частный случай статистической связи.

5.4 Корреляционный анализ – анализ, основной задачей которого является определение формы, направленности и тесноты взаимосвязи.

5.5 Корреляционное поле (или диаграмма рассеяния) является графической зависимостью между результатами измерений двух признаков. Для ее построения исходные данные наносят на график, отображая каждую пару значений (xi, yi) в виде точки с координатами xi и yi в прямоугольной системе координат.

5.6 Корреляционный анализ: форма зависимости. По форме взаимосвязи корреляционные зависимости принято разделять на линейные (см. рисунок 1) и нелинейные (см. рисунок 2).

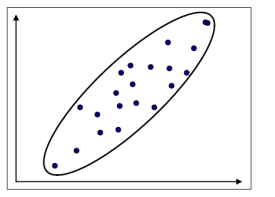


Рисунок 1 – Линейная статистическая связь

При линейной зависимости огибающая корреляционного поля близка к эллипсу. Линейная взаимосвязь двух случайных величин состоит в том, что при увеличении одной случайной величины другая случайная величина имеет тенденцию возрастать (или убывать) по линейному закону.

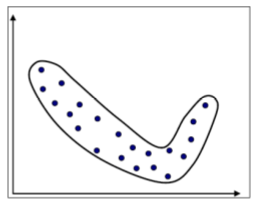


Рисунок 2 – Нелинейная статистическая связь

5.7 Корреляционный анализ: направленность взаимосвязи. Направленность взаимосвязи бывает отрицательная и положительная. Направленность отрицательная, если увеличение значения одного признака приводит к уменьшению значения второго (см. рисунок 3), положительная, если увеличение значения одного признака приводит к увеличению значения второго (см. рисунок 1).

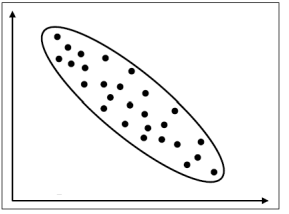


Рисунок 3 – Отрицательная направленность

5.8 Корреляционный анализ: теснота (сила) взаимосвязи. Теснота взаимосвязи может быть оценена качественно по ширине корреляционного поля – чем меньше его ширина, тем больше теснота и сильнее зависимость.

5.9 Коэффициент корреляции – это количественная оценка тесноты взаимосвязи двух случайных величин, характеризует только линейную взаимосвязь. Значения коэффициента корреляции находятся в диапазоне -1< r < 1. При r >0 связь оценивается, как прямая, при r < 0 – обратная. При r = 0 – связь отсутствует, при | r |=1 – связь функциональная.

Сила связи оценивается:

при | r |<0,3 – как слабая, при 0,3<[ r |<0,7 – умеренная, при | r |>0,7 – сильная.

5.10 Регрессионный анализ. Применяется, когда возникает необходимость аппроксимировать (описать приблизительно) диаграмму рассеяния математическим уравнением. То есть зависимость между переменными величинами Y и Х можно выразить аналитически с помощью формул и уравнений и графически в виде геометрического места точек в системе прямоугольных координат.

При помощи регрессионного анализа можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными.

5.11 Этапы регрессионного анализа.

1. Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений.

2. Определение зависимых и независимых (объясняющих) переменных.

3. Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель (гипотеза).

4. Формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная).

5. Определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии)

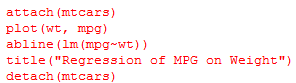
6. Оценка точности регрессионного анализа.

7. Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов.

8. Предсказание неизвестных значений зависимой переменной.

Лабораторная 2.2

1. Построим диаграмму рассеяния, на которой вес автомобиля отложен на горизонтальной оси, а расход топлива – на вертикальной (см. рисунок 4).



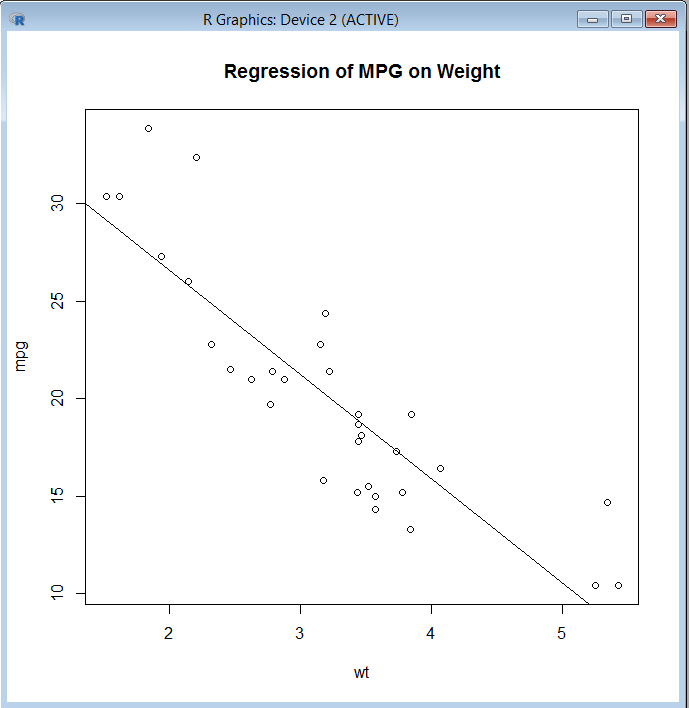


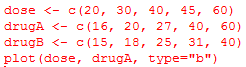
Рисунок 4 – Диаграмма рассеяния с регрессионной прямой

С помощью команды jpeg(“mygraph.pdf”) сохраним диаграмму в отдельный файл с именем “mygraph.pdf”.

