Linearna regresija: teža možganov pri sesalcih

1. Opis podatkov

Zbrali smo vzorec telesne teže in teže možganov 59 različnih 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, i 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, nato pa pogledali strukturo podatkov

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

```
## 'data.frame':    59 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 ...
```

2. Opisna statistika

Izračunali bomo opisno statistiko za naše podatke v obliki povzetka s petimi števili (minimum, maksimum, prvi in tretji kvartil, mediano), vzorčni povprečji in vzorčna standardna odklona telesne teže in teže možganov.

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

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.005    0.768    3.385  208.840   53.830 6654.180
```

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

```
## [1] 920.9927
```

Vidimo, da masa vzorca telesne teže sesalca varira med 0.005 in 6654.180kg. Povprečna telesna teža znaša 208.840kg s standardnim odklonom 920.9927 kg. Postopek računanja ponovimo še za vzorce teže možganov.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.14    5.60   21.00  297.41  172.00 5711.86
```

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

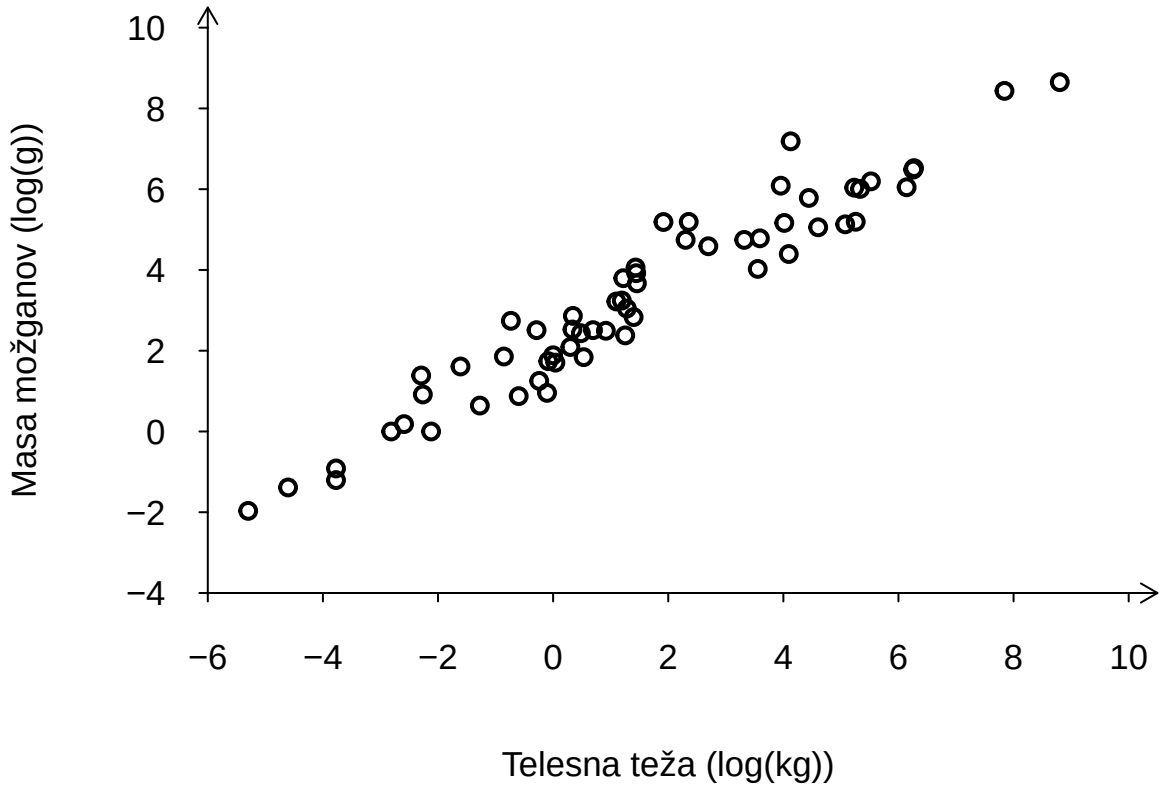
```
## [1] 951.7869
```

Vidimo, da teža možganov varira med 0.14 in 5711.86. Povrpečna teža možganov znaša 297.41g s standardnim odklonom 951.7869g. Razpon vrednosti telesnih tež sesalcev in teže njihovih možganov nam pomaga pri izbiri mej na oseh razsevnega diagrama.

