

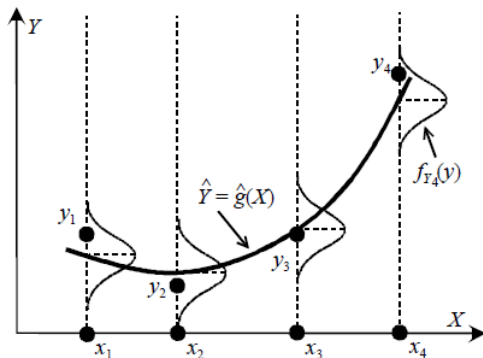
Нелинеарни регресиони модели и линеаризација

3.час

15. март 2016.

Регресиона анализа

Регресиона анализа је скуп статистичких метода којима се открива да ли постоје везе између посматраних појава. Основни задатак регресионе анализе је да предвиди понашање зависне променљиве (Y) помоћу познатих вредности једне или више независне променљиве (X), односно да одреди неслучајну функцију g тако да важи $g(X) = Y$



Полиномијална регресија

Полиномијални регресиони модел са једном зависном променљивом може се изразити релацијом:

$$Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \dots + \beta_k X^k + \epsilon.$$

где су:

- Y зависна променљива,
- X независна (објашњавајућа) променљива,
- $\beta_0, \beta_1, \dots, \beta_k$ непознате константе, регресиони параметри,
- ϵ стохастички члан, случајна грешка или резидуали.

Потребно је да случајне величине ϵ_i имају центрирану нормалну расподелу са константном дисперзијом и да су међусобно некорелисане.

Полиномијална регресија

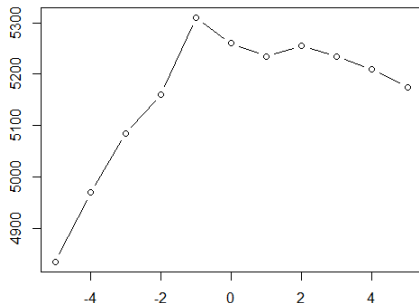
Као пример нелинеарне регресије облика полиномне функције може се размотрити промена броја становника у период од 1959. до 1969.године. Подаци који се користе у примеру дати су у табели:

Година	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
Популација (број становника)	4835	4970	5085	5160	5310	5260	5235	5255	5235	5210	5175

Полиномијална регресија

Да би се утврдио облик зависности броја становника током вренема, потребно је податке из табеле представити на одговарајућем дијаграму:

```
> plot(godina, Populacija, type='b')
```



Који степен полинома ће се применити зависи од степена прецизности који се тражи. Већи степен полинома, даје већу тачност модела, али и више тешкоћа у израчунавању.

Полиномијална регресија

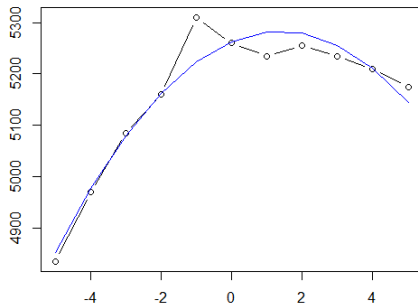
```
> summary(lm(Populacija~godina+I(godina^2)))
```

Coefficients:

(Intercept)	godina	I(godina^2)
5263.16	29.32	-10.59

```
> pol.model1 <- nls(Populacija ~ a + b*godina + c*godina^2, start = list(a = 0, b  
= 0, c = 0))
```

```
> lines(godina, a_ocena + b_ocena*godina + c_ocena*godina^2, col = "blue")
```



$$Populacija = 5263.16 + 29.32godina - 10.59godina^2$$

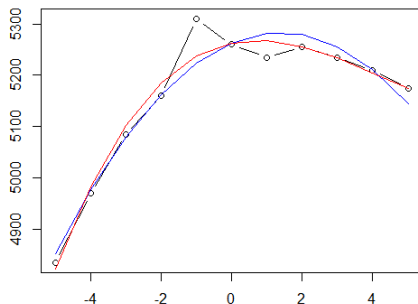
Полиномијална регресија

```
> summary(lm(Populacija ~ godina+I(godina^2)+I(godina^3)))
```

Coefficients:

(Intercept)	godina	I(godina^2)	I(godina^3)
5263.1585	14.3638	-10.5886	0.8401

```
> pol.model2 <- nls(Populacija ~ a + b*godina + c*godina^2 + d*godina^3, start =  
list(a = 0, b = 0, c = 0, d = 0))  
> lines(godina,a_ocena+b_ocena*godina+c_ocena*godina^2+d_ocena*godina^3,col='red')
```



$$Populacija = 5263.16 + 14.36godina - 10.59godina^2 + 0.84godina^3$$

Експоненцијални регресиони модел

У експоненцијалном регресионом моделу једначина регресије има облик:

$$Y = BA^X,$$

при чему важи $A, B > 0$; $A \neq 1$. Када је $A = 1$ добија се константна функција, која је, заправо, посебан случај просте линеарне регресије.