Таким образом, в указанную по дефолту директорию сохранится данная диаграмма:



Обозначим набор данных, содержащий дозировку и реакцию на два вида лекарств. Построим простой линейный график, изображающий зависимость реакции пациента от дозы лекарства A (см. рисунок 5):



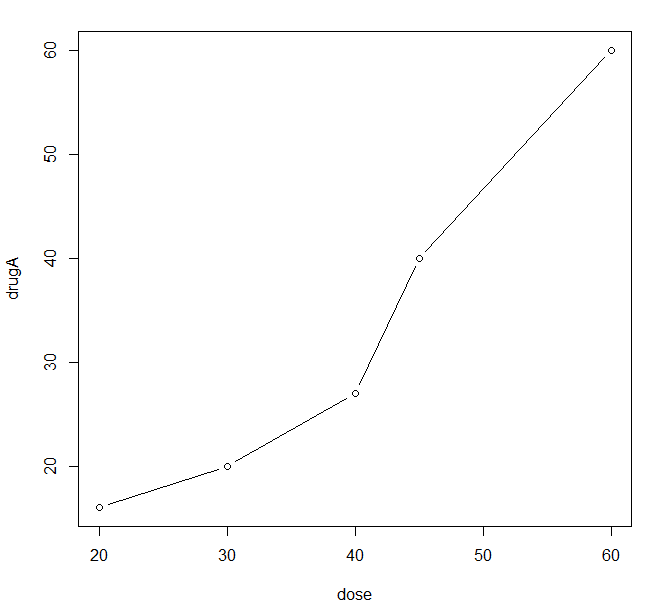


Рисунок 5 – Линейный график из точек и линий

С помощью функции par() можно изменять различные параметры графиков. Изменим тип линии со сплошной на пунктирный и тип символа звёздочку вместо кругов (см. рисунок 6).





Рисунок 6 – Видоизменённый по параметрам график

Также изменения параметров можно произвести одной командой:



График данного вида представлен на рисунке 7:



Рисунок 7 – График с объединёнными параметрами внутри одной команды

Создадим вектор из оттенков серого с помощью функции gray(0:20/20), которая создаст 20 оттенков (см. рисунок 8):





Рисунок 8 – Созданные оттенки серого цвета

Изменим параметры диаграммы с помощью параметров pin=с(4,3) и mai=с(1,.5, 1, .2)), позволяющие соответственно задать размер 4 дюйма в ширину и 3 дюйма в высоту с шириной полей сверху и снизу по одному дюйму, слева 0.5 дюйма и справа 0.2 дюйма (см. рисунок 9):

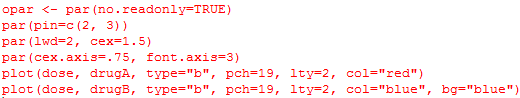




Рисунок 9 – Линейный график

Разместим на диаграмме заголовок (main), подзаголовок (sub) и подписи осей (xlab, ylab) с диапазонами значений на осях в параметрах (xlim, ylim) (см. рисунок 10):

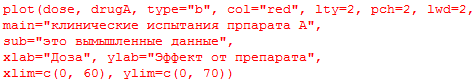




Рисунок 10 – Линейный график зависимости с заголовком, подзаголовком и модифицированными осями

1. Проанализируем код:

x <- c(1:10) – Создать вектор, упорядоченный от 1 до 10;

y <- x – присвоить вектору y вектор х;

z <- 10/x –присвоить вектору z вектор x, каждое значение которого было делителем числа 10;

opar <- par(no.readonly=TRUE) – вывод на экран графических параметров, которые можно изменять;

par(mar=c(5, 4, 4, 8) + 0.1) – вектор, задающий размеры полей: низ – 5, лево – 4, верх – 4 право – 8;

plot(x, y, type=”b”, pch=21, col=”red”, yaxt=”n”, lty=3, ann=FALSE) – изменить параметры графиков: на графике должны быть показаны точки и линии, с типом символа кругом, красного цвета, отменить создание у-осей, с типом линии – частыми точками, без вывода надписи и подписи на диаграмме;

lines(x, z, type=”b”, pch=22, col=”blue”, lty=2) – установить параметры линии графика: на графике должны быть показаны точки и линии, с типом символа полый квадрат, синего цвета и типом линии – пунктир;

axis(2, at=x, labels=x, col.axis=”red”, las=2) – изменить параметры осей: с левой стороны диаграммы рисовать ось, положение делений на осях = х, подписи под делениями осей – х, цвет значений осей – красный, положение подписей делений по отношению к оси – перпендикулярно;

axis(4, at=z, labels=round(z, digits=2), col.axis=”blue”, las=2, cex.axis=0.7, tck=-.01) – изменить параметры осей: с правой стороны диаграммы рисовать ось, положение делений на осях = z, подписи под делениями осей – округлённые до 2 значений после запятой элементы вектора z, цвет значений осей – синий, положение подписей делений по отношению к оси – перпендикулярно, масштабирование значений на осях 0,7 относительно размера по умолчанию, положение деления оси снаружи от рамки диаграммы;

mtext(”y=1/x”, side=4, line=3, cex.lab=1, las=2, col=”blue”) – записать в правом поле рисунка y=1/x, размер названий осей такой же, как заданный по умолчанию, положение подписей делений по отношению к оси – перпендикулярно, синего цвета;

title(“Пример осей”, xlab=”значение переменной X”, ylab=”Y=X”) – задать заголовок графика, подписать оси х и у;

par(opar) – восстановить исходные значения параметров.



