Napoved teže možganov pri sesalcih z linearno regresijo

Armen Hodža

1. Opis podatkov

Zbrali smo vzorec meritev telesne teže in teže možganov na vzorcu 58 vrst sesalcev. Podatke smo zapisali v dokument, ki ima štiri stolpce:

1. *vrsta* je nominalna spremenljivka, katere vrednosti so latinski nazivi vrste sesalcev.
2. *slovime* je nominalna spremenljivka, katere vrednosti so slovenski nazivi vrste sesalcev.
3. *telteza* je numerična zvezna spremenljivka, ki predstavlja telesno težo (v kilogramih).
4. *mozteza* je numerična zvezna spremenljivka, ki predstavlja težo možganov (v gramih).

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

```r
mozgani<-read.csv("C:/mozgani.csv", header=TRUE)
mozgani$logmteza<-log(mozgani$mozteza)
mozgani$logtteza<-log(mozgani$telteza)
str(mozgani)
```

```
## 'data.frame':    58 obs. of  6 variables:
##  $ vrsta   : chr  "Aotus trivirgatus" "Aplodontia rufa" "Blarina brevicauda" "Bos taurus" ...
##  $ slovime : chr  "Ponocna opica" "Planinski bober" "Rovka" "Krava" ...
##  $ telteza : num  0.48 1.35 0.005 464.983 36.328 ...
##  $ mozteza : num  15.5 8.1 0.14 423.01 119.5 ...
##  $ logmteza: num  2.74 2.09 -1.97 6.05 4.78 ...
##  $ logtteza: num  -0.734 0.3 -5.298 6.142 3.593 ...
```

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 telesne teže, logaritma telesne teže, teže možganov in logaritma teže možganov sesalcev.

```r
summary(mozgani$telteza)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.005    0.814    3.442  212.428   54.665 6654.180
```

```r
sd(mozgani$telteza)
```

```
## [1] 928.6204
```

Opazimo, da telesna teža vzorca sesalcev varira od 0.005 do 6654.18 kg, s povprečjem 212.428 in standardnim odklonom 928.6204 kg. Ponovimo postopek računanja za vzorec logaritma telesne teže.

```r
summary(mozgani$logtteza)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -5.2983 -0.2079  1.2361  1.5018  4.0009  8.8030
```

```r
sd(mozgani$logtteza)
```

```
## [1] 3.133862
```

Opazimo, da logaritem telesne teže vzorca sesalcev varira od -5.2983 do 8.8030, s povprečjem 1.5018 in standardnim odklonom 3.134. Ponovimo postopek računanja za vzorec teže možganov.

```r
summary(mozgani$mozteza)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.14    5.55   23.00  302.32  173.50 5711.86
```

```r
sd(mozgani$mozteza)
```

```
## [1] 959.3438
```

Opazimo, da teža možganov vzorca sesalcev varira od 0.14 do 5711.86 kg, s povprečjem 302.32 in standardnim odklonom 959.3438 kg. Ponovimo postopek računanja še za vzorec logaritma teže možganov.

```r
summary(mozgani$logmteza)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -1.966   1.714   3.132   3.290   5.156   8.650
```

```r
sd(mozgani$logmteza)
```

