# Statistična obdelava podatkov z linearno regresijo

February 7, 2021

# Contents

Konstrukcija linearnega regresijskega modela v programskem jeziku R med spremenljivkama $\sqrt{pot}$ in *hitrost*, kjer je $\sqrt{pot}$ odvisna spremenljivka.

# 1 Opis podatkov

Zbrali smo vzorec dolžin zavornih poti in hitrosti premikanja za 62 avtomobilov. Baza podatkov z 62 meritvami dveh spremenljivk

- *hitrost* je numerična zvezna spremenljivka, ki predstavlja hitrost avtomobila (v kilometrih na uro),
- *pot* je numerična spremenljivka, ki predstavlja zavorno pot (v metrih).

Bazo podatkov z imenom *zavor.csv* preberemo v R s pomočjo funkcije *read.csv*, in zatem pogledamo strukturo podatkov s pomočjo funkcije *str*.

```
zavor<-read.csv("Documents/FRI/2/VS/seminarska/zavor.csv", header = TRUE)
str(zavor)
'data.frame':   62 obs. of  2 variables:
 $ hitrost: int  6 8 8 8 8 11 11 13 13 13 ...
 $ pot    : num  1.22 0.61 1.22 2.44 2.44 2.13 2.13 2.44 2.74 3.35 ...
```

# 2 Opisna statistika

Pridobimo povzetek naših podatkov s petimi števili (minimum, maksimum, prvi in tretji kvartil in mediano), vzorčni povprečji in vzorčna standardna odklona hitrosti in poti. Povzetek s petimi števili pridobimo s funkcijo *summary*.

```
summary(zavor$hitrost)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   16.00   28.00   30.39   42.75   64.00
```

Izračunamo vzorčno povprečje in vzorčni standardni odklon za spremenljivko *hitrost*.

```
(mx<-mean(zavor$hitrost))
[1] 30.3871
(mx<-sd(zavor$hitrost))
[1] 16.01368
```

Opazimo, da hitrost avtomobilov variira od 6 do 64 km/h, s povprečjem 30.39 in standardnim odklonom 16.01.

Postopek računanja ponovimo še za spremenljivko *pot*.

```
summary(zavor$pot)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.610   4.037   8.990  11.980  17.295  42.060
(mx<-mean(zavor$pot))
[1] 11.98032
(mx<-sd(zavor$pot))
[1] 10.17246
```

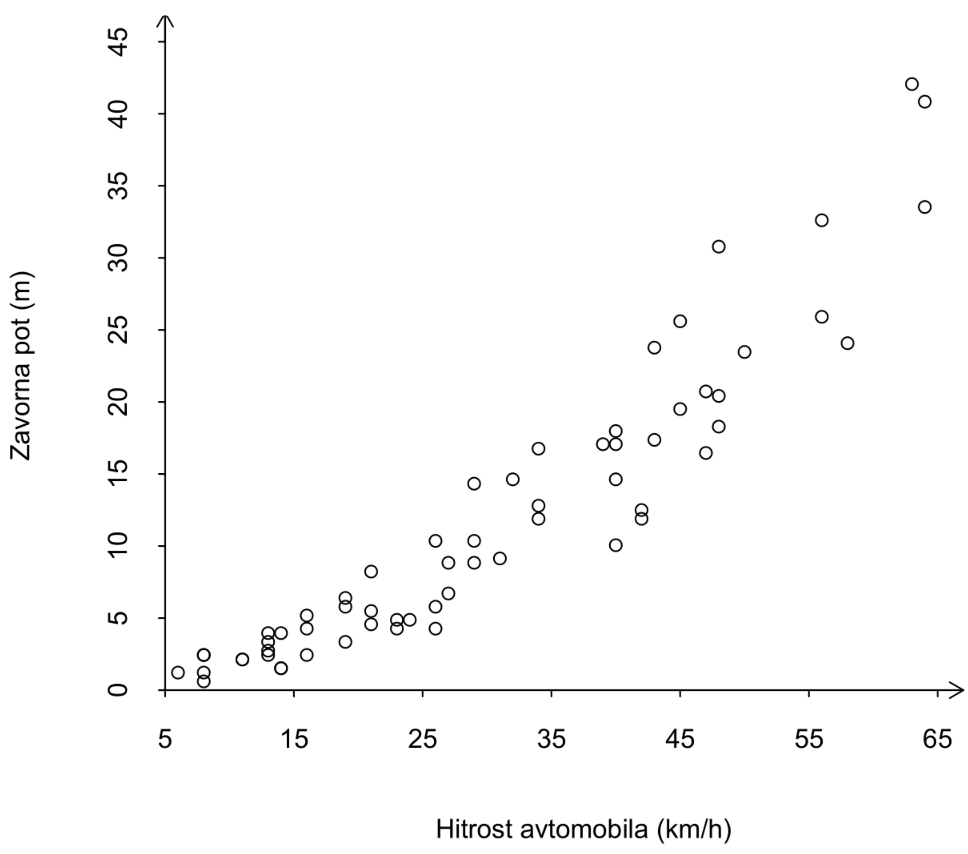
Opazimo, da pot zaviranja avtomobilov variira od 0.61 do 42.06 m, s povprečjem 11.98 in standardnim odklonom 10.17.

