Linearna regresija - 7. domača naloga

Mateja Gosenca

3. 5. 2009

0.1 1. naloga

0.1.1 Numerično reševanje

Najprej sem napisala kratek program, s pomočjo katerega sem določila k, n in χ^2:

```
#include<stdio.h>
#include<math.h>

int main (void){
  FILE *fin;

float x, y, erry, k, n;
float Hi2=0;
float hi2=0;
float sum1=0;
float sumx=0;
float sumy=0;
float sumx2=0;
float sumxy=0;
float sumy2=0;

fin = fopen("HitrostToka.dat", "r");
  while(fscanf(fin, "%f %f %f", &x, &y, &erry) == 3){
sumx+=(x/pow(erry,2));
sumy+=(y/pow(erry,2));
sumx2+=((x*x)/pow(erry,2));
sumxy+=((x*y)/pow(erry,2));
sum1+=(1/pow(erry,2));
    sumy2+=(y*y/pow(erry,2));
}
fclose(fin);

k=(sum1*sumxy - sumx*sumy)/(sum1*sumx2 - sumx*sumx);
n=(sumx2*sumy - sumx*sumxy)/(sum1*sumx2 - sumx*sumx);
hi2=sumy2+k*k*sumx2-2*k*sumxy-2*n*sumy+2*k*n*sumx+n*n*sum1;
```

```
fin = fopen("HitrostToka.dat", "r");
  while(fscanf(fin, "%f %f %f", &x, &y, &erry) == 3){
Hi2+=(pow((y-k*x-n),2)/pow(erry,2));
}
fclose(fin);

printf("k=%f\nn=%f\nHi2=%f\nhi2=%f\n", k, n, Hi2, hi2);

return 0;
}
```

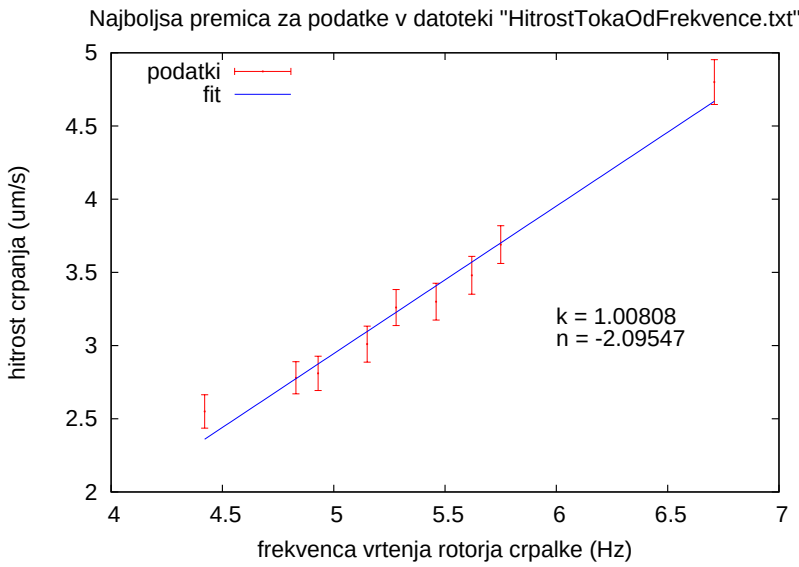
Pri tem sem χ^2 določila na dva različna načina, a so se odstopanja pojavila šele pri 3. decimalnem mestu, najverjetneje zaradi nenatančnosti numerične metode.

0.1.2 Rezultati

$k = 0.978108$ $n = -1.938723$ $\chi^2 = 5.434458$

0.1.3 Grafično reševanje

Nato sem koeficienta k in n poskusila določiti grafično, s pomočjo funkcije za fit v gnuplotu. Ker pri tem gnuplot in upošteval napake za hitrost črpanja, se rezultati, prikazani na sliki 1



Slika 1: Koeficienta k in n sta določena grafično.

