Teža možganov pri sesalcih

Tim Hrovat

1. Opis podatkov

V danem vzorcu imamo meritve telesne teže in teže možganov 58 vrst sesalcev. Podatki so zapisani v *.csv* dokumentu s 4 stolpci:

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*. V naslednji poizvedbi s pomočjo R preberemo datoteko in nato izpišemo njeno strukturo:

```
mozgani<-read.csv("./mozgani.csv", header=TRUE)
str(mozgani)
```

```
## 'data.frame':    58 obs. of  4 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 ...
```

Raziskovalna domneva predvideva, da med težo možganov in telesno težo sesalcev obstaja funkcijska zveza.

Za zgoraj opisane podatke bomo konstruirali regresijski model med transformiranima spremenljivkama $\log(mozteza)$ in $\log(telteza)$, kjer je $\log(mozteza)$ odvisna spremenljivka.

```
t_mozgani<-mozgani
t_mozgani$telteza <- log(mozgani$telteza)
t_mozgani$mozteza <- log(mozgani$mozteza)
```

2. Opisna statistika originalnih in transformiranih podatkov

(a) Originalni podatki

V tem delu bomo izračunali opisno statistiko za *originalne* podatke, ki vključuje povzetek s petimi števili (minimum, maksimum, prvi kvartil, mediano in tretji kvartil), vzorčno povprečje ter vzorčni standardni odklon za težo telesa in 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 kg do 6654.18 kg, s povprečjem 212.428 kg in standardnim odklonom 928.6204 kg.

Postopek ponovimo še za težo 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
```

Sedaj lahko opazimo, da teža možganov varira od 0.14 g do 5711.86 g, s povprečjem 302.32 g in standardnim odklonom 959.3438 g.

(b) Transformirani podatki

Enake postopke sedaj uporabimo še za *transformirane* podatke.

```r
summary(t_mozgani$telteza)
```

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

```r
sd(t_mozgani$telteza)
```

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

Opazimo, da transformirana telesna teža vzorca sesalcev varira od -5.2983 kg do 8.8030 kg, s povprečjem 1.5018 kg in standardnim odklonom 3.133862 kg.

Postopek ponovimo še za težo možganov:

```r
summary(t_mozgani$mozteza)
```

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

```r
sd(t_mozgani$mozteza)
```

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

Sedaj lahko opazimo, da teža možganov varira od -1.966 g do 8.650 g, s povprečjem 3.290 g in standardnim odklonom 2.435202 g.

Pridobljeni podatki nam bodo kasneje pomagali pri izbiri mej na oseh razsevnega diagrama.

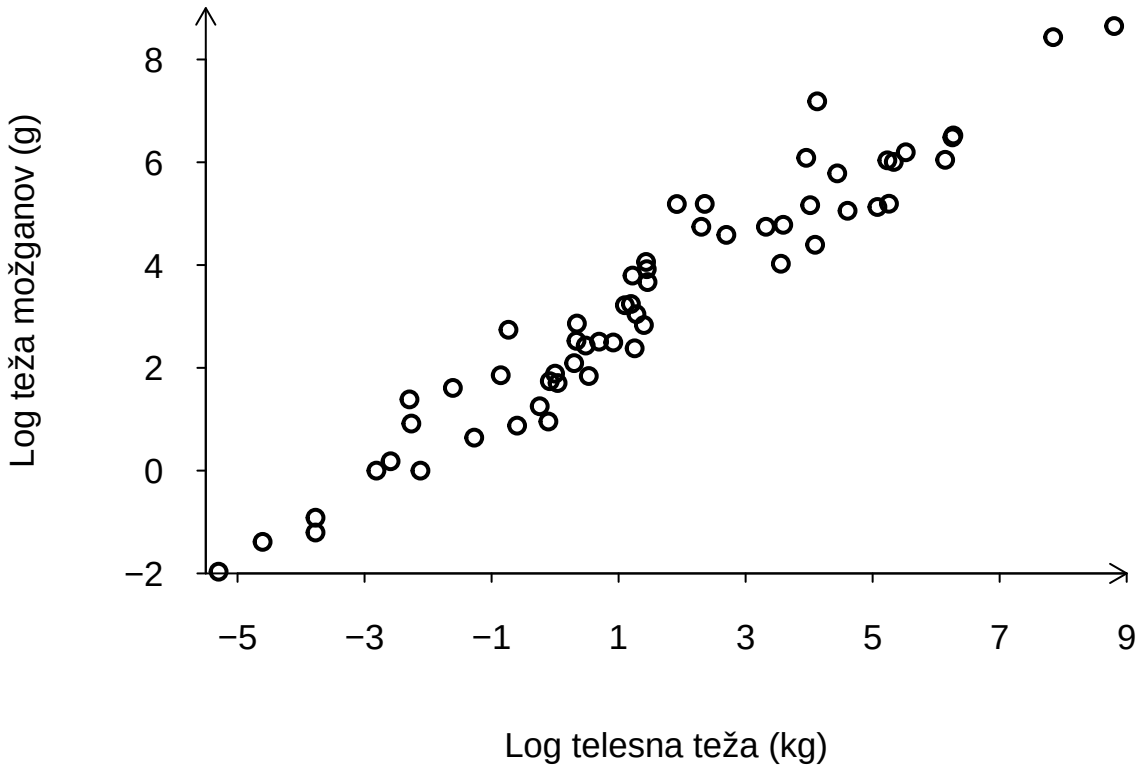
V nadaljevanju bomo za analizo uporabljali le transformirane podatke.

