Seminarska naloga

Gašper Oblak

Linearna regresija: poraba goriva

1. Opis podatkov

Zbrali smo meritve premera in visine na vzorcu 50 orjaskih klekov (lat. Thuja plicata). Podatke smo zapisali v dokument, ki ima dva stolpca:

1. *premer* je numericna zvezna spremenljivka, ki predstavlja premer debla, merjen na visini 1.37 m nad tlemi (v metrih).
2. *visina* je numericna zvezna spremenljivka, ki predstavlja visino drevesa (v metrih).

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

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

```
## 'data.frame':    50 obs. of  2 variables:
##  $ premer: num  3.75 1.51 2.3 3.2 5.25 7.5 8.95 4.25 8.3 3.9 ...
##  $ visina: num  29.8 15.5 20 22.5 29 32 35 22.5 35 24 ...
```

```
log2Premer <- log2(klek$premer)
```

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 premera in visine.

```
summary(klek$premer)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.110   2.803   3.815   4.183   4.945  10.150
```

```
sd(klek$premer)
```

```
## [1] 2.143167
```

Opazimo, da premer vzorca klekov varira od 1.110 do 10.150m, s povprečjem 4.183 in standardnim odklonom 2.143167 m. Ponovimo postopek računanja za visino.

```
summary(klek$visina)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.50   21.25   25.00   24.64   28.88   39.00
```

```
sd(klek$visina)
```

```
## [1] 6.622342
```

Opazimo, da visina klekov varira od 9.5 do 39.00 m s povprečjem 24.64 in standardnim odklonom 6.622342 m. Razpon vrednosti premera in visine nam pomaga pri izbiri mej na oseh razsevnega diagrama.

Ponovimo postopek racunanja za transformirano spremenljivko $\log_2(premer)$

```
summary(log2Premer)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1506  1.4849  1.9316  1.8786  2.3059  3.3434
```

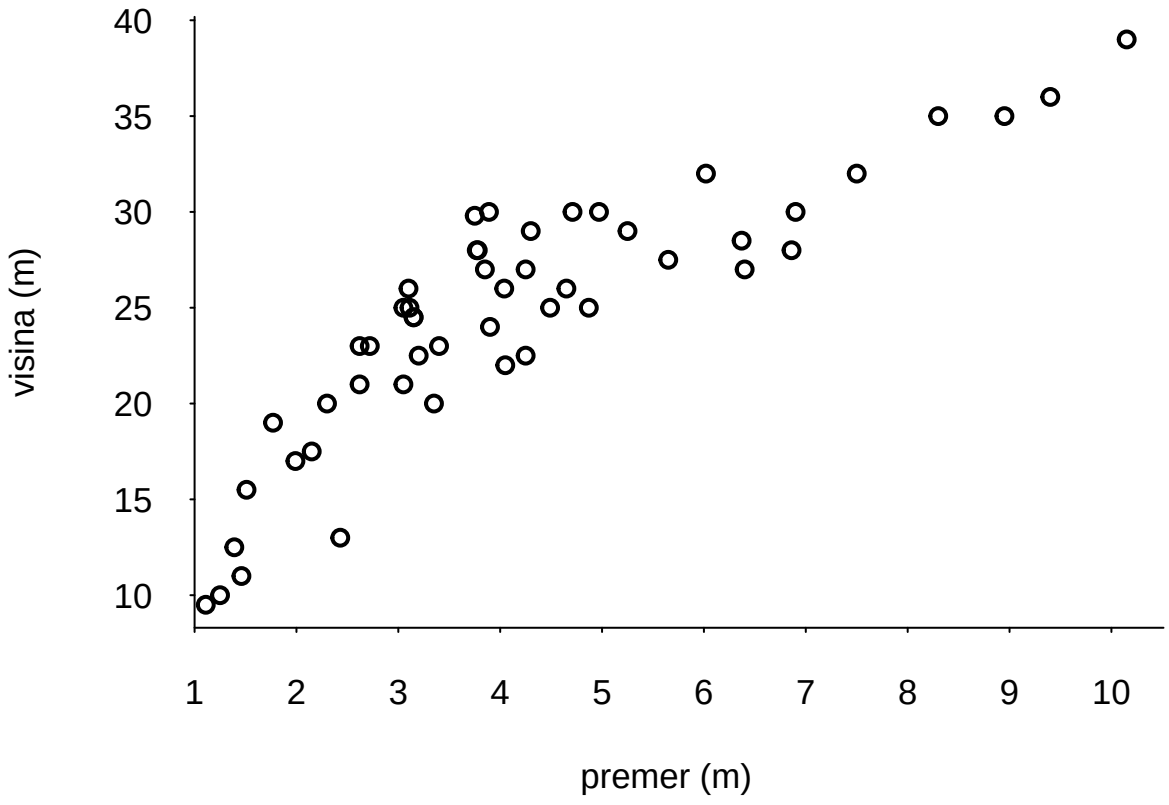
```
sd(log2Premer)
```

```
## [1] 0.758802
```