Рисунок 11 – Результат выполнения программы

1. Изучим добавление легенды к диаграмме. Легенда диаграммы – это заголовки рядов, которые по умолчанию располагаются в правой части этой самой диаграммы.

Легенда добавляется с помощью команды legend.text (см. рисунок 12).

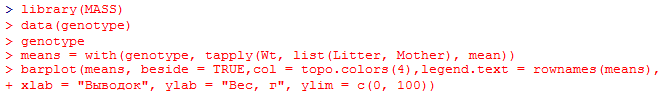




Рисунок 12 – Добавление легенды к диаграмме

1. По экспериментальным данным построим диаграммы.

С помощью пакета Rcmdr построим линейный график зависимости, взяв за х – долю граждан, ведущих здоровый образ жизни, а за у – уровень безработицы. Предварительно по значению х данные были отсортированы по возрастанию.

Изменим параметры графика, заданные по умолчанию:



График представлен на рисунке 13.



Рисунок 13 – Линейный график зависимости

Построим полосной график. Где по оси х будут регионы России, а по оси у – ожидаемая продолжительность здоровой жизни (см. рисунок 14).



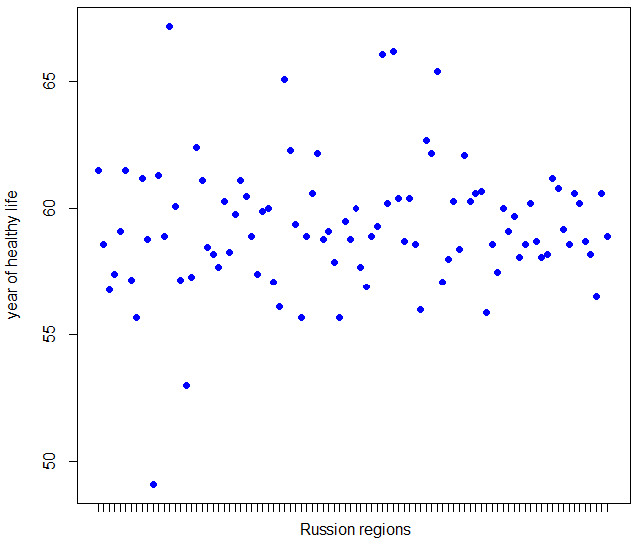


Рисунок 14 – Полосной график

Построим гистограмму, отображающую количество значений в столбце «Употребление алкоголя выше установленной нормы» (см. рисунок 15).





Рисунок 15 – Гистограмма

1. Ответим на контрольные вопросы.
   1. Принцип построения диаграмм в языке R. В стандартной интерактивной сессии создается диаграмма, вводом по одной команде и добавлением элементов диаграммы, пока не получится то, что необходимо.
   2. Функции изменения графических параметров диаграмм. C помощью функции par() без параметров можно узнать действующие значения графических параметров, которые можно изменить. На рисунке 16 представлен список некоторых функций изменения графических параметров диаграмм.

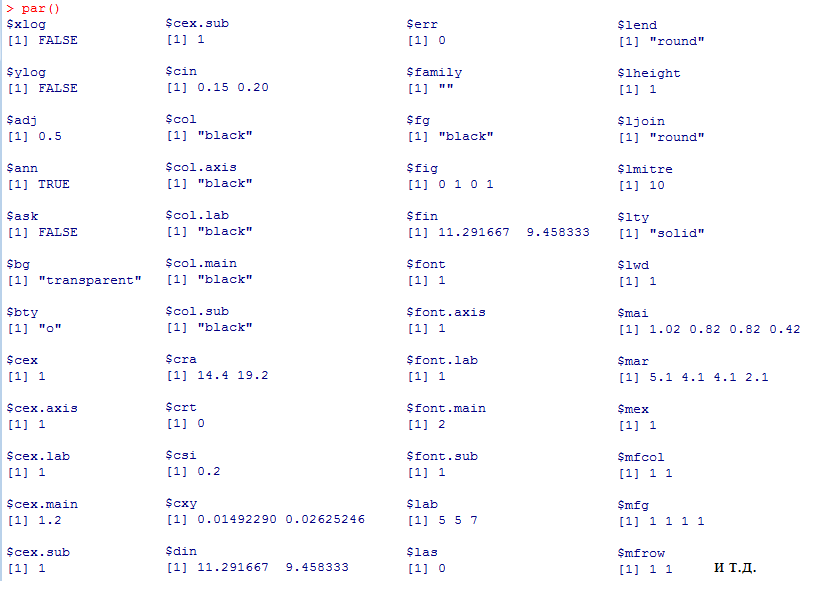


Рисунок 16 – Функции изменения графических параметров диаграмм языка R

Лабораторная работа 2.3

«Корреляционный и регрессионный анализ данных. Исследование тесноты взаимосвязей данных в среде R»

1. Загрузим экспериментальные данные для анализа в таблицу данных оболочки Rcmdr.
2. Исследуем элементы таблицы данных с помощью корреляционных матриц и корреляционных тестов по методам Пирсона (см. рисунок 17) и Спирмена (см. рисунок 18).