```r
rm(mozgani)
```

3. Razsevni diagram in vzorčni koeficient korelacije

Transformirane podatke bomo sedaj prikazali na razsevnem diagramu.

```r
par(las=1, cex=1.1, mar=c(4,4,2,2))
plot(t_mozgani$telteza,t_mozgani$mozteza,main="",
  xlim=c(-5.5, 9),ylim=c(-2, 9),
  xlab="Log telesna teža (kg)",ylab="Log teža možganov (g)",
  lwd=2,axes=FALSE
)
axis(1, pos=-2, at=seq(-5, 9, by=2), tcl=-0.2)
axis(2, pos=-5.5, at=seq(-2, 9, by=2), tcl=-0.2)
arrows(x0=-5.5, y0=-2, x1=9, y1=-2, length=0.1)
arrows(x0=-5.5, y0=-2, x1=-5.5, y1=9, length=0.1)
```



Kot lahko vidimo iz diagrama, so točke razporejene okoli namišljene premice, kar kaže na prisotnost linearne povezave med transformirano telesno težo in težo možganov.

Pearsonov koeficient korelacije lahko izračunamo z naslednjim ukazom:

```
(r<-cor(t_mozgani$telteza,t_mozgani$mozteza))
```

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

Opazimo, da je vrednost vzorčnega koeficienta korelacije zelo visoka (r = 0.963), kar kaže na močno linearno povezavo med težo možganov in telesno težo sesalcev. Poleg tega je koeficient korelacije pozitiven, kar pomeni, da ima sesalec z večjo telesno težo ponavadi tudi večjo težo možganov in sesalec z manjšo telesno težo tudi manjšo težo možganov.

4. Formiranje linearnega regresijskega modela

Za analizo povezave med težo možganov in telesno težo sesalcev smo oblikovali linearni regresijski model:

```
(vrsta<-lm(mozteza~telteza,data=t_mozgani))
```

```
##
## Call:
## lm(formula = mozteza ~ telteza, data = t_mozgani)
##
```

```
## Coefficients:
## (Intercept)      telteza
##      2.1661       0.7485
```

Rezultat modela nam daje ocenjeno regresijsko premico $\hat{y} = 2.1661 + 0.7485x$. Odsek in naklon sta enaka $\hat{a} = 2.1661$ 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 $\frac{4}{n}$.