## 2.1 Razlogi za transformacijo podatkov na podlagi razsevnega diagrama, koeficienta korelacije in diagnostičnih grafov originalnik (netransformiranih podatkov)

Zveza med originalnimi podatki obstaja, vendar ni linearna. Hitrost smo pustili nespremenjeno, *pot* pa smo korenili, da smo dobili linearno zvezo med spremenljivkama *hitrost* in $\sqrt{pot}$. Linearno zvezo med spremenljivkama potrebujemo, da lahko formiramo linearni regresijski model.

Narišemo razsevni diagram za originalne podatke, da se prepričamo, da je bila transformacija podatkov zares potrebna.

```
plot(zavor$hitrost, zavor$pot, xlab="Hitrost avtomobila (km/h)",
ylab="Zavorna pot (m)",xlim=c(5,70),ylim=c(0,45),axes=FALSE)
axis(1,pos=0,at=seq(5,70,by=10),tcl=-0.2)
axis(2,pos=5,at=seq(0,50,by=5),tcl=-0.2)
arrows(x0=65,y0=0,x1=67,y1=0,length=0.1)
arrows(x0=5,y0=45,x1=5,y1=47,length=0.1)
```




Slika 1: Razsevni diagram za originalne - netransformirane podatke

Ob opazovanju *Slike* 1 lahko vidimo, da točke na grafu niso porazdeljne linearno. Gre za parabolično obliko, z rahlo višjo koncentracijo točk med $x$ vrednostmi od 5 do 25.

Preverimo še korelacijo med originalnima netransformiranima spremenljivkama.

```
(orgR<-cor(zavor$hitrost,zavor$pot))
[1] 0.9356374
```

Opazimo močno in pozitivno korelacijo ($orgR = 0.9356374$) med netransformiranima spremenljivkama, vendar pa je ta nekoliko manjša kot med transformiranima spremenljivkama, ki znaša $r = 0.9615461$.

Preden si za originalne spremenljivke ogledamo diagnostične grafe, zanje konstruiramo linearni model $orgModel$ s pomočjo funkcije $lm$.

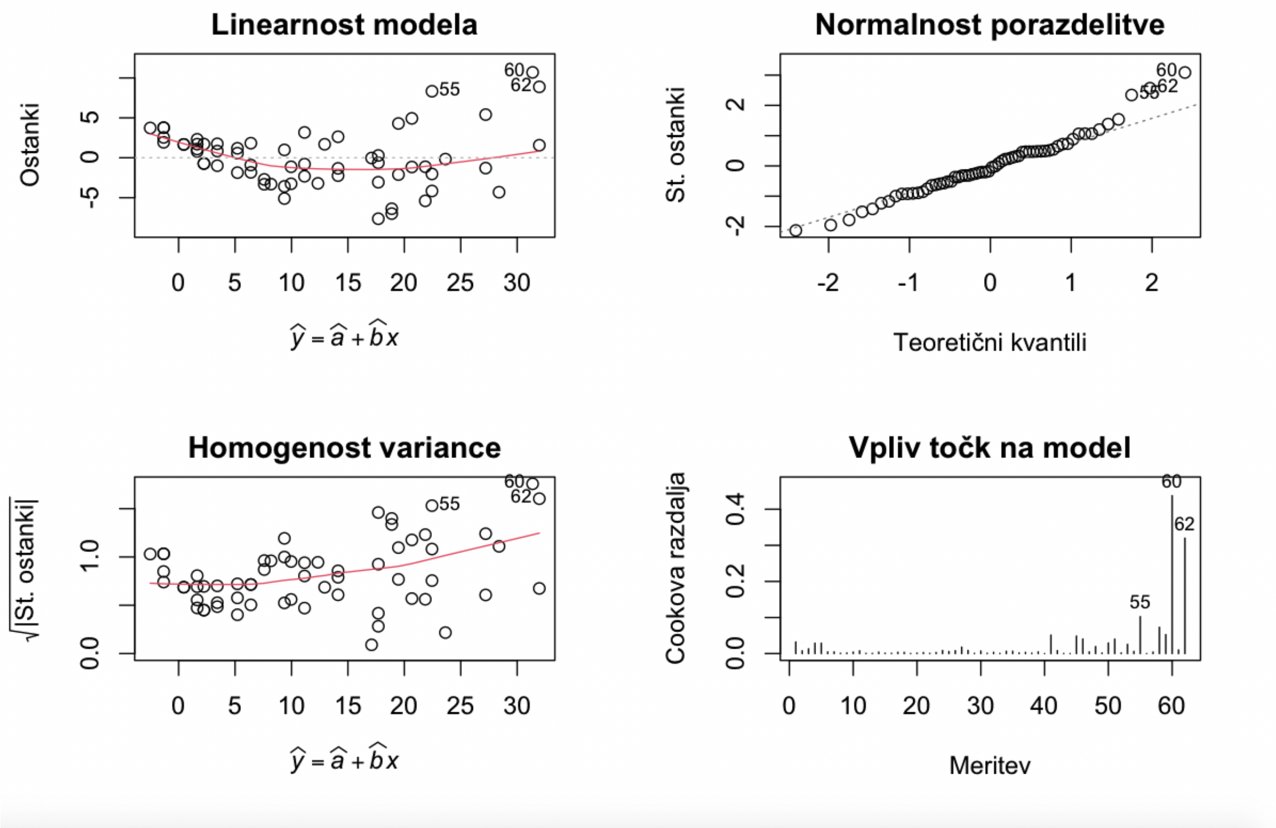
```
 (orgModel<-lm(formula = pot ~ hitrost, data = zavor))

Call:
lm(formula = pot ~ hitrost, data = zavor)

Coefficients:
(Intercept)      hitrost
    -6.0803       0.5944
```