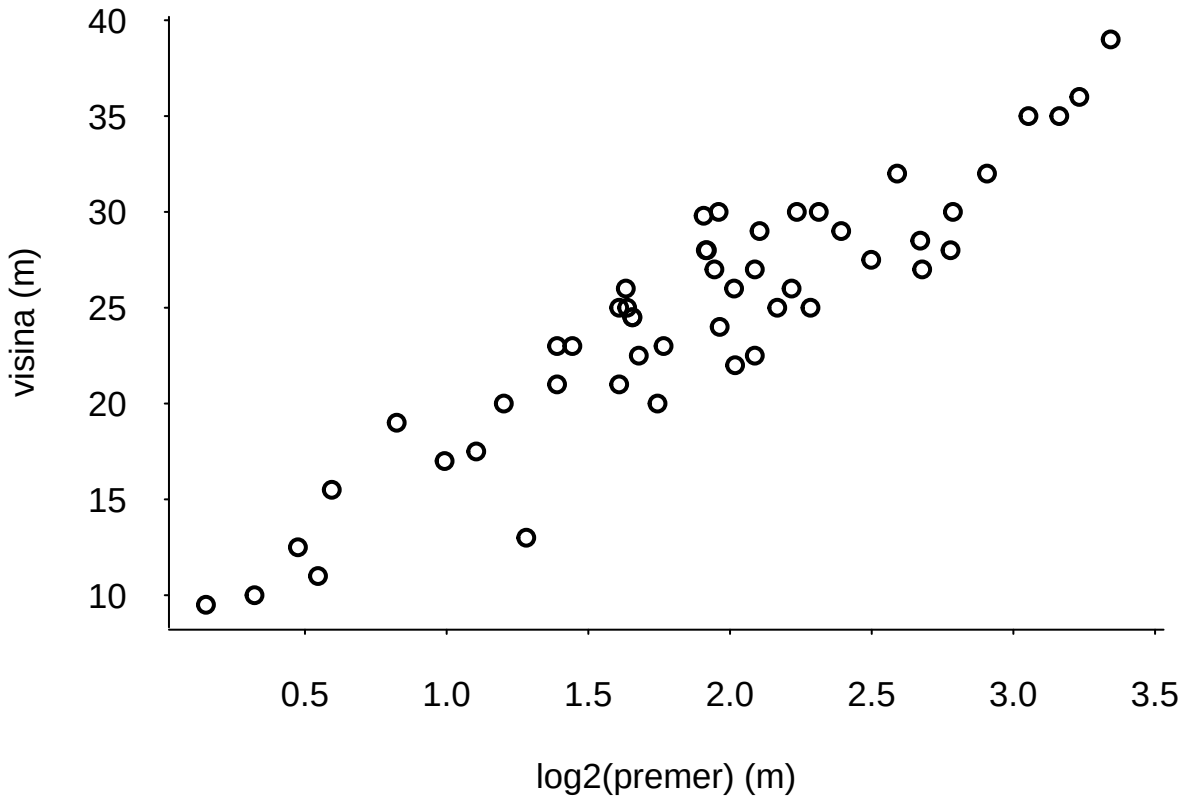
3. Razsevni diagram in vzorčni koeficient korealcije

Prikažimo dobljene podatke na razsevnem diagramu.

```
par(las=1, cex=1.1, mar=c(4,4,2,2))
plot(klek$premer, klek$visina, main="", xlim=c(1.110,10.150), ylim=c(9.5,39),
xlab="premer (m)", ylab="visina (m)", lwd=2, axes=FALSE)
axis(1,pos=8.3,at=seq(1,13,by=1),tcl=-0.1)
axis(2,pos=1,at=seq(0,45,by=5),tcl=-0.1)
```




```
par(las=1, cex=1.1, mar=c(4,4,2,2))
plot(log2Premer, klek$visina, main="", xlim=c(0.15, 3.4), ylim=c(9.5,39),
xlab="log2(premer) (m)", ylab="visina (m)", lwd=2, axes=FALSE)
axis(1,pos=8.2,at=seq(0,4,by=0.5),tcl=-0.1)
axis(2,pos=0.02,at=seq(0,45,by=5),tcl=-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(klek$premer,klek$visina))
```

```
## [1] 0.8650352
```

