МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«СЕВАСТОПОЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт информационных технологий

Кафедра «Информационные системы»

**ОТЧЕТ**

о выполнении лабораторной работы №2

по дисциплине “Интеллектуальный анализ данных”

Выполнил: ст. гр. ИС/б-20-2-о

Белик Г. М.

Проверил: ст. преп.

Сырых О. А.

Севастополь

2023

**ЛАБОРАТОРНАЯ РАБОТА №2**

**«Корреляционный и регрессионный анализ данных»**

**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

**Ход работы**

После запуска Rstudio исследуем функции языка R для работы с типами данных.

patientID <- c(1, 2, 3, 4, 5, 6)

age <- c(25, 34, 28, 52, 37, 18)

diabetes <- c("Type1", "Type2", "Type1", "Type2", "Type1", "Type2")

status <- c("Poor", "Improved", "Excellent", "Poor","Good", "Normal")

patientdata <- data.frame(patientID, age, diabetes, status)

patientdata

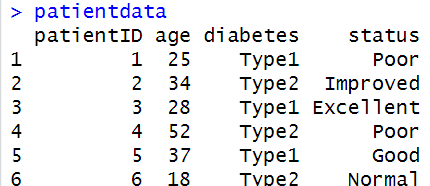


Рисунок 1 – Создание таблицы данных

a <- "My First List"

b <- c(25, 26, 18, 39)

c <- matrix(1:10, nrow=5)

d <- c("one", "two", "three")

mylist <- list(title=a, ages=b, c, d)

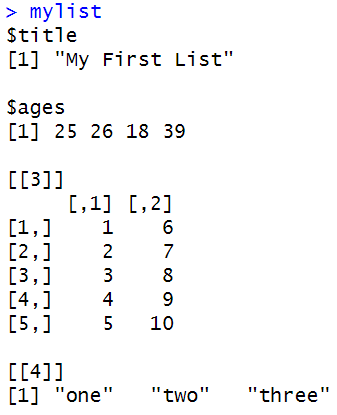


Рисунок 2 – Создание списка

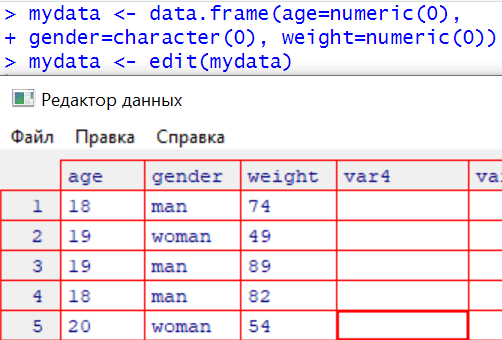


Рисунок 3 – Ввод данных с клавиатуры

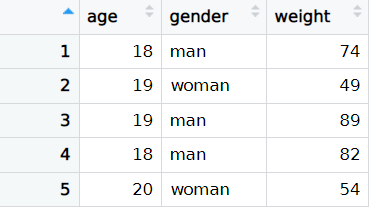


Рисунок 4 – Таблица, полученная с помощью edit()

В качестве исходных данных стала таблица замеров динамических характеристик автомобилей с указанием марки, модели авто, объема двигателя, мощности, крутящего момента, массы и разгона 0-100 км/ч.

Данные с изображения были скопированы в excel таблицу.

Далее был произведен импорт данных (таблицы) из Excel путем экспорта её в папку "C:/Users/Agold/OneDrive/Документы", полученную с помощью команды getwd() в текстовом формате .txt с раздлением /t табуляцией. Затем данные были импортированы из текстового документа в Rstudio.

Название таблицы в Rstudio “datatable”, первая строка – названия переменных, разделение элементов – таблуляция, название строк – столбец Car.

Исходными данными ослужила таблица с инфорацией о моделях автомобиля, объемом двигателя, мощностью, крутящим моментом, массой и разгоном.

****

Рисунок 5 – Таблица замеров автомобилей

Команда:

datatable <- read.table("datatable.txt",header = TRUE, sep = "\t",row.names = "Car")

