

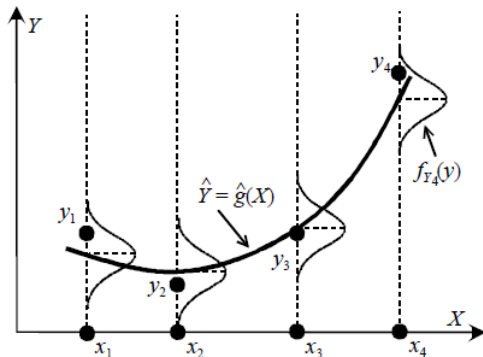
Линеарна регресија

1.час

1. 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 $g(X) = Y$



Проста линеарна регресија

Проста регресија је регресиони метод који разматра везу између једне зависно променљиве Y и једне независно променљиве X .

Прост линеарни регресиони модел може се изразити релацијом:

$$Y_i = a + bX_i + \epsilon_i, \quad i = 1, 2, \dots, N$$

где су:

- Y_i i -та зависна променљива,
- X_i i -та вредност независне (објашњавајуће) променљиве,
- a и b непознате константе, регресиони параметри,
- ϵ_i 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.

Оцене параметара просте линеарне регресије методом најмањих квадрата

Метод најмањих квадрата подразумева да се параметри модела оцењују тако да збир квадрата резидуала, тј. вертикално мерених одступања података узорка од тачака на регресионој правој оцењеној на основу узорка:

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (\hat{a} + \hat{b}X_i))^2$$

буде минималан.

За регресиону линеарну једначину $\mathbf{Y} = a + b\mathbf{X}$, оцене за параметре a и b су

$$\begin{aligned}\hat{a} &= \overline{\mathbf{Y}} - \hat{b}\overline{\mathbf{X}} \\ \hat{b} &= \frac{\text{cov}(\mathbf{X}, \mathbf{Y})}{S_{n\mathbf{X}}^2}\end{aligned}$$

Показује се да су оцене добијене овом методом најбоље линеарне оцене, непристрасне и постојане.

За проверу квалитета оцене може да се користи **коэффицијент детерминације**

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Ако је модел добар, онда је сума у броиоцу мала, па је R^2 близак 1.

Најпознатија мера линеарне корелације између случајних променљивих је **Пирсонов (Pearson) коэффициент линеарне корелације** $r = \sqrt{R^2}$. Вредност коэффицијента корелације креће се у интервалу $[-1, 1]$. У складу с величином овог коэффицијента може се закључити смер и интензитет линеарне корелације међу посматраним величинама.

Пирсонов коэффициент је бездимензиона величина која се може рачунати и по формули:

$$r = \frac{\text{cov}(X, Y)}{S_{n_X} S_{n_Y}}.$$

Пример простог линеарног модела

Посматрајмо базу података `faithful` (у пакету `car`) која садржи две случајне променљиве `waiting` и `eruptions`. Променљива `waiting` означава време чекања од следеће ерупције вулкана, а променљива `eruptions` означава трајање ерупције. Испитујемо зависност времена чекања и времена трајања ерупције.

Линеарни регресиони модел може се изразити као:

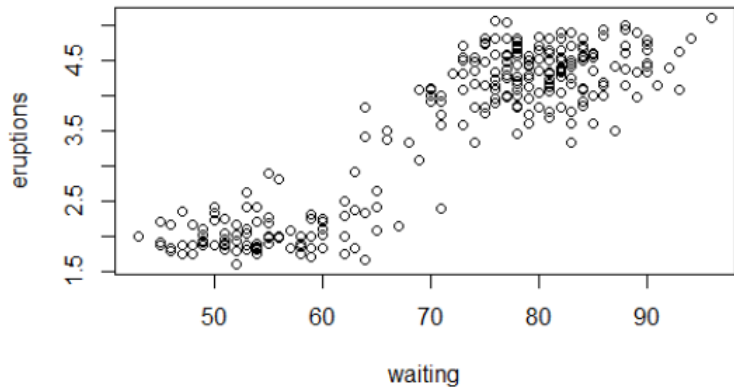
$$Eruptions = a + b * Waiting + \epsilon.$$

```
> head(faithful)
```

	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55

Дијаграм распршивања (растурања)

```
> plot(waiting,eruptions)
```



Потребно је израчунати параметре a и b по наведеној формули. Параметар се израчунава као разлика средње вредности времена трајања ерупције (зависна променљива) и узорачке средње вредности времена чекања (независна променљива). Параметар b се израчунава као количник коваријансе времена чекања и времена ерупције и дисперзије времена чекања.

```
> b_ocena <- cov(eruptions, waiting)/(sd(waiting)^2)
> a_ocena <- mean(eruptions)-b_ocena*mean(waiting)
> a_ocena
[1] -1.874016
> b_ocena
[1] 0.07562795
```