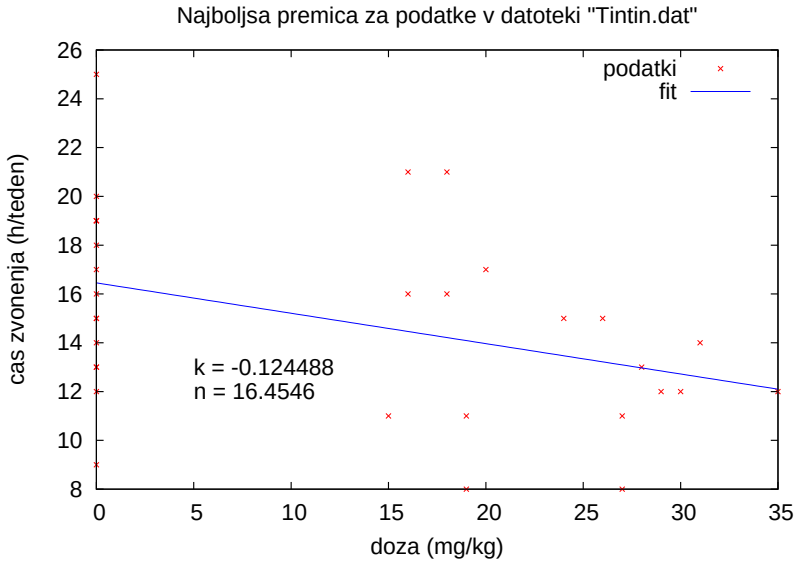
nekoliko razlikujejo od "numerično določenih". Podobne rezultate dobimo tudi, če k in n izračunamo brez da bi upoštevali napako: $k = 1.008077$ $n = -2.095434$.

0.2 2. naloga

Pri tej nalogi sem določila koeficienta linearne funkcije k in n za podatke v datoteki Tintin.dat.

0.2.1 Grafično reševanje

Rezultati so prikazani na sliki 2.



Slika 2: Prikazani so podatki ter najboljše prilegajoča premica za vse podatke.

Nato sem narisala podoben graf, le da sem tokrat izpustila podatke o pacientih, ki so dobili placebo. Ti rezultati so vidni na sliki 3.

0.2.2 Metoda najmanjših kvadratov

Koeficienta k in n sem določila tudi numerično, s pomočjo programa, napisanega v c in dobila rezultate

$k = -0.124488$
$n = 16.454575$
$\chi^2 = 419.574890$
ter
$k = -0.232646$
$n = 19.152527$
$\chi^2 = 194.507431$

le za paciente, ki so dobili zdravilo. Ker je pričakovana vrednost χ^2 za dobro prilagojeno premico enaka $n \pm \sqrt{2n}$, torej 32 ± 8 za vse paciente in 17 ± 5.8 za zdravljene, lahko zaljučimo, da ujemanje podatkov s premico ni najboljše.

0.2.3 Metoda s korelacijskim koeficientom in povprečnimi vrednostmi

Koeficienta k in n pa lahko določimo tudi drugače, tako da uporabimo rezultate, ki smo jih dobili pri nalogi 6.2:

$R = -0.394090$
$x_{pov} = 12.437500$

$y_{pov} = 14.906250$
$\sigma_x = 12.472313$
$\sigma_y = 3.939855$

Vemo, da gre premica skozi težišče oblaka točk, torej skozi x_{pov} in y_{pov} ter da ima naklon $\frac{R\sigma_y}{\sigma_x}$. Za k in n tako dobimo $k = -0.124$ ter $n = y_{pov} - kx_{pov} = 16.453$.

0.3 3. naloga

Pri tej nalogi je bilo potrebno določiti koleficienta A in λ v funkciji

$$y = Ae^{-\lambda x} \tag{1}$$

za eksponentno porazdeljene podatke na histogramu datoteke "Interval.dat". To sem naredila tako, da sem zgornjo formulo logaritmirala in dobila

$$ln(y) = ln(A) - \lambda x \tag{2}$$

kar pa lepo diši po linearni funkciji in zna gnuplot dobro fitati. Zato sem logaritmirala še podatke v drugem stolpcu datoteke, ki sem si jo pripravila za histogram (višine stolpcev), uporabljajoč ukaz v gnuplotu:

```
f(x) = a*x+b
fit f(x) 'IntervalRez100.dat' u ($1):(log($2)) via a, b
```

Tako dobljene podatke sem fitala z linearno fukncijo

$$y = kx + n \tag{3}$$

Če primerjamo enačbi (2) in (3), je očitno, da je $\lambda = -k$ in $A = e^n$. Tako dobljeni podatki s premico so prikazani na sliki 4

