**Прогнозирование.**

**Линейная однофакторная регрессионная модель**

**Однофакторная линейная регрессия в среде R**

[**Задание ОЛР1 -- нахождение, исследование и использование уравнения однофакторной линейной регрессии (50 баллов)**](https://edu.kpfu.ru/mod/assign/view.php?id=101626)

Выполнил студент 2 курса

Группы 09-715(1)

Санамян Артак Размикович

Подберите реальные данные для задачи однофакторной линейной регрессии. Выполните следующее:

1. Постройте на графике облако точек.
2. Найдите уравнение регрессии. (Укажите численные значения коэффициентов регрессии)
3. Совместите уравнение регрессии с облаком точек.
4. Сделайте выводы относительно зависимости исследуемых признаков.
5. Выберите произвольно несколько значений независимого признака ***x*** и вычислите ожидаемые (согласно полученному уравнению регрессии) значения признака ***y***.
6. Покажите на графике точки, соответствующие сделанному прогнозу.
7. Оформите отчёт. Включите в отчёт построенный график, запишите численно коэффициенты регрессии, коэффициент корреляции, значение критерия Стьюдента, расчётные значения прогнозируемого признака. Обоснуйте выводы относительно зависимости признаков и качества прогноза.
8. Прикрепите doc-файл c отчётом здесь.

Максимальный балл за задание - 50

Рассмотрим решение задачи построения, анализа и использования уравнения однофакторной линейной регрессии на примере данных о автомобилях, взятых из репозитория: <http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.names>

Файл содержит данные о 9 признаках автомобилей, это - расход милей на один галон, цилиндры, смещение и т.д. Число автомобилей, для которых были произведены измерения, равно 400. Файл с данными находится в рабочей директории в файле auto-mpg.data. Найдём зависимость расхода миль на галон топлива автомобиля от количества лошадиных сил в двигателе автомобиля (будем предполагать, что зависимость линейна и искать коэффициенты линейной функции).

# Прочитаем содержимое файла в переменную A

A=read.table("data1.txt",header=FALSE,sep=",")

# Сохраним данные по интересующим нас признакам в переменных

# Шестой столбец - длина моллюска

mpg=A[,6]

# Третий столбец - диаметр раковины diam=A[,3] # Построим облако точек

horse\_power=A[,4]

# Построим облако точек

plot(mpg, horse\_power, col="blue",type="p",pch=16,

xlab="horse power",ylab="mpg",main="Зависимость mpg (миль на галон топлива) от мощности (кол-ва лошадиных сил)")

Здесь мы указываем параметры функции plot:

col – цвет символов (синий);

type – тип символа («p» – (от англ.: «point»— точка);

pch – вид точки (от англ.: point character) (16 – закрашенная, 1 – пустой кружок и т.д.);

xlab, ylab – надписи на осях;

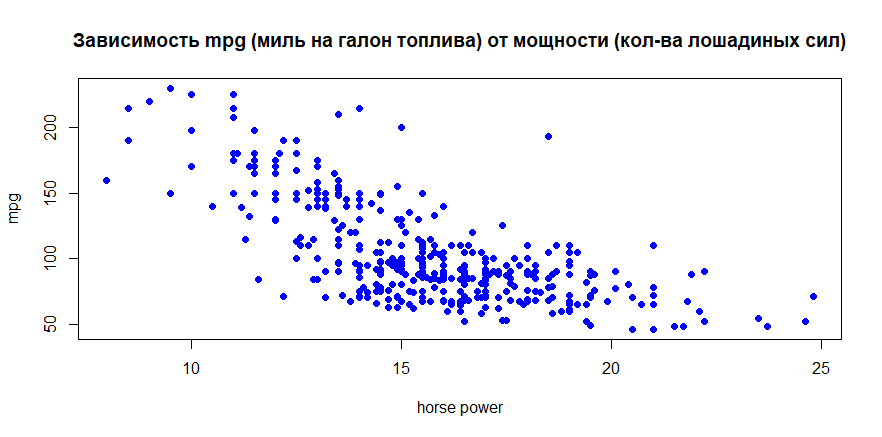
main – заголовок графика.

2 Все параметры функции plot можно посмотреть, например, здесь:

<http://www.statmethods.net/advgraphs/parameters.html>

Получим график, показанный на Рис.1.

Рис. 1 Облако точек, характеризующих зависимость mpg (миль на галон топлива) от мощности (кол-ва лошадиных сил)



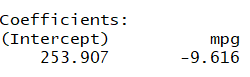
Как видно из Рис.1, есть основания полагать, что между признаками наблюдается положительная линейная зависимость. Для нахождения коэффициентов линейной регрессии воспользуемся встроенной функцией lm для нахождения коэффициентов этой зависимости. Заметим, что результатом вызова функции lm является некоторый объект. Назовём его, например, myregress. Обратите внимание на формат вызова функции lm – формулу нужно понимать так: линейная функция, отражающая зависимость переменной «mpg» от переменной «horse\_power»:

myregress=lm(formula = horse\_power ~ mpg)

myregress

Получим результат, показанный на Рис.2.