3. Razsevni diagram in vzorčni koeficient korealcije

Prikažimo dobljene podatke na razsevnem diagramu.

```
teltezaLog<-log(mozgani$telteza)
moztezaLog<-log(mozgani$mozteza)
par(las=1, cex=1.1, mar=c(4,4,2,2))
plot(teltezaLog, moztezaLog, main="", xlim=c(-6,10), ylim=c(-4,10),
     xlab="Telesna teža (log(kg))", ylab="Masa možganov (log(g))", lwd=2, axes=FALSE)
axis(1,pos=-4,at=seq(-6,10,by=2),tcl=-0.2)
axis(2,pos=-6,at=seq(-4,10,by=2),tcl=-0.2)
arrows(x0=10,y0=-4,x1=10.5,y1=-4,length=0.1)
arrows(x0=-6,y0=10,x1=-6,y1=10.5,length=0.1)
```



Točke na razsevnem diagramu se nahajajo okoli namišljene premice, torej lahko sklepamo, da je linearni model primeren.. Moč korelacije preverimo še z računanjem Pearsonovega koeficienta korelacije.

```
(r<-cor(teltezaLog,moztezaLog))
```

```
## [1] 0.9628862
```

Vrednost vzorčnega koeficienta korelacije je visoka ($r = 0.9628862$), kar pomeni, da je linearna povezanost telesne teže sesalcev z težo njihovih možganov visoka. Koeficient korelacije je pozitiven, kar nam pove, da imajo vrste sesalcev, ki imajo manjšo telesno težo, tudi manjšo težo možganov.

4. Formiranje linearnega regresijskega modela

Formirajmo linearni regresijski model.

```
(model<-lm(moztezaLog~teltezaLog,data=mozgani))
```

```
##
## Call:
## lm(formula = moztezaLog ~ teltezaLog, data = mozgani)
##
## Coefficients:
## (Intercept)   teltezaLog
##      2.1781       0.7468
```

Dobili smo ocenjeno regresijsko premico $\hat{y} = 2.1781 + 0.7468x$, oziroma oceni odseka in naklona sta enaki $\hat{a} = 2.1781$ in $\hat{b} = 0.7468$.

5. Točke visokega vzvoda in osamelci

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

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

```
##                  vrsta              slovime  telteza mozteza
## 3  Blarina brevicauda                Rovka    0.005    0.14
## 16     Elephas maximus         Azijski slon 2547.070 4603.17
## 30 Loxodonta africana          Afriski slon 6654.180 5711.86
## 36   Myotis lucifugus  Majhni rjavi netopir    0.010    0.25
```

Odkrili smo 4 točke visokega vzvoda. Dve vrsti imata zelo majhno telesno težo (pod 0.5kg), dve pa zelo veliko telesno težo (nad 2000kg).

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

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

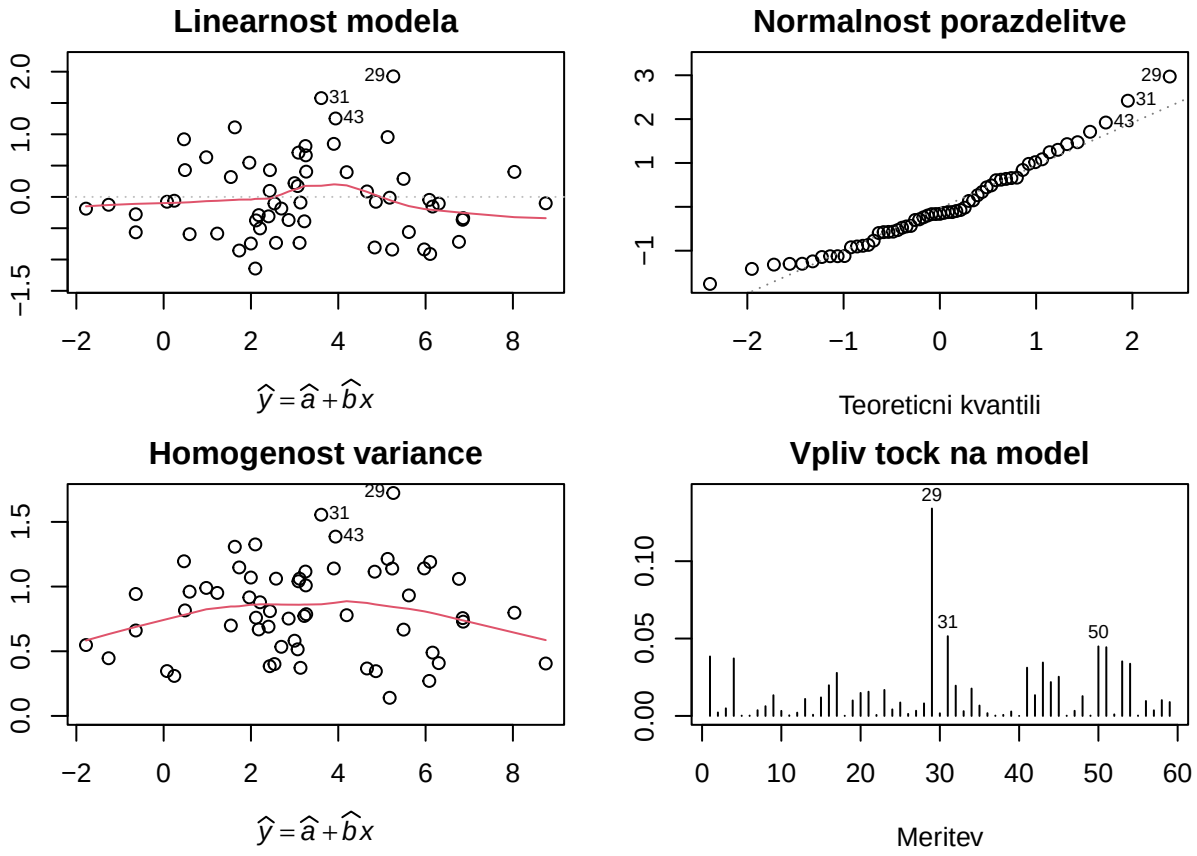
```
##                   vrsta      slovime telteza  mozteza
## 29 Homo sapiens sapiens       Clovek  61.998 1320.020
## 31       Macaca mulatta Rezus makaki   6.800  179.003
```

