Linearna regresija

Dominik Uršič

Zavorna pot

1. Opis podatkov

Zbrali smo vzorec meritev hitrosti in zavorne poti na vzorcu 62 avtomobilov. Podatke smo zapisali v dokument, ki ima dva stolpca:

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

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

```
zavor<-read.csv("C:/Users/domin/Desktop/VS/Seminarska/zavor.csv", header=TRUE)
zavor$sqrt_pot <- sqrt(zavor$pot) #da dobimo pot pod korenom
str(zavor)
```

```
## 'data.frame':    62 obs. of  3 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 ...
##  $ sqrt_pot: num  1.105 0.781 1.105 1.562 1.562 ...
```

2. Opisna statistika

Zdaj bomo izračunali opisno statistiko za naše podatke – povzetek s petimi vrednostmi (minimum, maksimum, prvi in tretji kvartil, mediano), vzorčni povprečji in vzorčna standardna odklona mase in porabe goriva.

```
summary(zavor$hitrost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   16.00   28.00   30.39   42.75   64.00
```

```
sd(zavor$hitrost)
```

```
## [1] 16.01368
```

Opazimo, da hitrost v vzorcu avtomobilov varira od 6.00 do 64.00 km/h, s povprečjem 30.39 in standardnim odklonom 16.01368 km/h. Ponovimo postopek računanja za vzorec zavorne poti.

```r
summary(zavor$pot)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.610   4.037   8.990  11.980  17.295  42.060
```

```r
sd(zavor$pot)
```

```
## [1] 10.17246
```

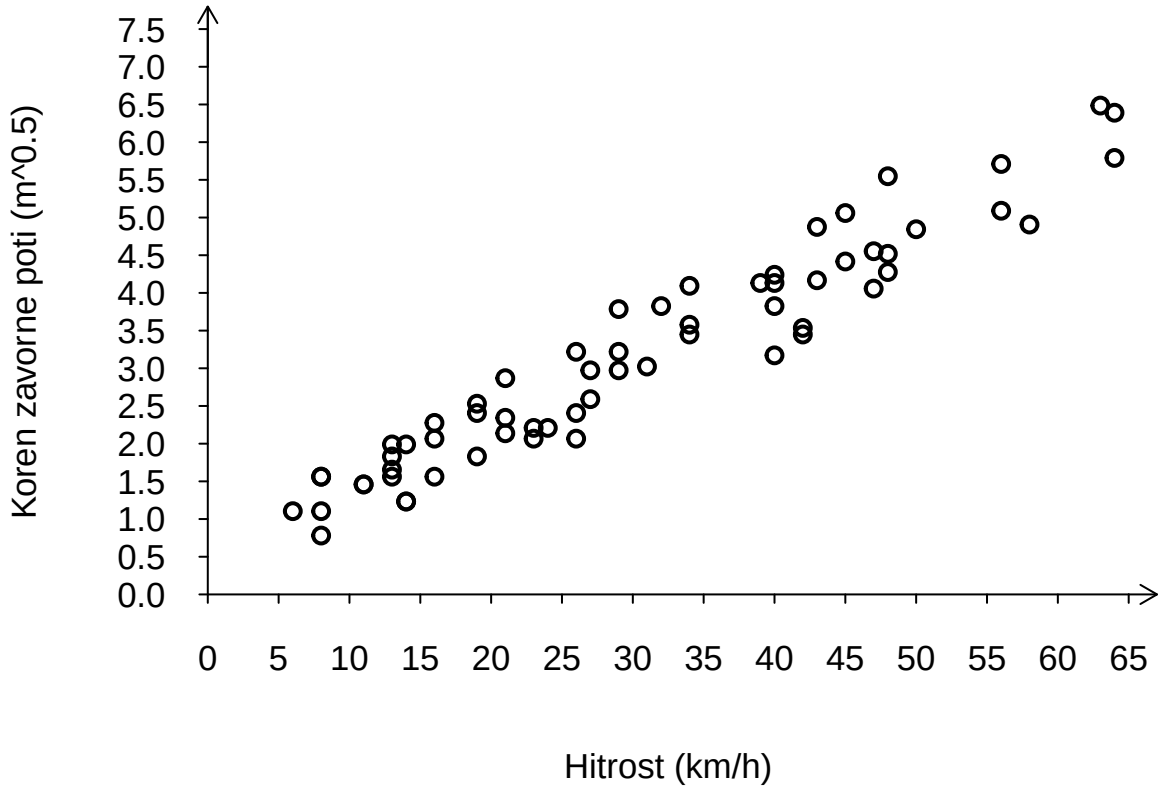
Opazimo, da zavorna pot vzorca avtomobilov varira od 0.610 do 42.06 metrov, s povprečjem 11.98m in standardnim odklonom 10.1724mm.