Vrednost vzorčnega koeficienta korelacije je visoka ($r = 0.865$), kar govori o visoki linearni povezanosti premera in visine kleka. Dalje, koeficient korelacije je pozitiven, kar pomeni, da imajo kleki z vecjim premerom visjo visino.

```
(r<-cor(log2Premer,klek$visina))
```

```
## [1] 0.9242486
```

Vrednost vzorčnega koeficienta korelacije je visoka ($r = 0.9242486$), kar govori o visoki linearni povezanosti $log_2(premera)$ in visine kleka. Dalje, koeficient korelacije je pozitiven, kar pomeni, da imajo kleki z vecjim $log_2(premerom)$ visjo visino.

4. Formiranje linearnega regresijskega modela

Formirajmo linearni regresijski model.

```
(model<-lm(visina~log2Premer,data=klek))
```

```
##
## Call:
## lm(formula = visina ~ log2Premer, data = klek)
##
## Coefficients:
## (Intercept)   log2Premer
##       9.483        8.066
```

Dobili smo ocenjeno regresijsko premico $\hat{y} = 9.483 + 8.066x$, oziroma oceni odseka in naklona sta enaki $\hat{a} = 9.483$ in $\hat{b} = 8.066$.

5. Točke visokega vzvoda in osamelci

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

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

```
##    premer visina
## 15   1.46   11.0
## 17   1.39   12.5
## 25   1.11    9.5
## 43   9.40   36.0
## 44  10.15   39.0
## 50   1.25   10.0
```

Odkrili smo 6 točk visokega vzvoda. Stiri kleki imajo majhen premer pod 2m in dva kleka najvisji premer nad 9.4.

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

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

```
##    premer visina
## 20   2.43     13
```

Ena podatkovna tocka je osamelec in sicer se nanasa na klek, ki ima dokaj velik premer ter majhno visino.

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")
```

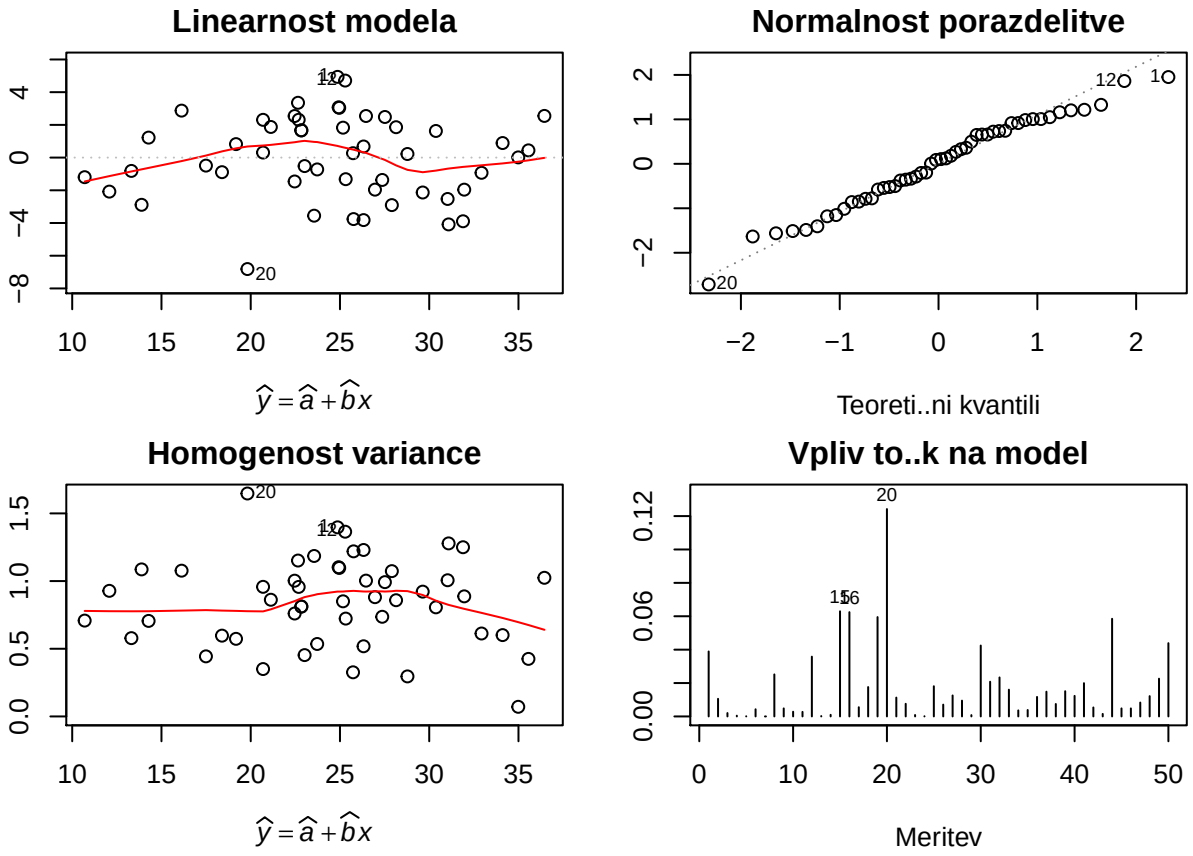
```
## Warning in title(xlab = "Teoretični kvantili", ylab = "St. ostanki", main =
## "Normalnost porazdelitve"): conversion failure on 'Teoretični kvantili' in
## 'mbcsToSbcs': dot substituted for <c4>
```