```
t_mozgani[hatvalues(vrsta)>4/nrow(t_mozgani),]
```

```
##                 vrsta              slovime   telteza   mozteza
## 3  Blarina brevicauda                Rovka -5.298317 -1.966113
## 16    Elephas maximus         Azijski slon  7.842699  8.434500
## 29 Loxodonta africana         Afriski slon  8.803001  8.650300
## 35   Myotis lucifugus Majhni rjavi netopir -4.605170 -1.386294
```

Pridobili smo 4 točke visokega vzvoda. Dve vrsti z visoko transformirano telesno težo (16, 29) in dve z nizko (3,35).

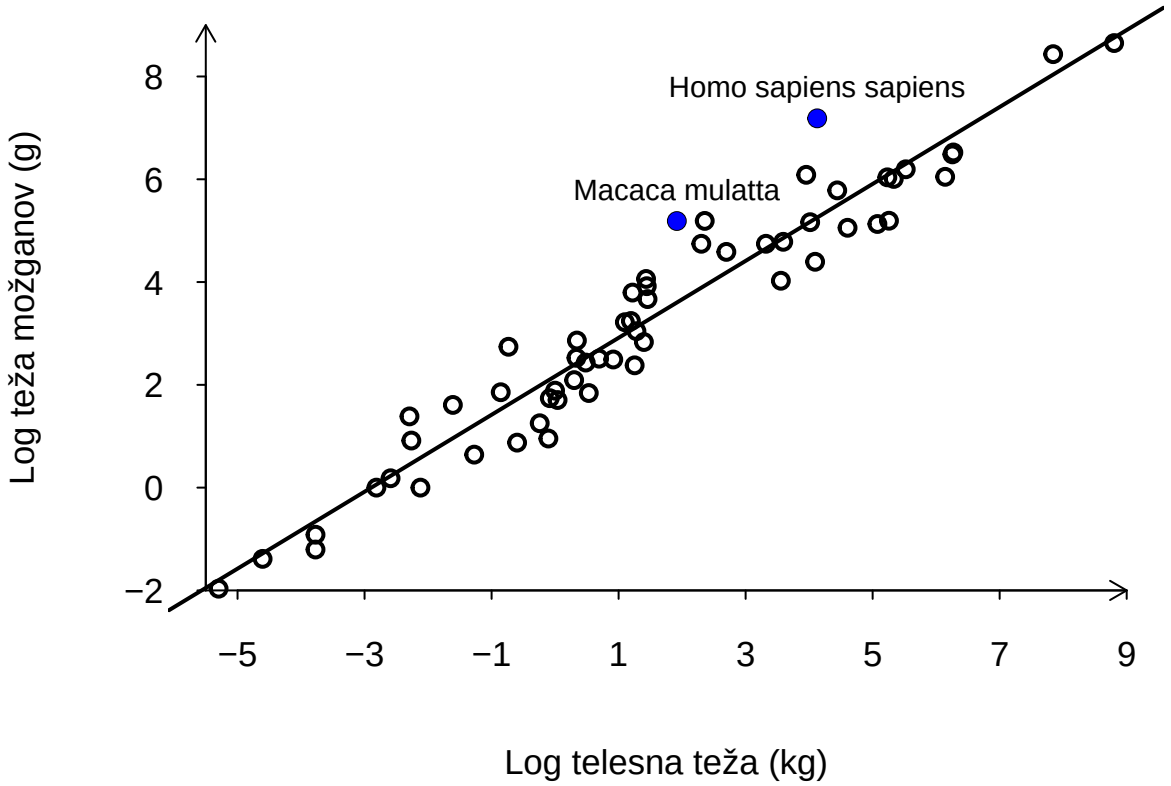
Za majhne in srednje velike vzorce je osamelec podatkovna točka, ki ustreza standardiziranemu ostanku zunaj intervala [-2, 2].

