Linearna regresija: zračni tlak

1. Opis podatkov

Dobili smo vzorec temperature vretja vode in zračnega tlaka na 17 lokacijah v Alpah. Podatke smo zapisali v dokument, ki ima 2 stolpca:

1. *temp* je numerična zvezna spremenljivka, ki predstavlja temperaturo vretja vode, merjeno v stopinjah Celzija.
2. *tlak* je numerična zvezna spremenljivka, ki predstavlja zračni tlak, merjeno v milibarih.

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

```
forbes<-read.csv("C:\\Users\\jovan\\Documents\\FRI\\forbes.csv", header=TRUE)
str(forbes)
```

```
## 'data.frame':    17 obs. of  2 variables:
##  $ temp: num  90.3 90.2 92.2 92.4 93 93.3 93.8 93.9 94.1 94.1 ...
##  $ tlak: num  704 704 759 768 784 ...
```

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 temperature in zračnega tlaka.

```
summary(forbes$temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   90.20   93.00   94.10   94.97   98.10  100.10
```

```
sd(forbes$temp)
```

```
## [1] 3.190369
```

Opazimo, da temperatura vretja vzorca vode varira od 90.20 do 100.10 stopnjah Celzija, s povprečjem 94.97 in standardnim odklonom 3.19 stopnjah Celzija. Ponovimo postopek računanja za vzorec zračnega tlaka.

```
summary(forbes$tlak)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   704.0   783.9   813.1   848.6   940.1  1017.9
```

```
sd(forbes$tlak)
```

```
## [1] 102.2813
```

Opazimo, da varira od 704.0 do 1017.9 milibarih, s povprečjem 848.6 in standardnim odklonom 102.28 milibarih. Razpon vrednosti temperature vretja vode in zračnega tlaka nam pomaga pri izbiri mej na oseh razsevnega diagrama.

3. Razsevni diagram in vzorčni koeficient korealcije

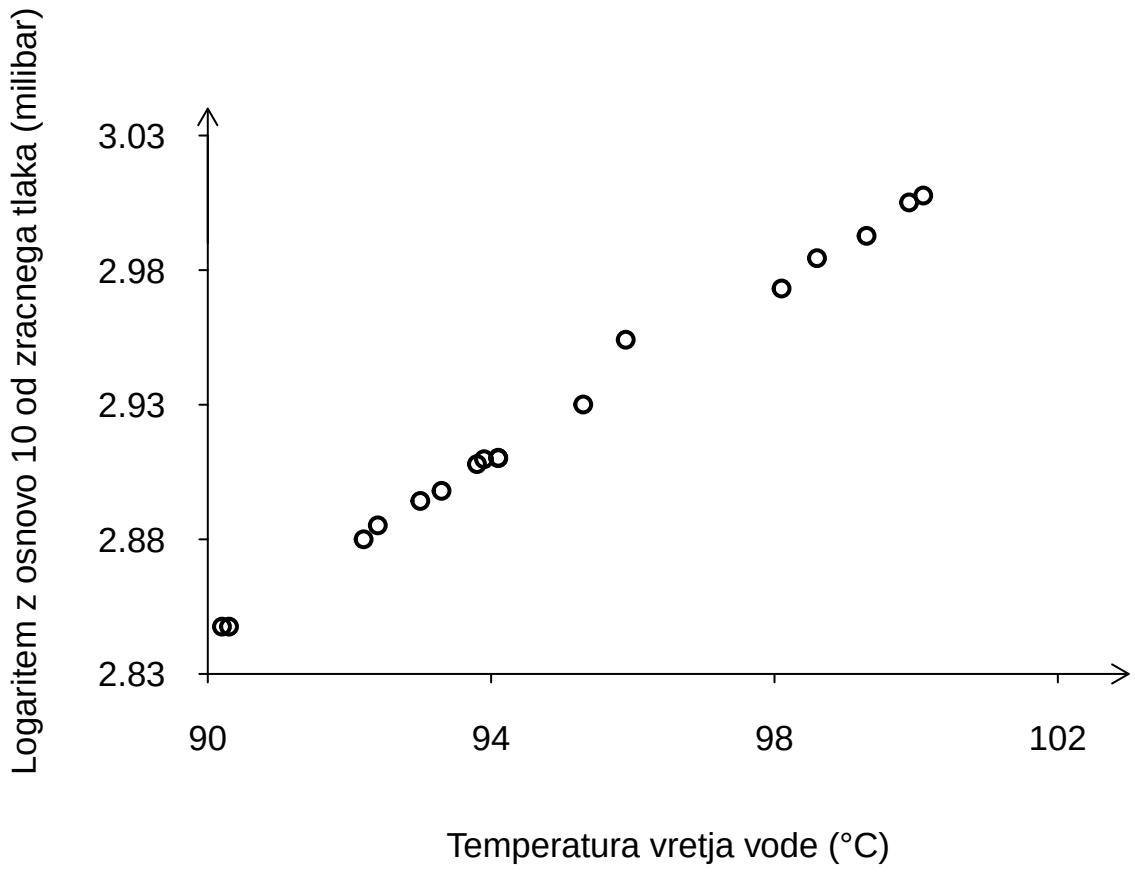
Prikažimo dobljene podatke na razsevnem diagramu. Za odvisno spremenljivko, zaradi boljšega modela, bomo vzeli log_10(forbes$tlak), za neodvisno forbes$temp.

```
par(las=1, cex=1.1, mar=c(4,4,2,2))
temp<-forbes$temp
log_tlak<-log10(forbes$tlak)
```

```
min_tlak <-signif(log10(680), 3)
max_tlak <-signif(log10(1100), 3)

min_temp <- 90
max_temp <- 103

plot(temp, log_tlak, main="", xlim=c(min_temp,max_temp), ylim=c(min_tlak, max_tlak),
xlab="Temperatura vretja vode (°C)",
ylab="Logaritem z osnovo 10 od zračnega tlaka (milibar)", lwd=2, axes=FALSE)
axis(1,pos=min_tlak,at=seq(min_temp,max_temp,by=4),tcl=-0.2)
axis(2,pos=min_temp,at=seq(min_tlak, max_tlak,by=0.05),tcl=-0.2)
arrows(x0=max_temp-1,y0=min_tlak,x1=max_temp,y1=min_tlak,length=0.1)
arrows(x0=min_temp,y0=max_tlak -0.05,x1=min_temp,y1=max_tlak,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(temp,log_tlak))
```

```
## [1] 0.9974508
```