Izrišemo diagnostične grafe za originalne spremenljivke.

```
par(mfrow=c(2,2),cex=1.1,mar=c(6,4.5,2,3))
plot(orgModel,which=1,caption="",ann=F)
title(xlab=expression(italic(widehat(y)==widehat(a)+widehat(b)*x))
ylab="Ostanki", main="Linearnost modela")
plot(orgModel,which=2,caption="",ann=F)
title(xlab="Teoreti ni kvantili", ylab="St. ostanki",
main="Normalnost porazdelitve")
plot(orgModel,which=3,caption="",ann=F)
title(xlab=expression(italic(widehat(y)==widehat(a)+widehat(b)*x))
ylab=expression(sqrt(paste("|St. ostanki|"))), main="Homogenost variance")
plot(orgModel,which=4,caption="",ann=F)
title(xlab="Meritev", ylab="Cookova razdalja", main="Vpliv to k na model")
```




Slika 2: Diagnostični diagram za originalne podatke

Na *Sliki* 2 prikazujemo linearnost modela, normalnost porazdelitve, homogenost variance in vpliv točk na model, konstruiran na podlagi orginalnih podatkov.

### 2.1.1 Graf za preverjanje linearnosti modela

Validnost linearnega regresijskega modela preverimo tako, da narišemo graf ostankov v odvisnosti od $x$ vrednosti ali od predvidenih vrednosti $\hat{y} = \hat{a}x + \hat{b}$. Opazimo, da se na grafu pojavi vzorec - točke niso enakomerno raztresene nad in pod premico $Ostanki = 0$, torej lahko zaključimo, da je linearni model ni validen.

### 2.1.2 Graf normalnosti porazdelitve naključnih napak

Normalnost porazdelitve naključnih napak preverjamo s pomočjo grafa porazdelitve standardnih ostankov. Ostanek standardiziramo tako, da ga delimo z oceno njegovega standardnega odklona. Na $x$-osi $Q-Q$ grafa normalne porazdelitve so podani teoretični kvantili, na $y$-osi pa kvantili standardnih ostankov. Ker dobljene točke na $Q-Q$ tvorijo premico le z manjšimi odstopanji, zaključimo, da je porazdelitev naključnih napak normalna, večja odstopanja od premice imajo le točke 55, 60 in 62.

### 2.1.3 Graf homogenosti variance: Breusch-Paganov test

Osnovna predpostavka linearnega regresijskega modela je, da imajo naključne napake konstantno varianco, gre za t. i. homogenost variance. Za prepoznavo nekonstantne variance je najbolj učinkovit graf korena standardiziranih ostankov v odvisnosti od $x$ ali od predvidenih vrednosti $\hat{y} = \hat{a}x + \hat{b}$.

