***Въведение в Статистиката с R – Мони – Bivariate Data***

1. **Работа с двумерни категорийни, качествени данни –**

В тази точка ще се интересуваме от начините за описване на зависимостта между два кетигорийни признака. От таблицата с данни tires категорийни променливи са

Сега ще демонстрираме основните възможности на R за обработка на връзката между две категорийни данни като използваме таблицата с данни tires.

**Групировка в абсолютни честоти**

По аналогичен начин на едномерния случай, в резултат от групировката на данните според наблюдаваните при тях значения на два признака се получават така наречените кръстосани таблици. Те могат да съдържат броят на наблюденията, които попадат в непресичащите се групи, определени от различните регистрирани значения на двата признака при наблюдаваните статистически единици.

До векторите, които съдържат резултатите от наблюденията можем да достигнем чрез функцията

> attach(tires)

или чрез оператора “$” като преди него напишем името на таблицата с данни. Т.к. ние ше работим изцяло с тази таблица с данни ще използваме функцията attach.

Сега в таблицата с данни tires кръстосаната таблица, която се получава след групировка по двата признака Х1 – Вид гума, за която ще бъде попълнена анкетата и X2 – Производител се извежда в R отново след използването на функцията ***table***, но този път като параметри имаме имената на двата категорийни признака.

> table(X1, X2)

X2

X1 P1 P2 P3 P4 P5 P6

T1 3 3 1 4 5 6

T2 3 6 5 4 2 4

T3 4 4 5 3 3 4

T4 5 5 3 3 4 6

T5 6 3 2 4 7 6

T6 6 8 13 13 11 12

T7 11 10 10 11 11 11

Вместо X1 и X2 можем да сменим имената на двата признака като е добре да ги зададем като отделен вектор

names = c("Вид", "Производителl")

table(X1, X2 , dnn = names)

Производител

Вид P1 P2 P3 P4 P5 P6

T1 3 3 1 4 5 6

T2 3 6 5 4 2 4

T3 4 4 5 3 3 4

T4 5 5 3 3 4 6

T5 6 3 2 4 7 6

T6 6 8 13 13 11 12

T7 11 10 10 11 11 11

Вместо етикетите на видовете коли можем да напишем имената на видомете, като е добре да ги зададем като отделен вектор. Това става с помощта на функцията ***rownames***.

rnames = c("Audi", "BMW","Honda", "Mercedes", "Nissan","Peugeot", "Citroen")

names = c("Вид", "Производител")

t = table(X1, X2 , dnn = names)

rownames(t) = rnames

t

Производител

Вид P1 P2 P3 P4 P5 P6

Audi 3 3 1 4 5 6

BMW 3 6 5 4 2 4

Honda 4 4 5 3 3 4

Mercedes 5 5 3 3 4 6

Nissan 6 3 2 4 7 6

Peugeot 6 8 13 13 11 12

Citroen 11 10 10 11 11 11

По аналогичен начин, с помощта на функцията ***colnames*** можем да сменим имената на колоните.

Функцията ***prop.table*** с втори параметър 1, в R ни позволява да определим какъв е процентът на единиците в съответната група от общата сума в реда.

prop.table(t,1)\*100

Производител

Вид P1 P2 P3 P4 P5 P6

Audi 13.636364 13.636364 4.545455 18.181818 22.727273 27.272727

BMW 12.500000 25.000000 20.833333 16.666667 8.333333 16.666667

Honda 17.391304 17.391304 21.739130 13.043478 13.043478 17.391304

Mercedes 19.230769 19.230769 11.538462 11.538462 15.384615 23.076923

Nissan 21.428571 10.714286 7.142857 14.285714 25.000000 21.428571

Peugeot 9.523810 12.698413 20.634921 20.634921 17.460317 19.047619

Citroen 17.187500 15.625000 15.625000 17.187500 17.187500 17.187500

Добре е да закръглим числата до втория знак след десетичната запетая

t1 = round(prop.table(t,1)\*100,2)

Производител

Вид P1 P2 P3 P4 P5 P6

Audi 13.64 13.64 4.55 18.18 22.73 27.27

BMW 12.50 25.00 20.83 16.67 8.33 16.67

Honda 17.39 17.39 21.74 13.04 13.04 17.39

Mercede s 19.23 19.23 11.54 11.54 15.38 23.08

Nissan 21.43 10.71 7.14 14.29 25.00 21.43

Peugeot 9.52 12.70 20.63 20.63 17.46 19.05

Citroen 17.19 15.62 15.62 17.19 17.19 17.19

Това означава например, че:

* 13.64% от Аудитата са произведени от Р1.
* 4.55% от Аудитата са произведени от Р3.
* 25% от BMW са произведени от Р2.
* Ако избраната кола е Нисан, с 25% шанс тя е произведена от Р5.

Проверяваме дали сумите в редовете са по 100.

apply(t1,1,sum)

Audi BMW Honda Mercedes Nissan Peugeot Citroen

100.01 100.00 99.99 100.00 100.00 99.99 100.00

По аналогичен начин, с помощта на функцията ***prop.table*** с втори параметър 2, в R ни позволява да определим какъв е процентът на единиците в съответната група от общата сума в колоната.

|  |
| --- |
| t2=round(prop.table(t,2)\*100,2)  Производител  Вид P1 P2 P3 P4 P5 P6  Audi 7.89 7.69 2.56 9.52 11.63 12.24  BMW 7.89 15.38 12.82 9.52 4.65 8.16  Honda 10.53 10.26 12.82 7.14 6.98 8.16  Mercedes 13.16 12.82 7.69 7.14 9.30 12.24  Nissan 15.79 7.69 5.13 9.52 16.28 12.24  Peugeot 15.79 20.51 33.33 30.95 25.58 24.49  Citroen 28.95 25.64 25.64 26.19 25.58 22.45  Това означава например, че:   * 7.89% от колите, в извадката, които са произведени от Р1 са Аудита. * 7.69% от колите, в извадката, които са произведени от Р2 са Аудита. * 22.45% от колите, в извадката, които са произведени от Р6 са Ситроени. * Ако избраната кола е произведена от Р1, най-голям е щансът тя да е Ситроен.   Проверяваме дали сумите в редовете са по 100.  apply(t2,2,sum)  P1 P2 P3 P4 P5 P6  100.00 99.99 99.99 99.98 100.00 99.98  Функцията ***prop.table*** с без втори параметър ни позволява да определим какъв е процентът на единиците в съответната група от общата сума в таблицата с абсолютни честоти, т.е. от обема на извадката.  t3 = round(prop.table(t)\*100,2)  Производител  Вид P1 P2 P3 P4 P5 P6  Audi 1.2 1.2 0.4 1.6 2.0 2.4  BMW 1.2 2.4 2.0 1.6 0.8 1.6  Honda 1.6 1.6 2.0 1.2 1.2 1.6  Mercedes 2.0 2.0 1.2 1.2 1.6 2.4  Nissan 2.4 1.2 0.8 1.6 2.8 2.4  Peugeot 2.4 3.2 5.2 5.2 4.4 4.8  Citroen 4.4 4.0 4.0 4.4 4.4 4.4  Това означава например, че:   * 1.2% от колите, в извадката, са произведени от Р1 и са Аудита. * 2.4% от колите, в извадката, са произведени от Р2 и са BMW. * 4.4% от колите, в извадката, са произведени от Р6 са Ситроени.   Проверяваме дали сумата в таблицата е 100.  sum(t3)  [1] 100 |

*Задачи за упражнение.* Направете кръстосани таблици, които да отразяват резултата от групировка по признаците

X2 - Производител

X3 – Доставчик.

Сменете имената на променливите. Сменете имената на редовете. Сменете имената на колоните. Представете групировката в

* абсолютни числа
* процент от общото
* процент от сумата в реда
* процент от сумата в колоната

и анализирайте резултатите.