Vrednost vzorčnega koeficienta korelacije je visoka ($r = 0.997$), kar govori o visoki linearni povezanosti temperature vretja vode in zračnega tlaka. Dalje, koeficient korelacije je pozitiven, kar pomeni, da ko se temperatura vretja vode poveča, se poveča tudi zračni tlak.

4. Formiranje linearnega regresijskega modela

Formirajmo linearni regresijski model.

```
(model<-lm(log_tlak~temp, data = forbes))
```

```
##
## Call:
## lm(formula = log_tlak ~ temp, data = forbes)
##
## Coefficients:
## (Intercept)         temp
##     1.39011      0.01617
```

Dobili smo ocenjeno regresijsko premico $\hat{y} = 1.39 + 0.01617x$, oziroma oceni odseka in naklona sta enaki $\hat{a} = 1.39$ in $\hat{b} = 0.01617$.

5. Točke visokega vzvoda in osamelci

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

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

```
## [1] temp tlak
## <0 rows> (or 0-length row.names)
```

Nimamo točke visokega vzvoda.

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

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

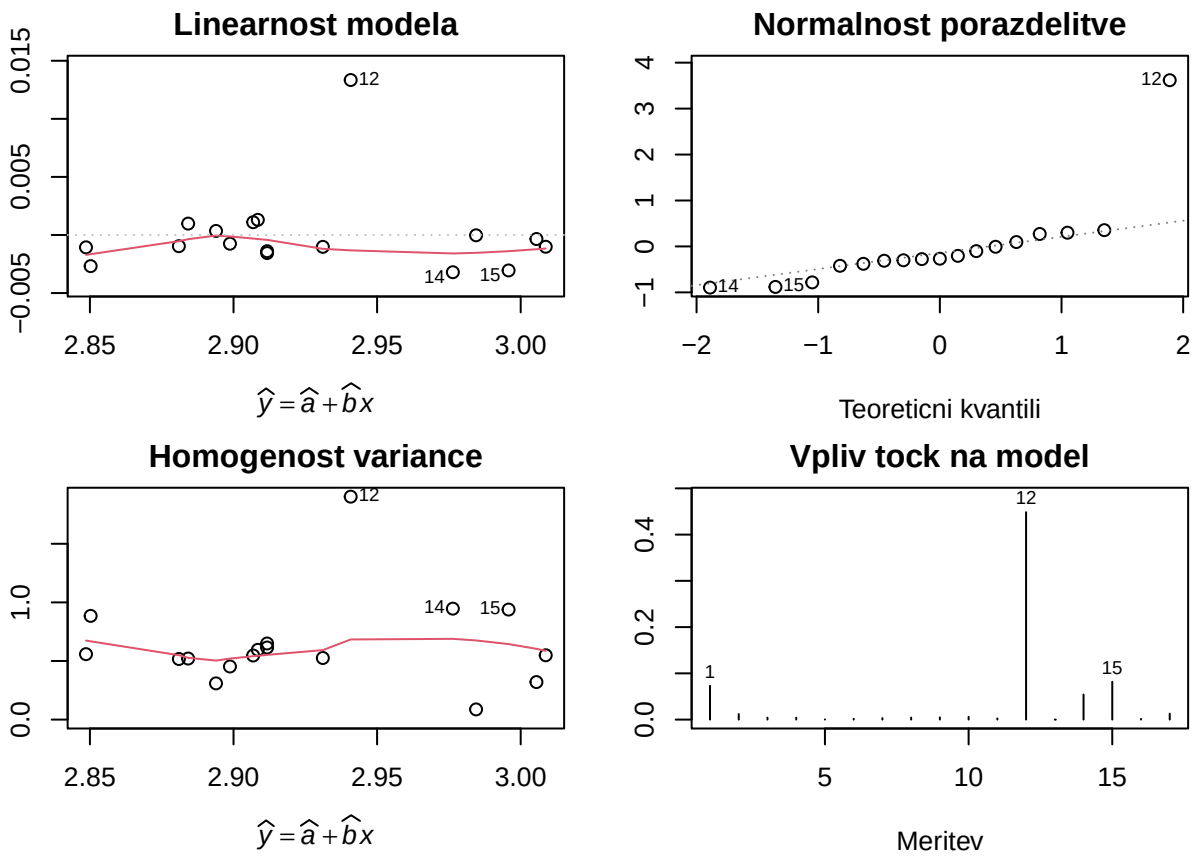