```
t_mozgani[abs(rstandard(vrsta))>2,]
```

```
##                   vrsta      slovime  telteza  mozteza
## 28 Homo sapiens sapiens       Clovek 4.127102 7.185402
## 30       Macaca mulatta Rezus makaki 1.916923 5.187403
```

Pridobili smo dve podatkovni točki - osamelca (28,30). Na sledečem razsevnem diagramu si bomo ogledali kako se ti dve točki razlikujeta od ostalih.

```
par(las=1, cex=1.1, mar=c(4,4,2,2))
plot(t_mozgani$telteza,t_mozgani$mozteza,main="",
  xlim=c(-5.5, 9),ylim=c(-2, 9),
  xlab="Log telesna teža (kg)",ylab="Log teža možganov (g)",
  lwd=2,axes=FALSE
)
axis(1, pos=-2, at=seq(-5, 9, by=2), tcl=-0.2)
axis(2, pos=-5.5, at=seq(-2, 9, by=2), tcl=-0.2)
arrows(x0=-5.5, y0=-2, x1=9, y1=-2, length=0.1)
arrows(x0=-5.5, y0=-2, x1=-5.5, y1=9, length=0.1)
abline(vrsta,lwd=2)
points(t_mozgani$telteza[c(28,30)],t_mozgani$mozteza[c(28,30)],col="blue",pch=19)
text(t_mozgani$telteza[c(28,30)],t_mozgani$mozteza[c(28,30)]+c(0,0),labels=
t_mozgani$vrsta[c(28,30)],pos=3,cex=0.8)
```

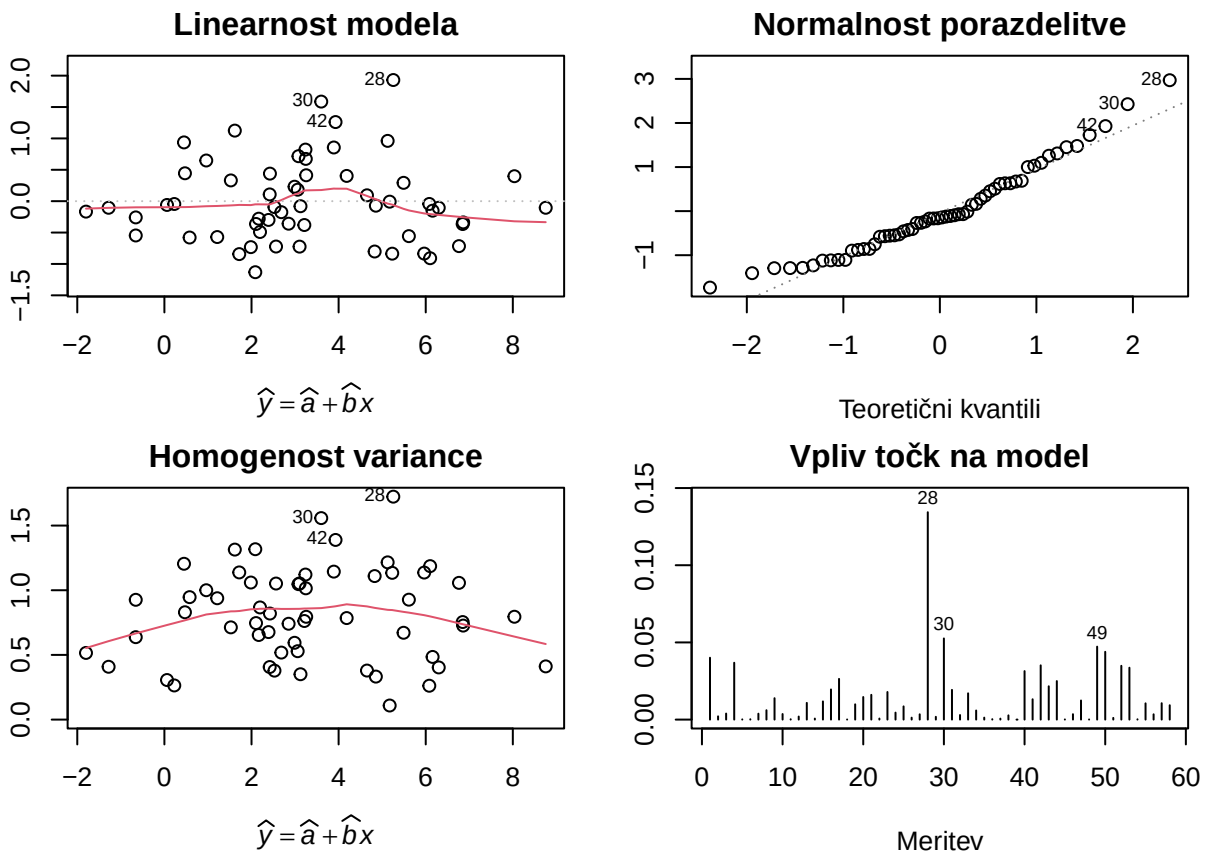


Opazujemo, da se na razsevnem diagramu nahajata dve točki označeni z modro barvo, ki ustrezata vrstam z nenavadno visoko transformirano težo možganov v primerjavi s telesno težo.

6. Preverjanje predpostavk linearnega regresijskega modela

Predpostavke linearnega regresijskega modela bomo preverili s štirimi diagnostičnimi grafi.Če predpostavke modela niso izpolnjene, lahko to vpliva na natančnost ocen neznanih parametrov, p-vrednosti testov, intervalov zaupanja in intervalov predikcije.

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



(a) Linearnost modela

Validnost linearnega regresijskega modela lahko preverimo s pomočjo grafičnih pregledov ostankov.Ena izmed najpogostejših metod je risanje grafov ostankov v odvisnosti od x-vrednosti ali predvidenih vrednosti $\hat{y} = a\hat{x} + \hat{b}$. Na podlagi tega pregleda lahko ocenimo, ali obstajajo vzorci ali nepravilnosti.

Če so točke raztresene nad in pod premico $Ostanki = 0$ in ne moremo zaznati nobenih žočitnih vzorcev, to pomeni, da je linearni model verjetno veljaven. Če pa opazimo prisotnost kakšnega vzorca, npr. točke tvorijo nelinearno strukturo, ta vzorec lahko kaže na funkcijo, ki manjka v modelu.

