Seminarska naloga iz Linearne Regresije

Verjetnost in Statistika

Luka Dragar

9.jan.2022

Linearna regresija: dolzina ribe

1. Opis podatkov

Zbrali smo meritve radija lusk in telesne dolzine na vzorcu 38 enoletnih maloustih basov (lat. Micropterus dolomieu). Podatke smo zapisali v dokument, ki ima 2 stolpca:

1. *dolzina* je numericna zvezna spremenljivka, ki predstavlja dolzino telesa (v milimetrih).
2. *rlusk* je numericna zvezna spremenljivka, ki predstavlja radij lusk (v milimetrih).

Baza podatkov se imenuje *jezero.csv*. Najprej bomo prebrali podatke v R in zatem pogledali strukturo podatkov

```
jezero<-read.csv("/Users/carbs/Downloads/jezero.csv", header=TRUE)
str(jezero)
```

```
## 'data.frame':    38 obs. of  2 variables:
##  $ rlusk  : num  1.9 1.9 1.1 1.3 1.6 1.9 1.1 1.5 1.6 1.5 ...
##  $ dolzina: int  71 64 57 68 72 80 55 75 75 71 ...
```

```
dolzina<-jezero$dolzina
rlusk<-jezero$rlusk
```

2. Opisna statistika

Zdaj bomo izračunali opisno statistiko za naše podatke – povzetek s petimi števili (minimum, maksimum, prvi in tretji kvartil, mediano), vzorčni povprečji in vzorčna standardna odklona dolzine in radija lusk.

```
summary(dolzina)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.00   71.25  106.50   98.34  120.00  165.00
```

```
sd(dolzina)
```

```
## [1] 28.42941
```

Opazimo, da dolzina vzorca rib varira od 55.00mm do 165.00mm, s povprečjem 98.34 in standardnim odklonom 28.42941 mm. Ponovimo postopek računanja za vzorec radija lusk.

```
summary(rlusk)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.100   1.650   2.450   2.387   2.900   4.900
```

```
sd(rlusk)
```

```
## [1] 0.8469886
```

Opazimo, da radija lusk vzorca rib varira od 1.100mm do 4.900mm, s povprečjem 2.387 in standardnim odklonom 0.8469886 mm$.

Summary izracunamo se za transformirane podatke log()

```
summary(log(dolzina))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.007   4.266   4.668   4.546   4.787   5.106
```

```
sd(log(dolzina))
```

```
## [1] 0.3008392
```

Opazimo, da log(dolzina) vzorca rib varira od 4.007mm do 5.106mm, s povprečjem 4.546 in standardnim odklonom $ 0.3008392$ mm.

Ponovimo postopek računanja za vzorec log(radija lusk).

```
summary(log(rlusk))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09531 0.49945 0.89588 0.80701 1.06471 1.58924
```

```
sd(log(rlusk))
```

```
## [1] 0.3654564
```