```
##    temp  tlak
## 12 95.9 899.8
```

Ena podatkovna točka je osamelec in se nanaša na 12. podatkovno točko, kjer je izmerjena vrednost temperature vretja vode 95.9 °C in vrednost zračnega tlaka 899.8.

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 okoli 2.90.

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 temperaturi vretja vode in zračnega tlaka lahko zaključimo, da so naključne napake normalno porazdeljene (ni večjih odstopanj od premice, razen za 12. 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. 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 = 0.6879209, Df = 1, p = 0.40687
```

Na osnovi rezultata Breusch-Paganovega testa (testna statistika $\chi^2 = 0.688$, $df = 1$, p-vrednost $p = 0.407 > 0.05$), 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 1., 12. in 15. podatkovne točka. Spomnimo se, da smo 12. točko identificirali kot osamelec. 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.

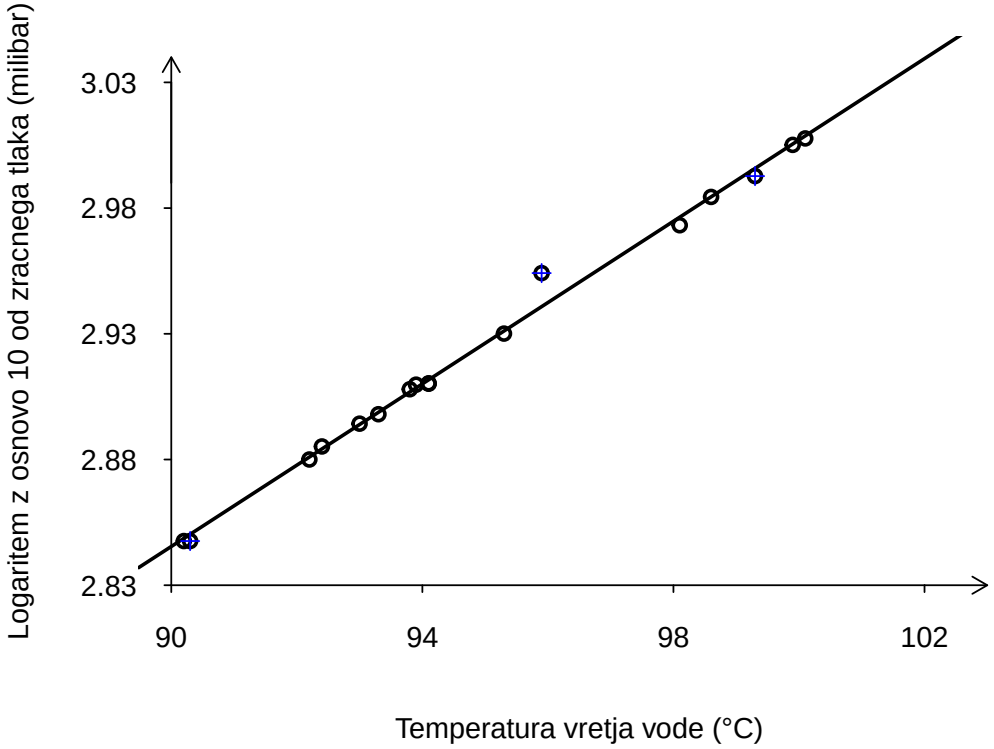
```
par(las=1, mar=c(4,4,2,3))
temp<-forbes$temp
log_tlak<-log10(forbes$tlak)