```
## Warning in title(xlab = "Teoretični kvantili", ylab = "St. ostanki", main =
## "Normalnost porazdelitve"): conversion failure on 'Teoretični kvantili' in
## 'mbcsToSbcs': dot substituted for <8d>
```

```
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")
```

```
## Warning in title(xlab = "Meritev", ylab = "Cookova razdalja", main = "Vpliv točk
## na model"): conversion failure on 'Vpliv točk na model' in 'mbcsToSbcs': dot
## substituted for <c4>

## Warning in title(xlab = "Meritev", ylab = "Cookova razdalja", main = "Vpliv točk
## na model"): conversion failure on 'Vpliv točk na model' in 'mbcsToSbcs': dot
## substituted for <8d>
```



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 20 do 30, kar je prisotno zaradi originalnih vrednosti v vzorcu klekov (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 $log2(premeru)$ in visini klekov lahko zaključimo, da so naključne napake normalno porazdeljene

(ni večjih odstopanj od premice, razen za 20., 12., in 1. 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 = 0.005928373, Df = 1, p = 0.93863
```

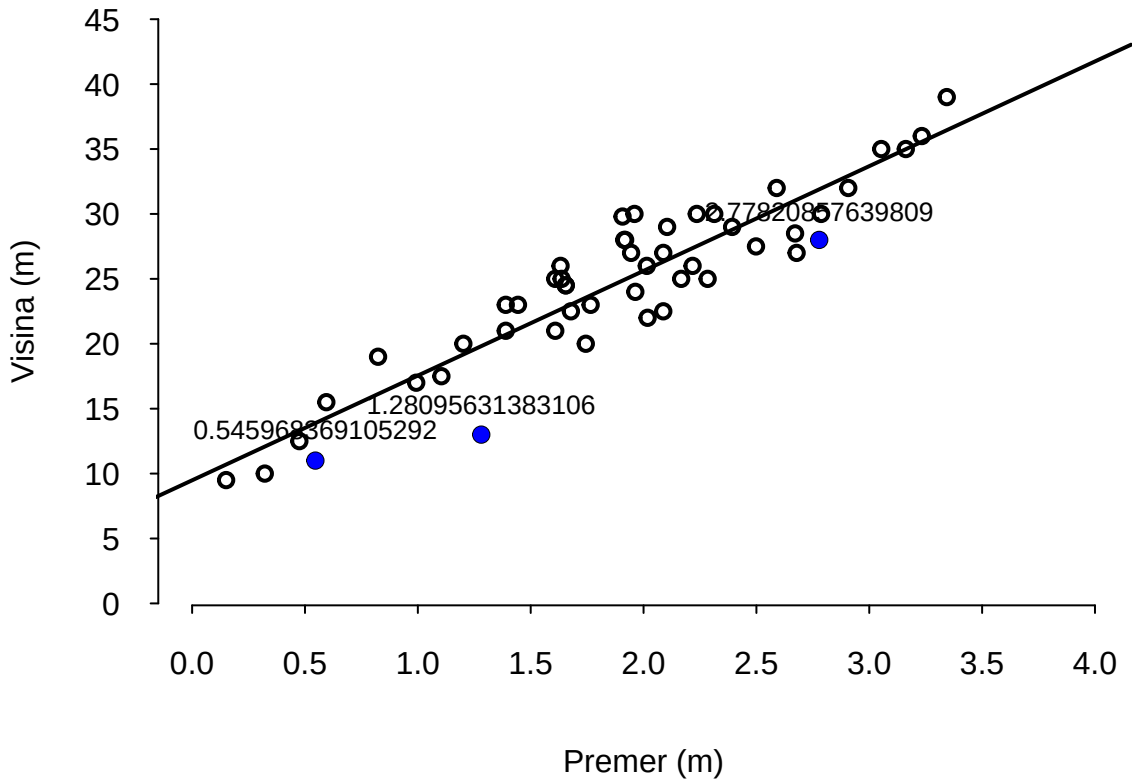
Na osnovi rezultata Breusch-Paganovega testa (testna statistika $\chi^2 = 0.005928373$, $df = 1$, p-vrednost $p = 0.93864 > 0.05$), ne zavrnemo ničelne domneve. Ni dovolj dokazov, da varianca naključnih napak ni homogena.