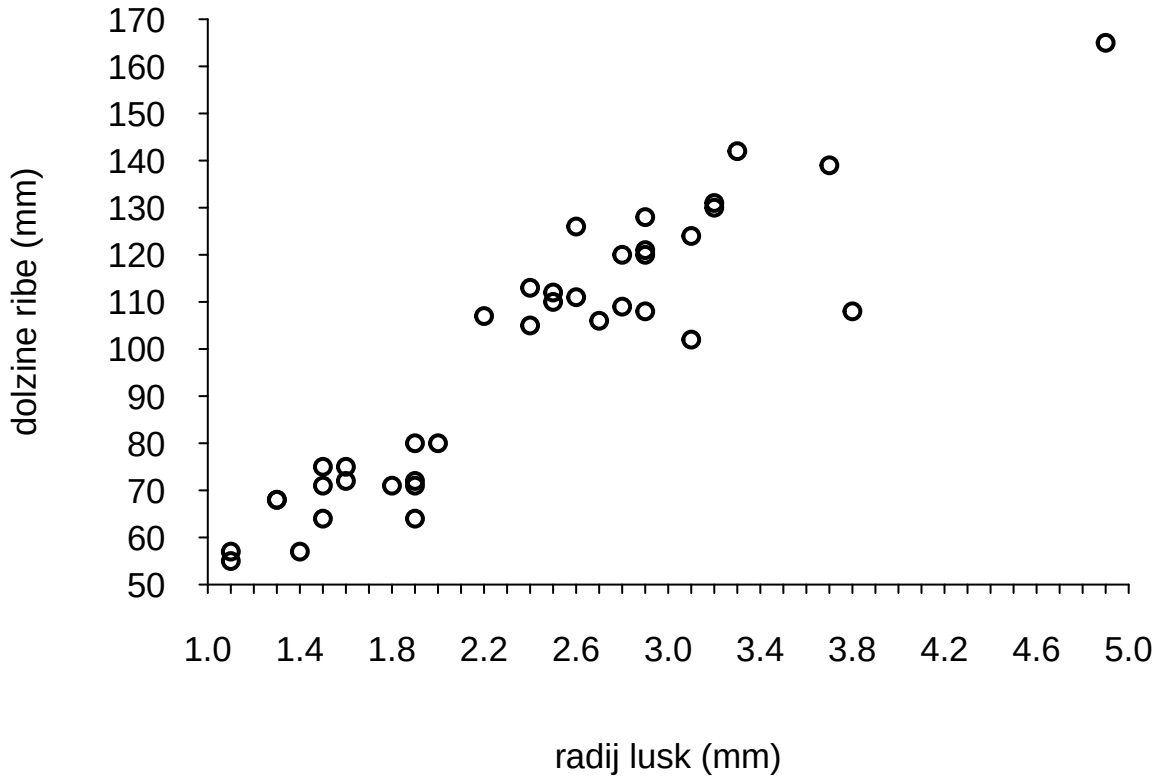
Opazimo, da log(radija lusk) vzorca rib varira od 0.09531mm do 1.58924mm, s povprečjem 0.80701 in standardnim odklonom 0.3654564 mm$.

Razpon vrednosti dolzine in radija lusk nam pomaga pri izbiri mej na oseh razsevnega diagrama.

3. Razsevni diagram in vzorčni koeficient korealcije

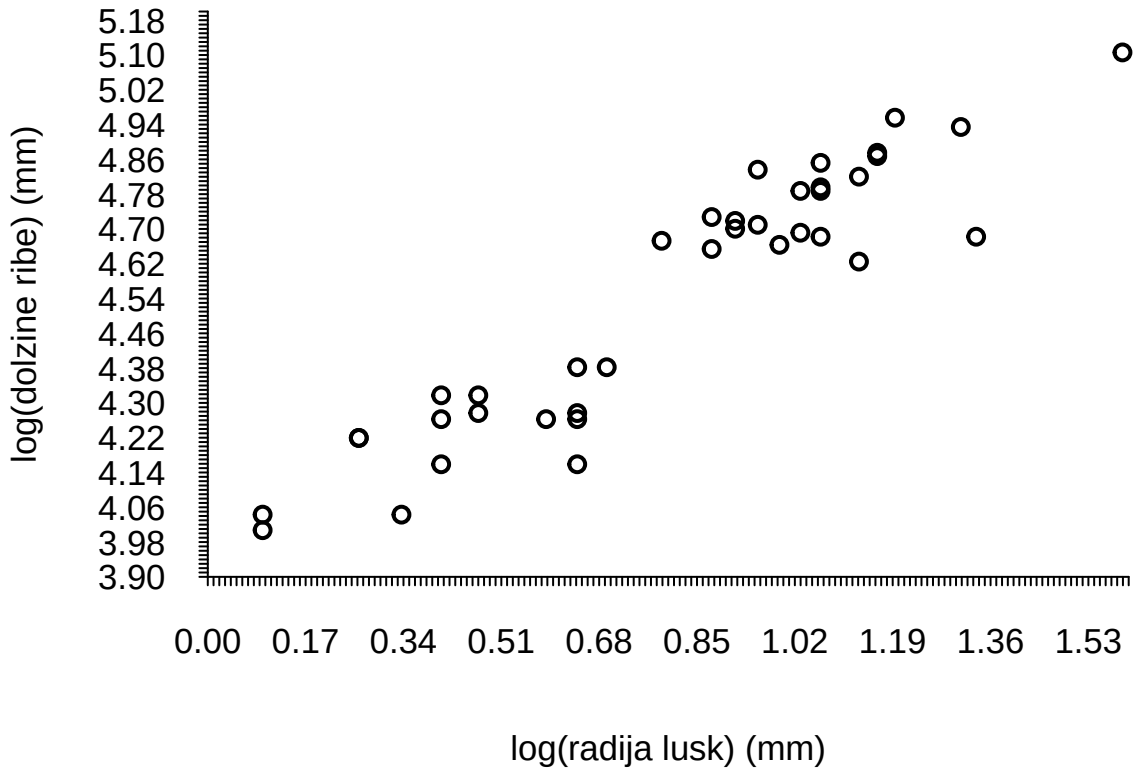
Prikažimo dobljene podatke na razsevnem diagramu.

```
par(las=1, cex=1.1, mar=c(4,4,2,2))
plot(rlusk,dolzina, main="", ylim=c(50,170), xlim=c(1,5),
ylab="dolzine ribe (mm)", xlab="radij lusk (mm)", lwd=2, axes=FALSE)
axis(2,pos=1,at=seq(50,170,by=10),tcl=-0.2)
axis(1,pos=50,at=seq(1,5,by=0.1),tcl=-0.2)
arrows(x0=5,y0=1,x1=5.1,y1=1,length=0.1)
arrows(x0=50,y0=5,x1=50,y1=5.1,length=0.1)
```