min_tlak <-signif(log10(680), 3)
max_tlak <-signif(log10(1100), 3)

min_temp <- 90
max_temp <- 103

plot(temp, log_tlak, main="", xlim=c(min_temp,max_temp), ylim=c(min_tlak, max_tlak),
xlab="Temperatura vretja vode (°C)", ylab="Logaritem z osnovo 10 od zračnega tlaka (milibar)", lwd=2, a
axis(1,pos=min_tlak,at=seq(min_temp,max_temp,by=4),tcl=-0.2)
axis(2,pos=min_temp,at=seq(min_tlak, max_tlak,by=0.05),tcl=-0.2)
arrows(x0=max_temp-1,y0=min_tlak,x1=max_temp,y1=min_tlak,length=0.1)
arrows(x0=min_temp,y0=max_tlak -0.05,x1=min_temp,y1=max_tlak,length=0.1)

abline(model, lwd=2)
points(temp[c(1,12,15)],log_tlak[c(1,12,15)],col="blue",pch=3)
text(temp[c(1,12,15)], log_tlak[c(1,12,15)]+c(0.2,0,0.1),labels=
forbes$model[c(1,12,15)],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(1,12,15)]>=qf(0.5,2,nrow(forbes)-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 = log_tlak ~ temp, data = forbes)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0032050 -0.0013950 -0.0009634  0.0003484  0.0133409
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.3901054  0.0283815   48.98   <2e-16 ***
## temp        0.0161700  0.0002987   54.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.003812 on 15 degrees of freedom
## Multiple R-squared:  0.9949, Adjusted R-squared:  0.9946
## F-statistic:  2931 on 1 and 15 DF,  p-value: < 2.2e-16
```

Vrednost testne statistike za preverjanje linearnosti modela je enaka $t = 54.14$, s $df = 15$ 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 $R^2 = 0.995$, kar pomeni, da 99.5% variabilnosti zračnega tlaka 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) 1.330  1.451
## temp        0.016  0.017
```

Interval zaupanja za odsek je enak $I_a = [1.330, 1.451]$ in interval zaupanja za naklon $I_b = [0.016, 0.017]$.

9. Interval predikcije za vrednost Y pri izbrani vrednosti X

Pri predvidevanju vrednosti zračnega tlaka nas zanima bodoča vrednost spremenljivke Y pri izbrani vrednosti spremenljivke $X = x_0$. Želimo oceniti spodnjo in zgornjo mejo, med katerima se verjetno nahaja vrednost zračnega tlaka teh temperatura vretja vode.

```
xtemp = data.frame(temp=c(90,95,100))
10 ^ predict(model, xtemp, interval="predict")
```

```
##         fit      lwr       upr
## 1  700.4906 686.2202  715.0578
## 2  843.8245 827.7367  860.2251
## 3 1016.4873 995.7489 1037.6577
```

Predvidena vrednost zračnega tlaka (na celi populaciji lokacij)

1. 90°C je 700.49 mbar, s 95% intervalom predikcije zračnega tlaka [686.22, 715.06],
2. 95°C je 843.82 mbar, s 95% intervalom predikcije zračnega tlaka [827.74, 860.23],
3. 100°C je 1016.49 mbar, s 95% intervalom predikcije zračnega tlaka [995.75, 1037.66]

10. Zaključek

Zanimala nas je funkcionalna odvisnost med temperaturo vretja vode in logaritma z osnovo 10 zračnega tlaka na isti lokaciji. Dobili smo vzorec temperature vretja vode in zračnega tlaka na 17 lokacijah v Alpah. Ugotovili smo, da je enostavni linearni model odvisnosti logaritma z osnovo 10 zračnega tlaka od temperature vretja vode dober. Diagnostični grafi in statistični testi niso pokazali na težave z linearnim regresijskim modelom. Koeficient determinacije je 99.5%, kar pomeni, da tolikšen delež variabilnosti zračnega tlaka zajamemo z linearnim modelom. Napoved vrednosti zračnega tlaka je dobra.