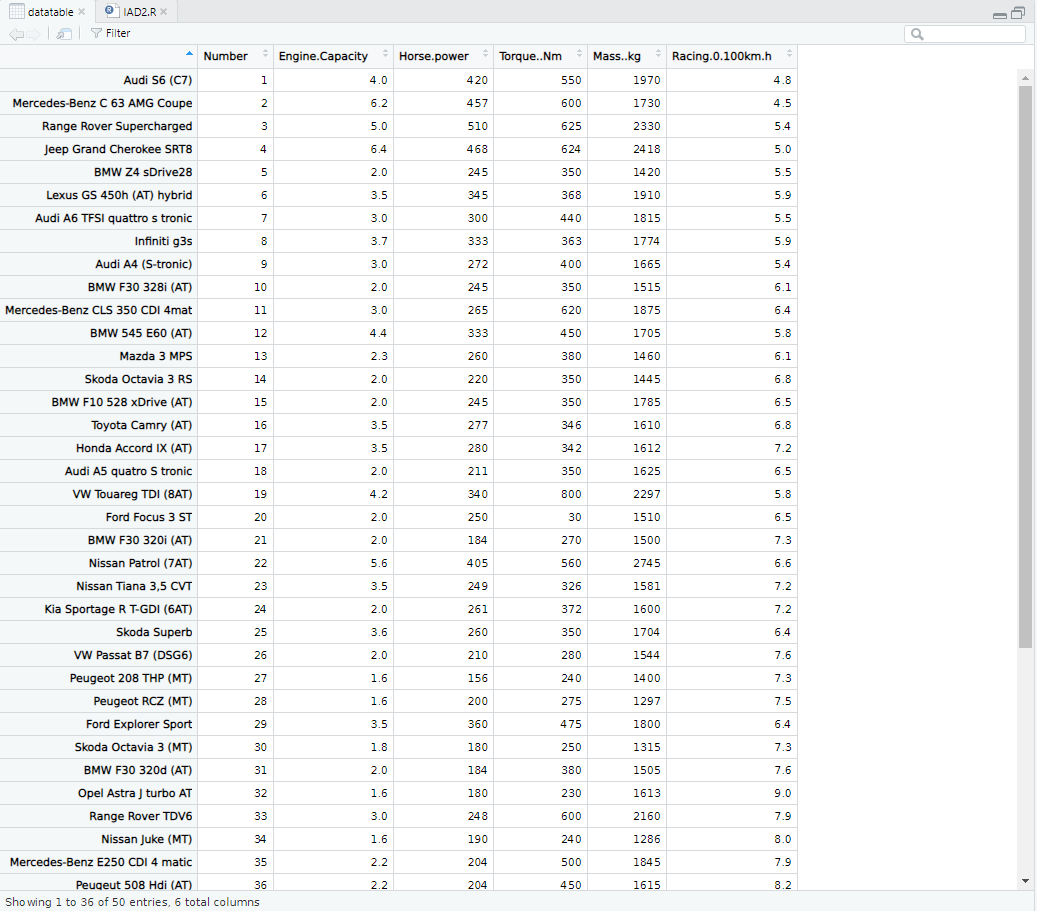


Рисунок 6 – Таблица datatable в Rstudio

При попытке импортировать таблицу из файла типа .xls с помощью библиотеки RODBC возникает ошибка.

library(RODBC)

channel <- odbcConnectExcel("database.xls")

mydataframe <- sqlFetch(channel, "sheet1")

odbcClose(channel)



Рисунок 7 – Ошибка при использовании library(ROBDC)

При попытке импортировать таблицу из файла типа .xlsx с помощью библиотеки xlsx также возникает ошибка.

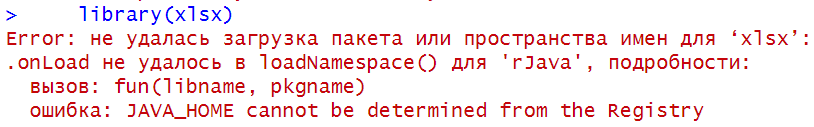


Рисунок 8 – Ошибка при использовании library(xlsx)

Затем построим диаграмму с указанием таблицы данных mtacrs, wt по горизонтальной оси, mpg по вертикальной, регрессионной прямой, названия.

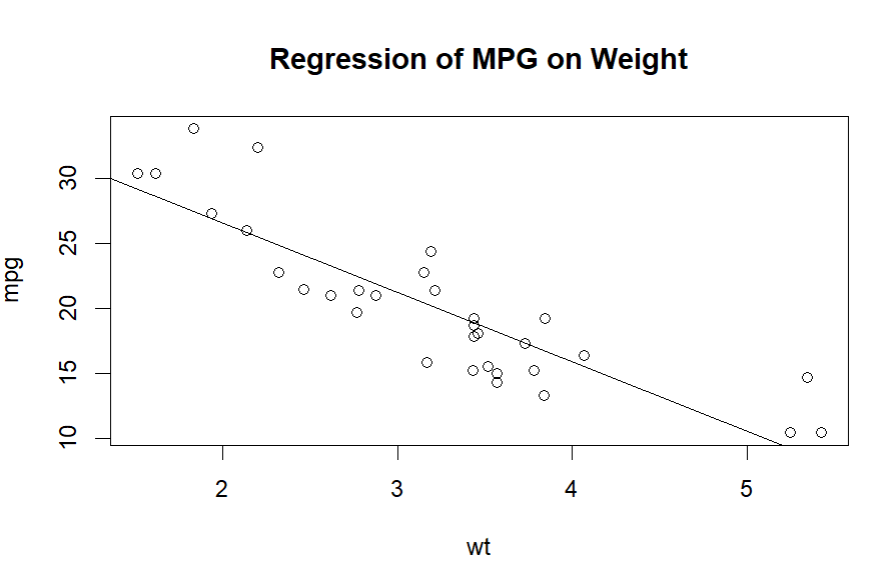
attach(mtcars)

plot(wt, mpg)

abline(lm(mpg~wt))

title("Regression of MPG on Weight")

detach(mtcars)

 Рисунок 9 – Диаграмма массы и расхода топлива автомобиля

С помощью следующего набора команд сохраним диаграмму в pdf файл.

pdf("mygraph.pdf")

attach(mtcars)

plot(wt, mpg)

abline(lm(mpg~wt))

title(" MPG on Weight")

detach(mtcars)

dev.off()