Для удобства напишем значения переменных:

healthyLifestyle – доля граждан, ведущих ЗОЖ;

moreAlcoholLimit – доля граждан, употребляющих алкоголь больше установленной нормы;

Unemployment – уровень безработицы;

yearOfHealthyLife – ожидаемая продолжительность здоровой жизни.

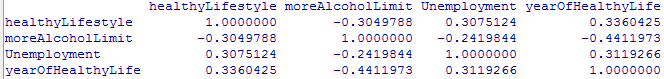


Рисунок 17 – Коэффициенты корреляции Пирсона

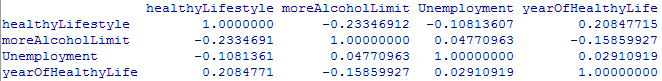
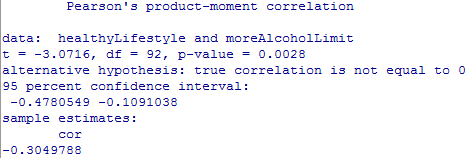


Рисунок 18 – Коэффициенты корреляции Спирмена

Коэффициенты корреляции в данных матрицах рассчитываются с помощью функции cor(). В первом случае рассчитывался коэффициент корреляции Пирсона, основанный на том, что обе анализируемые переменные распределены нормально и имеют линейную связь. А для ненормально распределенных переменных, а также при наличии нелинейной связи между ними, рассчитывался непараметрический коэффициент корреляции Спирмена. В отличие от коэффициента Пирсона, этот вариант коэффициента корреляции работает не с исходными значениями переменных, а с их рангами.

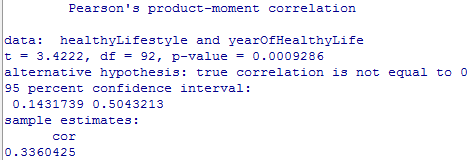
Проанализировав матрицы, можно сделать вывод, что связь между healthyLifestyle, Unemployment и yearOfHealthyLife прямая, а сила связи – умеренная. Связь между healthyLifestyle с moreAlcoholLimit и yearOfHealthyLife с moreAlcoholLimit обратная, а сила связи умеренная. А связь между Unemployment и moreAlcoholLimit также обратная и по силе связи слабая.

Выполним оценку статистической значимости коэффициентов корреляции с помощью функции cor.test(), которая вдобавок проверяет нулевую гипотезу о равенстве коэффициента нулю:

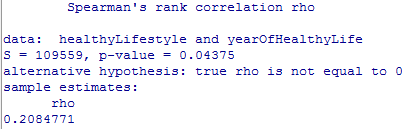


Рассчитанный коэффициент корреляции Пирсона, показывающий связь между здоровым образом жизни и употреблением алкоголя больше нормы, оказался равен -0,3049, что является не очень высоким показателем, но коэффициент статистически значимо отличается от нуля p-value = 0,0028.

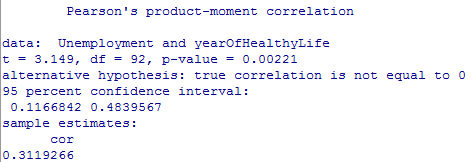
Результаты корреляционного теста для пары ЗОЖ и ожидаемая продолжительность здоровой жизни при коэффициенте Пирсона:



Корреляционный тест для рассмотренной выше пары коэффициента Спирмена:



Результаты корреляционного теста для пары ожидаемой продолжительности жизни и уровня безработицы при коэффициенте Пирсона:



1. Проиллюстрируем полученные результаты с помощью матрицы точечных графиков (см. рисунок 19).

С помощью данной диаграммы можно определить потенциальные взаимосвязи между количественными переменными.



Все переменные, указанные справа от знака “~”, показываются на графике. На рисунке 19 мы имеем 4 переменные с графиком взаимосвязи для каждой пары. Матрица содержит 12 графиков: 6 в верхней панели и 6 в нижней панели от диагонали с названиями переменных.

