# Linearna regresija: poraba goriva

## 1. Opis podatkov

Zbrali smo vzorec mase in porabe goriva 32 avtomobilov različnih znamk. Podatke smo zapisali v dokument, ki ima tri stolpce:

1. *model* je nominalna spremenljivka, katere vrednosti so nazivi modela avtomobila.
2. *masa* je numerična zvezna spremenljivka, ki predstavlja maso avtomobilov, merjeno v kilogramih.
3. *kml* je numerična zvezna spremenljivka, ki predstavlja porabo goriva avtomobila, merjeno v številu prevoženih kilometrov na liter goriva (na dve decimalki).

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

```
avto<-read.csv("C:/avto.csv", header=TRUE)
str(avto)
```

```
## 'data.frame':    32 obs. of  3 variables:
##  $ model: chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
##  $ masa : int  1188 1304 1052 1458 1560 1569 1619 1447 1429 1560 ...
##  $ kml  : num  8.93 8.93 9.69 9.1 7.95 ...
```

## 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 mase in porabe goriva.

```
summary(avto$masa)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     686    1170    1508    1459    1637    2460
```

```
sd(avto$masa)
```

```
## [1] 443.7458
```

Opazimo, da masa vzorca avtomobilov varira od 686 do 2460kg, s povprečjem 1459.2 in standardnim odklonom 443.7 kg. Ponovimo postopek računanja za vzorec porabe goriva.

```
summary(avto$kml)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.420   6.558   8.160   8.541   9.690  14.410
```

```
sd(avto$kml)
```

```
## [1] 2.561557
```

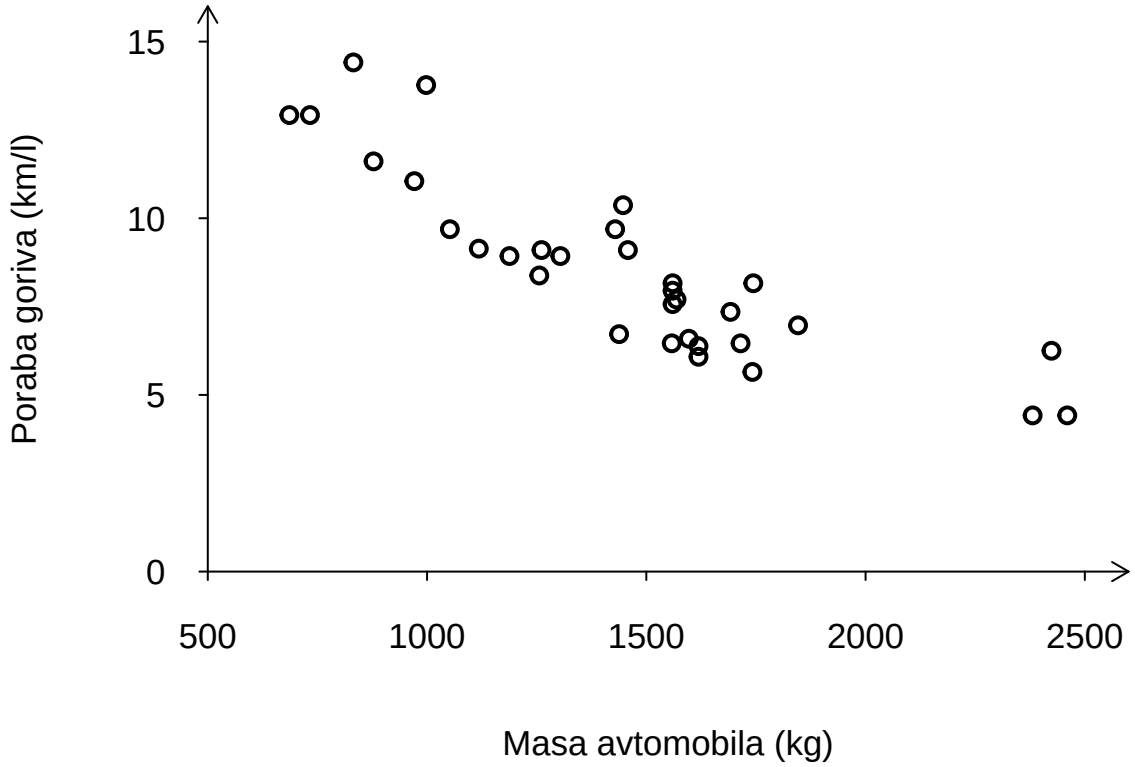
Opazimo, da poraba goriva vzorca avtomobilov varira od 4.42 do 14.41 prevoženih kilometrov na liter goriva, s povprečjem 8.541 in standardnim odklonom 2.562 km/l$. Razpon vrednosti mase avtomobilov in porabe goriva nam pomaga pri izbiri mej na oseh razsevnega diagrama.