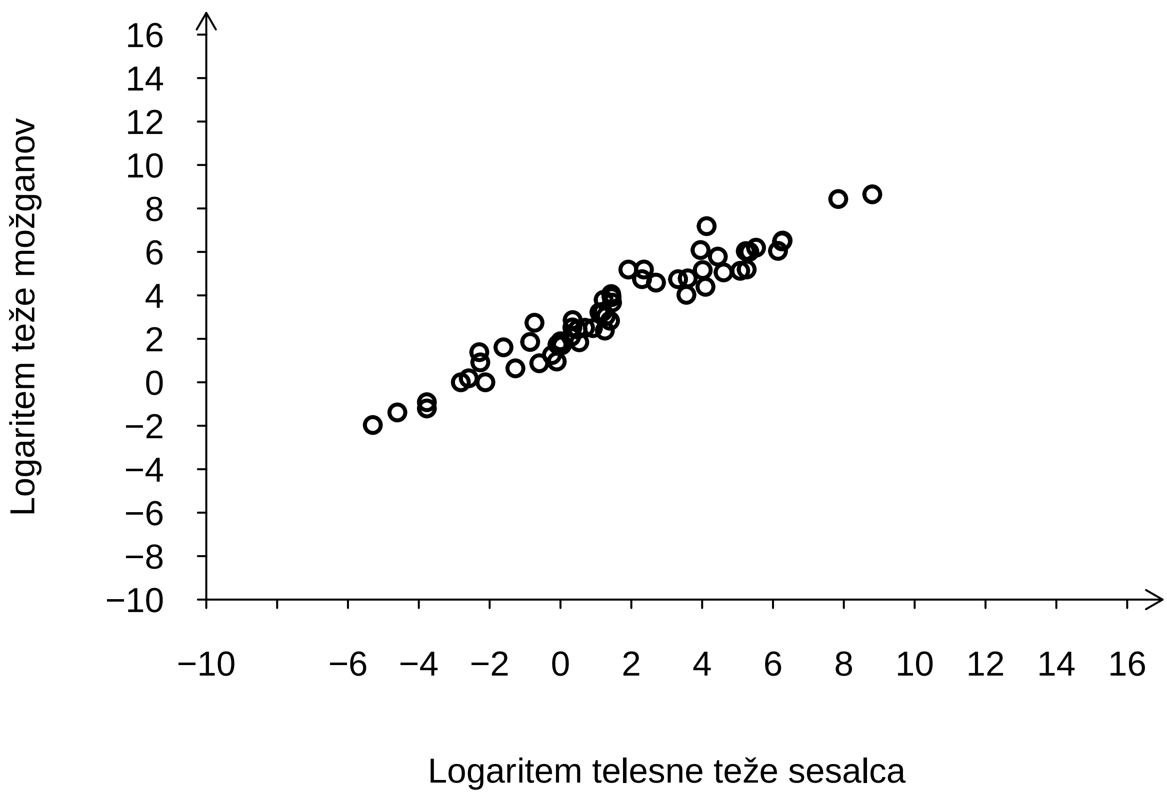
```
## [1] 2.435202
```

Opazimo, da teža možganov vzorca sesalcev varira od -1.966 do 8.650, s povprečjem 3.29 in standardnim odklonom 2.435. Razpon vrednosti logaritma telesne teže sesalcev in logaritma teže možganov nam pomaga pri izbiri mej na oseh razsevnega diagrama.

3. Razsevni diagram in vzorčni koeficient korelacije

Prikažimo dobljene podatke na razsevnem diagramu.

```
par(las=1, cex=1.1, mar=c(4,4,2,2))
plot(mozgani$logtteza, mozgani$logmteza, main="", xlim=c(-10,16), ylim=c(-10,16),
xlab="Logaritem telesne teže sesalca", ylab="Logaritem teže možganov", lwd=2, axes=FALSE)
axis(1,pos=-10,at=seq(-10,16,by=2),tcl=-0.2)
axis(2,pos=-10,at=seq(-10,16,by=2),tcl=-0.2)
arrows(x0=16,y0=-10,x1=17,y1=-10,length=0.1)
arrows(x0=-10,y0=16,x1=-10,y1=17,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(mozgani$logtteza,mozgani$logmteza))
```

```
## [1] 0.9632881
```

Vrednost vzorčnega koeficienta korelacije je visoka ($r = 0.963$), kar govori o visoki linearni povezanosti telesne teže sesalcev in njihove teže možganov. Dalje, koeficient korelacije je pozitiven, kar pomeni, da sesalci večje telesne teže običajno imajo večjo težo možganov.

4. Formiranje linearnega regresijskega modela

Formirajmo linearni regresijski model.

```r
(model<-lm(logmteza~logtteza,data=mozgani))
```

```
##
## Call:
## lm(formula = logmteza ~ logtteza, data = mozgani)
##
## Coefficients:
## (Intercept)     logtteza
##      2.1661       0.7485
```

Dobili smo ocenjeno regresijsko premico $\hat{y} = 2.1661 + 0.7485x$, oziroma oceni odseka in naklona sta enaki $\hat{a} = 2.1661$ in $\hat{b} = 0.7485$.