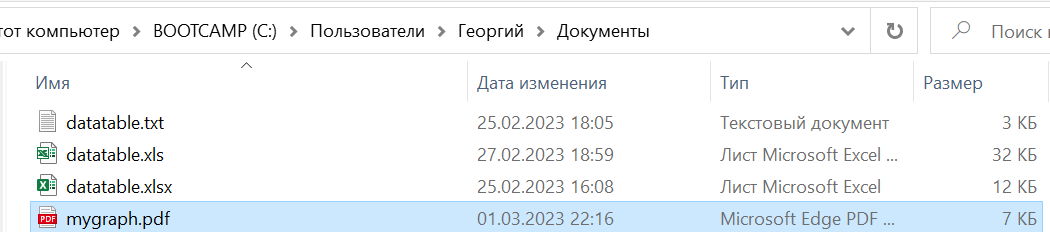
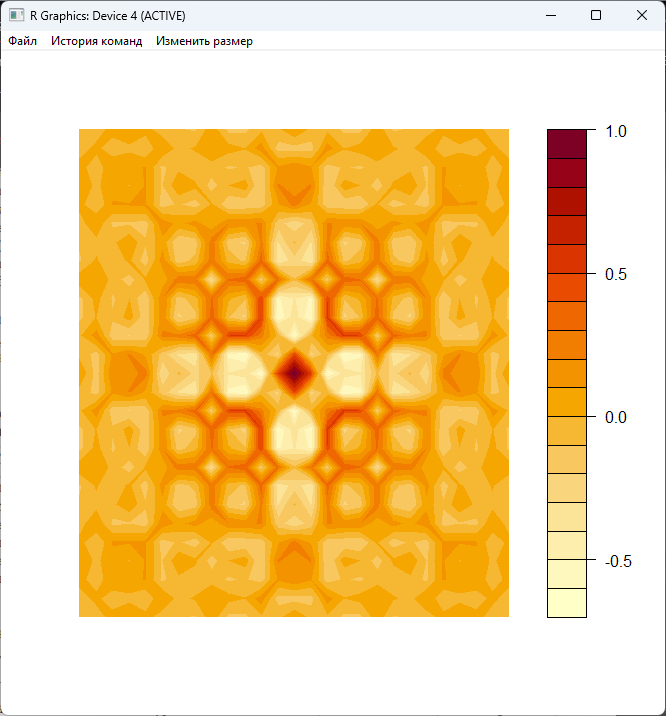


Рисунок 10 – Сохранение диаграммы в pdf формате

Далее создадим новое окно для изображения:

dev.new()

demo(image())

 Рисунок 11 – Изображение в новом графчиеском устройстве

На следующем шаге построим график, отображающий связь между объемом двигателя и лошадиными силами. Также указаны название графика, имена осей, и тип меток (точки).

dev.new()

plot(datatable$Engine.Capacity,datatable$Horse.power,xlab="Engine Capacity",ylab="Horse power",main="Graph",type="p")

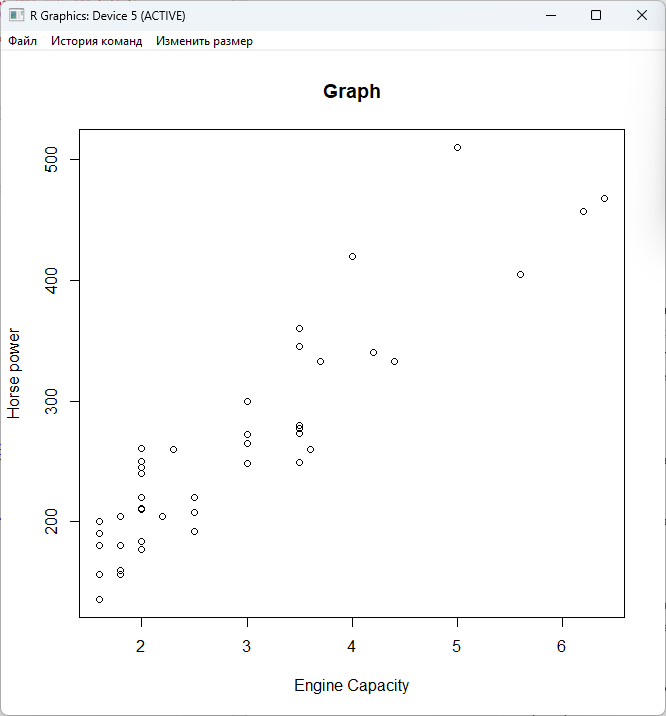


Рисунок 12 – Зависимость мощности от обьема двигателя

Анализируя рисунок, можно отметить явный равномерный рост мощности двигателя с увеличением объема.

На следующем шаге построим график, отображающий связь между объемом двигателя и крутящим моментом. Также указаны название графика, имена осей, и тип меток (точки).

Код:

plot(datatable$Engine.Capacity,datatable$Torque..Nm,xlab="Engine Capacity",ylab="Torque",main="Graph",type="p")

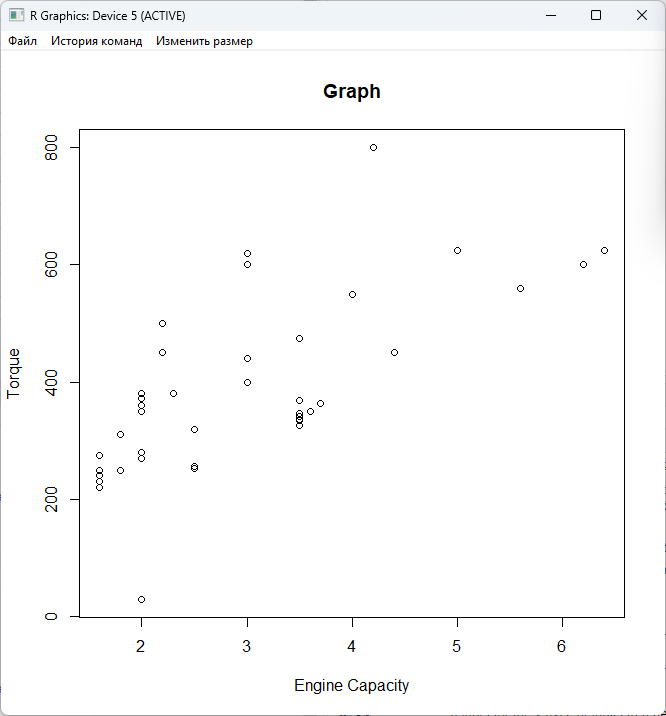


Рисунок 13 – Зависимость крутящего момента от обьема двигателя

Анализируя рисунок, можно отметить явный равномерный рост крутящего момента с увеличением объема двигателя.