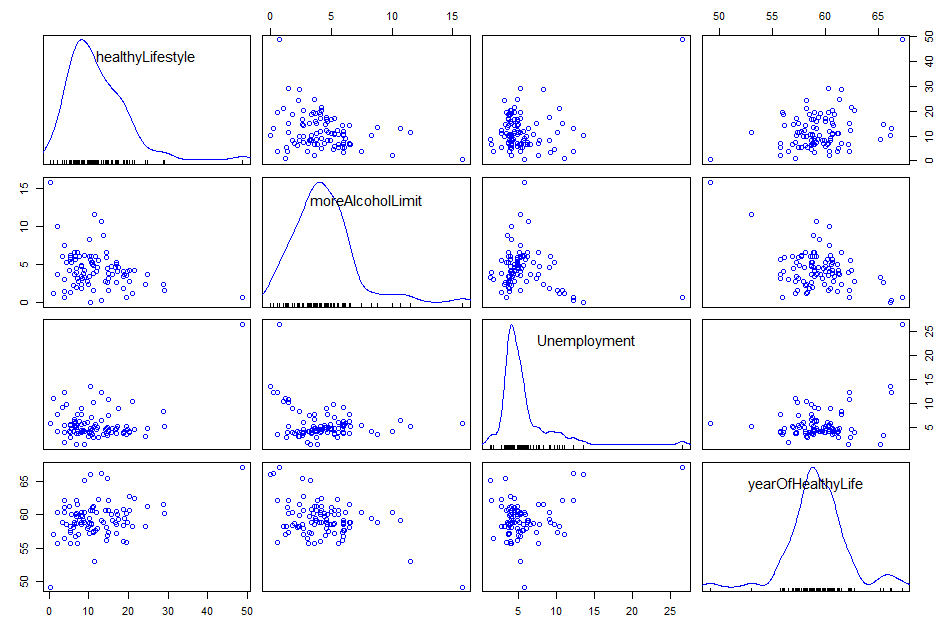


Рисунок 19 – Матрица точечных графиков

Так как панели идентичны, отключим для лучшего восприятия верхнюю панель (см. рисунок 20).



Рисунок 20 – Матрица точечных графиков без верхней панели

1. Используем функцию подгонки моделей для линейной регрессии и построим график остатков для двух элементов таблицы данных (см. рисунок 21).

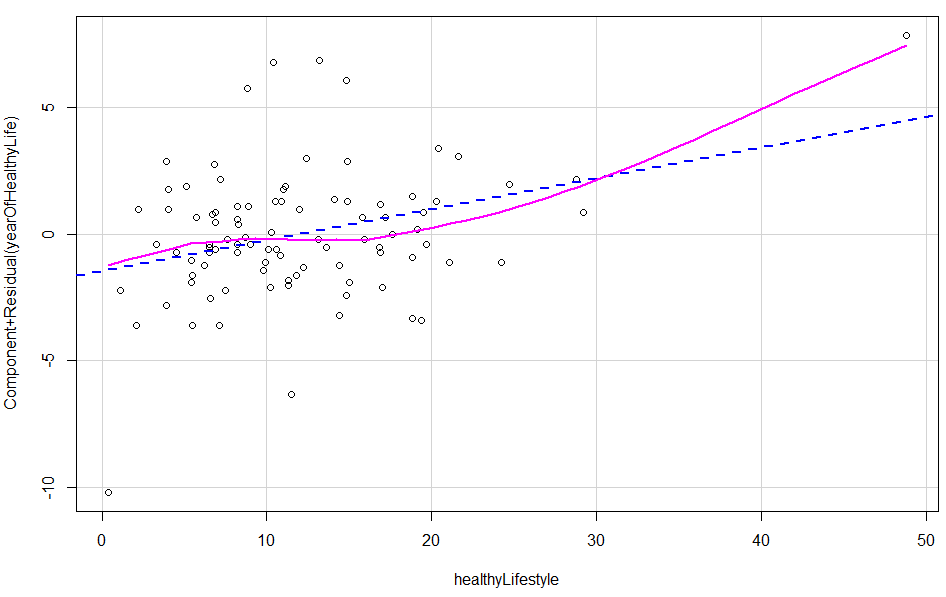


Рисунок 21 – График линейной регрессии

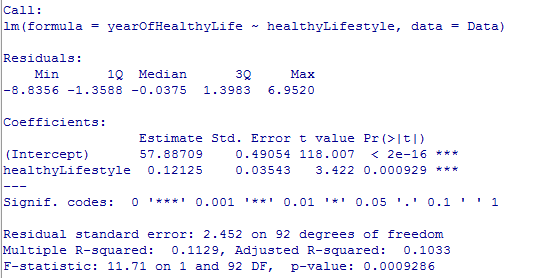
Из рисунка 21 видно, что аппроксимирующая и сглаживающая линии не совпадают.

Код для построения рисунка 21:



Команда lm(formula или “~”) – основная функция в R для подгонки регрессионных моделей. Формула означает, что значение y (ожидаемая продолжительность здоровой жизни) мы будем предсказывать по значению  x (доля граждан, ведущих ЗОЖ).

Результат функции crPlots():



Составляя уравнение регрессии данной модели, опишем её в виде уравнения (1):

(1)

Коэффициенты k и b – минимизирующие величину ошибки, соответственно равны 0,12125 и 57,88709.

Подставим коэффициенты в уравнение регрессии:

yearOfHealthyLife = 0,12125\*healthyLifesttyle + 57,88709

Далее проверим, насколько точно наша модель описывает данные. Для этого используется коэффициент R-squared. Чем ближе величина этих значений к 1, тем лучше. Единица – это идеальный результат, означающий, что модель на 100% описывает данные. Но в нашей модели результат 0,1129, следовательно, только на 10% модель описывает данные.

Чтобы проверить, насколько предсказываемая величина зависит от предикторов (исходных данных), смотрим на значение p-value. В нашем случае оно равно 0,0009286, то есть мы можем быть уверены на 99,90714%, что предсказываемая величина действительно зависит от предикторов.

Произведём такой же анализ для модели, где за X(исходные данные) возьмём долю безработных, а за Y(предсказываемая величина) – долю граждан, ведущих ЗОЖ. Остаточный график для модели изображён на рисунке 22.