Prikažimo transformirane podatke na razsevnem diagramu.

```
par(las=1, cex=1.1, mar=c(4,4,2,2))
plot(log(rlusk),log(dolzina), main="", ylim=c(3.9,5.2), xlim=c(0,1.6),
ylab="log(dolzine ribe) (mm)", xlab="log(radija lusk) (mm)", lwd=2, axes=FALSE)
axis(2,pos=0,at=seq(3.9,5.2,by=0.01),tcl=-0.2)
axis(1,pos=3.9,at=seq(0,1.6,by=0.01),tcl=-0.2)
arrows(x0=1.6,y0=0,x1=1.7,y1=0,length=0.1)
arrows(x0=3.9,y0=1.6,x1=3.9,y1=1.7,length=0.1)
```



Točke na razsevnem diagramu se nahajajo okoli namišljene premice, tako da linearni model zaenkrat izgleda kot primeren. Moč korelacije preverimo še z računanjem Pearsonovega koeficienta korelacije.

```
(r<-cor(rlusk,dolzina))
```

```
## [1] 0.9221425
```

```
(r<-cor(log(rlusk),log(dolzina)))
```

```
## [1] 0.9351461
```

Vrednost vzorčnega koeficienta korelacije je visoka ($r = 0.9221425$), kar govori o visoki linearni povezanosti radija lusk in dolzine ribe. Dalje, koeficient korelacije je pozitiven, kar pomeni, da imajo ribe z vecjo dolzino vecji radij lusk manjše dolzine pa manjsi radij lusk. Koeficient korelacije je sicer se kar visok, a v primerjavi s transformiranimi podatki($r = 0.9351461$) opazimo padec povezanosti nasih podatkov.

4. Formiranje linearnega regresijskega modela

Formirajmo linearni regresijski model.

```
(model<-lm(log(dolzina)~log(rlusk),data=jezero))
```

```
##
## Call:
## lm(formula = log(dolzina) ~ log(rlusk), data = jezero)
##
## Coefficients:
## (Intercept)   log(rlusk)
##      3.9244       0.7698
```

Dobili smo ocenjeno regresijsko premico $\hat{y} = 3.9244 + 0.7698x$, oziroma oceni odseka in naklona sta enaki $\hat{a} = 3.9244$ in $\hat{b} = 0.7698$.

5. Točke visokega vzvoda in osamelci

Identificirajmo točke visokega vzvoda in osamelce. Vrednost x je točka visokega vzvoda, če je njen vzvod večji od $\frac{4}{n}$.

```
jezero[hatvalues(model)>4/nrow(jezero),]
```

```
##    rlusk dolzina
## 3    1.1      57
## 7    1.1      55
## 19   4.9     165
```

Odkrili smo 3 točke visokega vzvoda. Dve ribi imata majhen radij lusk 1.1mm ena riba pa ima velik radij lusk 4.9mm.

Za podatke majhne in srednje velikosti vzorca je osamelec podatkovna točka, kateri ustreza standardizirani ostanek izven intervala $[-2, 2]$.

```
jezero[abs(rstandard(model))>2,]
```

```
##    rlusk dolzina
## 2    1.9      64
## 16   3.8     108
```

Dve podatkovni točki sta osamelca in se nanašata na dve ribi z nenavadno dolzino glede na njun radij lusk.