Za koeficiente premice sem dobila
$k = -0.00285642 \pm 0.0002256$
$n = 3.26403 \pm 0.2429$
Od koder sledi, da je
$\lambda = 0.00285642 \pm 0.0002256$
$A = 26.154728596$

Ta koeficienta sem uporabila za to, da sem narisala eksponentno krivuljo s tema koeficientoma v isti graf kot nelogaritmirani histogram. To je vidno na sliki 5

Nato pa sem poskusila eksponentno funkcijo narisati direktno, s tem da sem gnuplotu za izhodišče podala prej izračunana koeficienta:

```
f(x) = a*exp(-b*x)
a=26.15472859
b=0.0028564
fit f(x) 'IntervalRez.dat' u 1:2 via a, b
```

Na tak način sem dobila še malo bolj natančni vrednosti za λ in A: $\lambda = 0.00308158 \pm 0.0001667$
$A = 31.7182 \pm 1.223$ ter krivuljo, ki se še lepše prilega histogramu (slika 6)

0.4 4. naloga

Potrebno je bilo določiti koeficienta F_{max} in c za sigmoidno krivuljo

$$\frac{F}{F_{max}} = \frac{c}{c+a} \tag{4}$$

Najprej sem vpeljala recipročni spremenljjivki. Enačbo sem tako preoblikovala v

$$\frac{F_{max}}{F} = \frac{a}{c} + 1 \tag{5}$$

Narisala sem graf odvisnosti $\frac{F_{max}}{F}$ od $\frac{1}{c}$ in ga pofitala z linearno funkcijo uporabljajoč ukaze v gnuplotu

```
f(x) = k*x+n
f2(x) = l*x+1
fit f(x) 'Adrenalin.dat' u (1/$1):(1/(($2)/100)) via k, n
fit f2(x) 'Adrenalin.dat' u (1/$1):(1/(($2)/100)) via l
```

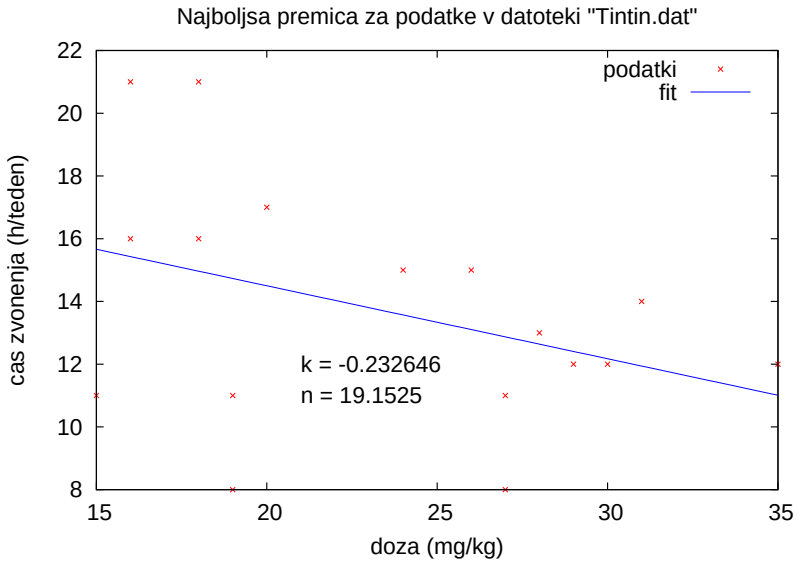
To je vidno na sliki 7

Tako sem dobila koeficienta k in n. Koeficient k približno ustreza koeficientu a iz enačbe (4). Čeprav bi koeficient n načeloma maral biti 1, sem narisala še eno premico, kjer sem n pustila kot prosti parameter. Za določitev F_{max} sem nato narisala še graf prvotnih podatkov skupaj s krivuljo, ki se najboljše prilega podatkom iz datoteke "Adrenalin.dat". Uporabila sem ukaze