Razpon vrednosti hitrosti avtomobilov, korena zavorne poti in zavorne poti nam pomaga pri izbiri mej na oseh razsevnega diagrama.

3. Razsevni in vzorčni koeficient korelacije

Prikažimo dobljene podatke na razsevnem diagramu.

```r
par(las=1, cex=1.1, mar=c(4,4,2,2))
plot(zavor$hitrost, zavor$sqrt_pot, main="", xlim=c(0,65), ylim=c(0,7.5),
xlab="Hitrost (km/h)", ylab="Koren zavorne poti (m^0.5)", lwd=2, axes=FALSE)
axis(1,pos=0,at=seq(0,65,by=5),tcl=-0.2)
axis(2,pos=0,at=seq(0,8,by=0.5),tcl=-0.2)
arrows(x0=65,y0=0,x1=67,y1=0,length=0.1)
arrows(x0=0,y0=7,x1=0,y1=7.8,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(zavor$hitrost, zavor$sqrt_pot))
```

```
## [1] 0.9615461
```

Vrednost vzorčnega koeficienta korelacije je visoka ($r = 0.962$), kar govori o visoki linearni povezanosti hitrosti avtomobilov in njihove zavorne poti. Dalje, koeficient korelacije je pozitiven, kar pomeni, da avtomobili z večjo hitrostjo imajo večjo zavorno pot.

4. Formiranje linearnega regresijskega modela

Formirajmo linearni regresijski model.

```
(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 $\hat{y} = 0.52 + 0.08661x$, oziroma oceni odseka in naklona sta enaki $\hat{a} = 0.52$ in $\hat{b} = 0.08661$.

5. Točke visokega vzvoda in osamelci

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

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

```
##    hitrost   pot sqrt_pot
## 59      58 24.08 4.907138
## 60      63 42.06 6.485368
## 61      64 33.53 5.790509
## 62      64 40.84 6.390618
```

Odkrili smo 4 točke visokega vzvoda. Te točke predstavljajo avtomobile z visoko hitrostjo, ki odstopajo od povprečja. Dva avtomobila imata visoko hitrost, preko 60 km/h. Te vrednosti močno vplivajo na regresijski model in so potencialni osamelci.

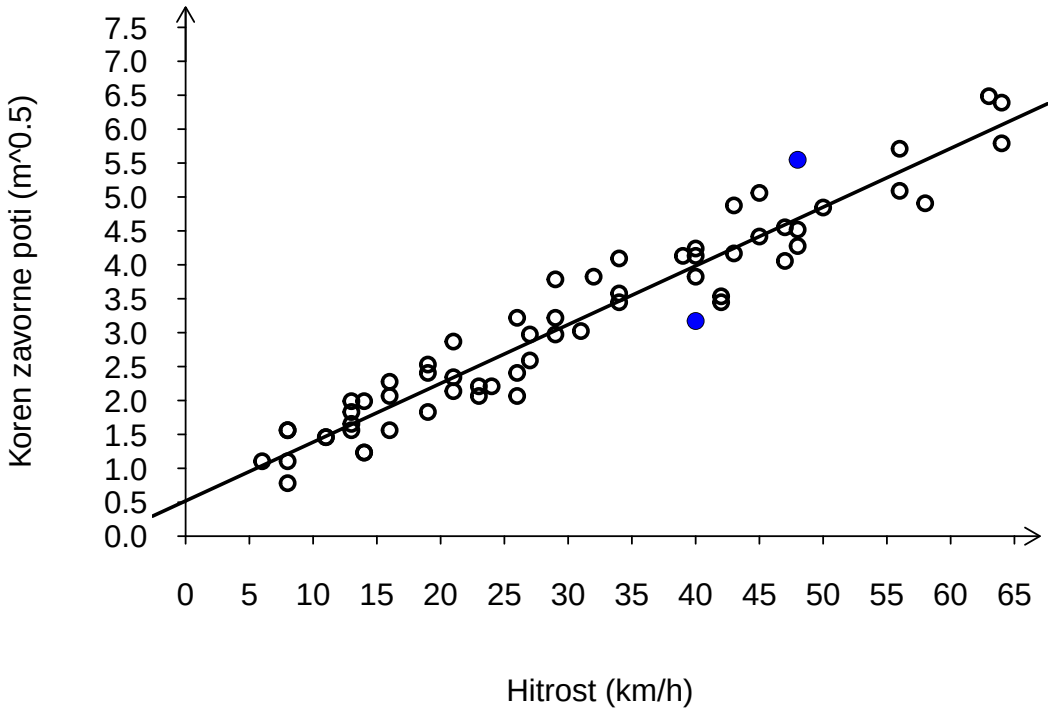
Za podatke majhne in srednje velikosti vzorca je osamelec podatkovna točka, kateri ustreza standardizirani ostanek izven intervala $[-2, 2]$.

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