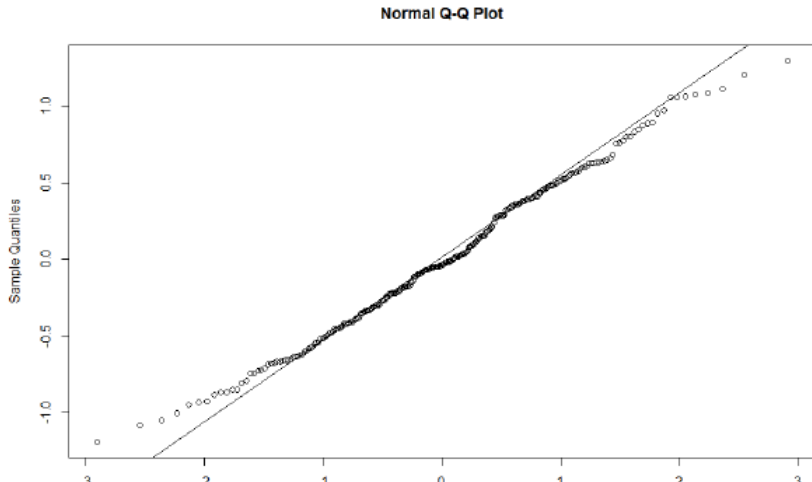
Параметри су $a = -1.874016$ и $b = 0.07562795$, а једначина регресије гласи:

$$y = -1.874016 + 0.07562795 * x.$$

Користећи регресиону једначину можемо проценити трајање наредне ерупције, ако нам је познато време протекло од последње ерупције.

Даље рачунамо резидуале регресионе једначине и проверавамо да ли резидуали имају нормалну расподелу.

```
> reziduali <- a_ocena + b_ocena*waiting-eruptions  
> qqnorm(reziduali)  
> qqline(reziduali)
```



Статистички програм R садржи уграђене функције које се користе за регресиони линеарни модел.

```
> lm(eruptions ~ waiting)
```

Call:

```
lm(formula = eruptions ~ waiting)
```

Coefficients:

```
(Intercept) waiting -1.87402 0.07563
```

```
> summary(lm(eruptions ~ waiting))
```

Call:

```
lm(formula = eruptions ~ waiting)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.29917	-0.37689	0.03508	0.34909	1.19329

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.874016	0.160143	-11.70	<2e-16 ***
waiting	0.075628	0.002219	34.09	<2e-16 ***

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

Residual standard error: 0.4965 on 270 degrees of freedom

Multiple R-squared: 0.8115, Adjusted R-squared: 0.8108

F-statistic: 1162 on 1 and 270 DF, p-value: < 2.2e-16

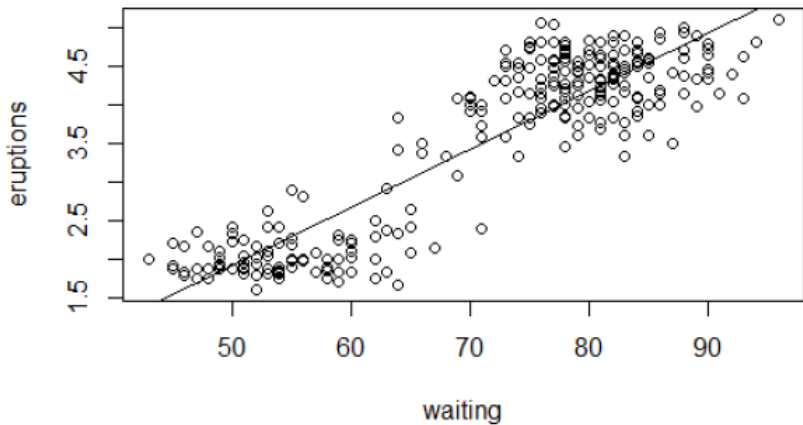
Помоћу функције `cor.test()` добијамо вредност тест статистике и p вредност за тестирање нулте хипотезе да не постоји линеарна корелација, интервал поверења и Пирсонов коефицијент.

```
> cor.test(eruptions, waiting)
```

```
Pearson's product-moment correlation
data: eruptions and waiting
t = 34.089, df = 270, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8756964 0.9210652
sample estimates:
      cor
0.9008112
```

С обзиром на добијену p вредност ($2.2e-16$) одбацујемо нулту хипотезу да не постоји линеарна корелација између времена ерупције и времена чекања.