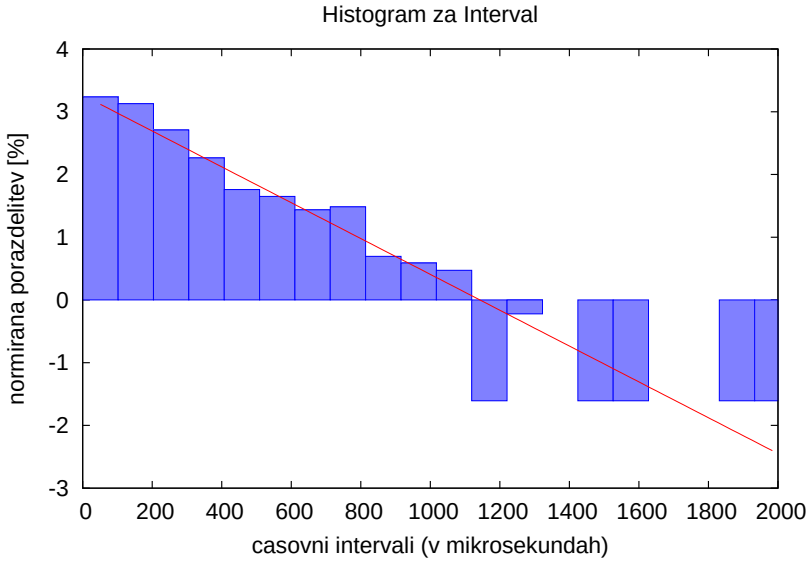
```
f(x) = b*x/(x+a)
a=34.5391
fit f(x) 'Adrenalin.dat' u 1:($2/100) via a, b
```

To je prikazano na sliki 8

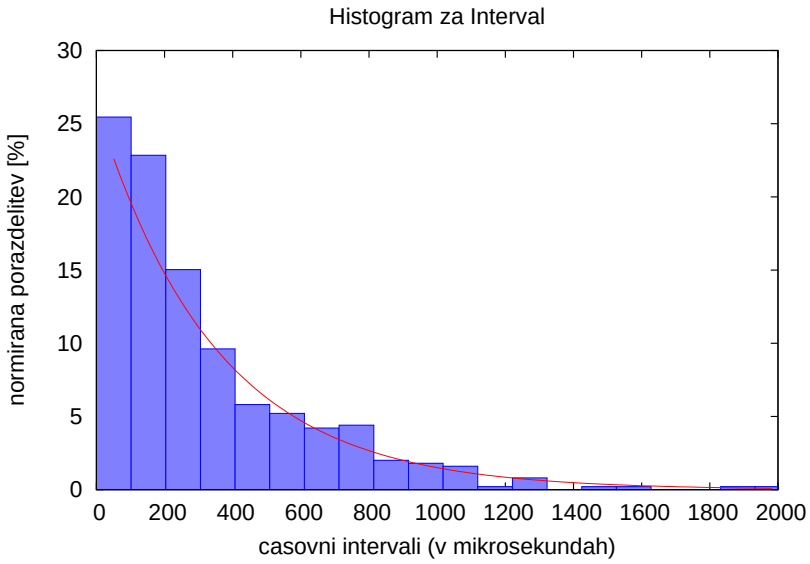
Za koeficiente sem dobila $a = 24.7597 \pm 4.166$ $F_{max} = 1.06318 \pm 0.0476$



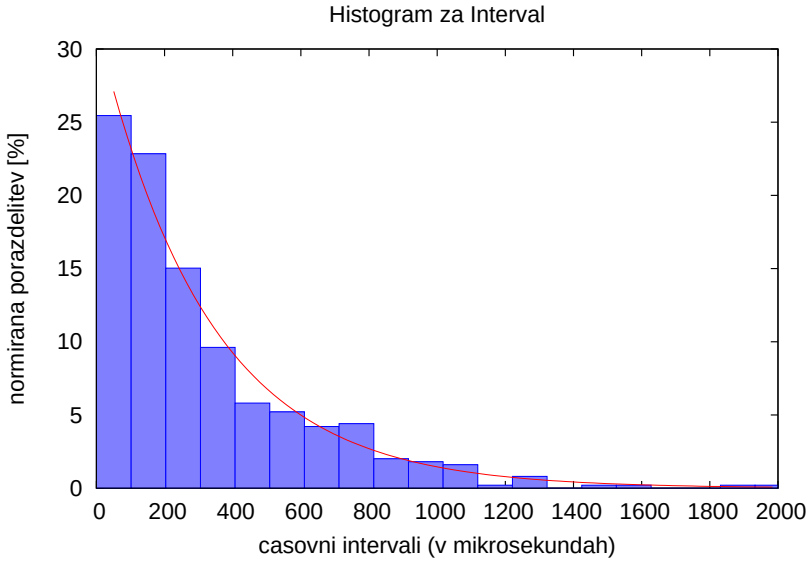
Slika 3: Prikazani so podatki ter najboljše prilegajoča premica za podatke brez primerjalnih pacientov