На следующем шаге построим график, отображающий связь между мощностью (лошадиными силами) и разгоном от 0 до 100 км/ч (секнуд). Также указаны название графика, имена осей, и тип меток (треугольники).

Код:

par(lty=2, pch=17)

plot(datatable$Horse.power,datatable$Racing.0.100km.h,xlab="Horse Power",ylab="Racing 0-100 km/h, seconds",main="Graph",type="p")

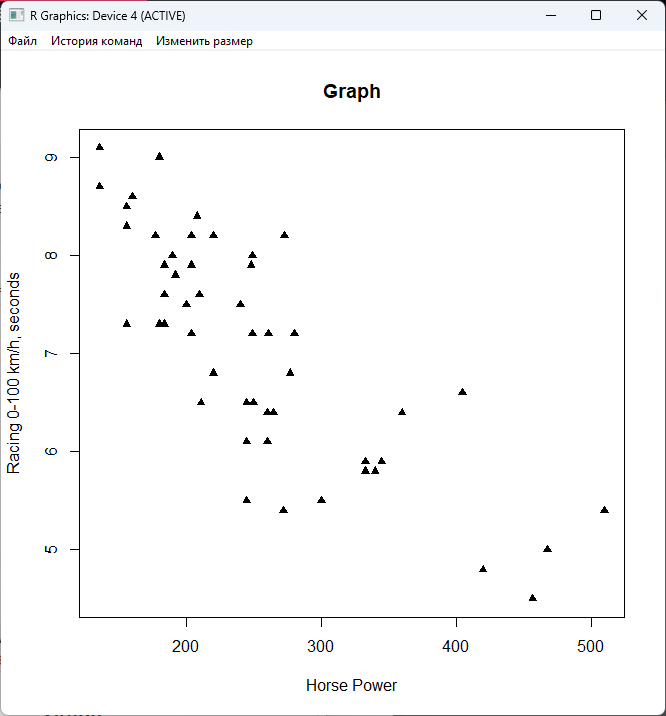


Рисунок 14 – Зависимость разгона от мощности двигателя

Анализируя рисунок, можно отметить явное уменьшение времени разгона от 0 до 100 км/ч с увеличением мощности двигателя.

Затем продемонстрируем возможности работы Rstudio с цветами на примере создания вектора из 30 цветов и демонстрацией в виде “пирога”.

Код:

n<-30

mycolors<-rainbow(n)

pie(rep(1, n), labels=mycolors, col=mycolors)

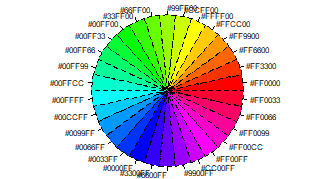


Рисунок 14 – Деление круга на 30 частей радужной палитры

Далее необходимо было выполнить данным в методических указаниях код и добавить комментарии к нему.

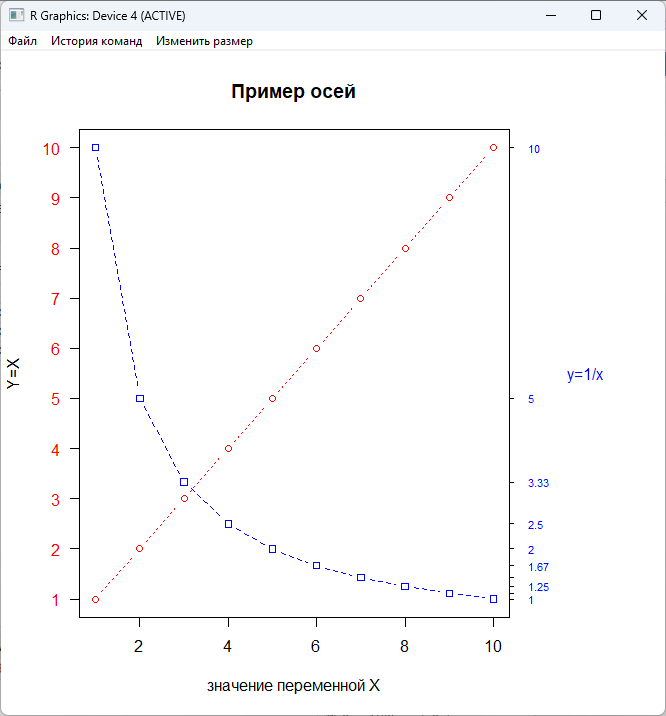


Рисунок 15 – График двух линий для комментрирования

Код:

# Создаём вектор `x` от 1 до 10.

x <- c(1:10)

# Копируем `x` в `y`.

y <- x

# Вычисляем `z` как `x / 10`.

z <- 10 / x

# Сохраняем текущие графические параметры в `opar`.

opar <- par(no.readonly = TRUE)

# Задаём поля графика (низ, лево, верх, право).

par(mar = c(5, 4, 4, 8) + 0.1)

# Рисуем график `x` и `y` как точечный график пунктрирными линиями красного цвета.

plot(x, y, type = "b", pch = 21, col = "red", yaxt = "n", lty = 3, ann = FALSE)

# Добавляем линейный график `x` и `z` с прямоугольниками.

lines(x, z, type = "b", pch = 22, col = "blue", lty = 2)

# Добавляем цвет y оси красный.

axis(2, at = x, labels = x, col.axis = "red", las = 2)