**Гафични изображения за онагледяване на зависимости между два категорийни признака.**

За да онагледим зависимостта между два категорийни признака можем да използваме функцията ***barplot***. Например:

rnames = c("Audi", "BMW","Honda", "Mercedes", "Nissan","Peugeot", "Citroen")

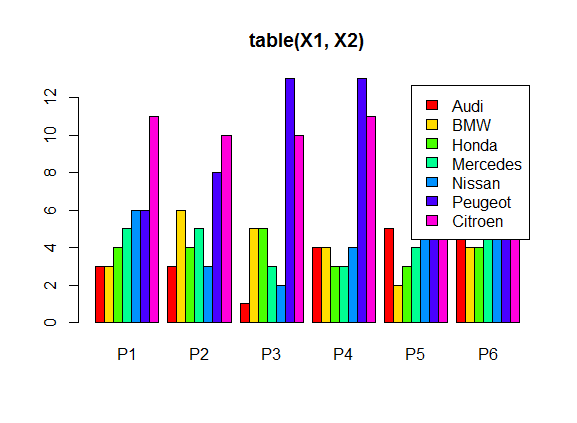
names = c("Вид", "Производител")

t = table(X1, X2 , dnn = names)

rownames(t) = rnames

t

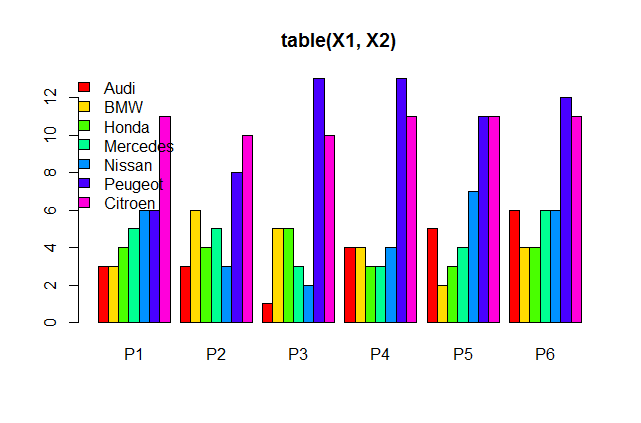
barplot(t , col=rainbow(7), main="table(X1, X2)", beside=TRUE, legend.text=rnames)



Можем да сложим легендата на мястото, където кликнем с мишката и да махнем кунтура й.

barplot(t , col=rainbow(7), main="table(X1, X2)" , beside=TRUE)

legend(locator(1), rnames, cex=1, fill= rainbow(7),bty="n")

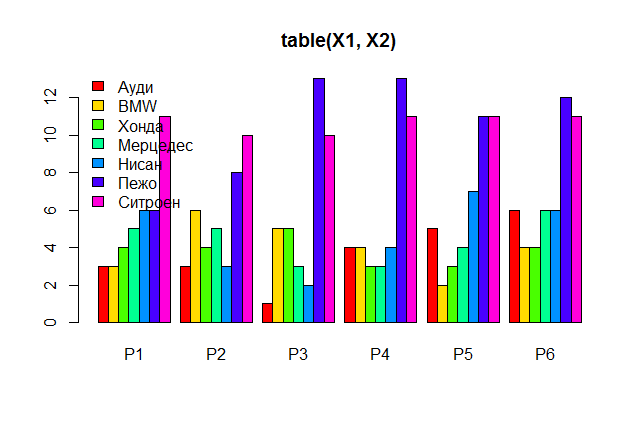


С командата legend.text можем да сменим текста в легендата

barplot(t , col=rainbow(7), main="table(X1, X2)" , beside=TRUE)

legend.text = c("Ауди","BMW","Хонда","Мерцедес","Нисан", "Пежо","Ситроен")

legend(locator(1),legend.text, cex=1, fill= rainbow(7),bty="n")



1. **Работа с един категориен и един метриран признак –**

В тази точка ще предполагаме, че имаме един категориен признак, като например Х1 – вид на колата или X2 – производител, и ще се интересуваме от начините за описване на зависимостта между този категориен признак от друг количествен признак.

От таблицата с данни tires количествени признаци са

X4 - Цена на закупуване, актуализирана към днешна дата

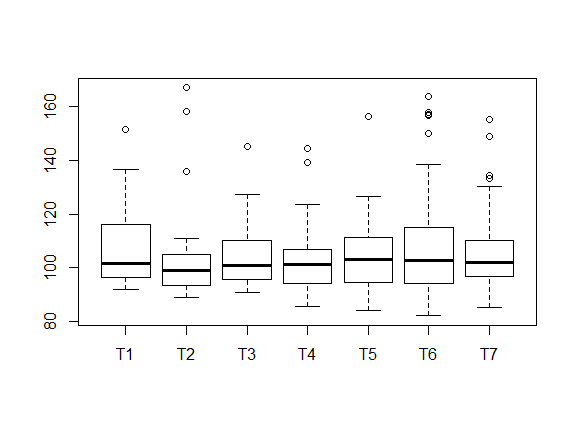
X5 - Пробег

X6 - Продължителност на живот в дни

X9 - Диаметър на джантата в цолове

За по-лесно установяване на влиянието на категорийния признак (Х1 – вид на колата) върху количествения признак (X4 - Цена на закупуване, актуализирана към днешна дата), може да направим резюме (например във формата на boxplot) на данните от метрирания признак, поотделно по подгрупите на категорийния признак. За целта използваме оператора “~”. От лявата му страна се пише зависимата променлива. По-долу това е метрираният признак Х4. От дясната страна на “~” се пише независимата променлива. По-долу това е категорийната променлива Х1. Т.е. така моделираме влиянието на Х1 върху Х4 или, което е все едно - зависимостта на Х4 от Х1.

> boxplot(X4~X1)



В предната тема вече видяхме, че преди да изчертаем тези графики може, с функциите ***sort*** и ***tapply*** да подредим групите в X1 например според минимумите в Х4. За целта виждаме реда на подгрупите в Х1 според медианите по признака Х4 в тях

> sort(tapply(X4, X1, min))

T6 T5 T7 T4 T2 T3 T1

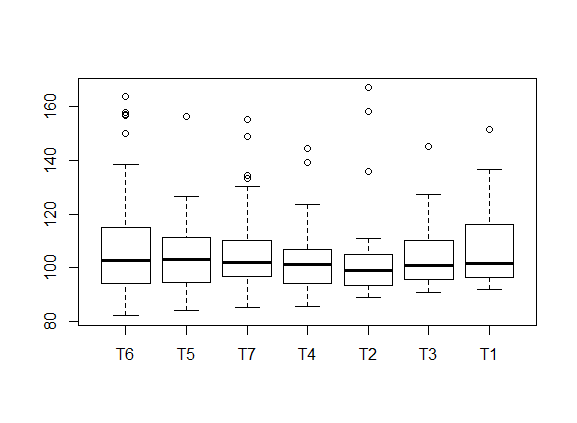
82.13 84.03 85.29 85.73 88.95 91.03 91.94

и дефинираме нова променлива оrd, която е таблица с групировката по Х1, но нивата са подредени според реда, който сме задали.

> ord = ordered(X1, levels = c("T6","T5","T7","T4","T2","T3","T1"))

Изчертаваме графики с мустачки на зависимостта на Х4 от новата, подредена по наше желание променлива.

> boxplot(X4~ord)



Можем да сменим имената за по-лесно разчитане на резултатите. При Т1 – Т7 те бяха съответно

"Audi", "BMW","Honda", "Mercedes", "Nissan","Peugeot", "Citroen"

Значи сега ще са

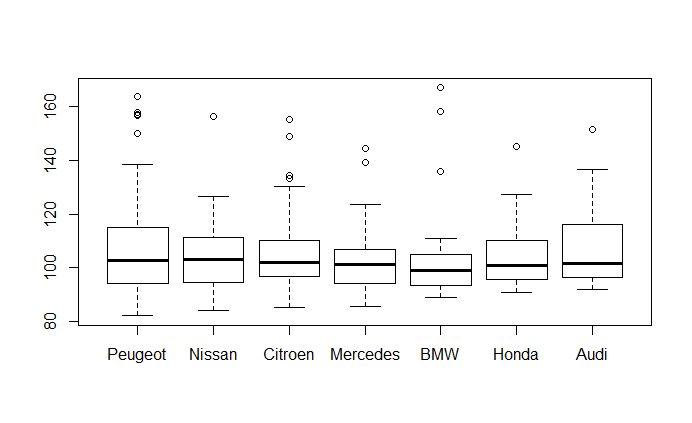
names = c("Peugeot", "Nissan", "Citroen", "Mercedes", "BMW", "Honda", "Audi")

Първо изтриваме етикетите по абсцисната ос на boxplot

> boxplot(X4~ord, xaxt="n")

След това добавяме новите имена

> axis(side=1,at=c(1,2,3,4,5,6,7), labels = names)



Виждаме, че колите от марката Пежо имат най-голямо разнообразие в цените на гумите си в извадката.

