Проста линейна регресия

Задача 1: Твърди се, че съществува следната връзка между максималния пулс на сърцето и възрастта:

Пулс=220-възраст

15 човека на различна възраст са тествани за максимален пулс. Получени са следните наблюдения:

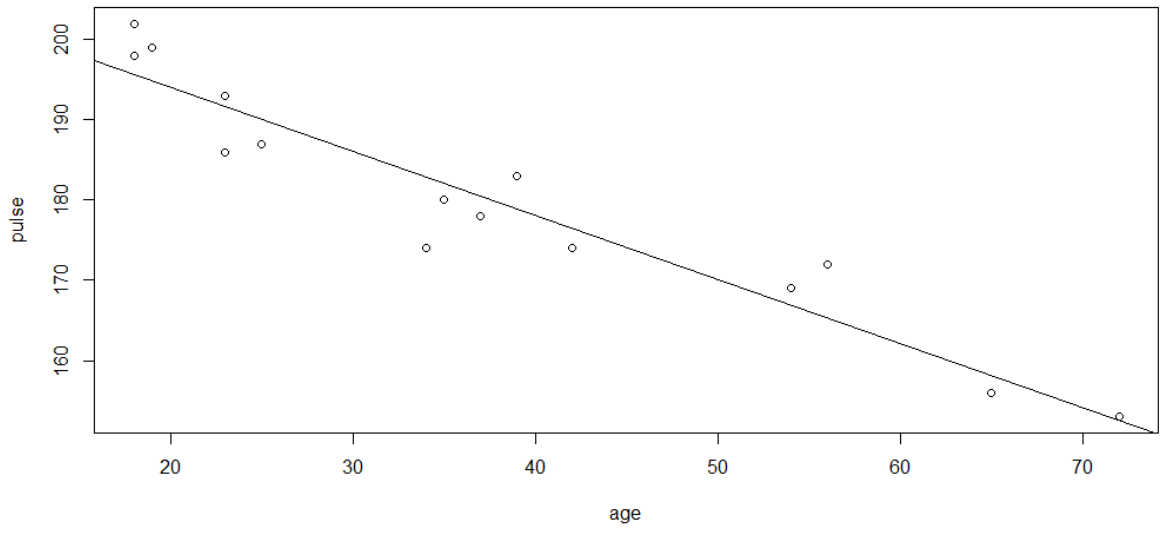
Възраст 18 23 25 35 65 54 34 56 72 19 23 42 18 39 37

Пулс 202 186 187 180 156 169 174 172 153 199 193 174 198 183 178

Постройте линеен модел, описващ данните. Проверете дали коефициентите в този модел съвпадат с първоначалната хипотеза. Намерете максималния пулс, който би достигнал човек на възраст 30, 40 или 50 год. Постройте 90% доверителен интервал на този пулс.

Решение:

```
> age=scan()
1: 18 23 25 35 65 54 34 56 72 19 23 42 18 39 37
16:
Read 15 items
> pulse=scan()
1: 202 186 187 180 156 169 174 172 153 199 193 174 198 183 178
16:
Read 15 items
> plot (x,y)
> plot (age,pulse)
> l=lm(pulse~age)
> abline(l)
>
```



```
> summary(l)

Call:
lm(formula = pulse ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9258 -2.5383  0.3879  3.1867  6.6242

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
age          -0.79773    0.06996  -11.40 3.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared:  0.9091,    Adjusted R-squared:  0.9021
F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
```

P-value е много малко и при 2-та коефициента, оттук изглежда моделът да бъде добър, R-squared също е голям.

Първоначалната ни хипотеза е: Пулс=220-1*възраст

Различават се коефициентите, но въпросът е да оценим статистически тази грешка. Трябва да проверим първоначалната хипотеза.

Ho: Bo=220

H1:Bo!=220

```
> s=summary(l)
> s$coefficients
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 210.0484584 2.86693893  73.26576 2.124074e-18
age          -0.7977266 0.06996281 -11.40215 3.847987e-08
> ts=(s$coefficients[1,1]-220) / s$coefficients[1,2]
> ts
[1] -3.471138
> 2*pt(ts,df=13)
[1] 0.004136843
```

P-value < 0.05. Отхвърляме хипотезата и приемаме алтернативата, тоест Bo е различен от 220

```
> t=(s$coefficients[2,1] + 1) / s$coefficients[2,2]
> t
[1] 2.891157
> 2*pt(t,df=13, lower.tail=F)
[1] 0.01262031
```

Ho: Bo=1

H1:Bo!=1

0.012 < 0.05. Отхвърляме хипотезата.

За да направим предикшън ние трябва да използваме новите стойности(30,40,50) за възрастта и трябва новите стойности да са оформени в data.frame , който се казва по същия начин като името на променливата в модела

```
> d=data.frame(age=c(30,40,50))
> d
  age
1  30
2  40
3  50
> predict(l,d)
       1        2        3
186.1167 178.1394 170.1621
>
```

Показва при 30, 40 и 50 години колко ще бъде максималният пулс

```
> predict(l,d,interval='confidence')
       fit      lwr      upr
1 186.1167 183.3330 188.9004
2 178.1394 175.5543 180.7245
3 170.1621 166.9706 173.3537
```

Стойността на 30 годишните, които ще имат максимален пулс е 186.1176, а доверителният интервал е lwr и upr

В predict имаме 2 типа доверителен интервал -„confidence“, “prediction”

При confidence, то дава доверителен интервал за средната стойност

prediction дава доверителен интервал за конкретна стойност

Пр: За 40 г. интервалът prediction Казва, ако вземем 1 човек от колко до колко би попаднала неговия пулс.

Пр: При confidence, тоест ако имаме много хора на възраст 40 и на тези всички хора сметнем средното аритметично на пулса