Dve podatkovni točki sta osamelci in se nanašata na dve vrste sesalcev. Ena vrsta ima veliko težo možganov glede na njeno telesno težo, druga pa majhno težo možganov glede na njeno 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, opazimo rahlo nagnetenost točk med 2 in 6.

2) Graf normalnosti porazdelitve naključnih napak

Normalnost porazdelitve naključnih napak preverjamo s pomočjo 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 sesalcev in njihovih možganov lahko sklenemo, da so naključne napake normalno porazdeljene (ni večjih odstopanj od premice, razen za 29., 31., in 43. 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.

Iz točk na grafu lahko sklepamo, da ni pretiranega padanja variance, naraste le za 29., 31. in 43. podatkovno točko. Ničelna domneva konstantne variance se lahko formalno preveri s Breusch-Paganovim testom. ,,

```
suppressWarnings(library(car))
```

```
## Loading required package: carData
```

```
ncvTest(model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.5732155, Df = 1, p = 0.44898
```

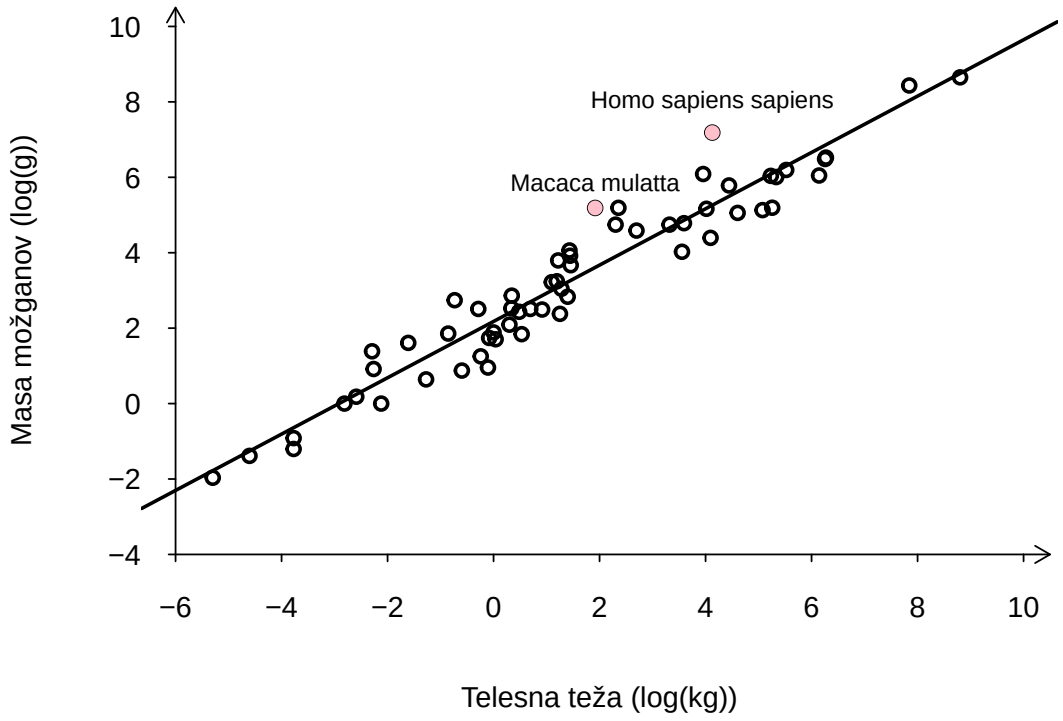
Na osnovi rezultata Breusch-Paganovega testa (testna statistika $\chi^2 = 0.5732155$, $df = 1$, je p-vrednost $p = 0.44898 > 0.05$), ne zavrnemo ničelne domneve. Ni dovolj dokazov, da varianca naključnih napak ni homogena.