# Добавляем ось с значениями z и подписями. Цвет оси синий, размер подписей - 0.7, длина черточек - -0.01.

axis(4, at = z, labels = round(z, digits = 2), col.axis = "blue", las = 2, cex.axis = 0.7, tck = -.01)

# Добавляем текст "y = 1/x" на правой стороне синего цвета, с коэффициентом расширения символов 1 и стилем 2.

mtext("y = 1/x", side = 4, line = 3, cex.lab = 1, las = 2, col = "blue")

# Добавляем заголовок "Пример осей" к графику. Подпись на оси x - "значение переменной X", на оси y - "Y=X".

title("Пример осей", xlab = "значение переменной X", ylab = "Y=X")

# Восстанавливаем предыдущие графические параметры.

par(opar)

**Корреляционный анализ**

Вычислим матрицу корреляции методами Пирсона и Спирмена:

cor ( datatable [ c("Engine.Capacity", "Horse.power", "Torque..Nm", "Mass..kg", "Racing.0.100km.h")],

method ="pearson",

use = "complete")

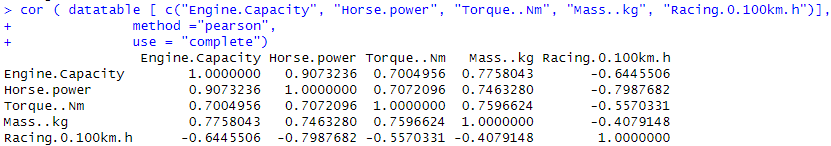


Рисунок 16 – Матрица корреляции, составленная методом Пирсона

cor ( datatable [ c("Engine.Capacity", "Horse.power", "Torque..Nm", "Mass..kg", "Racing.0.100km.h")],

method ="spearman",

use = "complete")

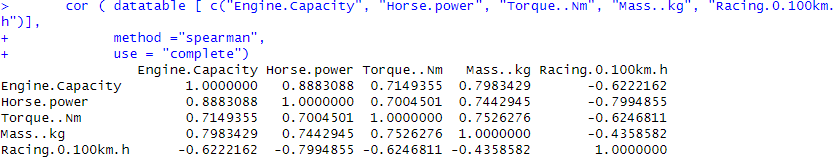


Рисунок 17 – Матрица корреляции, составленная методом Спирмена

Судя по коэффициентам, сильная корреляция наблюдается между объемом двигателя и лошадиными силами, между объемом двигателя и крутящим моментом, между объемом двигателя и массой, между лошадиными силами и крутящим моменом, междк лошадиными силами и массой, между лошадиными силами и разгоном, между крутящим моментом и массой. Средняя - между между объемом двигателя и разгоном, между крутящим моментом и разгоном, между массой и разгоном.

Выполним оценку уровня значимости коэффициента корреляции между объемом двигателя и лошадиными силами и между массой и разгоном.

Код:

with( datatable, cor.test (Engine.Capacity, Horse.power, method ="pearson"))

with( datatable, cor.test (Horse.power,Racing.0.100km.h , method ="pearson"))

with( datatable, cor.test (Mass..kg,Racing.0.100km.h, method ="pearson"))

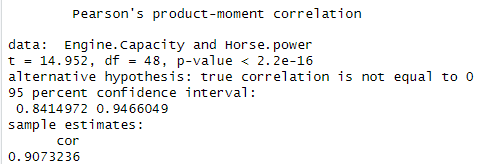


Рисунок 18 – Оценка уровня значимости коэффициента корреляции

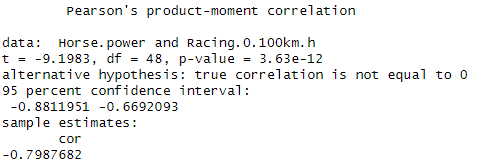


Рисунок 19 – Оценка уровня значимости коэффициента корреляции

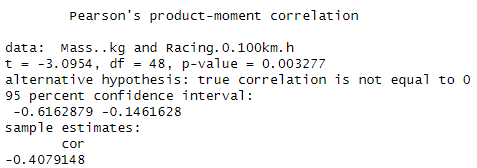


Рисунок 20 – Оценка уровня значимости коэффициента корреляции

Для первого случая сильная связь между переменными подтверждается, для второго не доказана.

Построим матрицу точечных графиков:

pairs(datatable [ c("Engine.Capacity", "Horse.power", "Torque..Nm", "Mass..kg", "Racing.0.100km.h")], pch = 21, bg = c("red", "yellow", "blue"),)

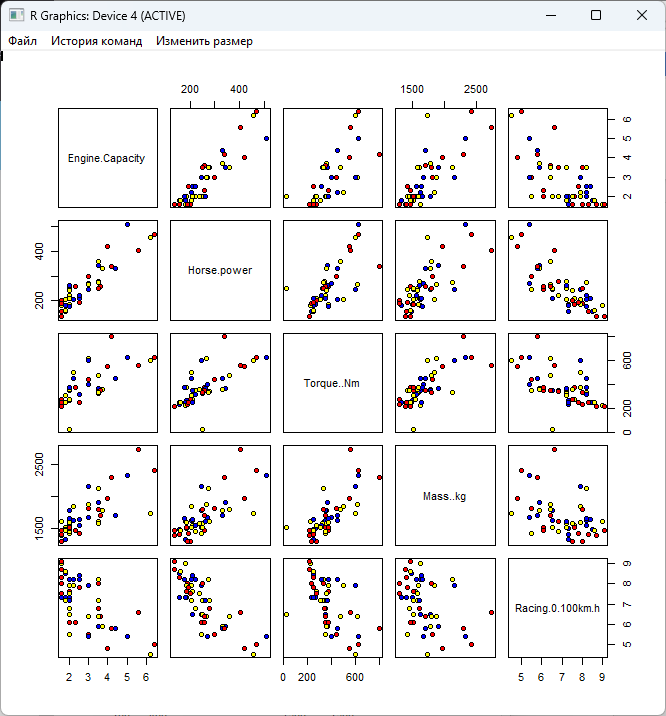


Рисунок 21 – Матрица точечных графиков

Далее определим уровень зависимости лошадиных сил и

разгона. Код:

RegModel.1 <-lm(Racing.0.100km.h~Horse.power,data = datatable)

summary(RegModel.1)

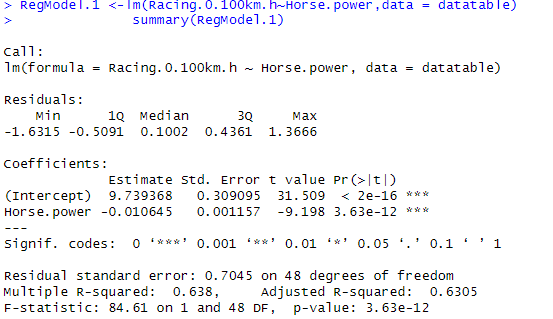


Рисунок 22 – Определение уравнения зависимости Horse Power и Racing 0-100

Racing 0-100 = -0.010645 \* Horse Power + 9.739368

Далее определим уровень зависимости объема двигателя и лошадиных сил.

Код:

RegModel.2 <-lm(Horse.power~Engine.Capacity,data = datatable)

summary(RegModel.2)

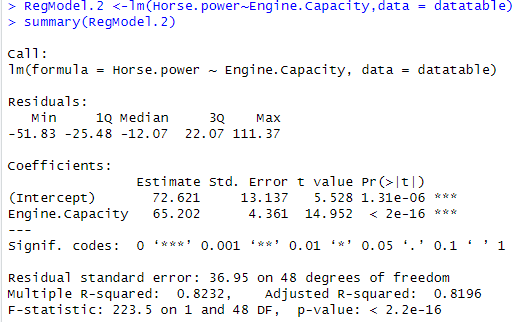


Рисунок 23 – Определение уравнения зависимости Engine.Capacity и Horse Power

Horse Power = 65.202 \* Engine.Capacity + 72.621

Далее определим уровень зависимости массы и

разгона. Код:

RegModel.3 <-lm(Racing.0.100km.h ~Mass..kg,data = datatable)

summary(RegModel.3)

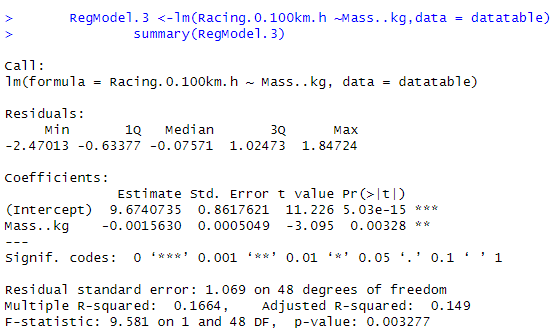


Рисунок 24 – Определение уравнения зависимости Mass и Racing 0-100

Racing 0-100 = -0.001563 \* Mass + 9.6740735

Код:

library(Rcmdr)

crPlots(RegModel.1, smooth=list(span=1))

crPlots(RegModel.2, smooth=list(span=1))

crPlots(RegModel.3, smooth=list(span=1))