Pri uporabljenih podatkih na tem grafu nismo opazili nobenih očitnih vzorcev ali manjkajočih funkcij. To nam omogoča, da zaključimo, da je linearni model veljaven in opazimo, da točke na grafu niso povsem naključno razporejene. Pojavlja se večja koncentracija točk za predvidene vrednosti med 1 in 4, kar je posledica originalnih vrednosti v vzorcu sesalcev, kot je razvidno iz razsevnega diagrama.

(b) Normalnost porazdelitve naključnih napak

Normalnost porazdelitve naključnih napak preverjamo s pomočjo Q-Q grafa (Quantile-Quantile). Na x-osi Q-Q grafa so podani teoretični kvantili normalne porazdelitve, na y-osi pa kvantili standardiziranih ostankov. Če dobljene točke na Q-Q grafu tvorijo približno premico (z manjšimi odstopanji), lahko 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, saj ni večjih odstopanj od premice razen za 28. in 30. podatkovno točko.

(c) Homogenost variance

Učinkovit graf za odkrivanje nekonstantne variance je graf korena standardiziranih ostankov v odvisnosti od x ali predvidenih vrednosti $\hat{y} = a\hat{x} + \hat{b}$. Če variabilnost korena standardiziranih ostankov narašča ali pada s povečanjem vrednosti $\hat{y}$, to kaže, da varianca naključnih napak ni konstantna. Pri naraščanju variance je graf pogosto oblike naraščajoče črte, medtem ko pri padanju variance graf izgleda kot padajoča črta.

Pri ocenjevanju lahko uporabimo funkcijo glajenja. V primeru konstantne variance pričakujemo horizontalno črto, okoli katere so točke enakomerno razporejene.

Za naš primer točke na grafu ne nakazujejo naraščanja ali padanja variance, kar kaže na približno konstantno varianco naključnih napak.

(d) Cookova razdalja: graf in analiza vpliva točk preko pogoja velikega vpliva točk

S Cookovo razdaljo D_i, $1 \leq i \leq n$ merimo vpliv i-te točke na regresijski model. D_i bo majhna, če i-ta točka ne vpliva močno na model. 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 so označene tri točke z najvišjo Cookovo razdaljo (28, 30, 49). Opazimo, da smo dve od teh točk identificirali kot osamelce. Za te točke lahko preverimo, ali močno vplivajo na regresijski model. Postopek deluje, tako da pogledamo ali je njihova Cookova razdalja večja ali enaka od mediane Fisherjeve porazdelitve z 2 in (58 - 2) = 56 prostostnima stopnjama.