## 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(avto$masa, avto$kml, main="", xlim=c(500,2600), ylim=c(0,16),
xlab="Masa avtomobila (kg)", ylab="Poraba goriva (km/l)", lwd=2, axes=FALSE)
```

```
axis(1,pos=0,at=seq(500,2500,by=500),tcl=-0.2)
axis(2,pos=500,at=seq(0,15,by=5),tcl=-0.2)
arrows(x0=2500,y0=0,x1=2600,y1=0,length=0.1)
arrows(x0=500,y0=15,x1=500,y1=16,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(avto$masa,avto$kml))
```

```
## [1] -0.8678461
```

Vrednost vzorčnega koeficienta korelacije je visoka ($r = -0.868$), kar govori o visoki linearni povezanosti mase avtomobilov in njihove porabe goriva. Dalje, koeficient korelacije je negativen, kar pomeni, da avtomobili manjše mase prevozijo več kilometrov na liter goriva.

## 4. Formiranje linearnega regresijskega modela

Formirajmo linearni regresijski model.

```
(model<-lm(kml~masa,data=avto))
```

```
##
## Call:
## lm(formula = kml ~ masa, data = avto)
##
## Coefficients:
## (Intercept)         masa
```

```
##    15.85089     -0.00501
```

Dobili smo ocenjeno regresijsko premico $\hat{y} = 15.851 - 0.005x$, oziroma oceni odseka in naklona sta enaki $\hat{a} = 15.851$ in $\hat{b} = -0.005$.

## 5. Točke visokega vzvoda in osamelci

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

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

```
##                   model masa   kml
## 15  Cadillac Fleetwood 2381  4.42
## 16 Lincoln Continental 2460  4.42
## 17   Chrysler Imperial 2424  6.25
## 28        Lotus Europa  686 12.92
```

Odkrili smo 4 točke visokega vzvoda. Tri znamke avtomobilov imajo visoko maso nad 2000 kg in ena znamka najnižjo maso, pod 700 kg.

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

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

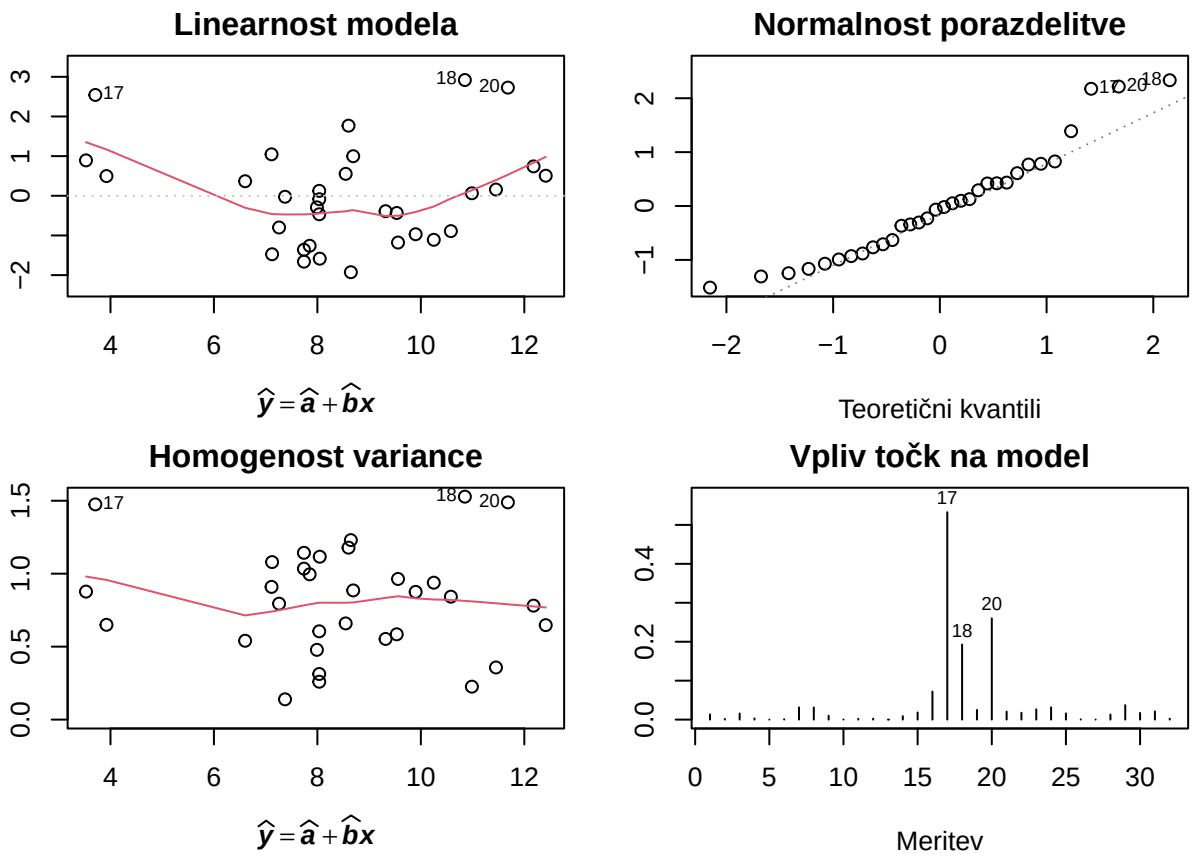
```
##                model masa   kml
## 17 Chrysler Imperial 2424  6.25
## 18          Fiat 128  998 13.77
## 20    Toyota Corolla  832 14.41
```

Tri podatkovne točke so osamelci in se nanašajo na tri znamke avtomobilov z nenavadno velikim številom prevoženih kilometrov na liter goriva glede na njihovo maso. Opazimo še, da je 17. podatkovna točka (Chrysler Imperial) 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 6 do 10, kar je prisotno zaradi originalnih vrednosti v vzorcu avtomobilov (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 masi in porabi goriva avtomobilov lahko zaključimo, da so naključne napake normalno porazdeljene (ni večjih odstopanj od premice, razen za 17., 18., in 20. 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 ◁, in pri padanju variance oblike ▷. 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)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.03700701, Df = 1, p = 0.84745
```