Рис.2. Результат вызова функции lm.



**Уравнение регрессии:**

y = -9.616x + 253.907

Как видно из Рис.2, свободный коэффициент модели (т.е. значение линейной функции в нуле) равен примерно 253,907, коэффициент при переменной «mpg» (т.е. угол наклона прямой) равен примерно -9,616. Построим прямую регрессии и совместим её с облаком точек:

# Построим график уравнения однофакторной линейной регрессии (с найденными коэффициентами)

abline(myregress,col="red",lwd="3",add=TRUE)

Здесь параметр lwd задаёт толщину линии, параметр add со значением TRUE указывает на то, что график должен быть построен на том же рисунке, что и предыдущий графический объект (т.е. в нашем случае он будет совмещён с облаком точек). Получим результат , показанный на Рис.3.

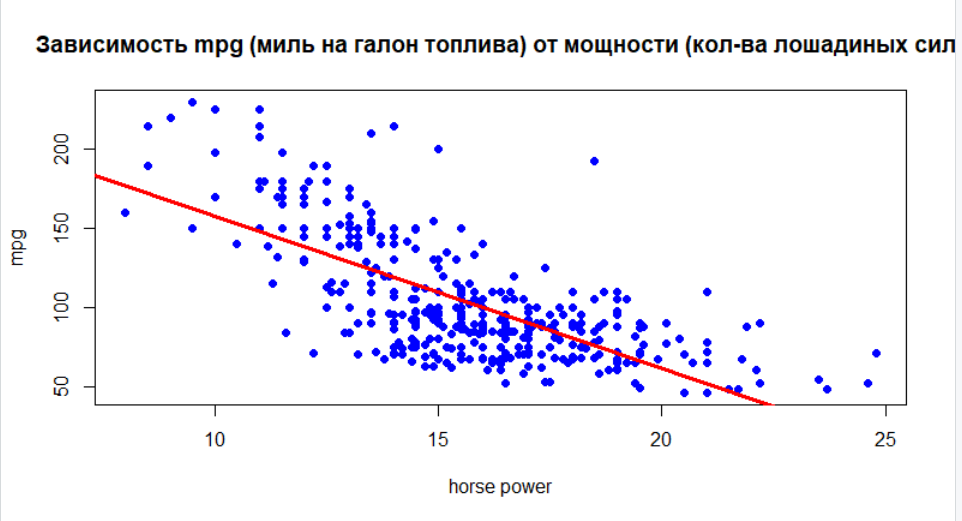


Рис.3. Облако точек, характеризующих зависимость mpg (миль на галлон топлива) от мощности (кол-ва лошадиных сил), и прямая регрессии.

4. Для исследования модели (т.е. для того, чтобы выяснить, насколько значима зависимость между признаками) вызовем функцию summary, выводящую всю информацию о модели: summary(myregress) Получим результат,

Call:

lm(formula = horse\_power ~ mpg)

Residuals:

Min 1Q Median 3Q Max

-65.598 -17.713 -2.463 17.873 116.980

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 253.9074 8.0796 31.43 <2e-16 \*\*\*

mpg -9.6155 0.5119 -18.78 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.93 on 390 degrees of freedom

Multiple R-squared: 0.475, Adjusted R-squared: 0.4736

F-statistic: 352.8 on 1 and 390 DF, p-value: < 2.2e-16

Выборочный коэффициент корреляции между фактором (мощность двигателя) и переменной отклика (мили на галлон топлива):

r = -0.6891955

Для нашего примера выборочный коэффициент корреляции r между фактором (мощность двигателя) и переменной отклика (мили на галлон топлива) равен -0.6891955. Поскольку r близок к -1, можем предположить наличие обратной линейной зависимости между признаками. Вычислим значение T-критерия по формуле (1):

# Сначала найдём объём выборки n

n=length(mpg)

# Вычислим и выведем значение T-критерия t=r/sqrt(1-r^2)\*sqrt(n-2)

t

Получим значение критерия Стьюдента равно -18.78413

Заметим, что это число присутствует в таблице «Coefficients» (столбец «t-value»), т.е. объект myregress содержит информацию о наблюдаемом значении критерия Стьюдента.

Поскольку наблюдаемое значение критерия Стьюдента (-18.78413) отличается от нуля, что мало вероятно, если считать, что наша гипотеза о независимости признаков верна (т.е. ρ=0), у нас есть основания полагать, что гипотеза 𝐻0: 𝜌 = 0 ложна (т.е. на самом деле 𝜌 ≠ 0, а значит, существует линейная зависимость между признаками).