5. Točke visokega vzvoda in osamelci

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

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

```
##                  vrsta              slovime  telteza mozteza  logmteza  logtteza
## 3  Blarina brevicauda                 Rovka    0.005    0.14 -1.966113 -5.298317
## 16    Elephas maximus         Azijski slon 2547.070 4603.17  8.434500  7.842699
## 29 Loxodonta africana         Afriski slon 6654.180 5711.86  8.650300  8.803001
## 35   Myotis lucifugus Majhni rjavi netopir    0.010    0.25 -1.386294 -4.605170
```

Odkrili smo 4 točke visokega vzvoda. Dve vrsti sesalca imajo visoko telesno težo nad 2500 kg in dve vrsti nizko telesno težo, pod 0.01 kg.

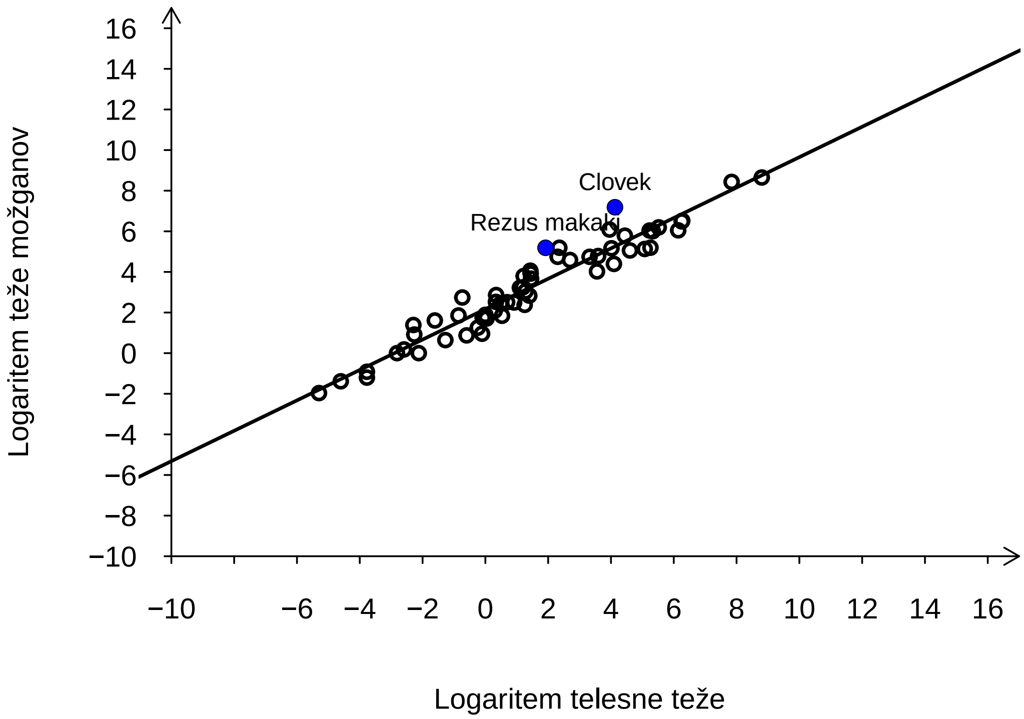
Za podatke majhne in srednje velikosti vzorca je osamelec podatkovna točka, kateri ustreza standardizirani ostanek izven intervala $[-2, 2]$.

```r
mozgani[abs(rstandard(model))>2,]
```

```
##                    vrsta      slovime telteza  mozteza logmteza logtteza
## 28 Homo sapiens sapiens       Clovek  61.998 1320.020 7.185402 4.127102
## 30       Macaca mulatta Rezus makaki   6.800  179.003 5.187403 1.916923
```

Identificirali smo dve podatkovni točki (28. in 30. točka) kot osamelce. Zdaj poglejmo na razsevnem diagramu po čem so te točke drugačne od ostalih. Kodi za razsevni diagram dodamo še dve vrstici, s katerima bomo dodali ocenjeno regresijsko premico in pobarvali te dve točki.

```r
par(las=1, mar=c(4,4,2,3))
plot(mozgani$logtteza, mozgani$logmteza, main="", xlim=c(-10,16), ylim=c(-10,16),
xlab="Logaritem telesne teže", ylab="Logaritem teže možganov", lwd=2, axes=FALSE)
axis(1,pos=-10,at=seq(-10,16,by=2),tcl=-0.2)
axis(2,pos=-10,at=seq(-10,16,by=2),tcl=-0.2)
arrows(x0=16,y0=-10,x1=17,y1=-10,length=0.1)
arrows(x0=-10,y0=16,x1=-10,y1=17,length=0.1)
abline(model,lwd=2)
points(log(mozgani$telteza[c(28,30)]),log(mozgani$mozteza[c(28,30)]),col="blue",pch=19)
text(log(mozgani$telteza[c(28,30)]),log(mozgani$mozteza[c(28,30)]),labels=mozgani$slovime[c(28,30)],pos=
```