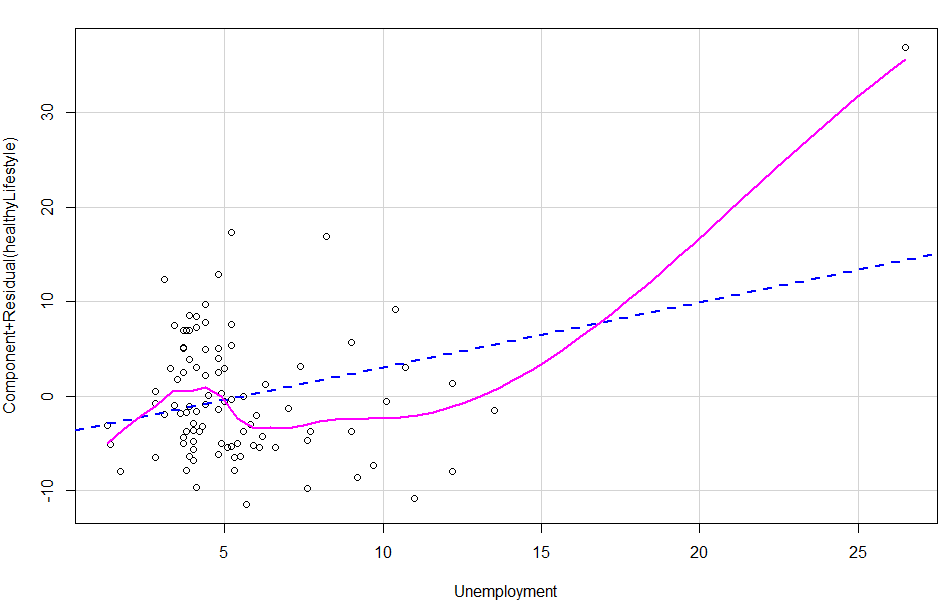
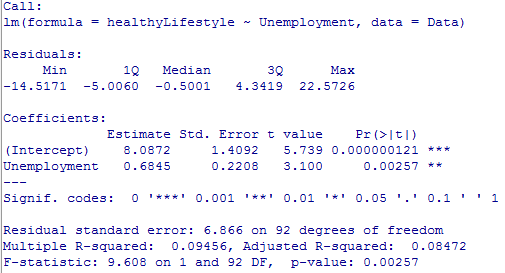


Рисунок 22 – График линейной регрессии для Unemployment и healthyLifestyle

Результат функции crPlots():



Составим уравнение регрессии данной модели:

healthyLifestyle = 0,6845\*Unemployment + 8,0872

По коэффициенту R-squared определим, что только на 9% модель описывает данные.

Далее смотрим на значение p-value. В нашем случае оно равно 0,00257, то есть мы можем быть уверены на 99,7%, что предсказываемая величина действительно зависит от исходных данных. Следовательно, чем p-значение меньше, тем лучше, поскольку при этом увеличивается «сила» отклонения нулевой гипотезы и увеличивается ожидаемая значимость результата.

Произведём ещё один анализ для модели, где за X(исходные данные) возьмём долю безработных, а за Y(предсказываемая величина) – долю граждан, употребляющих алкоголь больше установленной нормы. Остаточный график для модели изображён на рисунке 23.

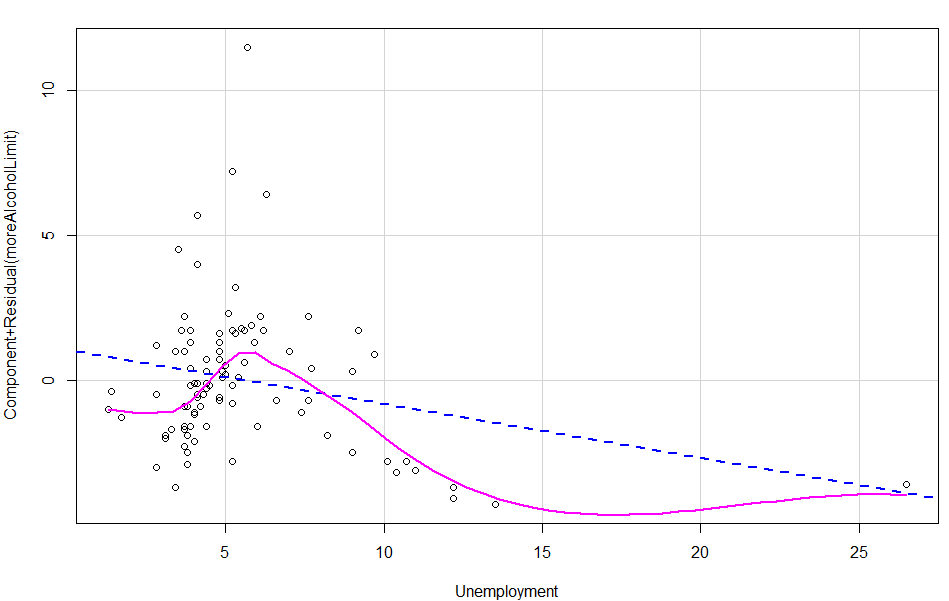
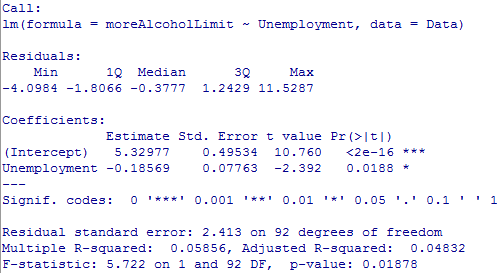


Рисунок 23 – График линейной регрессии для Unemployment и moreAlcoholLimit

Результат функции crPlots():



Составим уравнение регрессии данной модели:

moreAlcoholLimit = -0,18569\*Unemployment + 5,32977

По коэффициенту R-squared определим, что только на 5% модель описывает данные.