6. Preverjanje predpostavk linearnega regresijskega modela

Predpostavke linearnega regresijskega modela bomo preverili s štirimi grafi, ki se imenujejo diagnostični grafi (ali grafi za diagnostiko modela). Če neke predpostavke modela niso izpolnjene, so lahko ocene neznanih parametrov, p-vrednost testa, intervali zaupanja in intervali predikcije netočni.

```
par(mfrow=c(2,2),mar=c(4,3,2,1))
plot(model,which=1,caption="",ann=FALSE)
title(xlab=expression(italic(widehat(y)==widehat(a)+widehat(b)*x)),
ylab="Ostanki",main="Linearnost modela")
plot(model,which=2,caption="", ann=FALSE)
title(xlab="Teoretični kvantili", ylab= "St. ostanki",
main="Normalnost porazdelitve")
```

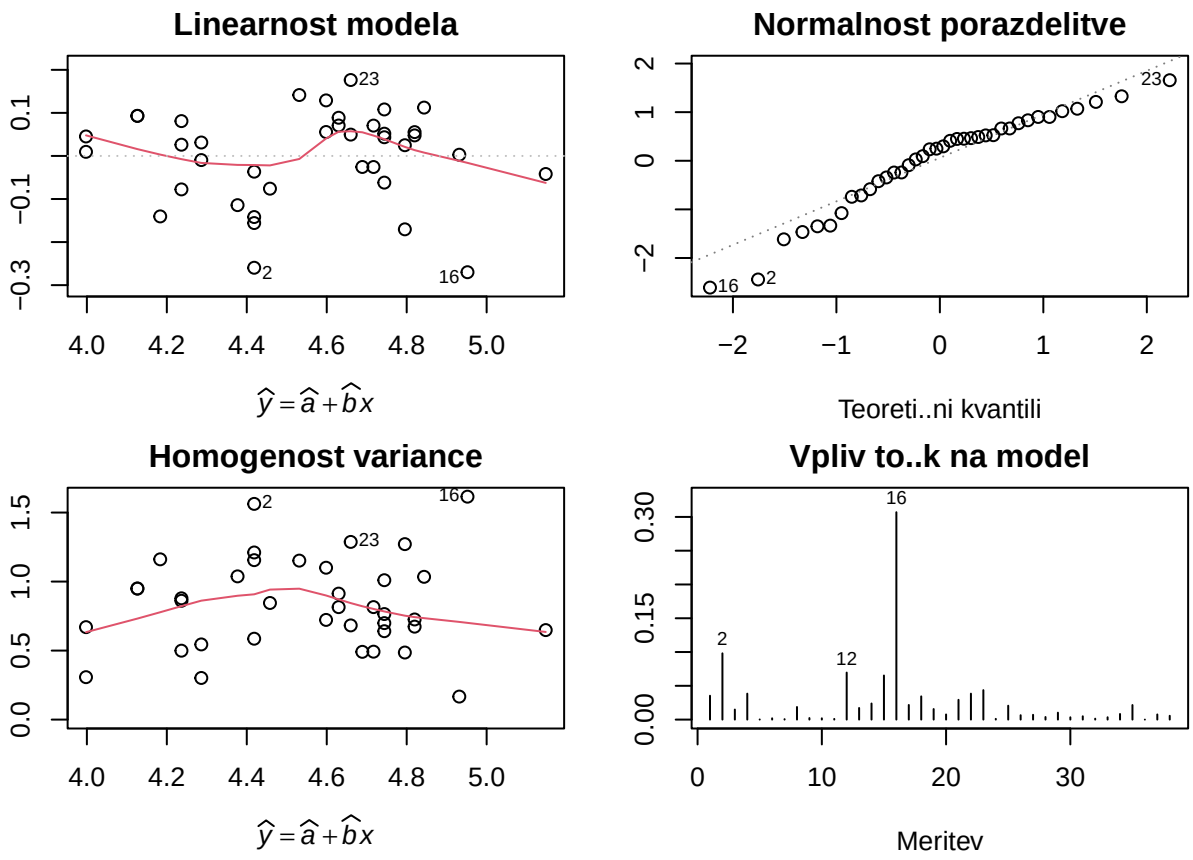
```
## Warning in title(xlab = "Teoretični kvantili", ylab = "St. ostanki", main =
## "Normalnost porazdelitve"): conversion failure on 'Teoretični kvantili' in
## 'mbcsToSbcs': dot substituted for <c4>
```

```
## Warning in title(xlab = "Teoretični kvantili", ylab = "St. ostanki", main =
## "Normalnost porazdelitve"): conversion failure on 'Teoretični kvantili' in
## 'mbcsToSbcs': dot substituted for <8d>
```

```
plot(model,which=3,caption="",ann=FALSE)
title(xlab=expression(italic(widehat(y)==widehat(a)+widehat(b)*x)),
ylab=expression(sqrt(paste("|Stevilo ostankov|"))), main="Homogenost variance")
plot(model,which=4,caption="", ann=FALSE)
title(xlab="Meritev",ylab="Cookova razdalja", main="Vpliv točk na model")
```

```
## Warning in title(xlab = "Meritev", ylab = "Cookova razdalja", main = "Vpliv točk
## na model"): conversion failure on 'Vpliv točk na model' in 'mbcsToSbcs': dot
## substituted for <c4>
```

```
## Warning in title(xlab = "Meritev", ylab = "Cookova razdalja", main = "Vpliv točk
## na model"): conversion failure on 'Vpliv točk na model' in 'mbcsToSbcs': dot
## substituted for <8d>
```



