Napoved teže možganov z linearno regresijo

Tim Hajdinjak

1. Opis podatkov

Zbrali smo vzorec telesnih tež in tež možganov 58 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, ki predstavlja telesno težo sesalcev, merjeno v kilogramih.
4. *mozteza* je numerična zvezna spremenljivka, ki predstavlja težo možganov sesalcev, merjeno v gramih.

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

```
sesalci<-read.csv("C:/Users/Tim Hajdinjak/Desktop/mozgani.csv", header=TRUE)
str(sesalci)
```

```
## '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 ...
```

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 in teže možganov.

```
summary(sesalci$telteza)
```

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

```
sd(sesalci$telteza)
```

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

Opazimo, da telesna teža vzorca sesalcev varira od 0.005 do 6654.180kg, s povprečjem 212.428 in standardnim odklonom 928.6204 kg. Ponovimo postopek računanja za teže možganov sesalcev.

```
summary(sesalci$mozteza)
```

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

```
sd(sesalci$mozteza)
```

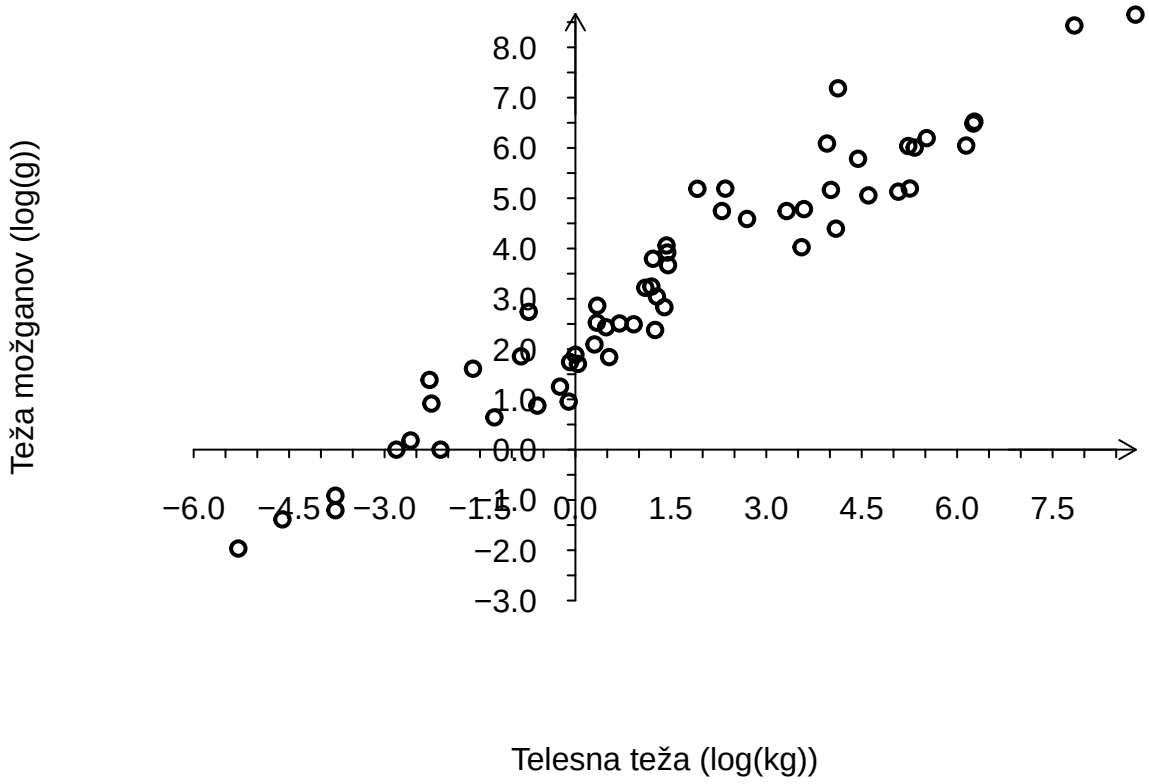
```
## [1] 959.3438
```

Opazimo, da teža možganov vzorca sesalcev varira od 0.14 do 5711.86 gramov, s povprečjem 302.32 in standardnim odklonom 959.3438 gramov. Razpon vrednosti telesnih tež in tež možganov vzorca sesalcev 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, mar=c(5,5,1,1))
plot(log(sesalci$telteza), log(sesalci$mozteza), main="", xlim=c(-6,log(6700)),ylim=c(-3,
log(5800)), xlab="Telesna teža (log(kg))", ylab="Teža možganov (log(g))", lwd=2, axes=FALSE)
axis(1,pos=0,at=seq(-6,log(6700),by=0.5),tcl=-0.2)
axis(2,pos=0,at=seq(-3,log(5800),by=0.5),tcl=-0.2)
arrows(x0=log(6700)-2,y0=0,x1=log(6700),y1=0,length=0.1)
arrows(x0=0,y0=log(5800)-2,x1=0,y1=log(5800),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(log(sesalci$telteza),log(sesalci$mozteza)))
```

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

Vrednost vzorčnega koeficienta korelacije je visoka ($r = 0.9632881$), kar govori o visoki linearni povezanosti telesnih tež sesalcev in njihovih tež možganov. Dalje, koeficient korelacije je pozitiven, kar pomeni, da sesalci visokih telesnih tež imajo visoke teže svojih možganov.

4. Formiranje linearnega regresijskega modela

Formirajmo linearni regresijski model.