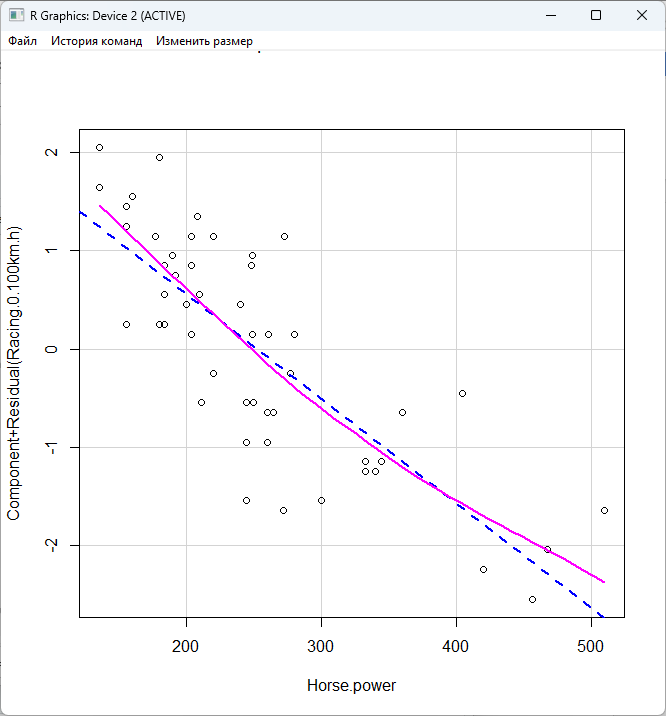


Рисунок 25 – График остатков для признаков Racing и Horse Power

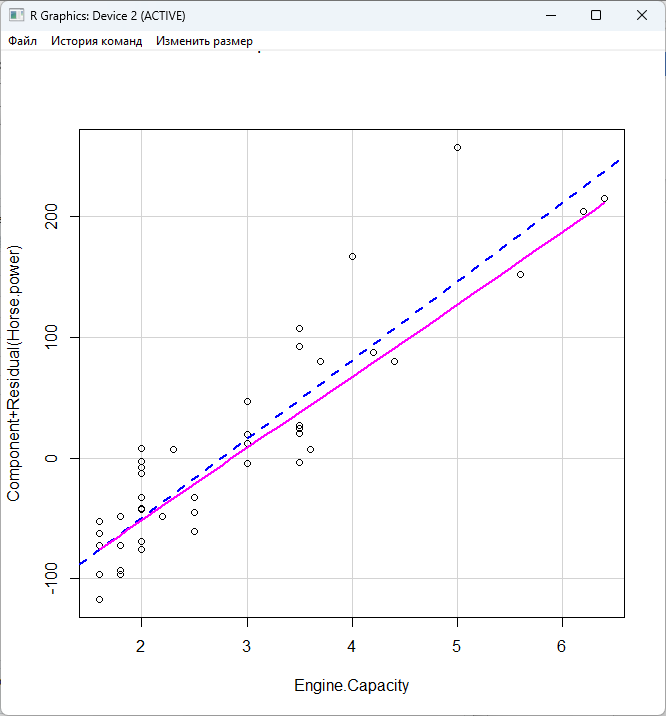


Рисунок 26 – График остатков для признаков Engine Capacity и Racing

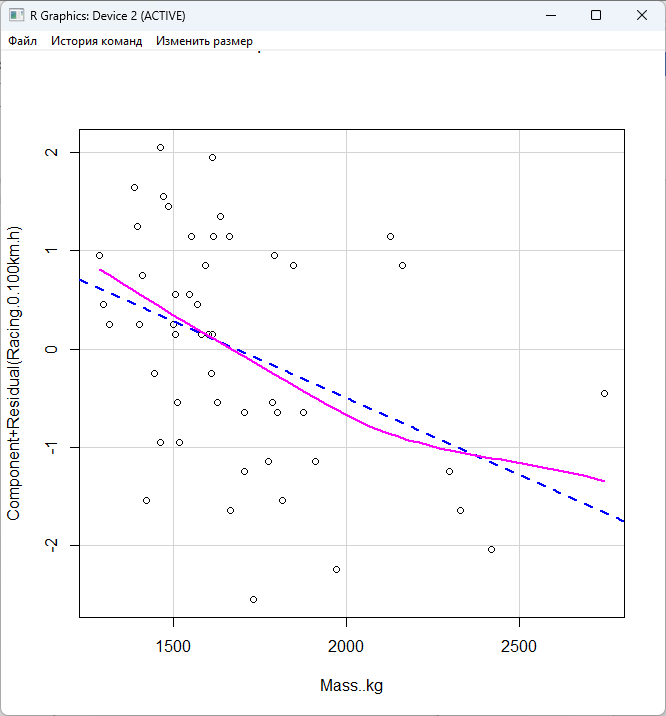


Рисунок 27 – График остатков для признаков Mass и Racing

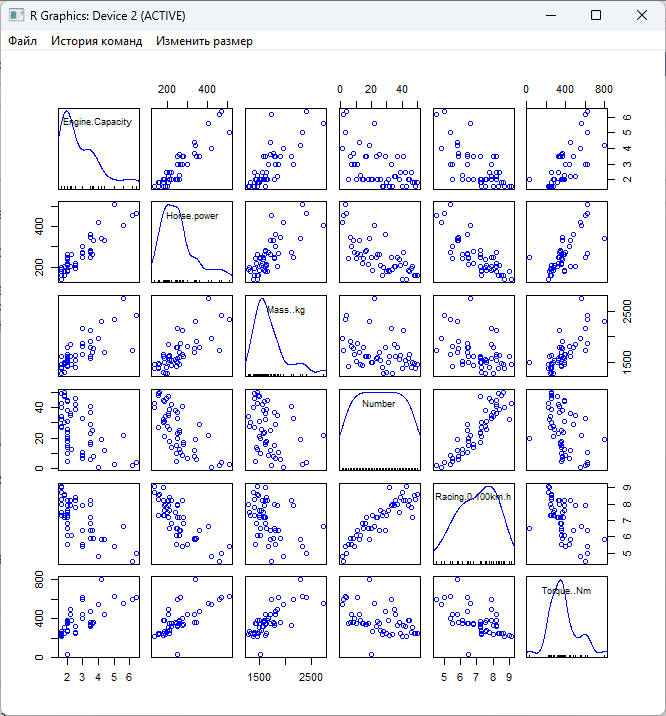


Рисунок 28 – Построенная диаграмма рассеяния для всех переменных

RegModel.4 <-

Rcmdr+ lm(Racing.0.100km.h~Engine.Capacity+Horse.power+Mass..kg+Torque..Nm,

Rcmdr+ data=datatable)

Rcmdr> summary(RegModel.4)