1) Graf za preverjanje linearnosti modela

Validnost linearnega regresijskega modela lahko preverimo tako, da narišemo graf ostankov v odvisnosti od x vrednosti ali od predvidenih vrednosti $\hat{y} = \hat{a}x + \hat{b}$ in preverimo, če obstaja kakšen vzorec. Če so točke dokaj enakomerno raztresene nad in pod premico $Ostanki = 0$ in ne moremo zaznati neke oblike, je linearni model validen. Če na grafu opazimo kakšen vzorec (npr. točke formirajo nelinearno funkcijo), nam sama oblika vzorca daje informacijo o funkciji od x, ki manjka v modelu.

Za uporabljene podatke na grafu linearnosti modela ne opazimo vzorca ali manjkajoče funkcije in lahko zaključimo, da je linearni model validen. Točke na grafu izgledajo popolnoma naključno razporejene.

2) Graf normalnosti porazdelitve naključnih napak

Normalnost porazdelitve naključnih napak preverjamo preko grafa porazdelitve standardiziranih ostankov. Na x-osi Q - Q grafa normalne porazdelitve so podani teoretični kvantili, na y - osi pa kvantili standardiziranih ostankov. Če dobljene točke na Q-Q grafu tvorijo premico (z manjšimi odstopanji), zaključimo, da je porazdelitev naključnih napak (vsaj približno) normalna.

Za podatke o masi in porabi goriva avtomobilov lahko zaključimo, da so naključne napake normalno porazdeljene (ni večjih odstopanj od premice, razen za 2., 16., in 23. podatkovno točko).

3) Graf homogenosti variance

Učinkovit graf za registriranje nekonstantne variance je graf korena standardiziranih ostankov v odvisnosti od x ali od predvidenih vrednosti $\hat{y} = \hat{a}x + \hat{b}$. Če variabilnost korena standardiziranih ostankov narašča

ali pada s povečanjem vrednosti $\hat{y}$, je to znak, da varianca naključnih napak ni konstantna. Pri naraščanju variance je graf pogosto oblike $\triangleleft$, in pri padanju variance oblike $\triangleright$. Pri ocenjevanju lahko pomaga funkcija glajenja, v primeru konstantne variance se pričakuje horizontalna črta, okoli katere so točke enakomerno razporejene.

Za naš primer, točke na grafu sugerirajo, da ni naraščanja ali padanja variance. Ničelna domneva konstantne variance se lahko formalno preveri s Breusch-Paganovim testom.