При тях е наблюдавана най-ниската цена на гуми, но и най-високата. Според горните графики, минималните цени нарастват от ляво на дясно.

Можем за превърнем метрирания признак в качествен, посредством функцията ***cut*** и след това да продължим анализа, така все едно имаме два категорийни признака. Например:

> table(cut(X4, quantile(X4)), X2)

X2

P1 P2 P3 P4 P5 P6

(82.1,94.5] 14 12 11 6 9 10

(94.5,102] 10 10 7 11 13 11

(102,111] 7 7 7 14 9 18

(111,167] 7 10 13 11 12 10

Този начин, обаче не винаги е за предпочитане, защото при групировката по метрирания признак се губи част от информацията и по-късно това ограничава възможностите за анализ.

1. **Работа с два количествени признака –**

При извършване на анализи най-удобно се работи с количествени признаци т.к. можем да използваме различни видове мерки.

*Сравняване чрез подреждане на наблюденията и визуализирането им върху реалната права.*

Най-лесно е до отбележим наредените наблюдения и да ги сравним. Това става с помощта на функцията ***stripchart***

|  |
| --- |
| > stripchart(scale(X4),cex=0.2, main="Според скалираната цена в лв. ", xlab = "лв.")  > stripchart(scale(X5),cex=0.2, main="Според пробега в км. ", xlab = "лв.") |

*Сравняване чрез графики с мустачки.*

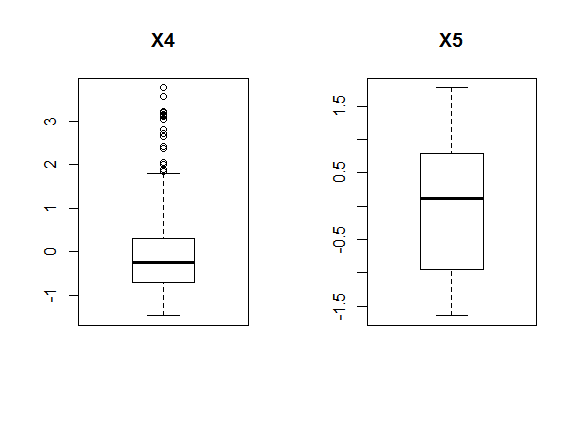
И с просто око се вижда, че при разпределението на гумите според цената им имаме дясна асиметрия, т.е. струпване на наблюденията около по-малките значения на признака. При разпределението на гумите според пробега им имаме почти асиметрия и наблюденията са почти равномерно разпределения в целия интервал от наблюдавани стойности.

Вече знаем, че единият начин за сравняване на разпределения е посредством техните графики с мустачки. Добре е преди това да сме центрирали и нормирали случайните величини. Т.е. от всяко наблюдение да сме извадили средното на извадката и да разделили на стандартното отклонение. Това може да бъде направено с функцията ***scale***. Например:

> par(mfrow = c(1, 2)) # построяваме си матрица от 1 ред и два колони, в която ще разположим графиките си.

> boxplot(scale(X4), main="X4")

> boxplot(scale(X5), main="X5")



Тук виждаме, че разпределението на гумите, според цената им е много по-дясносиметрично от разпределението на гумите според техния пробег.

*Сравняване чрез емпирични функции на разпределение.*

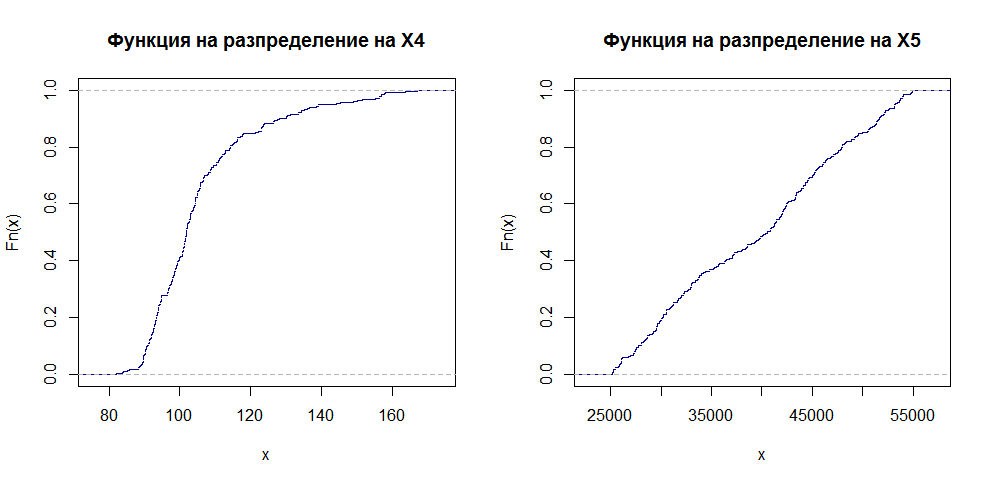
По-точно сравняване на разпределения става чрез изчертаването на емпиричните им функции на разпределение. Да припомним, че емпирична функция на разпределение наричаме

Стойностите на емпиричната функция на разпределение се определяха с помощта на функцията ***ecdf***.

> par(mfrow = c(2, 1))

> plot(ecdf(X4), verticals = FALSE, col="darkblue", do.points = FALSE, lwd = 1, main = "Функция на разпределение на X4")

> plot(ecdf(X5), verticals = FALSE, col="darkblue", do.points = FALSE, lwd = 1, main = "Функция на разпределение на X5")

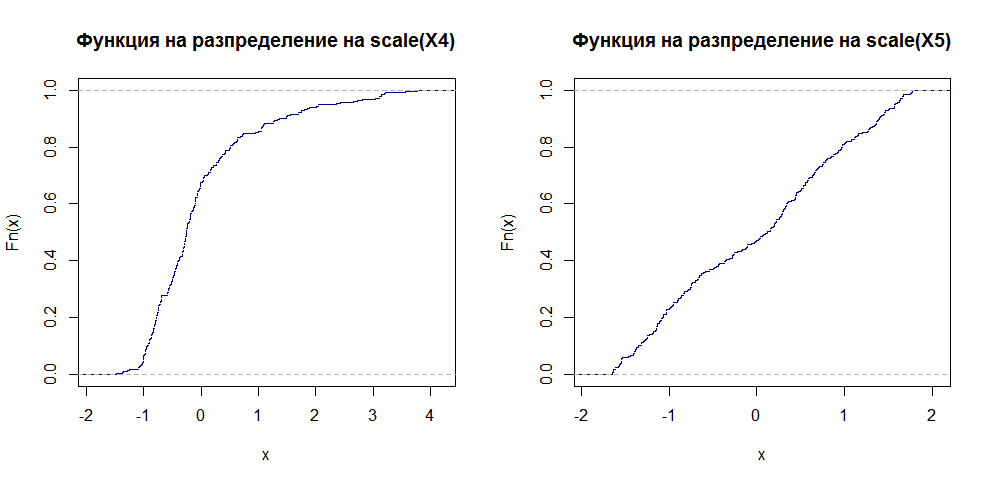


На тези графики не сме центрирали и нормирали случайните величини, за които те са изчертани. Ако ще сравняваме разпределенията е добре преди това да центрираме и нормираме случайните си величини. Така по абсцисната ос ще получим по-близки значения на наблюдаваните признаци.

> par(mfrow = c(1, 2))

> plot(ecdf(scale(X4)), verticals = FALSE, col="darkblue", do.points = FALSE, lwd = 1, main = "Функция на разпределение на scale(X4)")

> plot(ecdf(scale(X5)), verticals = FALSE, col="darkblue", do.points = FALSE, lwd = 1, main = "Функция на разпределение на scale(X5)")



Забелязва се, че нарастването на емпиричната функция на разпределение на цената на гумите е много по-бързо при малките стойности и много по-бавно при големите стойности. Това показва, че наблюдаваните гуми имат много по-често срещани по-малки цени, в сравнение със средното и по-рядко се случват по-големи цени, които обаче значително се отличават от средното за да компенсират струпванията около по-малките цени. Т.е. имаме дясно асиметрично разпределение. Същия извод направихме и чрез графиката с мустачки.