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 stiri točke z najvišjo Cookovo razdaljo. Za naše podatke, to so 15., 16., 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(log2Premer, klek$visina, main="", xlim=c(0,4), ylim=c(0,45),xlab=
"Premer (m)", ylab="Visina (m)",lwd=2, axes=FALSE)
axis(1,pos=-0.15,at=seq(0,4,by=0.5),tcl=-0.2)
axis(2,pos=-0.15,at=seq(0,45,by=5),tcl=-0.2)
arrows(x0=2500,y0=0,x1=2600,y1=0,length=0.1)
arrows(x0=500,y0=15,x1=500,y1=16,length=0.1)
abline(model,lwd=2)
points(log2Premer[c(15,16,20)],klek$visina[c(15,16,20)],col="blue",pch=19)
text(log2Premer[c(15,16,20)],klek$visina[c(15,16,20)]+c(0.2,0,0.1),labels=
log2Premer[c(15,16,20)],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(15,16,20)]>=qf(0.5,2,nrow(klek)-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 = visina ~ log2Premer, data = klek)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8155 -1.8352  0.2443  1.8688  4.9355
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4830     0.9730   9.746 5.84e-13 ***
## log2Premer    8.0663     0.4809  16.772  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.555 on 48 degrees of freedom
## Multiple R-squared:  0.8542, Adjusted R-squared:  0.8512
## F-statistic: 281.3 on 1 and 48 DF,  p-value: < 2.2e-16
```

Vrednost testne statistike za preverjanje linearnosti modela je enaka $t = 16.772$, s $df = 48$ 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 $R^2 = 0.85$, kar pomeni, da 85% variabilnosti visine 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) 7.527 11.439
## log2Premer  7.099  9.033
```

Interval zaupanja za odsek je enak $I_a = [7.527, 11.439]$ in interval zaupanja za naklon $I_b = [7.099, 9.033]$.

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 $\hat{y} = 9.483 + 8.066x$ klekov z določenim premerom x_0, ampak želimo tudi oceniti spodnjo in zgornjo mejo, med katerima se verjetno nahaja visina različnih modelov avtomobilov teh mas.

```
xlog2Premer = data.frame(log2Premer=c(3,6,9))
predict(model, xlog2Premer, interval="predict")
```

```
##        fit      lwr      upr
## 1 33.68177 28.38225 38.98128
## 2 57.88053 51.33896 64.42210
## 3 82.07930 73.45777 90.70082
```

Predvidena visina za klek premera (na celi populaciji avtomobilov)

1. 3 m je 33 m, s 95% intervalom predikcije premera [2.1, 9.4],
2. 6 m je 58 m, s 95% intervalom predikcije premera [10.1], 23.0],
3. 9 m kg 82 m, s 95% intervalom predikcije premera [24.0, 42.8]

10. Zaključek

Zanimala nas je funkcionalna odvisnost med premerom in visino klekov v metrih. Zbrali smo vzorec 50 klekov, jim izmerili premer in zabeležili visino. Ugotovili smo, da je enostavni linearni model odvisnosti premera od visine dober. Diagnostični grafi in statistični testi niso pokazali na težave z linearnim regresijskim modelom. Koeficient determinacije je 85%, kar pomeni, da toliksen delez variabilnsoti dolzine visine zajamemo z linearnim modelom. Napoved visine na osnovi njegovega premera je zadovoljiva, vendar bi vključevanje dodatnih neodvisnih spremenljivk zagotovo dala še boljši model in bolj zanesljivo napoved.