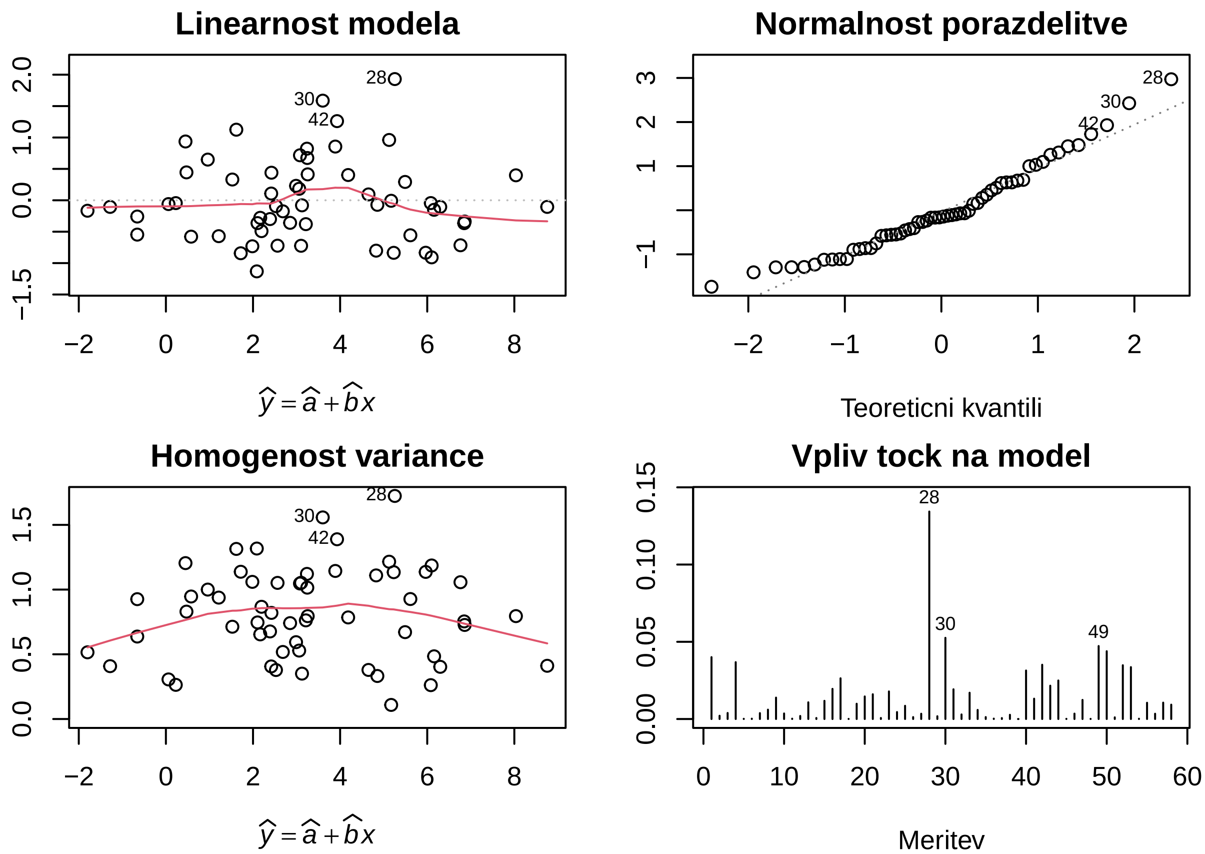


Na razsevnem diagramu opazimo, da se omenjeni osamelci nanašajo na dve vrsti sesalcev z nenavadno veliko težo možganov glede na njihovo telesno težo.

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



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 ne izgledajo popolnoma naključno razporejene, opažamo večjo koncentracijo točk za predvidene vrednosti od 2 do 6, kar je prisotno zaradi originalnih vrednosti v vzorcu vrst sesalcev (poglej razsevni diagram).