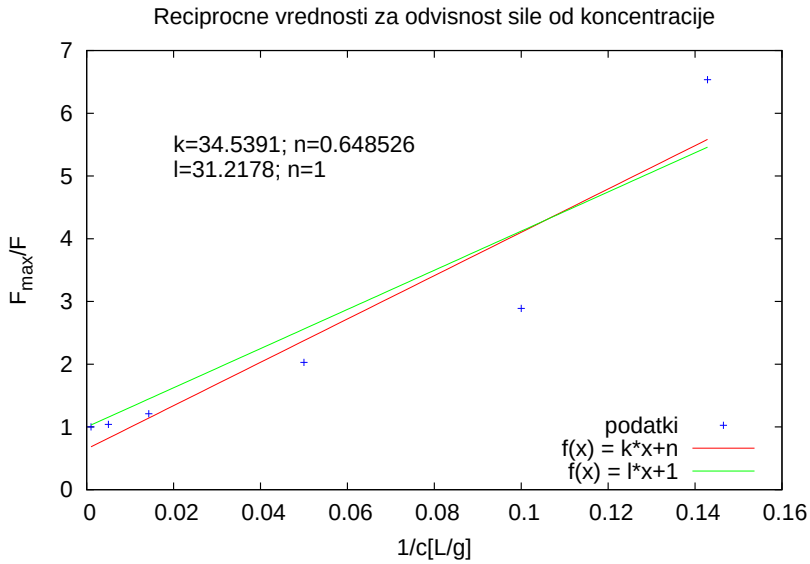
Slika 4: Prikazani so logaritmirani podatki histograma in najboljše prilegajoča premica. Podatki so razvrščeni v 20 predalčkov.



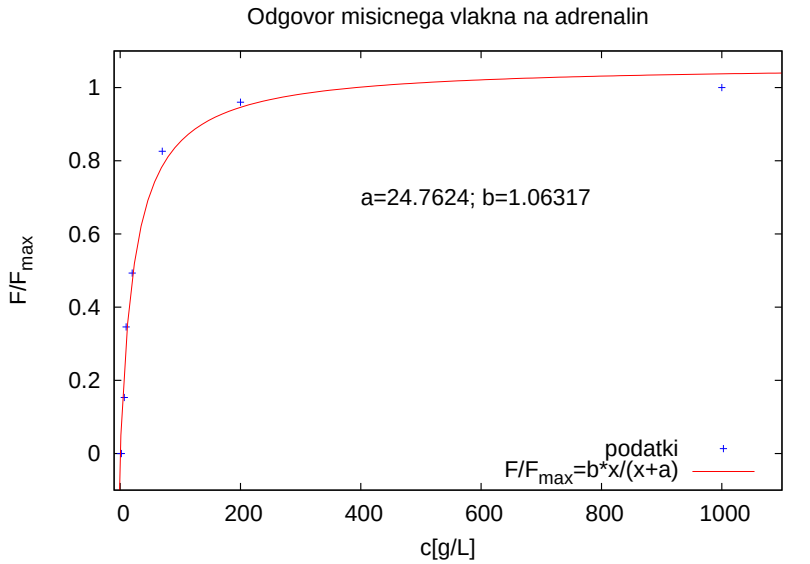
Slika 5: Prikazan je histogram za podatke v datoteki "Interval.dat"in eksponenta krivulja s podanima koeficientoma. Podatki so razvrščeni v 20 predalčkov.



Slika 6: Prikazan je histogram za podatke v datoteki "Interval.dat"in eksponenta krivulja, ki jo je gnuplot narisal kot rezultat fita. Podatki so razvrščeni v 20 predalčkov.



Slika 7: Podatke na y osi sem delila s 100, ker so bili prej v procentih.



Slika 8: Prikazani so podatki iz datoteke "Adrenalin.dat"in krivulja, ki se tem podatkom najlepše prilega.