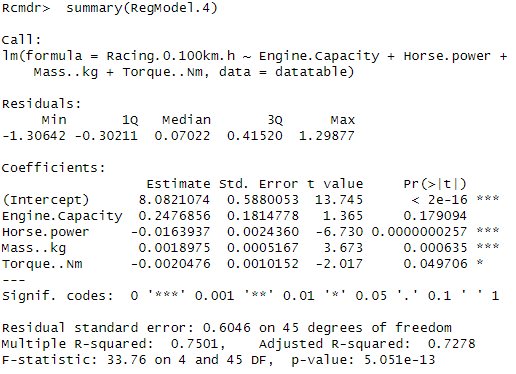


Рисунок 29 – Результат подгонки множественной линейной регрессии

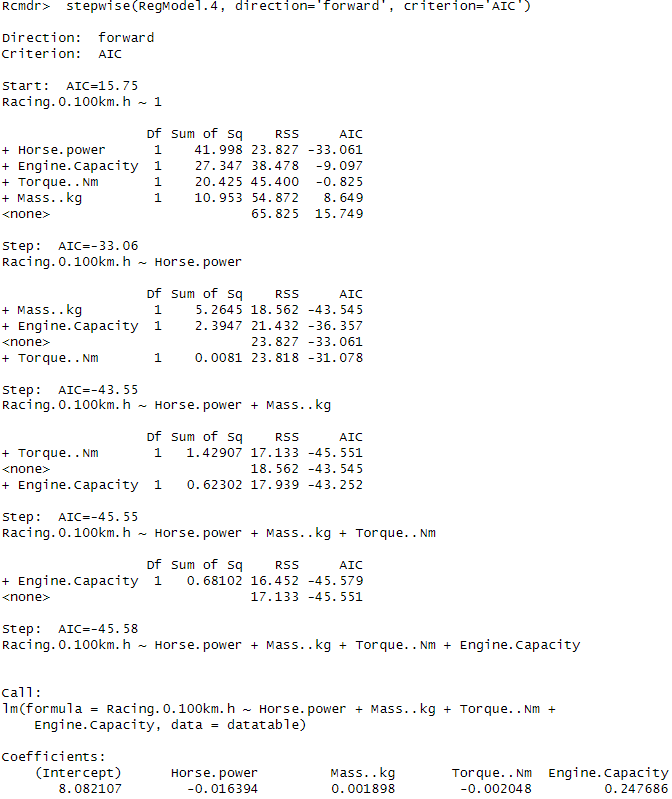


Рисунок 30 – Коэффициент AIC. Направление вперед

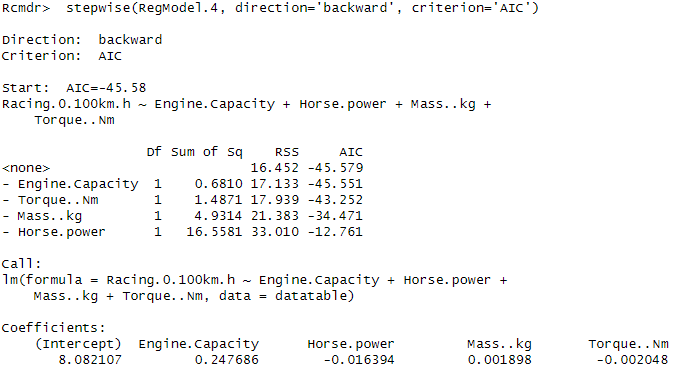


Рисунок 30 – Коэффициент AIC. Направление назад

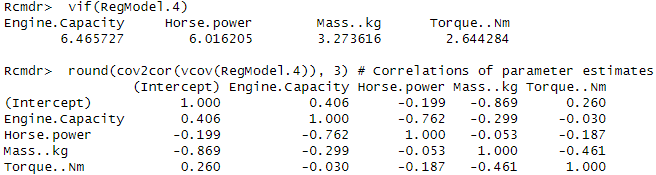


Рисунок 31 – Коэффициент VIF. Факторы, влияющие на дисперсию

**Выводы**

При выполнении данной работы были исследованы возможности языка R для построения графиков и диаграмм. Изучено проведение корреляционного и регрессионного анализа данных. Получены практические навыки определения тесноты взаимосвязей эксперементальных данных. В результате выполнения были получены следующие уравнения регрессии:

Racing 0-100 = -0.010645 \* Horse Power + 9.739368

Horse Power = 65.202 \* Engine.Capacity + 72.621

Racing 0-100 = -0.001563 \* Mass + 9.6740735

**Контрольные вопросы**

1. Функциональная связь.

При функциональной связи заданному значению фактора X соответствует строго определенные значение параметра Y.

1. Статистическая связь.

Статистическая связь соотношение между двумя переменными, при котором изменение значения одной переменной влечет изменение распределения другой переменной.

1. Корреляционная связь.

При корреляционной связи заданному значению фактора X может соответствовать множество возможных значений параметра Y.

1. Корреляционный анализ.

Основной задачей корреляционного анализа является определение формы, направленности и тесноты взаимосвязи. При исследования корреляции используются графический и аналитический подходы.

1. Корреляционное поле.

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.

1. Корреляционный анализ: форма зависимости.

Визуальный анализ корреляционного поля позволяет сделать предположение о форме взаимосвязи двух исследуемых показателей. По форме взаимосвязи корреляционные зависимости принято разделять на линейные и нелинейные. 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.

1. Корреляционный анализ: направленность взаимосвязи.

Выявление формы статистической зависимости необходимо для выбора метода оценки тесноты (силы) взаимосвязи. Направленность является положительной, если увеличение значения одного признака приводит к увеличению значения. Направленность является отрицательной, если увеличение значения одного признака приводит к уменьшению значения второго.

1. 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. Количественная оценка тесноты взаимосвязи двух случайных величин осуществляется с помощью коэффициента корреляции r.

1. Коэффициент корреляции. Его свойства.

Коэффициент корреляции характеризует только линейную взаимосвязь. Коэффициент корреляции величина относительная, он принимает значение от минус единицы до плюс единицы, т.е. -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, 0,3 < | r | < 0,7 – умеренная, при | r |0,7 – сильная.

1. Регрессионный анализ.

Основная особенность регрессионного анализа: при его помощи можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными. Под регрессией понимается функциональная зависимость между независимыми переменными и средним значением зависимой переменной, которая строится с целью предсказания этого среднего значения при фиксированных значениях переменной.

1. Этапы регрессионного анализа.

* Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений.
* Определение зависимых и независимых переменных.
* Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель (гипотеза).
* Формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная).
* Определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии).
* Оценка точности регрессионного анализа.
* Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов.
* Предсказание неизвестных значений зависимой переменной.