```
##    hitrost   pot sqrt_pot
## 41      40 10.06 3.171750
## 55      48 30.78 5.547973
```

Identificirali smo dve podatkovne točke (41. in 55. točka) kot osamelce. Zdaj poglejmo na razsevnem diagramu po čem so te točke drugačne od ostalih. Kodi za razsevni diagram dodamo še dve vrstici, s katerima bomo dodali ocenjeno regresijsko premico in pobarvali te dve točki.

```
par(las=1, cex=1.1, mar=c(4,4,2,2))
plot(zavor$hitrost, zavor$sqrt_pot, main="", xlim=c(0,65), ylim=c(0,7.5),
xlab="Hitrost (km/h)", ylab="Koren zavorne poti (m^0.5)", lwd=2, axes=FALSE)
axis(1,pos=0,at=seq(0,65,by=5),tcl=-0.2)
axis(2,pos=0,at=seq(0,8,by=0.5),tcl=-0.2)
arrows(x0=65,y0=0,x1=67,y1=0,length=0.1)
arrows(x0=0,y0=7,x1=0,y1=7.8,length=0.1)
abline(model,lwd=2)
points(zavor$hitrost[c(41,55)],zavor$sqrt_pot[c(41,55)],col="blue",pch=19)
```

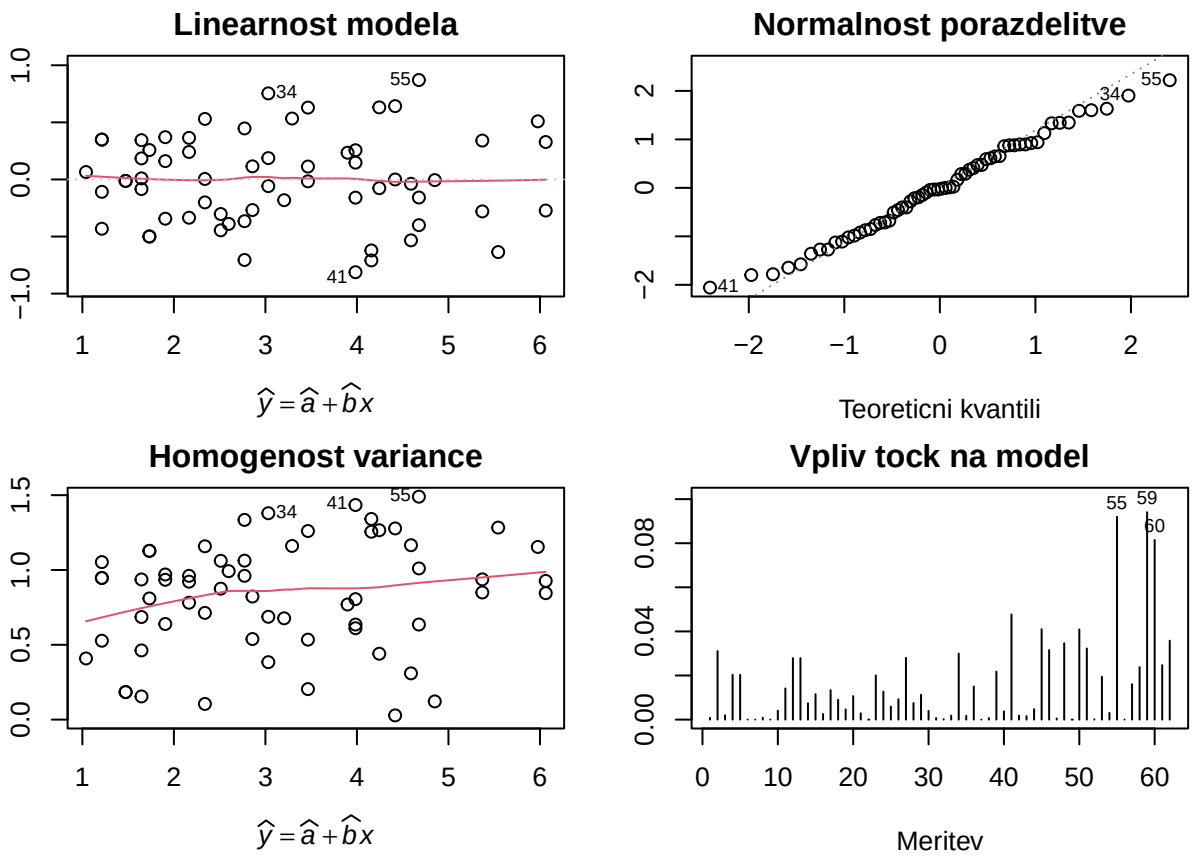


Na razsevnem diagramu opazimo, da se omenjeni osamelci nanšajo na avtomobile, ki imajo visoko zavorno pot glede na hitrost ali pa obratno, nizko zavorno pot glede na hitrost. Pri 41. ima hitrost 40km/h zavorna pot pa je samo $3.17m^{0.5}$ (10.06m). Pri 55. točki pa je hitrost le 48km/h zavorna pot pa zelo velika $5.55m^{0.5}$ (30.78m). Opazimo pa lahko, da ni nobena točka hkrati osamelec in točka visokega vzvoda.

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, točke so enakomerno razporejene nad in pod premico in lahko zaključimo, da je linearni model validen. Točke na grafu izgledajo precej naključno razporejene, opažamo le manjšo koncentracijo točk za višje vrednosti večje od 5, kar je prisotno, zaradi originalnih vrednosti v vzorcu.

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 hitrosti in zavorni poti avtomobilov lahko zaključimo, da so naključne napake normalno porazdeljene (ni večjih odstopanj od premice, razen za 41., 34. in 55. 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 kažejo, da ni naraščanja ali padanja variance. Ničelna domneva konstantne variance se lahko formalno preveri še s Breusch-Paganovim testom.