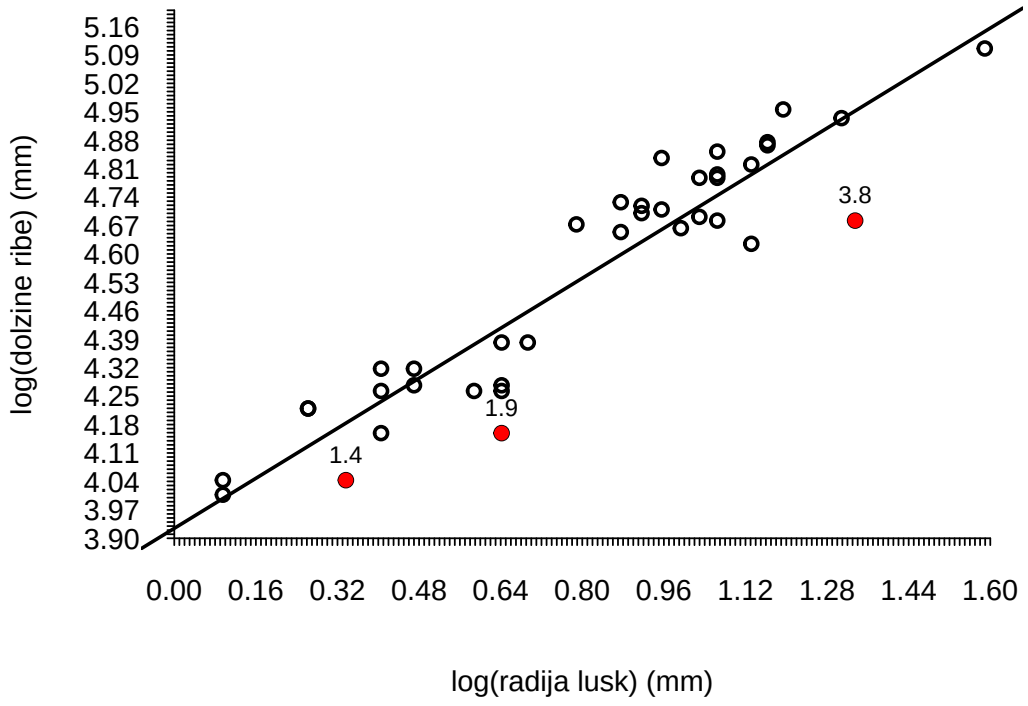
```
#suppressWarnings(library(car))
#ncvTest(model)
```

Na osnovi rezultata Breusch-Paganovega testa (testna statistika $\chi^2 = 0.4018547$, $df = 1$, p-vrednost $p = 0.52613 > 0.05$), ne zavrnemo ničelne domneve. Ni dovolj dokazov, da varianca naključnih napak ni homogena.

4) Graf vpliva posameznih točk na model

Vpliv i-te točke na linearni regresijski model merimo s Cookovo razdaljo D_i, $1 \leq i \leq n$. Če i-ta točka ne vpliva močno na model, bo D_i majhna vrednost. Če je $D_i \geq c$, kjer je $c = F_{2,n-2;0.5}$ mediana Fisherjeve porazdelitve z 2 in $n-2$ prostostnima stopnjama, i-ta točka močno vpliva na regresijski model. Na grafu vpliva točk na linearni regresijski model so vedno označene tri točke z najvišjo Cookovo razdaljo. Za naše podatke, to so 2., 12., in 16. podatkovne točka. Spomnimo se, da smo te točke identificirali kot osamelce razen 12.. Zdaj poglejmo na razsevnem diagramu po čem so te tri točke drugačne od ostalih. Kodi za razsevni diagram dodamo če dve vrstici, s katerima bomo dodali ocenjeno regresijsko premico in pobarvali te tri točke.

```
ld<-log(dolzina)
pd<-log(rlusk)
par(las=1, mar=c(4,4,2,3))
plot(log(rlusk),log(dolzina), main="", ylim=c(3.9,5.2), xlim=c(0,1.6),
ylab="log(dolzine ribe) (mm)", xlab="log(radija lusk) (mm)", lwd=2, axes=FALSE)
axis(2,pos=0,at=seq(3.9,5.2,by=0.01),tcl=-0.2)
axis(1,pos=3.9,at=seq(0,1.6,by=0.01),tcl=-0.2)
arrows(x0=1.6,y0=0,x1=1.7,y1=0,length=0.1)
arrows(x0=3.9,y0=1.6,x1=3.9,y1=1.7,length=0.1)
abline(model,lwd=2)
points(log(rlusk)[c(2,12,16)],log(dolzina)[c(2,12,16)],col="red",pch=19)
text(log(rlusk)[c(2,12,16)],log(dolzina)[c(2,12,16)],labels=rlusk[c(2,12,16)],pos=3,cex=0.8)
```



Na razsevnem diagramu opazimo, da so vse tri točke najbolj oddaljene od ocenjene regresijske premice (oziroma jim ustrezajo največji ostanki). Lahko preverimo še, ali je njihov vpliv velik, oziroma ali je njihova Cookova razdalja večja ali enaka od mediane Fisherjeve porazdelitve z 2 in 30 prostostnimi stopnjami.

```
any(cooks.distance(model)[c(2,12,16)]>=qf(0.5,2,nrow(jezero)-2))
```

```
## [1] FALSE
```

Nobena od teh točk nima velikega vpliva na linearni regresijski model, zato jih ni potrebno odstraniti.