Функцията на разпределение на пробега нараства много по-равномерно. Това означава, че колите са почти равномерно разпределени според своя пробег в интервала от минимума до максимума.

*Сравняване чрез хистограми.*

Сравняване на два количествени признака може да стане и чрез тяхната хистограма, като е добре да добавим и маркери за точните стойности на наблюденията.

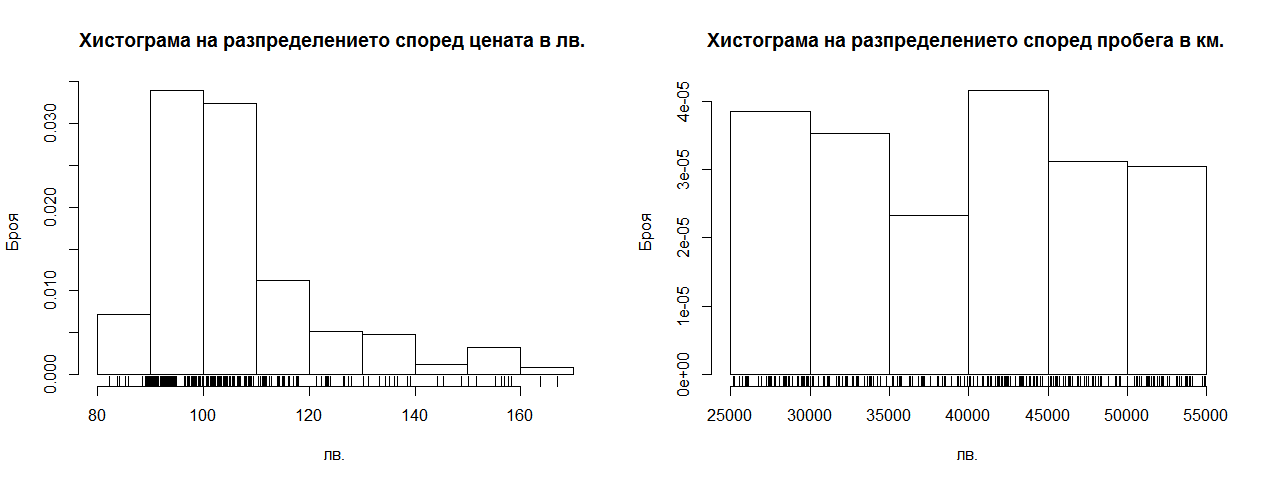
par(mfrow = c(1, 2))

hist(X4, probability=TRUE, right = FALSE, main = "Хистограма на разпределението според цената в лв. ", xlab = "лв.", ylab = "Броя")

rug(jitter(X4))

hist(X5, probability=TRUE, right = FALSE, main = "Хистограма на разпределението според пробега в км. ", xlab = "лв.", ylab = "Броя")

rug(jitter(X5))



Отново при забелязваме дясна асиметрия при разпределението според цената и почти равномерно разпределение на гумите според техния пробег.

За да имаме еднакви скали по абсцисната ос първо е добре да центрираме и нормираме данните с функцията ***scale***.

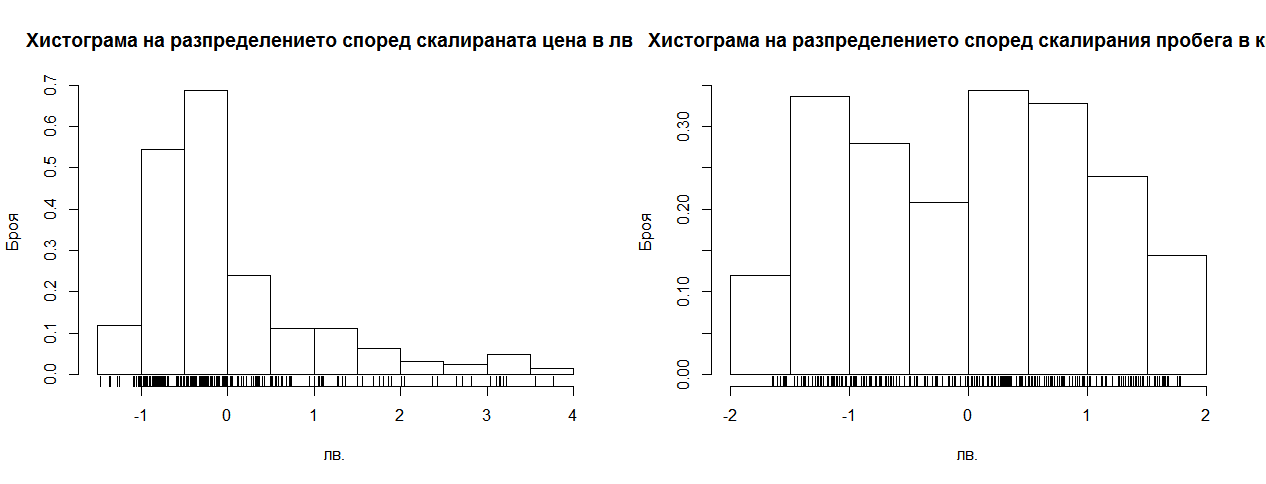
par(mfrow = c(1, 2))

hist(scale(X4), probability=TRUE, right = FALSE, main = "Хистограма на разпределението според скалираната цена в лв. ", xlab = "лв.", ylab = "Броя")

rug(jitter(scale(X4)))

hist(scale(X5), probability=TRUE, right = FALSE, main = "Хистограма на разпределението според скалирания пробега в км. ", xlab = "лв.", ylab = "Броя")

rug(jitter(scale(X5)))



Така успяваме да забележим не само дясната асиметрия на разпределението на гумите според цената, но и да оформим хипотезата, че това разпределение ще има тежка дясна опашка. При пробегът разпределението е по-скоро симетрично и почти равномерно.

*Сравняване чрез полигони.*

Можем да сравнявним две емпирични разпределения е чрез сравняване на техните полигони.

> par(mfrow = c(1, 2))

> tmp = hist(scale(X4), probability=TRUE, right = FALSE, main = "Хистограма на разпределението според скалираната цена в лв. ", xlab = "лв.", ylab = "Броя")

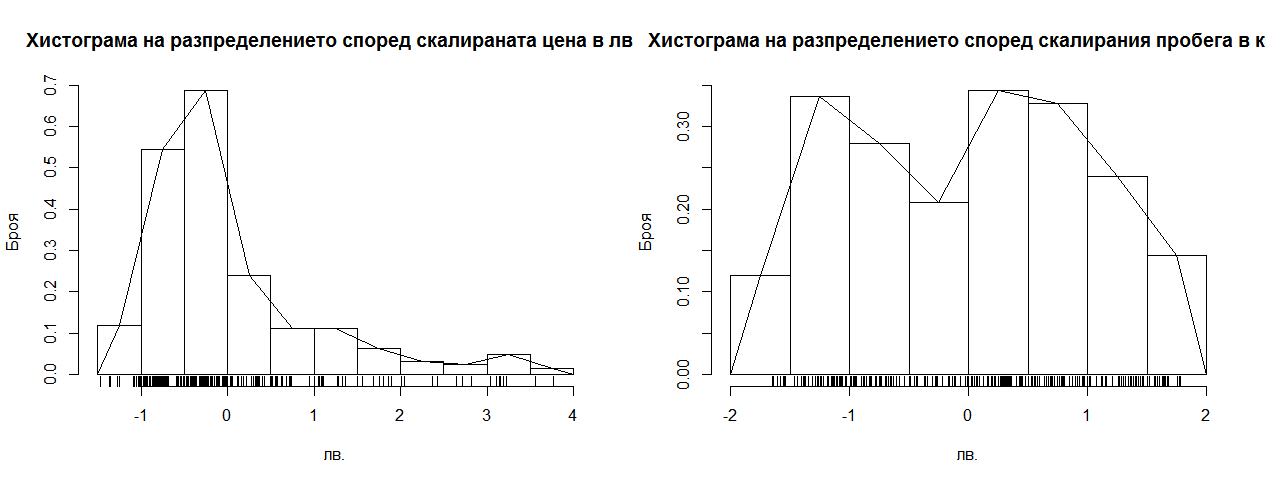
> rug(jitter(scale(X4)))