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 telesni teži in teži možganov sesalcev lahko zaključimo, da so naključne napake normalno porazdeljene (ni večjih odstopanj od premice, razen za 28. in 30. 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.

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 28. in 30. podatkovna točka. Spomnimo se, da smo te točke identificirali kot osamelce. Na razsevnem diagramu opazimo, da so obe dve točki 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.

```r
any(cooks.distance(model)[c(28,30)]>=qf(0.5,2,nrow(mozgani)-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.

```r
summary(model)
```

```
##
## Call:
## lm(formula = logmteza ~ logtteza, data = mozgani)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13170 -0.46298 -0.09914  0.40122  1.93005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.16608    0.09620   22.52   <2e-16 ***
## logtteza     0.74853    0.02788   26.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6596 on 56 degrees of freedom
## Multiple R-squared:  0.9279, Adjusted R-squared:  0.9266
## F-statistic:   721 on 1 and 56 DF,  p-value: < 2.2e-16
```

Vrednost testne statistike za preverjanje linearnosti modela je enaka $t = 26.85$, s $df = 56$ prostostnimi stopnjami in s p-vrednostjo $p = 2.2 \cdot e^{-16}$, 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 $R^2 = 0.928$, kar pomeni, da 93% variabilnosti teže možganov pojasnjuje linearni regresijski model.

8. Interval predikcije za vrednost Y pri izbrani vrednosti X

Pri predvidevanju vrednosti teže možganov nas zanima bodoča vrednost spremenljivke Y pri izbrani vrednosti spremenljivke $X = x_0$. Ne zanima nas le predvidena vrednost $\hat{y} = 2.166 + 0.749x_0$ sesalcev določene telesne teže x_0, ampak želimo tudi oceniti spodnjo in zgornjo mejo, med katerima se verjetno nahaja teža možganov različnih vrst sesalcev teh telesnih tež.

```
xtelteza <- data.frame(logtteza=log(c(10, 100, 1000)))
predict(model, newdata=xtelteza, interval="predict")
```

```
##        fit      lwr      upr
## 1 3.889640 2.556230 5.223049
## 2 5.613202 4.269320 6.957084
## 3 7.336764 5.970336 8.703192
```

Predvidena vrednost logaritma teže možganov za logaritma telesne teže sesalca (na celi populaciji sesalcev)

1. Log(10) je 3.89, s 95% intervalom predikcije teže možganov $[2.56, 5.22]$,
2. Log(100) je 5.61, s 95% intervalom predikcije teže možganov $[4.27, 6.96]$,
3. Log(1000) je 7.34, s 95% intervalom predikcije teže možganov $[5.97, 8.70]$.

```
xtelteza <- data.frame(logtteza=log(c(10, 100, 1000)))
exp(predict(model, newdata=xtelteza, interval="predict"))
```

```
##          fit       lwr      upr
## 1   48.89327  12.88714  185.499
## 2  274.02027  71.47304 1050.565
## 3 1535.73489 391.63737 6022.106
```

Predvidena vrednost teže možganov za telesno težo sesalca (na celi populaciji sesalcev)

1. 10 kg je 48.89 g, s 95% intervalom predikcije teže možganov $[12.89, 185.5]$,
2. 100 kg je 274.02 g, s 95% intervalom predikcije teže možganov $[71.47, 1050.57]$,
3. 1000 kg je 1535.73 g, s 95% intervalom predikcije teže možganov $[391.63, 6022.11]$.

9. Zaključek

Zanimala nas je funkcionalna odvisnost med telesno težo sesalcev in njihovo težo možganov, merjeno v gramih. Zbrali smo vzorec 58 vrst sesalcev, jim izmerili telesno težo in njihovo težo možganov. Ugotovili smo, da je enostavni linearni model odvisnosti teže možganov od telesne teže dober. Diagnostični grafi in statistični testi niso pokazali na težave z linearnim regresijskim modelom. Koeficient determinacije je 93%, kar pomeni, da tolikšen delež variabilnosti teže možganov zajamemo z linearnim modelom. Napoved teže možganov na osnovi telesne teže sesalcev je zadovoljiva, vendar bi vključevanje dodatnih neodvisnih spremenljivk zagotovo dala še boljši model in bolj zanesljivo napoved.