Да бисмо пронашли параметре помоћу методе најмањих квадрата, потребно је модел трансформисати да буде линеаран у параметрима, а то ћемо урадити логаритамском трансформацијом. Добијамо једначину:

$$\log Y = X \log A + \log B.$$

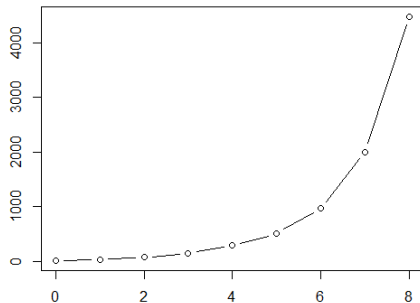
Експоненцијални регресиони модел

Откривена је бактерија која се врло брзо размножава. У табели су дати подаци о броју бактерија у узорку сваки сат након што је посматрање узорка почело.

Сати	Бр. бактерија
0	20
1	40
2	75
3	150
4	298
5	510
6	971
7	2004
8	4474

Експоненцијални регресиони модел

```
> plot(sati, bakterije, main = "Zavisnost broja bakterija od vremena",  
       xlab = "sati", ylab = "br. bakterija", type = "b")
```



Експоненцијални регресиони модел

```
> e.model <- nls(log(bakterije) ~ log(A)*sati + log(B), start = list(A = 1, B = 1))
> summary(e.model)
```

Formula: $\log(\text{bakterije}) \sim \log(A) * \text{sati} + \log(B)$

Parameters:

	Estimate	Std. Error	t value	Pr(> t)
A	1.93895	0.01676	115.72	9.49e-13 ***
B	20.07273	0.82582	24.31	5.08e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06694 on 7 degrees of freedom

Number of iterations to convergence: 6

Achieved convergence tolerance: 2.35e-08

```
> summary(lm(log(bakterije)~sati))
```

Call:

lm(formula = $\log(\text{bakterije}) \sim \text{sati}$)

Residuals:

	Min	1Q	Median	3Q	Max
	-0.09391	-0.03148	-0.00363	0.02737	0.10951

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.999362	0.041141	72.90	2.4e-11 ***
sati	0.662146	0.008641	76.62	1.7e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

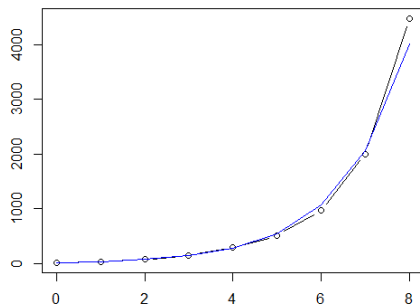
Residual standard error: 0.06694 on 7 degrees of freedom

Multiple R-squared: 0.9988, Adjusted R-squared: 0.9986

F-statistic: 5871 on 1 and 7 DF, p-value: 1.697e-11

Експоненцијални регресиони модел

```
> lines(sati, B_ocena*A_ocena^sati, col = "blue")
```



$$bakterije = 20.07 \cdot 1.94^{sati}$$

Степени регресиони модел

Једначина просте криволинијске регресије облика степене функције, гласи:

$$Y = AX^B,$$

а у логаритамском облику

$$\log Y = \log A + B \cdot \log X.$$

Параметар B показује за колико се процената промени вредност зависне променљиве Y при повећању независне променљиве X за 1%. Ако је $B > 0$, зависна променљива Y се у просеку приближно повећа за толико процената колико износи вредност параметра B .

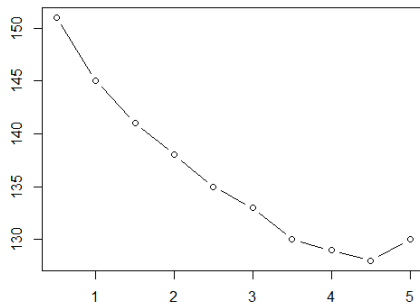
Степени регресиони модел

У табели су дати подаци о цени уласка у музеј која се мењала и кренула од 0.5 евра, а достигла цену од 5 евра. Зависна променљива је просечан дневни број посетилаца у зависности од цене улаза.

Цена	Посетиоци
0.5	151
1	145
1.5	141
2	138
2.5	135
3	133
3.5	130
4	129
4.5	128
5	130

Степени регресиони модел

```
> plot(cena, posetioci, main = "Zavisnost broja posetilaca od,  
+ cene ulaznice", type='b')
```



Степени регресиони модел

```
> stepeni.model <- nls(log(posetioci) ~ log(A) + log(cena)*B, start = list(A = 1,
B = 0))
> summary(stepeni.model)
```

Formula: $\log(\text{posetioci}) \sim \log(A) + \log(\text{cena}) * B$

Parameters:

	Estimate	Std. Error	t value	Pr(> t)
A	144.358217	0.642387	224.72	< 2e-16 ***
B	-0.074687	0.004147	-18.01	9.26e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009119 on 8 degrees of freedom