> lines(c(min(tmp$breaks), tmp$mids, max(tmp$breaks)), c(0, tmp$density, 0), type="l")

> tmp = hist(scale(X5), probability=TRUE, right = FALSE, main = "Хистограма на разпределението според скалирания пробега в км. ", xlab = "лв.", ylab = "Броя")

> rug(jitter(scale(X5)))

> lines(c(min(tmp$breaks), tmp$mids, max(tmp$breaks)), c(0, tmp$density, 0), type="l")



*Сравняване чрез плътности.*

Подобно сравнение може да бъде направено и с помощта на плътностите. Те се изчертаваха чрез функциите ***lines*** и ***density***, като преди това трябва да сме използвали функцията ***hist***. При оценяването на плътността на наблюдаваната величина се използва осредняване по подинтервали. Ширините на тези подинтервали се задават в параметъра ***bw***, чието съкращение идва от bandwidth. Колкото ширината на подинтервалите е по-голяма, толкова приближаващата плътността крива е по-гладка. Когато правим сравнения трябва да използваме един и същ параметър ***bw***. (Самият алгоритъм за построяването на гладката крива, понякога е доста сложен).

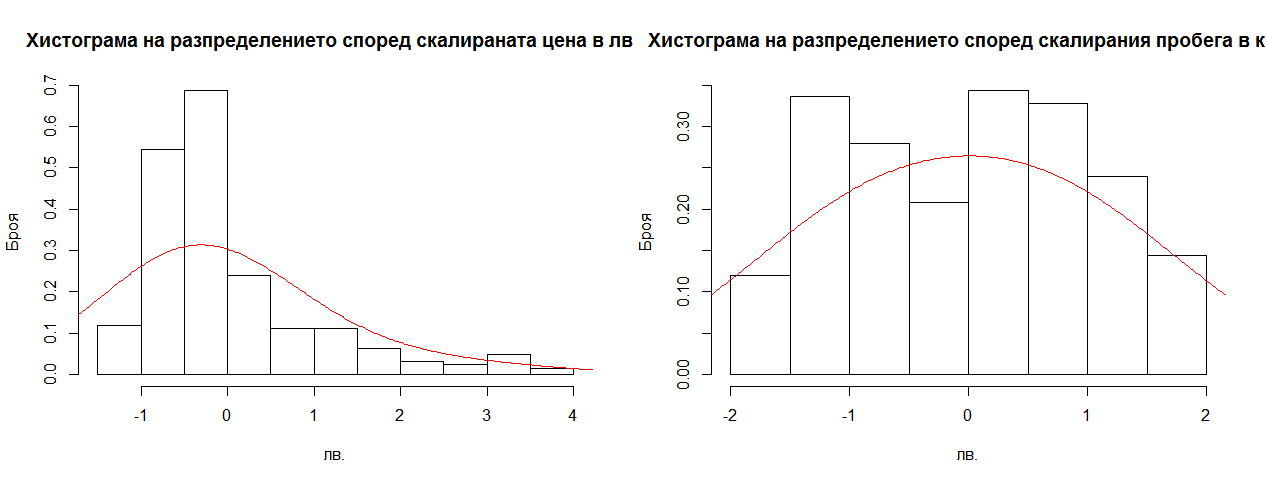
> par(mfrow = c(1, 2))

> hist(scale(X4),prob = T, main = "Хистограма на разпределението според скалираната цена в лв. ", xlab = "лв.", ylab = "Броя")

> lines(density(scale(X4), bw=1), col='red')

> hist(scale(X5),prob = T, main = "Хистограма на разпределението според скалирания пробега в км. ", xlab = "лв.", ylab = "Броя")

> lines(density(scale(X5),bw=1), col='red')



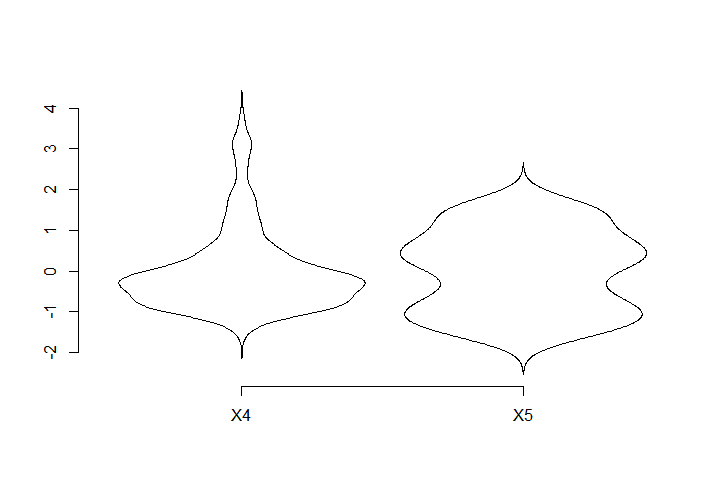
Към края на нашия курс ще научим и начини за сравняване на теоретични и емпирични разпределения.

За да се подчертаят различията във плътността тя може да бъде изчертана огледално поотделно за всяка наблюдавана величина. Това става с функцията ***simple.violinplot.***

> simple.violinplot(scale(X4),3 xaxt="n")

След това добавяме новите имена

> axis(side=1,at=c(1,2), labels = c("X4","X5"))



1. **Линейна регресия**

При наблюдаване на два количествени признака често възниква въпроса за моделиране на формата на зависимост между тях. Той може да бъде решен със средствата на Регресионния анализ. Това е метод, при който ако знаем Х можем да предскажем Y с известна грешка.

Ако факторпризнакът е един, говорим за единична регресия. Иначе говорим за множествена регресия.

Регресионният анализ започва с избор на линия на регресия. Да приемем, че нейното аналитично представяне е

Y = f(X, β) + ε,

където β е r+1-мерен вектор, чиито координати са неизвестни параметри на функцията f, а ε e стохастичната грешка с Еε = 0.

Х се нарича независима променлива или фактор-признак, а Y се нарича резултативна величина или зависима променлива. При различните наблюдения имаме реализации на тези случайни величини.

Наблюденията трябва да са независими и извършени при непроменени условия на експеримента. Грешките трябва да са случайни величини, които са некорелирани с Еε = 0 и Dε = σ2.

Ако функцията f е линейна относно неизвестните параметри, но не обезателно линейна относно независимите променливи, говорим за линеен регресионен модел. Иначе моделът се нарича нелинеен.

**Оценката на линията на регресията се прави в клас от функции.** И това е онази функция g от разглеждания клас, която минимизира средноквадратичната грешка на Y относно g(X) в класа от функции G.

Да припомним, че ***средноквадратична грешка на η относно g(ξ)*** означаваме с ***MSE(g)*** и това е

***MSE(g)*** *= Е(η - g(ξ))2.*

По тази причина първата задача на регресионния анализ е да се построят най-добри точкови и интервални оценки на параметрите на регресия така, че измежду всички линии с това аналитично представяне при получените оценки на параметрите да имаме най-малка сума от квадратите на грешките.

По данните от извадката, използвайки метода на най-малките квадрати, правим оценка на

вектора *β*. Ще я означаваме с . Тя е такава, че да минимизира



Намира се като решим относно *β*, следната система

,

наречена ***система нормални уравнения***.

От полученото уравнение на регресия пресмятаме оценки на стойностите на зависимата

променлива. Тези оценки ще означаваме с , т.е.



След като се определят оценките на параметрите в избрания модел се прави анализ на остатъците. Това става най-бързо от тяхната диаграма на разсейване. По-точно проверява се дали отклоненията на фактическите стойности от техните оценки имат случаен характер. Дали тези остатъци са еднакво разпределени. Дали имат равни дисперсии. С някои от критериите за съгласие, които ще разгледаме по-нататък се проверява дали разпределението им е нормално. Проверява се хипотезата за липса на корелация в остатъчния компонент.

Ако тези условия са удовлетворени се прави проверка на хипотезата за статистическата значимост на коефициентите в уравнението на регресия.