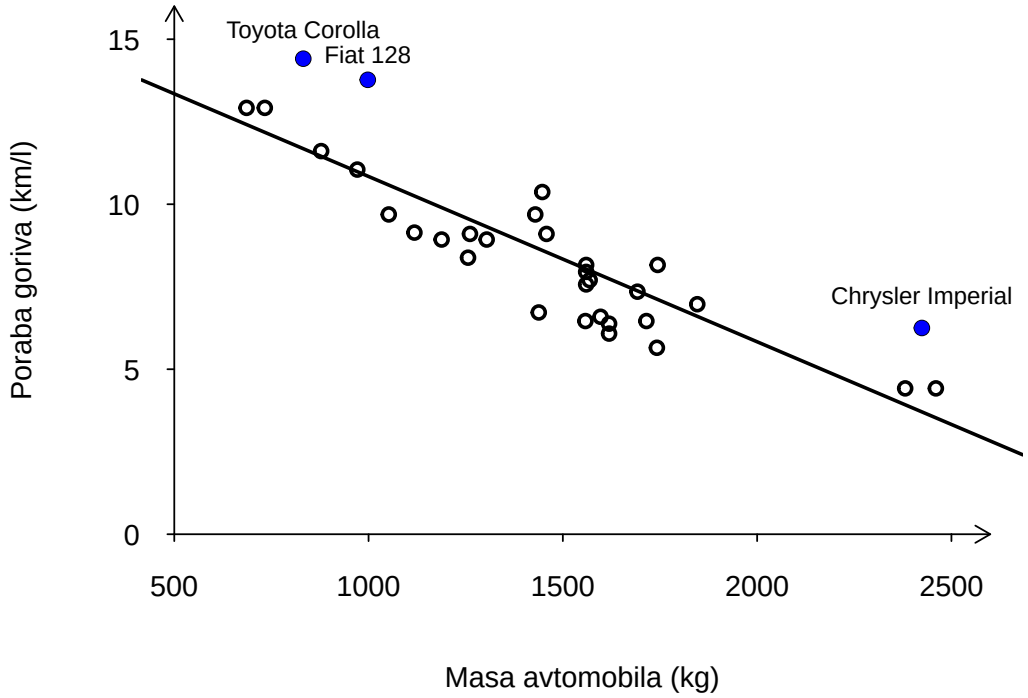
Na osnovi rezultata Breusch-Paganovega testa (testna statistika $\chi^2 = 0.037$, $df = 1$, p-vrednost $p = 0.847 > 0.05$), ne zavrnemo ničelne domneve. Ni dovolj dokazov, da varianca naključnih napak ni homogena.