Далее смотрим на значение p-value. В нашем случае оно равно 0,01878, то есть предсказываемая величина значительно зависит от исходных данных.

ВЫВОДЫ

В данной части лабораторной работы были исследованы возможности языка R для определения тесноты взаимосвязей экспериментальных данных. С помощью кнопочного интерфейса пакета Rcmdr были рассчитаны коэффициенты корреляции Пирсона, Спирмена, была оценена статистическая значимость коэффициентов корреляции, полученные результаты были проанализированы с помощью построенных матрицы точечных графиков и графика линейной регрессии.

Лабораторная работа 2.4

«Корреляционный и регрессионный анализ данных. Множественная линейная регрессия»

1. Загрузим в оболочку Rcmdr свои экспериментальные данные.
2. Построим множественную линейную регрессию (см. рисунок 24)

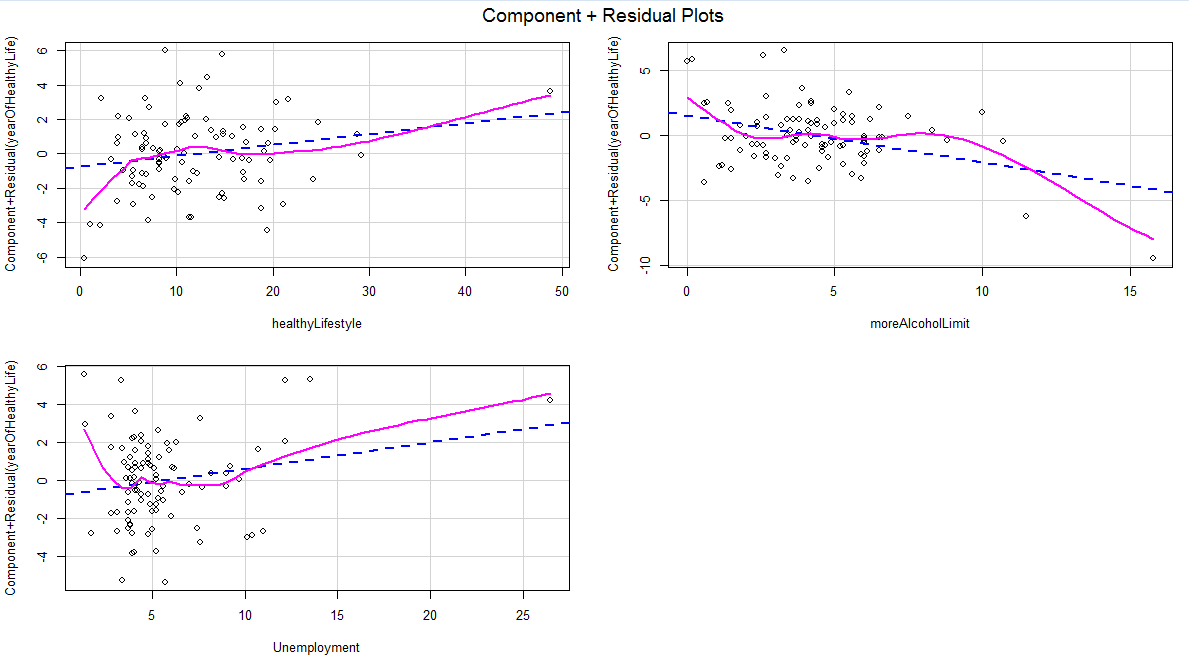
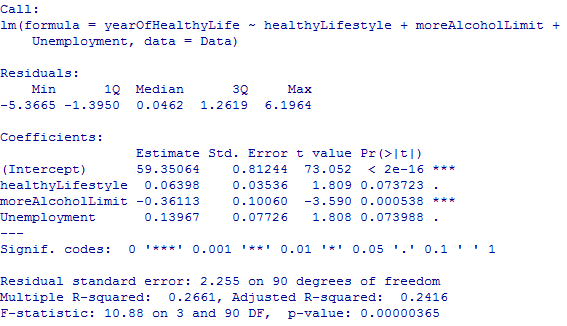


Рисунок 24 – Множественная линейная регрессия

За X(исходные данные) возьмём долю безработных, долю ведущих ЗОЖ и употребляющих алкоголь больше установленной нормы а за Y(предсказываемая величина) – ожидаемую продолжительность здоровой жизни.