Величината  се нарича ***стандартна грешка на модела***.

Може да тестваме повече от една функция *f*. При всяка от тях ще получаваме различни оценки . Най-добър модел за съответните данни ни дава тази линия, за която сумата от квадратите на отклоненията на фактическите (измерените) значения на резултативната величина *Y* от техните оценки  е минимална. Т.е. моделът с най-малка стандартна грешка е най-подходящ за нашите данни.

След намирането на уравнението на регресия можем да получим най-добра оценка за *Y* по зададено значение на *X*.

Този анализ не се отчита, че изменението на разглежданите величини може да се дължи на външни, невключени в модела признаци, но измерва силата на зависимост между включените в модела фактори.

***Проста линейна регресия – в учебника на този етап има само тази регресия***

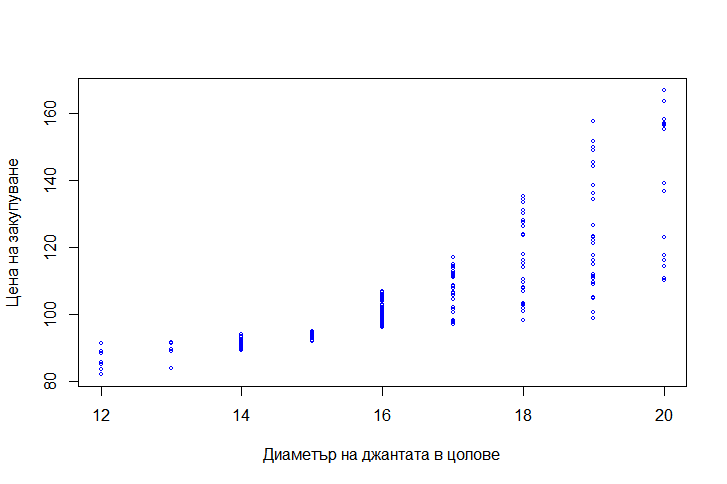
Нека изследваме влиянието на фактора *Х* върху резултативния признак *Y*.

*Изчертаване на корелационно поле на данните*

Обикновено се започва с изчертаване на корелационно поле на данните. По абсцисната ос се нанасят измерените значения на факторпризнака *X*, а по ординатната, измерените значения на резултативния признак *Y*. Да предположим, че разполагаме с *n* на брой двойки от наблюдения *(X1,Y1), (X2,Y2),… ,(Xn,Yn).*

Нека за по-голяма конкретност да използваме данните за от таблицата tires и да моделираме влиянието на диаметъра(X9) на джантата в цолове върху цената на гумата (Х4). Изчертаваме на корелационно поле на данните

> plot(X9, X4, xlab=" Диаметър на джантата в цолове ", ylab=" Цена на закупуване ", col="blue", cex =0.5)



*Избор на регресионен модел*

Т.к. точките се групират около права линия избираме регресионен модел

*Yi = β0 + β1 Xi + εi,*

където *β0* и *β1* са неизвестни параметри, а *εi* са стохастичните грешки, които трябва да са некорелирани с *Еεi = 0* и *Dεi = σ2*.

Да отбележим, че линията на регресия е *Yi = β0 + β1 Xi.*

*Забележка:* В случая, от графиката е ясно, че последното условие не е изпълнено, защото дисперсията на цената нараства с нарастването на диаметъра. Нататък алгоритъмът дава добри резултати само ако предните условия са удовлетворени. Т.е. за да имаме добри резултати трябва да трансформираме по подходящ начин данните. Това обаче ще направим на по-късен етап от нашето обучение. Т.е. за моите данни простата линейна регресия не е подходящ модел.

*Определяне на коефициентите в линията на регресия*

Най-добрите оценки на *β0* и *β1* се получават по метода на най-малките квадрати. Те са такива, че минимизират сумата от квадратите на отклоненията

.

Ако оценъчните стойности са , то те трябва да са такива, че минимума на горната сума да е



Да означим тези оценки с  и . Намираме ги от системата нормални уравнения, която в случая има вида

.

Решението на тази система е:



Ако гледаме на тези оценки като на случайни величини, те са неизместени и имат минимална дисперсия в сравнение с всички неизместени оценки, линейно зависещи от *Y1, Y2, …, Yn*.

Оценката на уравнението на регресия е



Тя минава през точката с координати (, ).

Коефициентът  в това уравнение показва, с колко единици средно, в приетата за резулта-тивния признак *Y* мярка, би се изменил той, ако изменим факторпризнака *Х* с една единица в прие-тата за него мярка. Когато зависимостта на резултативния признак от факторпризнака е правопро-порционална, коефициентът  е положителен. Обратно, ако тази зависимост е обратнопропор-ционална, този коефициент е отрицателен. Коефициентът  е равен на ординатата на точката, в която линията на регресия пресича ординатната ос. Линията на регресия ще е успоредна на абсцисната ос ако значенията на резултативния признак не се влияят от тези на факторпризнака.

R изчертава тези графики по следния начин. Първо, с помощта на функцията ***plot*** изчертава точките от корелационното поле както по-горе. После с помощта на функцията ***lm*** намира стойностите на коефициентите  и . Тази функция дава възможност с помощта на оператора “~” да се зададе функция, която описва регресионния модел. Отново от ляво на този оператор стои зависимата променлива (в случая Х9), а отдясно независимата променлива (в случая това е Х4). По-подробна информация за оценките на коефициентите можем да получим чрез функцията ***summary***. След това се добавя линията на регресия към графиката с корелационното поле на данните. Последното може да стане с помощта на функциите ***abline*** или ***lines***.

> plot(X9, X4, xlab=" Диаметър на джантата в цолове ", ylab=" Цена на закупуване ", col="blue", cex =0.5)

> My\_model = lm(X4 ~ X9)

> summary(My\_model)

Call:

lm(formula = X4 ~ X9)

Residuals:

Min 1Q Median 3Q Max #това са мин, I кв., мед., II кв., макс. на остатъците епсилон

-24.660 -4.166 -1.341 2.591 36.530

Coefficients:

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

(Intercept) -5.3225 5.2314 -1.017 0.31 #това са оценката , стандартната

#й грешка (обяснена е по-долу), и

# p-value то, т.е. лицето под

# съответната нормалната крива

# и извън ± t value то.

X9 6.7896 0.3171 21.415 <2e-16 \*\*\*#това са оценката , стандартната

#й грешка(обяснена е по-долу), и p-

# value то, т.е. лицето

#под съответната нормалната крива

# и извън ± t value то.

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 # това са кодовете, с които се четат

# резултатите от проверките за ста-

# тистическите значимости на

# коефициентите по-горе.

Residual standard error: 9.624 on 248 degrees of freedom #Това е стандартната грешка на оста-

# тъците, т.е. S на епсилоните

# (обяснена е по-долу)

Multiple R-squared: 0.649, Adjusted R-squared: 0.6476 #Toва е коефициентът на детерми-

# нация(определеност), т.е. квадра-

# тът на корелационния коефициент

# между X4 и X9, обяснен по-долу и

# Показваш каква част от вариацията

# на зависимата променлива X4 се

# определя от изменения в незави-

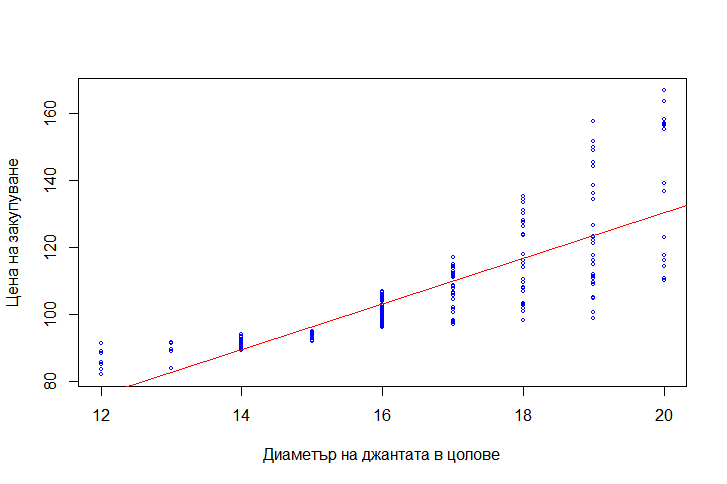
# симата променлива X9.