```
(model<-lm(log(mozteza)~log(telteza),data=sesalci))
```

```
##
## Call:
## lm(formula = log(mozteza) ~ log(telteza), data = sesalci)
##
## Coefficients:
##  (Intercept)  log(telteza)
##       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 $\frac{4}{n}$.

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

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

Odkrili smo 4 točke visokega vzvoda. Dve vrsti sesalcev imajo visoko telesno težo nad 2500 kg, druge dve vrsti pa telesno težo pod 0.1 kg.

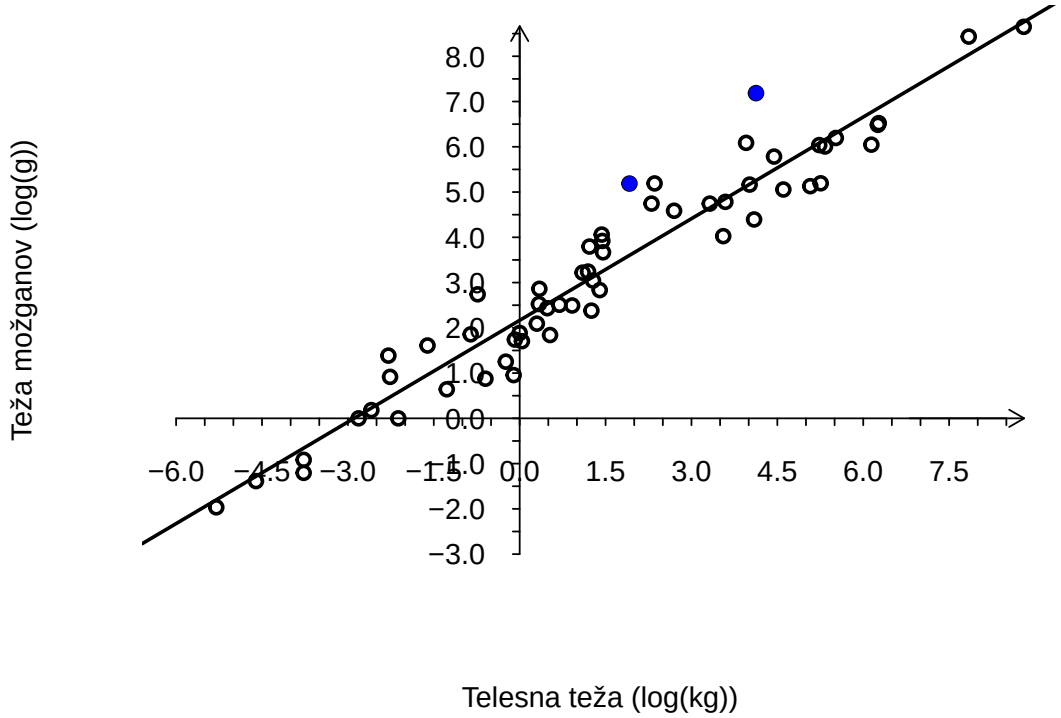
Za podatke majhne in srednje velikosti vzorca je osamelec podatkovna točka, kateri ustreza standardizirani ostanek izven intervala $[-2, 2]$.