```
suppressWarnings(library(car))
```

```
## Loading required package: carData
```

```
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 $\chi^2 = 2.4742$, $df = 1$, p-vrednost $p = 0.116 > 0.05$), 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 55., 59., in 60. podatkovne točka. 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(55,59,60)]>=qf(0.5,2,nrow(zavor)-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 = 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
```

Vrednost testne statistike za preverjanje linearnosti modela je enaka $t = 27.119$, s $df = 60$ 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 $R^2 = 0.9246$, kar pomeni, da 92% variabilnosti zavorne poti 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) 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 $I_b = [0.080, 0.093]$.

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 $\hat{y} = 0.52 + 0.08661x_0$ 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.

```
xhitrost = data.frame(hitrost=c(20,40,60))
predict(model, xhitrost, interval="predict")
```

```
##        fit      lwr      upr
## 1 2.252144 1.444033 3.060255
## 2 3.984286 3.176566 4.792006
## 3 5.716428 4.889128 6.543729
```

Predvidena vrednost zavorne poti za avtomobil z hitrostjo (na celi populaciji avtomobilov)

1. 20km/h je 2.25 $m^{0.5}$, s 95% intervalom predikcije zavorne poti $[1.44, 3.060]$,
2. 40km/h je 3.98 $m^{0.5}$, s 95% intervalom predikcije zavorne poti $[3.18, 4.79]$,
3. 60km/h je 5.72 $m^{0.5}$, s 95% intervalom predikcije zavorne poti $[4.89, 6.54]$.

10. Zaključek

Zanimala nas je funkcionalna odvisnost med hitrostjo avtomobilov in njihovo zavorno potjo merjeno v metrih. Zbrali smo vzorec 62 avtomobilov, jim izmerili hitrost in zabeležili njihovo zavorno pot pri tej hitrosti. Ugotovili smo, da je enostavni linearni model odvisnosti zavorne poti od hitrosti dober. Diagnostični grafi in statistični testi niso pokazali na težave z linearnim regresijskim modelom. Koeficient determinacije je 92%, kar pomeni, da tolikšen delež variabilnosti zavorne poti zajamemo z linearnim modelom. Napoved dolžine zavorne poti na osnovi njegove izmerjene hitrosti je zadovoljiva, vendar bi vključevanje dodatnih neodvisnih spremenljivk zagotovo dala še boljši model in bolj zanesljivo napoved.