Na osnovi grafa homogenosti variance opazimo, da varianca narašča, torej varianca ostankov ni homogena.

```
ncvTest(orgModel)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 20.27289, Df = 1, p = 6.7145e-06
```

Na osnovi rezultata Breusch-Paganovega testa (testna statistika $\chi^2 = 20.27289$, $df = 1$, $p$-vrednost $p = 6.7145\text{e-}06 < 0.05$) lahko potrdimo, da varianca naključnih napak ni konstantna.

#### 2.1.4 Cookova razdalja: graf in analiza vpliva točk preko osnovnega pogoja, razsevnega diagrama in pogoja velikega vpliva

Vpliv $i$-te točke na linearni regresijski model merimo s pomočjo Cookove razdalje

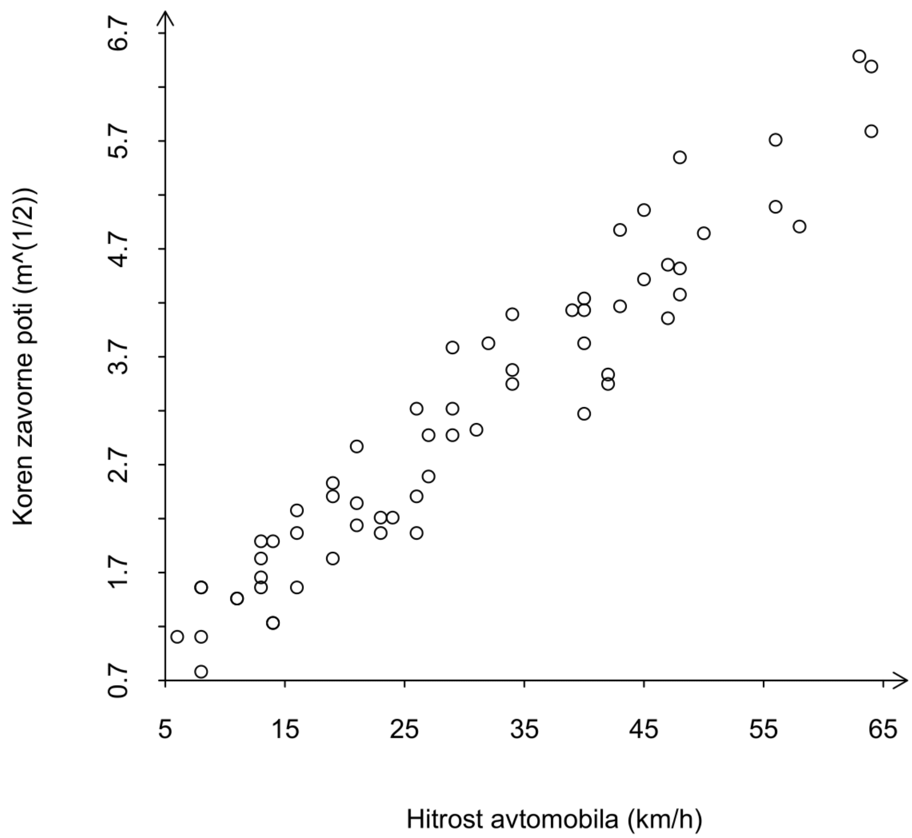
$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_{j(i)} - \hat{Y}_j)^2}{S^2}, \tag{1}$$

kjer je $\hat{Y} = \hat{a} + \hat{b}X_j$ $j$-ta predvidena vrednost, $S$ ocena standardnega odklona napak, $\hat{Y}_{j(i)}$ $j$-ta predvidena vrednost linearnega modela, ki je narejen brez $i$-te točke na osnovi preostalih $n-1$ točk. Na ta način torej merimo razliko med modelom, ki vsebuje $i$-to točko in modelom, ki je ne vsebuje.

```
which(cooks.distance(orgModel)>4/60)
55 58 60 62
55 58 60 62
```

## 3 Razsevni diagram in vzorčni koeficient korelacije

Transformirane podatke prikažemo z razsevnim diagramom na *Sliki* 3.




Slika 3: Razsevni diagram za transformirane podatke

```
plot(zavor$hitrost, sqrt(zavor$pot), xlab="Hitrost avtomobila (km/h)",
ylab="Koren zavorne poti (m^(1/2))",xlim=c(5,70),ylim=c(0.7,6.7),axes=FALSE)
axis(1,pos=0.7,at=seq(5,70,by=10),tcl=-0.2)
axis(2,pos=5,at=seq(0.7,6.7,by=0.5),tcl=-0.2)
arrows(x0=5,y0=6.7,x1=5,y1=6.9,length=0.1)
arrows(x0=65,y0=0.7,x1=67,y1=0.7,length=0.1)
```

Moč korelacije preverimo z računanjem Pearsonovega koeficienta korelacije.

```
(r<-cor(zavor$hitrost,sqrt(zavor$pot)))
[1] 0.9615461
```

Koeficient korelacije $r > 0.7$ pomeni, da je korelacija močna. Koeficient je pozitiven, kar pomeni, da z naraščanjem korena zavorne poti naraščajo tudi vrednosti hitrosti avtomobila in s padanjem korena zavorne poti padajo tudi vrednosti hitrosti avtomobila.