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 29., 31., in 50. podatkovne točka. Spomnimo se, da smo dve izmed točk identificirali kot osamelce (29. in 31. podatkovna točka). Zdaj poglejmo na razsevnem diagramu po čem so ti dve točki drugačni od ostalih. Kodi za razsevni diagram dodamo če dve vrstici, s katerima bomo dodali ocenjeno regresijsko premico in pobarvali ti dve točki.

```
par(las=1, mar=c(4,4,2,2))
plot(teltezaLog, moztezaLog, main="", xlim=c(-6, 10), ylim=c(-4, 10),xlab=
"Telesna teža (log(kg))", ylab="Masa možganov (log(g))",lwd=2, axes=FALSE)
axis(1,pos=-4,at=seq(-6,10,by=2),tcl=-0.2)
axis(2,pos=-6,at=seq(-4,10,by=2),tcl=-0.2)
arrows(x0=10,y0=-4,x1=10.5,y1=-4,length=0.1)
arrows(x0=-6,y0=10,x1=-6,y1=10.5,length=0.1)
abline(model,lwd=2)
points(teltezaLog[c(29,31)],moztezaLog[c(29,31)],col="pink",pch=19)
text(teltezaLog[c(29,31)],moztezaLog[c(29,31)]+c(0.2,0),labels=
mozgani$vrsta[c(29,31)],pos=3,cex=0.8)
```



Na razsevnem diagramu opazimo, da sta obe točki najbolj oddaljene od ocenjene regresijske premice (oziroma jim ustrezajo največji ostanki). Lahko preverimo še, ali je njun vpliv velik, oziroma ali je njuna Cookova razdalja večja ali enaka od mediane Fisherjeve porazdelitve z 2 in 30 prostostnimi stopnjami.

```
any(cooks.distance(model)[c(29,31)]>=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.

```
summary(model)
```

```
##
## Call:
## lm(formula = moztezaLog ~ teltezaLog, data = mozgani)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1439 -0.4461 -0.1018  0.4015  1.9252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.17812    0.09487   22.96   <2e-16 ***
## teltezaLog   0.74679    0.02773   26.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6579 on 57 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9259
## F-statistic: 725.4 on 1 and 57 DF,  p-value: < 2.2e-16
```

Vrednost testne statistike za preverjanje linearnosti modela je enaka $t = 26.93$, s $df = 57$ 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 $R^2 = 0.9271$, kar pomeni, da 92.7% variabilnosti teže možganov pojasnjuje linearni 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) 1.988  2.368
## teltezaLog  0.691  0.802
```

Interval zaupanja za odsek je enak $I_a = [1.988, 2.368]$ in interval zaupanja za naklon $I_b = [0.691, 0.802]$.

9. 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 $\hat{y} = 2.1781 + 0.7468x_0$ vrst določene teže x_0, ampak želimo tudi oceniti spodnjo in zgornjo mejo, med katerima se verjetno nahaja masa možganov različnih vrst teh telesnih tež.

```
xtelteza = data.frame(teltezaLog=log(c(10, 50, 100, 500)))
exp(predict(model, xtelteza, interval="predict"))
```

```
##         fit       lwr       upr
## 1  49.28701  13.04488  186.2193
## 2 163.95111  43.12970  623.2357
## 3 275.11845  72.05234 1050.4885
## 4 915.16959 236.22735 3545.4631
```

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

1. 10 kg je 49.28701 g, s 95% intervalom predikcije teže možganov [13.04488, 186.2193],
2. 50 kg je 163.95111 g, s 95% intervalom predikcije teže možganov [43.12970, 623.2357],
3. 100 kg je 275.11845 g, s 95% intervalom predikcije teže možganov [72.05234, 1050.4885],
4. 500 kg je 915.16959 g, s 95% intervalom predikacije teže možganov [236.22735, 3545.4631]

10. Zaključek

Zanimala nas je funkcionalna odvisnost med telesno težo sesalcev in njihovo težo možganov, merjeno v gramih. Zbrali smo vzorec 59 vrst sesalcev, jim izmerili telesno težo in zabeležili težo možganov. Ugotovili smo, da je enostavni linearni model odvisnosti teže možganov od telesne teže dober. Diagnostični grafi in statistični testi niso pokazali težav 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 njegove telesne teže je zadovoljiva, vendar bi vključevanje dodatnih neodvisnih spremenljivk zagotovo dala še boljši model in bolj zanesljivo napoved.