Чтобы сделать вывод о том, насколько сильно наблюдаемое значение T-критерия (т.е. -18.78413) отлично от нуля, найдём вероятность получить (в условиях нулевой гипотезы) ещё большее значение, т.е. вероятность P(T>-18.78413). Для этого воспользуемся тем известным фактом, что с.в. T имеет распределение Стьюдента с n-2 степенями свободы: P(T>-18.78413) = 1 – F(-18.78413), где F – функция распределения Стьюдента с (n-2) степенями свободы.

Заметим, что в summary вероятность превысить наблюдаемое значение критерия Стьюдента (в условиях истинности нулевой гипотезы) находится в столбце Pr(>|t|). Оно меньше, чем 2·10−16, т.е. практически равно нулю.

5) Выбираем следующие значения независимого признака x (миль на галлон топлива): 10,15,8,17,21,12, 19

Спрогнозированные значения переменной y (столбец fit) и границы 90%-го доверительного интервала для них:

fit lwr upr

1 157.75217 152.52903 162.97530

2 109.67453 107.30457 112.04449

3 176.98322 170.20682 183.75962

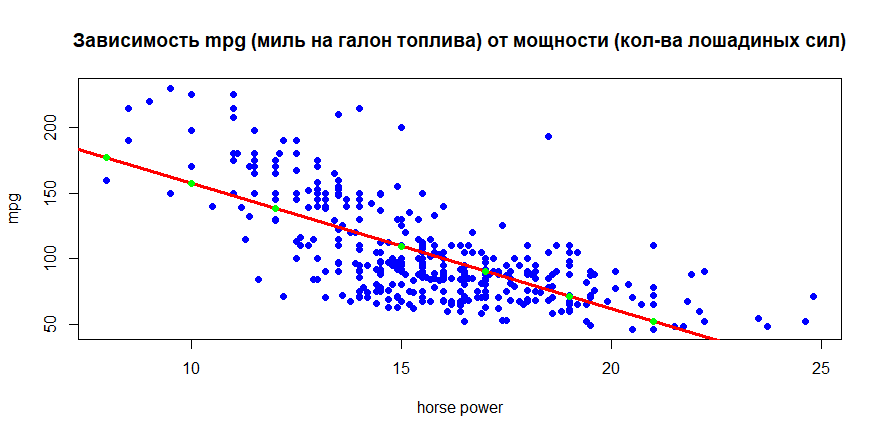
4 90.44347 87.81219 93.07475

5 51.98136 46.82060 57.14213

6 138.52111 134.73411 142.30811

7 71.21242 67.48023 74.94461

Первый столбец (fit) содержит прогнозируемые значения миль на галлон топлива, столбцы lwr (сокращение от англ.: «lower» – «нижний») и upr (сокращение от англ.: «upper» – «верхний») содержат границы 90% доверительного интервала для прогнозируемого значения mpg. Для наглядности извлечём из объекта элементы первого столбца и нанесём на график соответствующие точки:



**Листинг**

setwd("C://Users/pc/Documents/Магистратура/2 семестр/R/task1 ОЛР1")

getwd()

A=read.table("data1.txt",header=FALSE,sep=",")

mpg=A[,6]

horse\_power=A[,4]

plot(mpg, horse\_power, col="blue",type="p",pch=16,

#xlim = c(-10, 25), ylim = c(0, 300),

xlab="horse power",ylab="mpg",main="Зависимость mpg (миль на галон топлива) от мощности (кол-ва лошадиных сил)")

myregress=lm(formula = horse\_power ~ mpg)

myregress

abline(myregress,col="red",lwd="3")

# выводим информацию о модели

summary(myregress)

# выборочный коэффициент корреляции

r=cor(mpg,horse\_power)

r

# найдём объём выборки n

n=length(mpg)

# вычислим и выведем значение T-критерия

t=r/sqrt(1-r^2)\*sqrt(n-2)

t

# найдём вероятность получить (в условиях нулевой гипотезы) ещё большее значение p(abs(t) > -18.78)

p = 1 - pt(abs(t), n - 2)

p

# выбираем следующие значения независимого признака x (мощность двигателя):

newmpg=c(10,15,8,17,21,12, 19)

newmpg

# построение прогноза:

newHorsePower=predict(myregress,data.frame(mpg = newmpg), level = 0.9, interval = "confidence")

newHorsePower

newHP1=newHorsePower[,1]

# наносим на график соответствующие точки:

points(newmpg, newHP1, col="green", type="p", pch=16, add=TRUE)

Вопрос

При добавлении add = TRUE происходит ошибка

Warning message:

In plot.xy(xy.coords(x, y), type = type, ...) :

"add" -- не графический параметр