# 4 Formiranje linearnega regresijskega modela, prikaz računanja ocen naklona in odseka, ter enačba vzorčne regresijske premice

Formiramo linearni regresijski model s pomočjo funkcije *lm*.

```
(model<-lm(sqrt(pot)~hitrost ,data=zavor))

Call:
lm(formula = sqrt(pot) ~ hitrost , data = zavor)

Coefficients:
(Intercept)       hitrost
    0.52000       0.08661
```

Dobili smo ocenjeno regresijsko premico $\hat{y} = 0.52000 + 0.08661x$, oziroma oceni odseka in naklona sta enaki $\hat{a} = 0.52000$ in $\hat{b} = 0.08661$. Rezultate lahko preverimo z ročnim izračunom.

```
sy<-sd(sqrt(zavor$pot))
sx<-sd(zavor$hitrost)
(b<-r*sy/sx)
[1] 0.0866071

my<-mean(sqrt(zavor$pot))
mx<-mean(zavor$hitrost)
(a<-my-b*mx)
[1] 0.5200021
```

## 4.1 Točke visokega vzvoda in osamelci

Identificirajmo točke visokega vzvoda in osamelce. Vrednost $x$ je točka visokega vzvoda, če je njen vzvod večji od $\frac{4}{n}$, kjer $n = 62$. Vzvod točke izračunamo s pomočjo funkcije *hatvlaues*.

```
zavor[hatvalues(model)>4/62,]
    hitrost   pot
59       58 24.08
60       63 42.06
61       64 33.53
62       64 40.84
```

Odkrili smo 4 točke visokega vzvoda - 59, 60, 61 in 62. Pri vseh opazimo, da gre za točke z visoko hitrostjo. Za podatke majhne in srednje velikosti vzorca je osamelec podatkovna točka, kateri ustreza standardizirani ostanek izven intervala [-2,2]. Povprečje ostankov je enako 0, potem ostanke standardiziramo, ko jih delimo z njihovim standardnim odklonom (funkcija *rstandard*).Točke visokega vzvoda so točke, ki imajo višjo hitrost v primerjavi z ostalimi točkami.

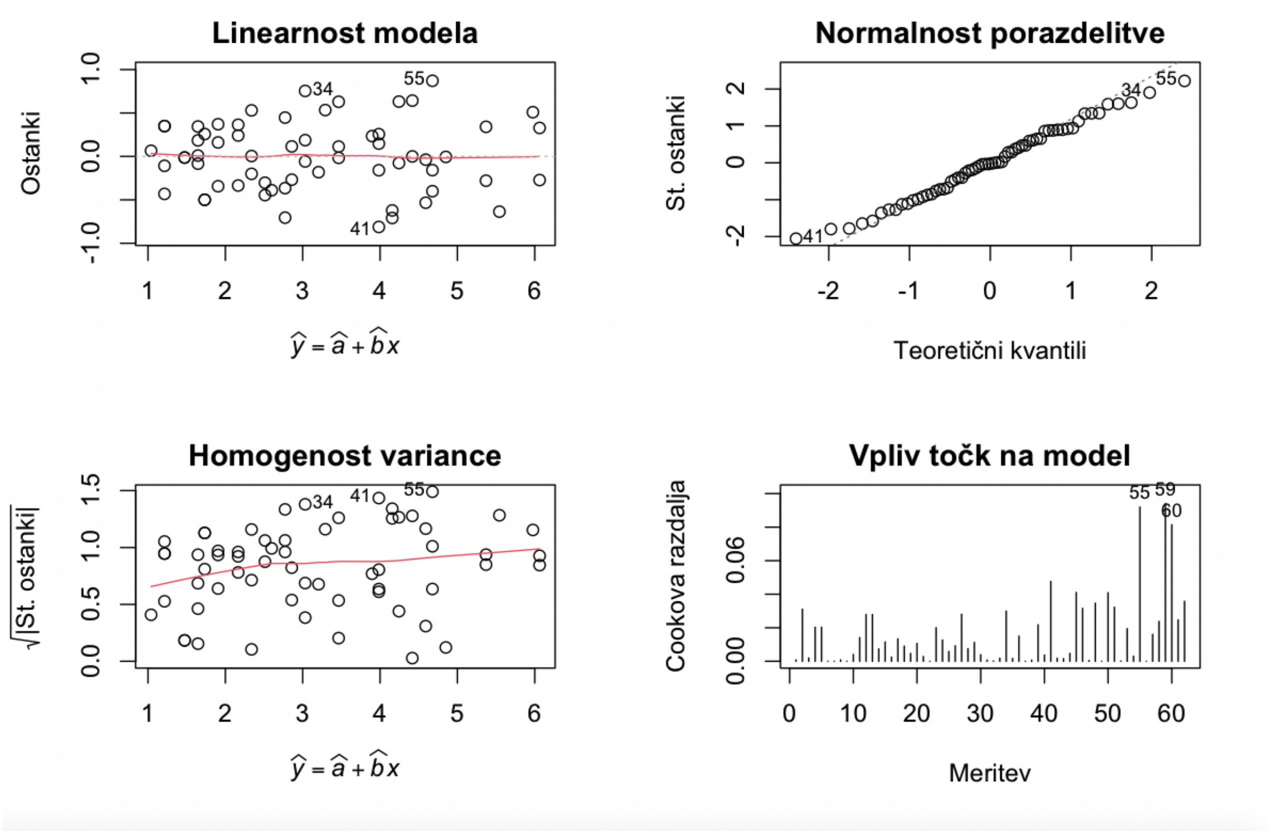
```
zavor[abs(rstandard(model))>2,]
    hitrost   pot
41       40 10.06
55       48 30.78
```