F-statistic: 458.6 on 1 and 248 DF, p-value: < 2.2e-16 #Това е резултатът от критерия на

# Фишер, който също е обяснен по-

# долу

>  abline(lm(X4 ~ X9), col="red")



Като алтернатива на съвкупността от горни функции може да бъде използвана функцията ***simple.lm*** от библиотеката ***UsingR.*** По подразбиране тя връща оценките на коефициентите и горната графика в черно.

> require(UsingR)

> simple.lm(X9,X4)

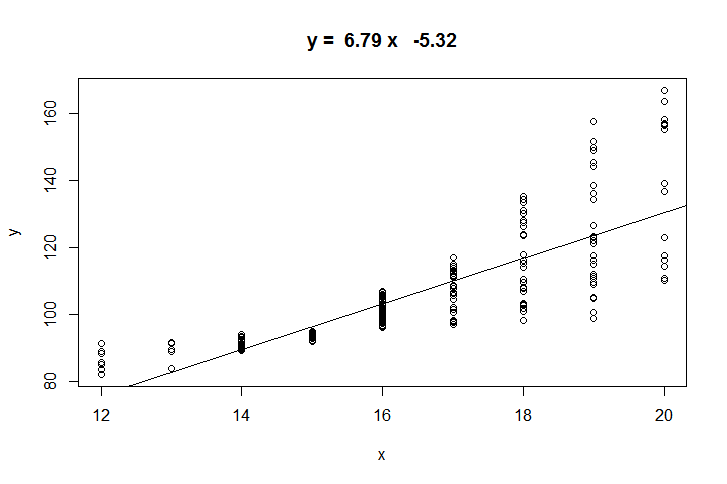
Call:

lm(formula = y ~ x)

Coefficients:

(Intercept) x

-5.323 6.790



За да извлечем например коефициентите в уравнението на регресия първо трябва да присвоим резултата от тази функция на променлива и после да вземем само компонентата, която съдържа коефициентите. Това можем да направим с $coefficients или с помощта на функцията ***coef***. Всички компоненти на резултата могат да бъдат разгледани с функцията ***ls.***

> Myresult = simple.lm(X9,X4)

> coef(Myresult) # извеждаме коефициентите от уравнението на регресия

(Intercept) x

-5.322501 6.789611

> ls(Myresult) # разглеждаме компонентите на резултата от регр. анализ, направен с simple.lm

[1] "assign" "call" "coefficients" "df.residual"

[5] "effects" "fitted.values" "model" "qr"

[9] "rank" "residuals" "terms" "xlevels"

> Myresult$coefficients # извеждаме коефициентите от уравнението на регресия

(Intercept) x

-5.322501 6.789611

> Myresult$coefficients[1] # извеждаме първия от коефициентите от уравнението на регресия

(Intercept)

-5.322501

> Myresult$coefficients[2] # извеждаме втория от коефициентите от уравнението на регресия

x

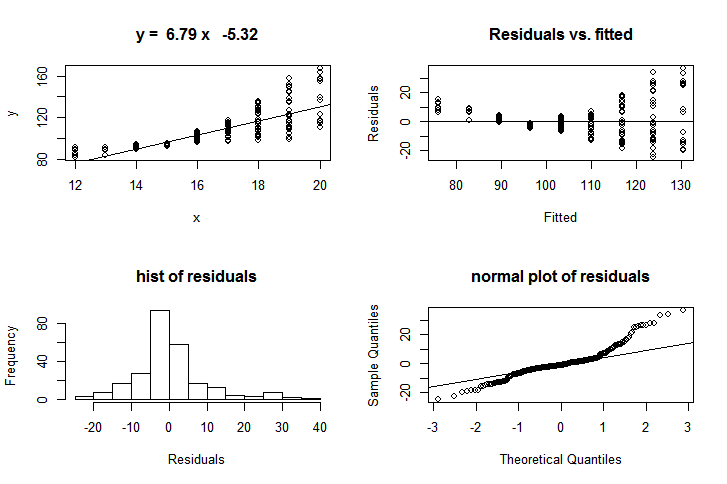
6.789611

Като заместим измерените значения на факторпризнака *Х*, в уравнението на регресия, намираме съответните оценки  за значенията на резултативния признак *Y*. Сумата и съответно средната аритметична на тези оценки е равна на съответната характеристика на изходните данни.

*Графика и анализ на остатъците*

Една твърде полезна графика, по която можем да преценим дали един регресионен модел е подходящ за данните е графиката на остатъците. Тя, заедно с корелационното поле на данните, хистограмата на остатъците и normal plot на остатъците може да бъде получена, когато във функцията ***simple.lm*** зададем параметърът ***show.residuals*** да бъде TRUE.

> Myresult = simple.lm(X9,X4,show.residuals=TRUE)



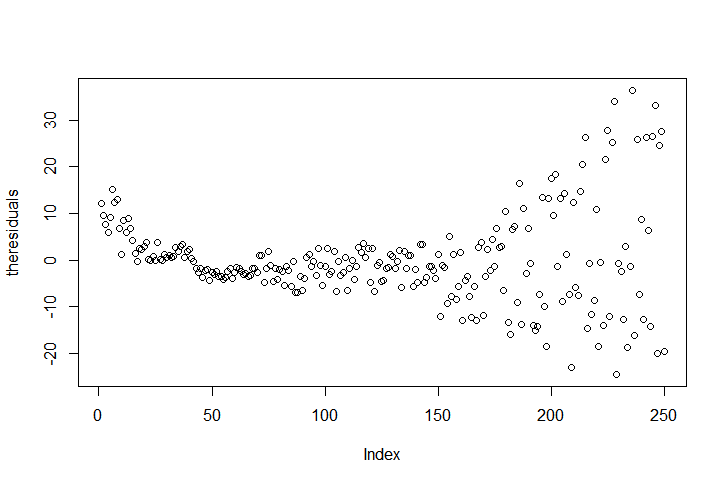
Първата графика е корелационното поле на данните заедно с линията на регресия. Втората графика на първия ред е графиката на остатъците и тук наблюдаваме, че те не са независими и нямат постоянна дисперсия, т.е. би трябвало да трансформираме данните по подходящ начин и тогава да приложим простата линейна регресия или изцяло да сменим модела. Първата графика на втория ред е хистограмата на остатъците. В идеалният случай тя би била близка до съответната нормална крива. В случай имаме малко по-голяма изостреност, която води до по-тежки опашки на разпределението. Това още веднъж показва, че преди да приложим простата линейна регресия е добре да трансформираме данните или да сменим простата линейна регресия с друг модел. Втората графика на втория ред е normal plot на остатъците. Ако остатъците са нормално разпределени точките от тази графика лежат точно върху ъглополовящата на първи квадрант. Непрекъснатата линия е регресията за normal plot на остатъците. В случая тя не съвпада с ъглополовящата на първи квадрант. Това още веднъж показва, че регресионният модел не е добър за данните. Отклоненията в началото и в края показват, че разпределението на остатъците има опашки, които са по-тежки от нормалните. Простият линеен регресионен модел не е най-добрият възможен за нашите данни.

Достъпът до остатъците на регресионния модел и тяхната графика могат да бъдат постигнати и с последователно използване на функциите ***resid*** и ***plot***.

> Myresult = simple.lm(X9,X4)

> theresiduals = resid(Myresult) # достига до остатъците

> plot(theresiduals)



Отново наблюдаваме изменяща се дисперсия и зависими данни, които говорят, че може да се построи по-добър модел. При нарастваща дисперсия е добре да се логаритмува зависимата променлива (в случай цената) преди да се приложи простият линеен регресионен анализ.

*Обща стандартна грешка на модела*

За да можем да направим статистически изводи за *β0* и *β1* и *Y* първо трябва да оценим дисперсията *σ2* на грешката и после да опишем разпределението на грешката. От теорията на общите линейни модели, най-добра неизместена оценка за *σ2* е

.

Тук r е броят на неизвестните параметри в модела. В случая те са два *β0* и *β1.*

Тази величина се нарича ***среден квадрат на грешката***.

,

се нарича ***обща стандартна грешка*** на модела.

Ако  значи имаме пълно съвпадение на изходните данни с техните оценки.