### 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 17., 18., in 20. podatkovne točka. Spomnimo se, da smo te točke identificirali kot osamelce. 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.

```
par(las=1, mar=c(4,4,2,3))
plot(avto$masa, avto$kml, main="", xlim=c(500,2600), ylim=c(0,16),xlab=
"Masa avtomobila (kg)", ylab="Poraba goriva (km/l)",lwd=2, axes=FALSE)
axis(1,pos=0,at=seq(500,2500,by=500),tcl=-0.2)
axis(2,pos=500,at=seq(0,15,by=5),tcl=-0.2)
arrows(x0=2500,y0=0,x1=2600,y1=0,length=0.1)
arrows(x0=500,y0=15,x1=500,y1=16,length=0.1)
abline(model,lwd=2)
points(avto$masa[c(17,18,20)],avto$kml[c(17,18,20)],col="blue",pch=19)
text(avto$masa[c(17,18,20)],avto$kml[c(17,18,20)]+c(0.2,0,0.1),labels=
avto$model[c(17,18,20)],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(17,18,20)]>=qf(0.5,2,nrow(avto)-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 = kml ~ masa, data = avto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9269 -1.0045 -0.0551  0.6003  2.9188
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.8508866  0.7975736  19.874  < 2e-16 ***
## masa        -0.0050097  0.0005236  -9.567 1.27e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.294 on 30 degrees of freedom
## Multiple R-squared:  0.7532, Adjusted R-squared:  0.7449
## F-statistic: 91.53 on 1 and 30 DF,  p-value: 1.268e-10
```

Vrednost testne statistike za preverjanje linearnosti modela je enaka $t = -9.567$, s $df = 30$ prostostnimi stopnjami in s p-vrednostjo $p = 1.27 \cdot 10^{-10}$, 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 $R^2 = 0.753$, kar pomeni, da 75% variabilnosti porabe goriva 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) 14.222 17.480
## masa        -0.006 -0.004
```

Interval zaupanja za odsek je enak $I_a = [14.222, 17.480]$ in interval zaupanja za naklon $I_b = [-0.006, -0.004]$.

## 9. Interval predikcije za vrednost $Y$ pri izbrani vrednosti $X$

Pri predvidevanju vrednosti porabe goriva nas zanima bodoča vrednost spremenljivke Y pri izbrani vrednosti spremenljivke $X = x_0$. Ne zanima nas le predvidena vrednost $\hat{y} = 15.851 - 0.005x_0$ avtomobilov določene mase $x_0$, ampak želimo tudi oceniti spodnjo in zgornjo mejo, med katerima se verjetno nahaja poraba goriva različnih modelov avtomobilov teh mas.

```
xmasa = data.frame(masa=c(1000,1500,2000))
predict(model, xmasa, interval="predict")
```

```
##         fit      lwr       upr
## 1 10.841177 8.113546 13.568808
## 2  8.336323 5.652908 11.019737
## 3  5.831468 3.086793  8.576144
```

Predvidena vrednost porabe goriva za avtomobil mase (na celi populaciji avtomobilov)

1. 1000 kg je 10.84 km/l, s 95% intervalom predikcije porabe goriva [8.11, 13.57],
2. 1500 kg je 8.34 km/l, s 95% intervalom predikcije porabe goriva [5.65, 10.02],
3. 2000 kg je 5.83 km/l, s 95% intervalom predikcije porabe goriva [3.09, 8.58]

## 10. Zaključek

Zanimala nas je funkcionalna odvisnost med maso avtomobilov in njihovo porabo goriva, merjeno kot število prevoženih kilometrov na liter goriva. Zbrali smo vzorec 32 znamk avtomobilov, jim izmerili maso in zabeležili porabo goriva. Ugotovili smo, da je enostavni linearni model odvisnosti porabe goriva od mase dober. Diagnostični grafi in statistični testi niso pokazali na težave z linearnim regresijskim modelom. Koeficient determinacije je 75%, kar pomeni, da tolikšen delež variabilnosti porabe goriva zajamemo z linearnim modelom. Napoved porabe mase na osnovi njegove mase je zadovoljiva, vendar bi vključevanje dodatnih neodvisnih spremenljivk zagotovo dala še boljši model in bolj zanesljivo napoved.