Dve podatkovni točki sta osamelca, oziroma imajo nenavadno nizko (en avto) ali nenavadno visoko (en avto) zavorno pot v primerjavi z ostalimi podatki.

# 5 Preverjanje predpostavk linearnega modela (diagnostični grafi in njihova obrazložitev)

Predpostavke linearnega regresijskega modela preverimo s pomočjo 4 diagnostičnih grafov. Če neke predpostavke niso izpolnjene, so lahko ocene neznanih parametrov, $p$-vrednost testa, intervali zaupanja in intervali predikcije netočni.

```
par(mfrow=c(2,2),cex=1.1,mar=c(6,4.5,2,3))
plot(model,which=1,caption="",ann=F)
title(xlab=expression(italic(widehat(y)==widehat(a)+widehat(b)*x)),
ylab="Ostanki", main="Linearnost modela")
plot(model,which=2,caption="",ann=F)
title(xlab="Teoreti ni kvantili", ylab="St. ostanki", main="Normalnost porazdelitve")
plot(model,which=3,caption="",ann=F)
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=F)
title(xlab="Meritev", ylab="Cookova razdalja", main="Vpliv to k na model")
```




Slika 4: Diagnostični grafi za transformirane podatke

Na *Sliki* 4 prikazujemo linearnost modela, normalnost porazdelitve, homogenost variance in vpliv točk na model.

## 5.1 Graf za preverjanje linearnosti modela

Validnost linearnega regresijskega modela preverimo tako, da narišemo graf ostankov v odvisnosti od $x$ vrednosti ali od predvidenih vrednosti $\hat{y} = \hat{a}x + \hat{b}$. Opazimo, da se na grafu ne pojavi nikakršen vzorec - točke so enakomerno raztresene nad in pod premico $Ostanki = 0$, torej lahko zaključimo, da je linearni model validen.

## 5.2 Graf normalnosti porazdelitve naključnih napak

Normalnost porazdelitve naključnih napak preverjamo s pomočjo grafa porazdelitve standardnih ostankov. Ostanek standardiziramo tako, da ga delimo z oceno njegovega standardnega odklona. Na $x$-osi $Q-Q$ grafa normalne porazdelitve so podani teoretični kvantili, na $y$-osi pa kvantili standardnih ostankov. Ker dobljene točke na $Q-Q$ tvorijo premico le z manjšimi odstopanji, zaključimo, da je porazdelitev naključnih napak normalna, večja odstopanja od premice imajo le točke 34, 41, 55.

## 5.3 Graf homogenosti variance: Breusch-Paganov test

Osnovna predpostavka linearnega regresijskega modela je, da imajo naključne napake konstantno varianco, gre za t. i. homogenost variance. Za prepoznavo nekonstantne variance je najbolj učinkovit graf korena standardiziranih ostankov v odvisnosti od $x$ ali od predvidenih vrednosti $\hat{y} = \hat{a}x + \hat{b}$.