В нашия пример, с помощта на функцията ***summary*** получихме, че

Residual standard error: 9.624 on 248 degrees of freedom #Това е стандартната грешка на оста-

# тъците, т.е. S на епсилоните

Мярката на грешката е същата както на зависимата променлива, в случая лева. Тук стандартната грешка на остатъците е 9.624 лв.

*Проверка за хипотезата за адекватност на модела (Това е само за теб. В учебника още го няма)*

Често пъти резултатите от изследването се оформят в таблица от вида:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Източник на дисперсията* | *Сума от квадратите*  *(Девиация)* | *Степени на свобода* | *Дисперсия* | *F - критерий* |
| *Регресия* |  | *1* |  | *F емп =* |
| *Отклонение от регресията* |  | *n - 2* |  |  |
| *Общо:* |  | *n - 1* |  |  |

За сумата от квадратите, обословена от регресията е вярно следното съотношение

.

В случая на нормално разпределени грешки можем да направим проверка на хипотезата за адекватност на тествания модел. Един от начините е да проверим хипотезата

Н0: *β1 = 0*

срещу алтернативата

Н1: *β1 ≠ 0.*

Избираме риска за грешка от първи род *α ∈ (0, 1)*.

Критичната област за нулевата хипотеза има вида



където *Сα* е *1- α*  квантилът на *F(1, n - 2).*

Когато пресметнатото от данните отношение е по-голямо от теоретичния квантил *Сα*

това означава, че по-горното неравенство е удовлетворено, т.е. извадката е в критичната област за нулевата хипотеза, т.е. отхвърляме нулевата хипотеза и в случая това означава, че приемаме алтелнативната *β1 ≠ 0* и моделът е адекватен.

В нашия пример, с помощта на функцията ***summary*** получихме, че

F-statistic: 458.6 on 1 and 248 DF, p-value: < 2.2e-16

Т.е. стайността на p-value, която съответства на F-statistic е много малка и е по-малка от риска за грешка от първи род α, т.е. трябва да отхвърлим нулевата хипотеза и моделът е адекватен в смисъл, че *β1* е статистически значимо различен от 0, т.е. зависимостта на зависимата променлива (Цена) от независимата променлива (размер в цолове) е статистически значима.

*Проверка на други хипотези, свързани с модела и построяване на доверителни интервали на коефициентите и на оценките на зависимата променлива.*

Вече можем да проверим допълнителни хипотези. Например можем ли да закръглим кофициентите си. В този случай в числителя на емпиричната характеристика имаме разликата на оценката и тестватната константа, а в знаменателя стои стандартната грешка на оценката, която може да бъде намерена в следващата таблица. Костантата *Сα* е *1 - α /2* квантилът на *t(n - 2).*

Можем да построим и доверителни интервали на *β0* и *β1* и *Y.*

|  |  |  |  |
| --- | --- | --- | --- |
| *Величина* | *Стандартна грешка*  Std. Error | *Степени на свобода* | *Граници на доверителния интервал* |
| *β0* |  | *n - 2* |  |
| *β1* |  | *n - 2* |  |
| *Средното значение на Y (т.е. ако независимите променливи не са случайни)* |  | *n - 2* |  |
| *Средното значение на Y (т.е. ако независимите променливи са случайни)* |  | *n - 2* |  |

В нашия пример, с помощта на функцията ***summary*** получихме, че

Coefficients:

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

(Intercept) -5.3225 5.2314 -1.017 0.31 #това са оценката , стандартната

#й грешка (обяснена е по-долу), и

# p-value то, т.е. лицето под

# съответната нормалната крива

# и извън ± t value то.

X9 6.7896 0.3171 21.415 <2e-16 \*\*\*#това са оценката , стандартната

#й грешка(обяснена е по-долу), и p-

# value то, т.е. лицето

#под съответната нормалната крива

# и извън ± t value то.

Замествайки оценките и техните стандартни грешки лесно можем да получим границите на доверителните интервали за коефициентите.

При 5% реск за грешка доверителният интервал за *β0 е*

-5.3225 ± 1.96\*5.2314

-5.3225 ± 10.25354

(-15.57604; 4.93104)

Т.е. този доверителен интервал е доста широк и е удачно да се провери хипотезата дали не можем да приемем, че този коефициент е нула.

При 5% реск за грешка доверителният интервал за *β1 е*

6.7896 ± 1.96\*0.3171

6.7896 ± 0.621516

(6.168084; 7.411116)

Т.е. можем да очакваме, че с нарастване на диаметъра на гумите с 1 цол, тахното цена ще нарасне от 6.168084 лв. до 7.411116 лв.

*Корелационен коефициент*

При достатъчно голям брой опити **. Ето защо *Sε* често се приема за стандартна грешка на оценката . Ако отнесем тази грешка към стандартното отклонение на данните от извадката, отнасящи се за резултативния признак, ще получим величина, която е *0* при пълно съвпадение, т.е. при функционална зависимост между *Х* и *Y* и е *1* ако оценките на *Y* не се влияят от *Х*. В последния случай всички оценки на резултативния признак ще са равни помежду си и по тази причина ще са равни на своята средна аритметична и на средната аритметична на изходните данни за този признак. На основата на тези разсъждения е образуван корелационния коефициент на Пирсън



Той се изменя от *0* до *1*. За посоката на зависимостта се съди по знака на регресионния коефициент .

В нашия пример, с помощта на функцията ***summary*** получихме, че

Multiple R-squared: 0.649, Adjusted R-squared: 0.6476

Т.е. 64,76% от вариацията на зависимата променлива – Цена е определена от вариацията на независимата променлива – Размер на гумите. Т.е. изменението на размера на гумите значително влияе на изменението на цената им. Това е в синхрон с извода направен при проверката за адекватност.

cor(X4,X9)

[1] 0.8056144

cor(X4,X9)^2

> cor(X4,X9)^2

[1] 0.6490146

***Въпроси:***

1. По какво се различават факторпризнака и резултативния признак? Ще се промени ли извода от регресионния анализ ако сменим местата им? Винаги ли можем да сменим местата им?

2. Какъв е смисълът на коефициента  в уравнението на изглаждащата права и как се намира самия коефициент?

3. С какво се различава изглаждащата права от всички останали прави, които можем да прекараме между точките от корелационното поле?

4. Кои са логическите обосновки, които ни дават основание да използваме корелационния коефициент на Пирсън за измерител на силата на зависимостта между наблюдаваните признаци?

*Рангов корелационен коефициент на Спирмън*

Да предположим, че над единиците от съвкупността са извършени наблюдения, върху два признака, измерени на рангова скала. Т.е. наблюденията са подредени по големина и на най-голямата наблюдение е даден ранг 1, на следващото ранг 2 и т.н.

Например

> rank(c(6,5,7,2,3))

[1] 4 3 5 1 2

Ако имаме две равни по големина наблюдения те имат равни рангове и стойността им е средното аритметично на ранговете, които биха имали ако наблюденията бяха различни по стойност.

> rank(c(6,6,5,7,2,3))

[1] 4.5 4.5 3.0 6.0 1.0 2.0

***Спирмън*** *използва като измерител на силата на зависимостта между наблюдаваните признаци - близостта на ранговете, и по-точно сумата от квадратите на разликите им.* Ако съществува силна положителна зависимост между ранговете на единиците, те би трябвало да съвпадат и сумата от квадратите на разликите им би била нула. Ако зависимостта е силна отрицателна, ранговете ще са подредени в обратен ред. Разликите им в този случай, ако *n* е четно, ще образуват редица само от нечетните числа от – *(n-1)* до *(n-1)* или ако *n* е нечетно, само от четните числа в този интервал. Тогава сумата от квадратите им ще е *.*

При липсата на каквато и да е зависимост можем да приемем, че тази сума ще е средното аритметично на двете крайни възможности, т.е. 

Като отнесем тази величина към действителната сума от квадратите на разликите, т.е. 

получаваме измерител на зависимостта, който би бил нула при силна правопропорционална зависимост между ранговете. Ето защо ***ранговият коефициент на корелация на Спирмън*** се пресмята по формулата



Той може да се определи с помощта на функциите ***cor*** и ***rank*** в R.

> cor(rank(X4),rank(X9))

[1] 0.9004327