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

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

Identificirali smo dve podatkovni točki (28. in 30. točka) kot osamelca. Zdaj poglejmo na razsevnem diagramu po čem sta te 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, cex=1, mar=c(5,5,1,1))
plot(log(sesalci$telteza), log(sesalci$mozteza), main="", xlim=c(-6,log(6700)),ylim=c(-3,
log(5800)), xlab="Telesna teža (log(kg))", ylab="Teža možganov (log(g))", lwd=2, axes=FALSE)
axis(1,pos=0,at=seq(-6,log(6700),by=0.5),tcl=-0.2)
axis(2,pos=0,at=seq(-3,log(5800),by=0.5),tcl=-0.2)
arrows(x0=log(6700)-2,y0=0,x1=log(6700),y1=0,length=0.1)
arrows(x0=0,y0=log(5800)-2,x1=0,y1=log(5800),length=0.1)
abline(model,lwd=2)
points(log(sesalci$telteza[c(28,30)]),log(sesalci$mozteza[c(28,30)]),col="blue",pch=19)
text(sesalci$telteza[c(28,30)],sesalci$mozteza[c(28,30)]+c(0.2,0),labels=
sesalci$slovime[c(28,30)],pos=3,cex=0.7)
```

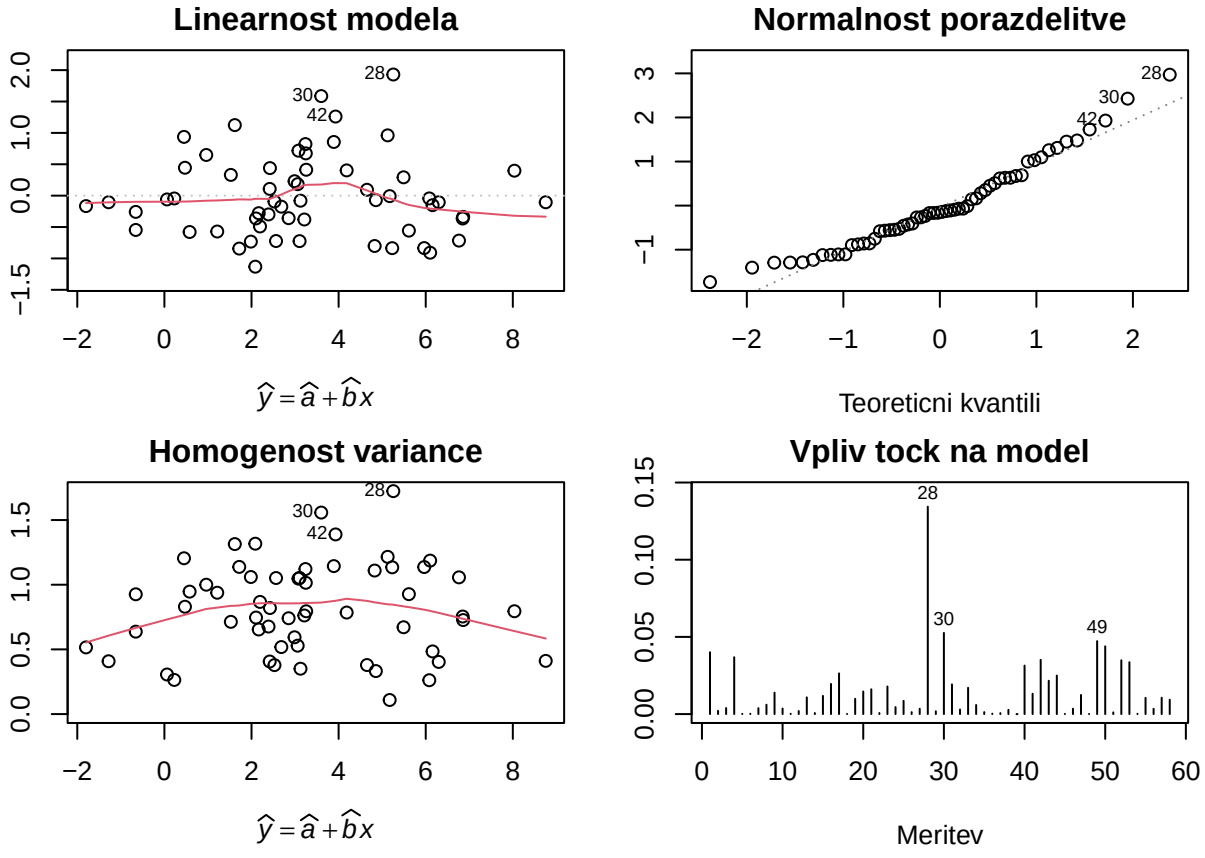


Na razsevnem diagramu opazimo, da se omenjena osamelca nanašata na dve vrsti sesalcev z nenavadno visoko težo možganov glede na telesno težo. Opazimo še, da nobena podatkovna točka ni hkrati točka visokega vzvoda in osamelec.

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 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 sesalcev in njihovih tež možganov lahko zaključimo, da so naključne napake normalno porazdeljene (ni večjih odstopanj od premice, razen za 28., 30., in 42. 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., 30., in 49. podatkovna točka. Spomnimo se, da smo dve od teh treh točk (28. in 30.) identificirali kot osamelca. 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(28, 30, 49)]>=qf(0.5,2,nrow(sesalci)-2))
```

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

Nobena od teh točk nima velik vpliv 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(mozteza) ~ log(telteza), data = sesalci)
##
## 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 ***
## log(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 $p < 2 \cdot 10^{-16}$, 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 $R^2 = 0.9279$, kar pomeni, da 92% 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 $\hat{y} = 97.341 + 0.965x_0$ sesalcev določene telesne teže x_0, ampak želimo tudi oceniti spodnjo in zgornjo mejo, med katerima se verjetno nahaja teža možganov različnih sesalcev teh telesnih tež.

```
xmasa = data.frame(telteza=c(40,250,1200))
logs<-predict(model, xmasa, interval="predict")
predictions<-exp(logs)
predictions
```

```
##         fit       lwr       upr
## 1  138.0102  36.20086  526.1428
## 2  544.0692 140.84255 2101.7179
## 3 1760.2975 447.87963 6918.4822
```

Predvidena vrednost teže možganov za sesalce s telesno težo (na celi populaciji sesalcev):

1. 40 kg je 138.0102 g, s 95% intervalom predikcije teže možganov $[36.20084, 526.1428]$,
2. 250 kg je 544.0692 g, s 95% intervalom predikcije teže možganov $[140.84255, 2101.7179]$,
3. 1200 kg je 1760.2975 g, s 95% intervalom predikcije teže možganov $[447.87963, 6918.4822]$.

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 zabeležili 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 92%, 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.