По таблице коэффициентов запишем полученное уравнение регрессии:

yearOfHealthyLife = 0,06398 \* healthyLifestyle – 0,36113\* moreAlcoholLimit + 0,13967\*Unemployment +59,35064

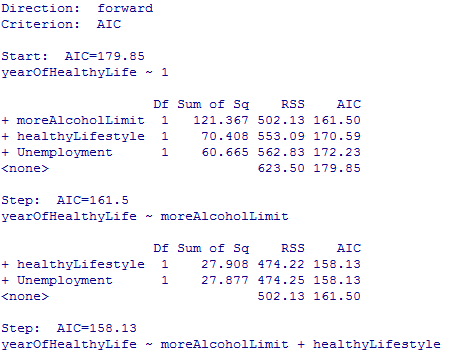
По коэффициенту R-squared определим, что модель описывает данные на 20%.

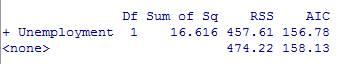
Далее смотрим на значение p-value. В нашем случае оно равно 0,00000365, то есть предсказываемая величина значительно зависит от исходных данных.

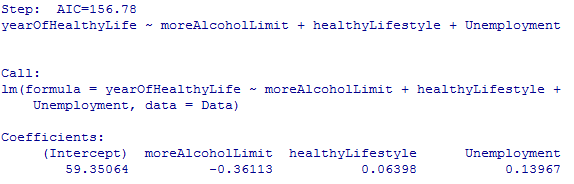
1. Произведём пошаговую линейную регрессию с нашими экспериментальными данными.

Для начала выберем прямую пошаговую регрессию (метод прямого включения), заключающейся в том, что первоначально строится модель с одной экзогенной переменной, затем добавляется следующая и строится новая модель.

С помощью функции stepwise() производятся следующие вычисления:



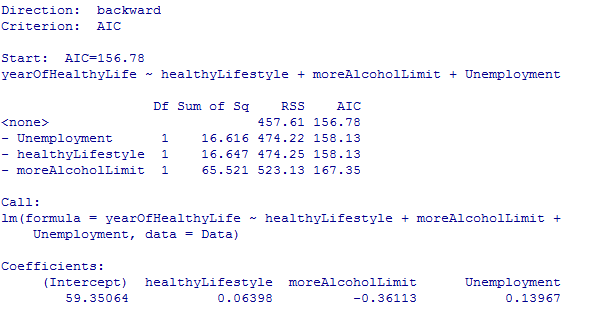




Запишем уравнение оптимальной модели:

yearOfHealthyLife = 59,35064 – 0,36113\* moreAlcoholLimit + 0,06398 \* healthyLifestyle + 0,13967\*Unemployment

Аналогично сделаем пошаговое построение по направлению назад, заключающееся в исключении переменной. Вначале все предикаты входят в уравнение регрессии. Затем по очереди выводятся из уравнения исходя из их соответствия критерию.

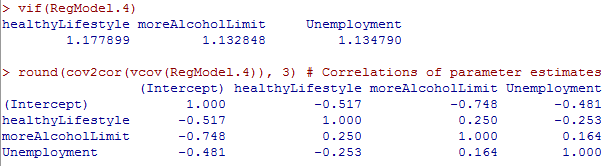


Запишем уравнение оптимальной модели:

yearOfHealthyLife = 59,35064 + 0,06398 \* healthyLifestyle – 0,36113\* moreAlcoholLimit + 0,13967\*Unemployment

В обоих случаях коэффициенты пошаговой множественной регрессии получились такие же, как и в предыдущем пункте.

Проверим коэффициенты VIF (фактор инфляции дисперсии), характеризующий силу мультиколлинеарности, то есть чем он выше для *j*-го предиктора, тем сильнее линейная связь между этим и остальными предикторами.



Проанализировав данные коэффициенты, можно сделать вывод, что они невысокие, то есть значения <10.

1. Ответим на контрольные вопросы.
   1. Множественная регрессия. Применяется, если существует больше одной независимой переменной. позволяет изучить совместное воздействие нескольких независимых переменных на переменную отклика. Уравнение множественной регрессии в общем случае:



Практическое применение двоякое: для предсказания переменной отклика и для определения интенсивности, с которой каждая независимая переменная линейно связана с зависимой.

* 1. Пошаговая множественная регрессия. Это методика построения множественной регрессии, бывает двух видов: пошаговая вперёд и пошаговая назад.

Пошаговая вперед заключается в том, что первоначально строится модель с одной экзогенной переменной. Затем добавляется следующая и строится новая модель. Модели сравниваются и, в зависимости от того ухудшилась или улучшилась модель, введенная переменная либо остается в модели, либо заменяется на другую. Таким образом, перебираются различные комбинации экзогенных переменных, в результате получается наилучшая модель.

По направлению назад заключается в исключении переменной. Вначале все предикаты входят в уравнение регрессии. Затем по очереди выводятся из уравнения исходя из их соответствия критерию.

* 1. Мультиколлинеарностью для линейной множественной регрессии называется наличие линейной зависимости между факторными переменными, включёнными в модель.

Мультиколлинеарность – коррелированность объясняющих переменных между собой. Считается, что явление мультиколлинеарности наблюдается тогда, когда коэффициент корреляции между объясняющими переменными превышает по модулю 0,7.

ВЫВОДЫ

В данной лабораторной работе были исследованы возможности языка R для построения множественной линейной регрессии. Были проанализированы экспериментальные данные на мультиколлинеарность, построено уравнение множественной регрессии, а также произведена пошаговая множественная регрессия с направлением вперёд и назад.