Number of iterations to convergence: 7

Achieved convergence tolerance: 8.631e-07

```
> summary(lm(log(posetioci)~log(cena)))
```

Call:

lm(formula = $\log(\text{posetioci}) \sim \log(\text{cena})$)

Residuals:

	Min	1Q	Median	3Q	Max
	-0.0111978	-0.0076458	0.0007581	0.0061529	0.0154414

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.972298	0.004450	1117.38	< 2e-16 ***
log(cena)	-0.074687	0.004147	-18.01	9.26e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

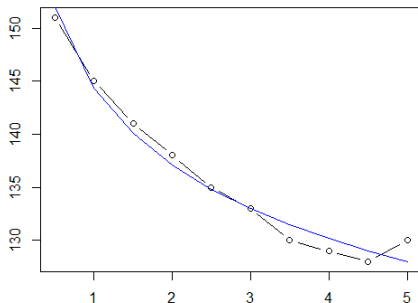
Residual standard error: 0.009119 on 8 degrees of freedom

Multiple R-squared: 0.9759, Adjusted R-squared: 0.9729

F-statistic: 324.4 on 1 and 8 DF, p-value: 9.265e-08

Степени регресиони модел

```
> lines(cena, a*(cena)^b, col = "blue")
```



$$posetioci = 144.358217 \cdot cena^{0.074687}.$$

Логаритамски регресиони модел

Стандардни облик једначине оваквог облика регресије је:

$$Y = A \cdot \ln X + B,$$

при чему је неопходно да вредности независне променљиве X буду строго позитивне, да би логаритамска функција била добро дефинисана.

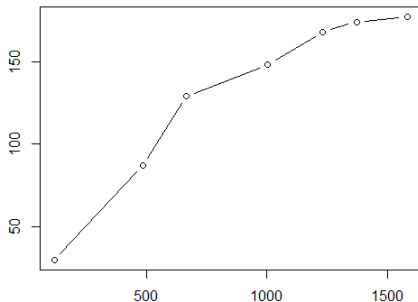
Логаритамски регресиони модел

Забележена је просечна дебљина седам стабла седам пута у току 4 године. Дат је просечни обим стабла при старости дрвећа од 177, 484, 664, 1004, 1231, 1372 и 1582 дана. Потребно је наћи одговарајућу регресиону једначину. Подаци у дати у наредној табели.

Старост	Обим стабла
117	30
484	87
664	129
1004	148
1231	168
1372	174
1582	177

Логаритамски регресиони модел

```
> plot(starost, obim, xlab='dani', ylab='obim (mm)',  
main = 'Obim stabla u zavisnosti od starosti', type='b')
```



Логаритамски регресиони модел

```
log.model <- nls(obim ~ A + B*log(starost), start = list(A = 0, B = 0))  
> summary(log.model)
```

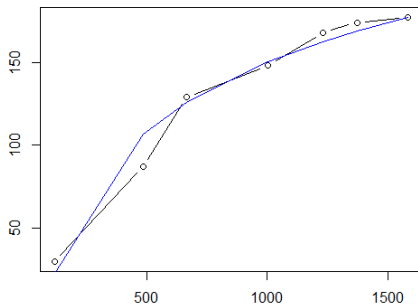
```
Formula: obim ~ A + B * log(starost)  
Parameters:  
      Estimate Std. Error t value Pr(>|t|)  
A -259.832      30.748   -8.45 0.000381 ***  
B   59.310       4.636   12.79 5.19e-05 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
Residual standard error: 10.26 on 5 degrees of freedom  
Number of iterations to convergence: 1  
Achieved convergence tolerance: 2.683e-08
```

```
> summary(lm(obim ~ log(starost)))
```

```
Call:  
lm(formula = obim ~ log(starost))  
Coefficients:  
      Estimate Std. Error t value Pr(>|t|)  
(Intercept) -259.832      30.748   -8.45 0.000381 ***  
log(starost)  59.310       4.636   12.79 5.19e-05 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
Residual standard error: 10.26 on 5 degrees of freedom  
Multiple R-squared:  0.9704, Adjusted R-squared:  0.9644  
F-statistic: 163.7 on 1 and 5 DF, p-value: 5.19e-05
```

Логаритамски регресиони модел

```
> lines(starost, -259.832 + 59.31*log(starost), col = "blue")
```



$$obim = -259.83 + 59.31 \cdot \log(starost)$$

Задаци

1. База података `pressure` садржи две променљиве температуру и притисак. Наћи одговарајући регресиони модел који представља зависност притиска од температуре. Израчунати вредности притиска ако је температура $250^{\circ}C$, $310^{\circ}C$ и $400^{\circ}C$.
2. У бази `baza.txt` се налазе подаци мерења висине пене пива у различитим тренуцима на $20^{\circ}C$. Наћи одговарајући регресиони модел који представља зависност висине пене од времена. Колика је висина после 100, 200 и 400 секунди?