```
> plot(waiting,eruptions)
> abline(faithful.lm)
```



Вишеструка линеарна регресија

Када се испитује зависност једне појаве од две или више независних променљивих појава, тада се говори о вишеструкој или мултиплој регресији.

У статистичкој пракси најчешће се користе подаци из узорка, који се формирају из основног скупа. На основу узорка се врши оцењивање у основном скупу, а тада једначина вишеструке линеарне регресије гласи:

$$\hat{Y}_i = a_i + b_{1i}X_{1i} + b_{2i}X_{2i} + \dots + b_{mi}X_{mi} + \epsilon_i, \quad i = 1, 2, \dots, n$$

где је m број независних променљивих.

Пример вишеструког линеарног модела

За пример вишеструке регресије посматраћемо податке фабрике за оксидацију амонијака на азотне киселине. База података `stackloss` садржи четири случајне променљиве `Air.Flow`, `Water.Temp`, `Acid.Conc` и `stack.loss` која је добијена након 21 дан рада фабрике за оксидацију амонијака (NH_3) до азотне киселине (HNO_3).

- `Air.Flow` (X_1) представља проток ваздуха или стопу рада постројења
- `Water.Temp` (X_2) је температура воде за хлађење која пролази кроз цеви у апсорпционом торњу
- `Acid.Conc` (X_3) је концентрација киселине која циркулише
- `Stack.loss` (Y) је зависна променљива и представља губитак у димњаку или инверзну меру ефикасности фабрике

База stackloss:

```
> head(stackloss)
```

	Air.Flow	Water.Temp	Acid.Conc.	stack.loss
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18

Рачунамо параметре вишеструке линеарне регресије

```
lm(stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.)
```

Call:

```
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.)
```

Coefficients:

(Intercept)	Air.Flow	Water.Temp	Acid.Conc.
-39.9197	0.7156	1.2953	-0.1521

Коефицијенти су: $a = -39.9197$, $b_1 = 0.7156$, $b_2 = 1.953$ и $b_3 = -0.1521$. Регресиона једначина је:

$$y = -39.9197 + 0.7156X_1 + 1.953X_2 - 0.1521X_3.$$

Логично је да зависна променљива `stack.loss` има негативну регресију са зависном променљивом `Acid Conc` (концентрација киселине, X_3), високу позитивну са зависном променљивом `Water.Temp` (температура воде, X_2).

Да би одредили репрезентативност дефинисаног модела вишеструке линеарне регресије потребно је израчунати коефицијент детерминације.

```
> summary(lm(stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.))$r.squared  
[1] 0.9135769
```

Коефицијент детерминације дефинисаног модела регресије је: $R^2 = 0.91358$. То значи да је 91,36% променљива `stack.loss` објашњена променљивима `Air.Flow` (X_1), `Water.Temp` (X_2) и `Acid.Conc`, а 8,64% другим утицајима.

Статистички програм R садржи уграђене функције које се користе за регресиони линеарни модел.

```
> summary(lm(stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.))
```

Call:

```
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.2377	-1.7117	-0.4551	2.3614	5.6978

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-39.9197	11.8960	-3.356	0.00375 **
Air.Flow	0.7156	0.1349	5.307	5.8e-05 ***
Water.Temp	1.2953	0.3680	3.520	0.00263 **
Acid.Conc.	-0.1521	0.1563	-0.973	0.34405

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

Residual standard error: 3.243 on 17 degrees of freedom

Multiple R-squared: 0.9136, Adjusted R-squared: 0.8983

F-statistic: 59.9 on 3 and 17 DF, p-value: 3.016e-09

Предвиђање на основу висеструке линеарне регресије

```
> newdata <- data.frame(Air.Flow=72, Water.Temp=20, Acid.Conc.=85)
> predict(lm(stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.), newdata)
24.58173
```

На основу добијеног модела вишеструке линеарне регресије и задатих параметара добијамо да је очекивана вредност зависне променљиве `stack.loss` једнака 24.582.

Задаци

1. Користећи базу података `rmr` нацртати график распршености променљивих `metabolic rate` и `body weight`. Наћи одговарајући линеарни регресиони модел и користећи тај модел одредити `metabolic rate` за тежину од 70kg.

Задаци

2. Користећи базу података `juch1` наћи одговарајући линеарни модел за квадратни корен концентрације *igf1* у односу старост за групу старију од 25 година и одредити коефицијент детерминације. Користећи добијени модел предвидети концентрацију *igf1* за особе са 50 година.