7. Testiranje linearnosti modela in koeficient determinacije

Poglejmo R-jevo poročilo o modelu.

```
summary(model)
```

```
##
## Call:
## lm(formula = log(dolzina) ~ log(rlusk), data = jezero)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26999 -0.05691  0.02868  0.07063  0.17629
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.92444    0.04296   91.35   <2e-16 ***
## log(rlusk)   0.76980    0.04860   15.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.108 on 36 degrees of freedom
## Multiple R-squared:  0.8745, Adjusted R-squared:  0.871
## F-statistic: 250.8 on 1 and 36 DF,  p-value: < 2.2e-16
```

Vrednost testne statistike za preverjanje linearnosti modela je enaka $t = 15.84$, s $df = 36$ prostostnimi stopnjami in s p-vrednostjo $p = 2.2 \cdot 10^{-16}$, ki je manjša od dane stopnje značilnosti 0.05. Na osnovi rezultatov t-testa zavrnemo ničelno domnevo $H_0 : b = 0$, za dano stopnjo značilnosti in dobljeni vzorec. Drugače rečeno, s formalnim statističnim testiranjem smo pritrdili, da linearni model ustreza podatkom.

Koeficient determinacije je enak $R^2 = 0.8745$, kar pomeni, da 87% variabilnosti dolzine rib pojasnjuje regresijski model.

8. Intervala zaupanja za naklon in odsek regresijske premice

Izračunajmo 95% interval zaupanja za neznani naklon in odsek regresijske premice.

```
round(confint(model),3)
```

```
##             2.5 % 97.5 %
## (Intercept) 3.837  4.012
## log(rlusk)  0.671  0.868
```

Interval zaupanja za odsek je enak $I_a = [3.837, 4.012]$ in interval zaupanja za naklon $I_b = [0.671, 0.868]$.

9. Interval predikcije za vrednost Y pri izbrani vrednosti X

Pri predvidevanju vrednosti dolzine ribe nas zanima bodoča vrednost spremenljivke Y pri izbrani vrednosti spremenljivke $X = x_0$. Ne zanima nas le predvidena vrednost $\hat{y} = 3.9244 + 0.7698x_0$ dolocenega radija lusk x_0, ampak želimo tudi oceniti spodnjo in zgornjo mejo, med katerima se verjetno nahaja dolzina različnih rib z luskami takega radija.

```
xrlusk = data.frame(rlusk=c(2,3,4))
exp(predict(model, xrlusk, interval="predict"))
```

```
##         fit       lwr      upr
## 1  86.31656  69.11321 107.8021
## 2 117.93679  94.28325 147.5245
## 3 147.17269 117.02508 185.0868
```

Predvidena vrednost dolzine ribe z dolocenim radijem lusk (na celi populaciji rib)

1. 2mm je 86.31656mm, s 95% intervalom predikcije dolzine ribe [69.11321, 107.8021],
2. 3mm je 117.93679mm, s 95% intervalom predikcije dolzine ribe [94.28325, 147.5245],
3. 4mm je 147.17269mm, s 95% intervalom predikcije dolzine ribe [117.02508, 185.0868]

10. Zaključek

Zanimala nas je funkcijska zveza med dolzino in radijem lusk enoletnih maloustih basov, merjeno kot dolzina ribe glede na radij lusk. Zbrali smo vzorec 38 rib, jim izmerili radij lusk in zabelezili dolzino. Ugotovili smo, da je enostavni linearni model odvisnosti med dolzino in radijem lusk dober. Diagnostični grafi in statistični testi niso pokazali na težave z linearnim regresijskim modelom. Koeficient determinacije je 87%, kar pomeni, da tolikšen delež variabilnosti dolzine rib zajamemo z linearnim modelom. Napoved dolzine ribe na osnovi radija lusk je zadovoljiva, vendar bi vključevanje dodatnih neodvisnih spremenljivk zagotovo dala še boljši model in bolj zanesljivo napoved.