Na podlagi grafa opazimo, da pri višjih vrednostih odvisne spremenljivke varianca rahlo narašča, kar pomeni, da v tem primeru regresijska premica ni najbolj primerna.

```
ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.474247, Df = 1, p = 0.11572
```

Na osnovi rezultata Breusch-Paganovega testa (testna statistika $\chi^2 = 2.474247$, $df = 1$, $p$-vrednost $p = 0.11572 > 0.05$) lahko sprejmemo ničelno domnevo, da je varianca naključnih napak konstantna.

## 5.4 Cookova razdalja: graf in analiza vpliva točk preko osnovnega pogoja, razsevnega diagrama in pogoja velikega vpliva

Vpliv $i$-te točke na linearni regresijski model merimo s pomočjo Cookove razdalje

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_{j(i)} - \hat{Y}_j)^2}{S^2}, \tag{2}$$

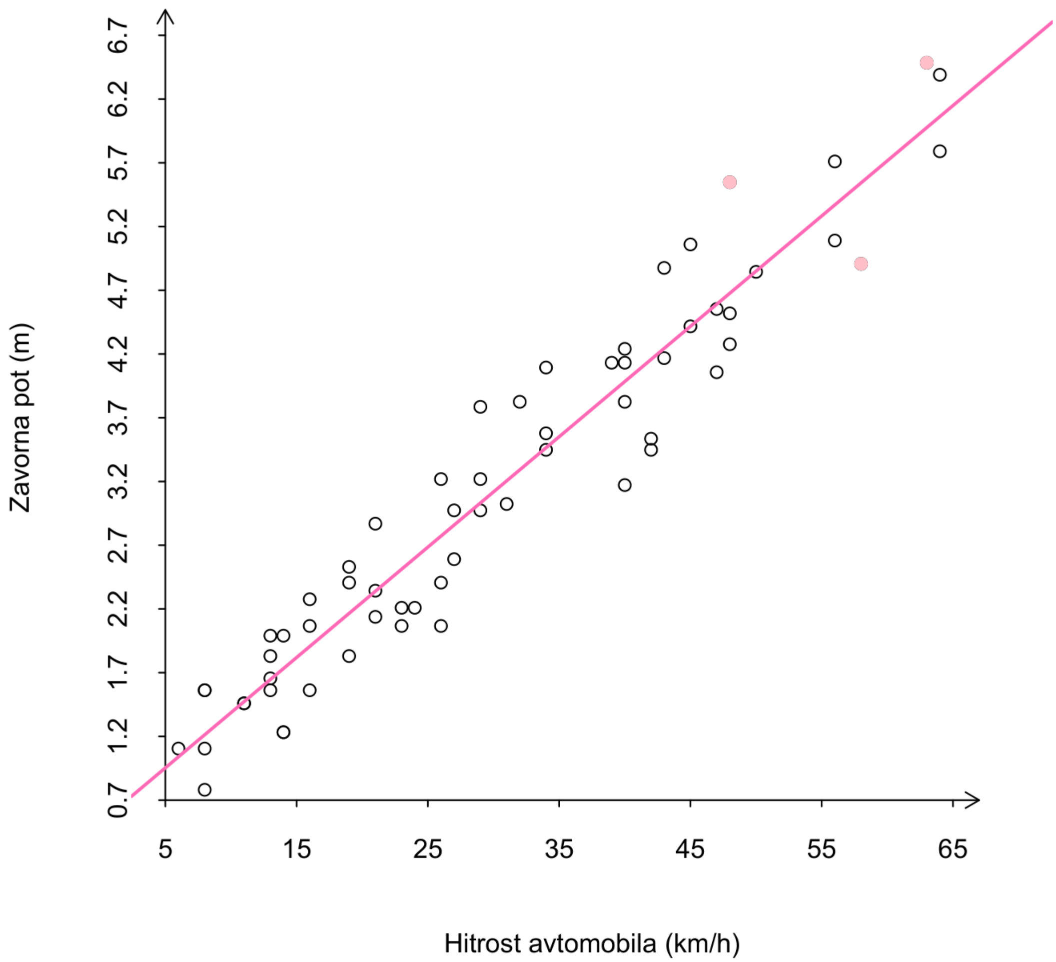
kjer je $\hat{Y} = \hat{a} + \hat{b}X_j$ $j$-ta predvidena vrednost, $S$ ocena standardnega odklona napak, $\hat{Y}_{j(i)}$ $j$-ta predvidena vrednost linearnega modela, ki je narejen brez $i$-te točke na osnovi preostalih $n-1$ točk. Na ta način torej merimo razliko med modelom, ki vsebuje $i$-to točko in modelom, ki je ne vsebuje.

```
which(cooks.distance(model)>4/60)
55 59 60
55 59 60
```