```
any(cooks.distance(vrsta)[c(28,30, 49)] >= qf(0.5, 2, nrow(t_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

Prikažemo R-jevo poročilo o modelu.

```
summary(vrsta)
```

```
##
## Call:
## lm(formula = mozteza ~ telteza, data = t_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 ***
## telteza      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 $p < 2.2 \cdot 10^{-16}$, 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 $R^2 = 0.9279$, kar pomeni, da linearni regresijski model pojasnjuje približno 93% variabilnosti teže možganov.

8. Interval predikcije za vrednost Y pri izbrani vrednosti X

Pri napovedovanju teže možganov nas zanima prihodnja vrednost spremenljivke Y za izbrano vrednost $X = x_0$. Poleg predvidene vrednosti $\hat{y} = 2.166 + 0.749x_0$, ki predstavlja povprečno težo možganov sesalca določene telesne teže x_0, želimo oceniti tudi spodnjo in zgornjo mejo, znotraj katerih se verjetno nahaja teža možganov različnih vrst sesalcev te telesne teže.

V nadaljevanju bomo izračunali predvideno težo možganov treh sesalcev s telesnimi težami 1 kg, 50 kg in 200 kg.

Ker smo originalne podatke transformirali z uporabo logaritemske funkcije bomo vhodne telesne teže tudi logaritmirali. Rezultat bomo nato pretvorili nazaj z uporabo inverzne (v tem primeru eksponentne) funkcije.

```
xtelteza <- data.frame(telteza=c(log(1), log(50), log(200)))
exp(predict(vrsta, xtelteza, interval="predict"))
```

```
##          fit        lwr        upr
## 1   8.723998   2.295112   33.16097
## 2 163.099126  42.730732  622.53380
## 3 460.377132 119.417546 1774.84056
```

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

1. 1 kg je 8.72 g, s 95% intervalom predikcije teže možganov $[2.295, 33.161]$,
2. 50 kg je 163.1 g, s 95% intervalom predikcije teže možganov $[42.731, 622.534]$,
3. 200 kg je 460.38 g, s 95% intervalom predikcije teže možganov $[119.418, 1774.841]$.

9. Zaključek

Na začetku smo definirali raziskovalno domnevo, da med telesno težo (v kg) in težo možganov (v g) sesalcev obstaja funkcijska povezava. Za preverjanje domneve smo zbrali podatke za vzorec 58 sesalcev, pri čemer smo zabeležili njihovo vrsto, slovensko ime, telesno težo in težo možganov.

Z analizo podatkov smo ugotovili, da obstaja linearna povezava med logaritemsko transformirano telesno težo in logaritemsko transformirano težo možganov. Diagnostični grafi in statistični testi niso razkrili nobenih težav z linearno regresijo, kar potrjuje ustreznost uporabljenega modela.

Koeficient determinacije meri 93%, kar pomeni, da smo tolikšen delež variabilnosti teže možganov zajeli z linearnim modelom. To kaže, da je model za napovedovanje teže možganov dobro prilagojen, vendar bi ga bilo mogoče izboljšati za še natančnejše napovedi z uporabo večjega vzorca, dodajanjem neodvisnih spremenljivk itd.