Narišemo razsevni diagram z dodano premico in označimo točke z najvišjo Cookovo razdaljo.

```
plot(zavor$hitrost, sqrt(zavor$pot), xlab="Hitrost avtomobila (km/h)",
ylab="Zavorna pot (m)",xlim=c(5,70),ylim=c(0.7,6.7),axes=FALSE)
axis(1,pos=0.7,at=seq(5,70,by=10),tcl=-0.2)
axis(2,pos=5,at=seq(0.7,6.7,by=0.5),tcl=-0.2)
```

```
arrows(x0=65,y0=0.7,x1=67,y1=0.7,length=0.1)
arrows(x0=5,y0=6.7,x1=5,y1=6.9,length=0.1)
abline(model,lwd=2,col="hotpink")
points(zavor$hitrost[c(55,59,60)],sqrt(zavor$pot)[c(55,59,60)],
labels=zavor$model[c(55,59,60)],col="pink",pch=19)
```




Slika 5: Razsevni diagram transformiranih podatkov z ocenjeno regresijsko premico in pobarvanimi točkami, ki vplivajo na model.

Na razsevnem diagramu obarvamo točke, ki so najbolj različne od ostalih. Gre za točke 50, 59 in 60, ki imajo najvišjo Cookovo razdaljo. Zanima nas, če so točke 55, 59, 60 neobičajne oz. drugačne od ostalih podatkov. To preverimo s sledečim ukazom:

```
any(cooks.distance(model)[c(55,59,60)]>=qf(0.5,2,60))
[1] FALSE
```

Odgovor ni pritrdilen, to pomeni, da podatkovne točke nimajo velikega vpliva na linearni regresijski model in jih ni potrebno odstraniti in konstruirati novega linearnega modela brez njih.

## 6 Testiranje linearnosti regresijskega modela in koeficient determinacije

S pomočjo funkcije *summary* dobimo poročilo o modelu in rezultate t-testa za testiranje linearnosti modela in koeficient determinacije.

```
summary(model)

Call:
lm(formula = sqrt(pot) ~ hitrost, data = zavor)

Residuals:
     Min       1Q   Median       3Q      Max
-0.8125  -0.2971  -0.0095   0.3102   0.8708

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.520002   0.109502   4.749 1.31e-05 ***
hitrost     0.086607   0.003194  27.119  < 2e-16 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1

Residual standard error: 0.3994 on 60 degrees of freedom
Multiple R-squared:  0.9246,    Adjusted R-squared:  0.9233
F-statistic: 735.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

Testna statistika za testiranje linearnosti modela je $T = 27.119$, s $df = 60$ prostostnimi stopnjami in s p-vrednostno $p = 2e-16$, 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. S formalnim statističnim testiranjem smo potrdili, da linearni model ustreza podatkom. Standardni odklon najključnih napak je ocenjen s $S = 0.3994$. Koeficient determinacije je kvadrat vzorčnega koeficienta korelacije in je enak $R^2 = 0.9246$, torej 92 % variabilnosti korena zavorne poti pojasnjuje linearni regresijski model.

## 7 Intervala zaupanja za naklon in odsek regresijske premice

Izračunajmo 95 % interval zaupanja za neznani naklon in odsek regresijske premice s pomočjo funkcije *confint*.

```
round(confint(model),3)
            2.5 % 97.5 %
(Intercept) 0.301  0.739
hitrost     0.080  0.093
```

Interval zaupanja za odsek je enak $I_a = [0.301, 0.739]$ in interval zaupanja za naklon $I_b = [0.080, 0.093]$.

# 8 Interval predikcije za vrednost Y pri izbrani vrednosti X

Pri predvidevanju zavorne poti nas zanima vrednost spremenljivke $Y$ pri izbrani vrednosti $X = x_0$. Želimo oceniti spodnjo in zgornjo mejo, med katerima se nahaja zavorna pot vseh avtomobilov, ki se vozijo s hitrostmi 25, 35, 50. Interval predikcije najdemo s pomočjo funkcije *predict*.

```
xhitrost<-data.frame(hitrost=c(25,35,50))

predict(model,newdata=xhitrost ,interval="predict")^2
        fit       lwr      upr
1  7.210189  3.530875 12.18915
2 12.611381  7.536833 18.98495
3 23.525964 16.283547 32.09706
```

Predvidena dolžina zavorne poti na celi populaciji avtomobilov, ki vozijo s hitrostjo:
1. 25 km/h je 7.21 m s 95 % intervalom predikcije zavorne poti [3.53, 12.19],
2. 35 km/h je 12.61 m s 95 % intervalom predikcije zavorne poti [7.53, 18.98],
3. 50 km/h je 23.53 m s 95 % intervalom predikcije zavorne poti [16.28, 32.10].