3. Napredne statističke metode za analizu podataka

3.1. Nelinearni regresijski model

3.1.1. Metoda najmanjih kvadrata

Neka je funkcija f zadana na diskretnom skupu točaka $x_0, \ldots, x_n$ i neka je za $n \geq m$ zadana aproksimacijska funkcija

$$\varphi(x, a_0, \ldots, a_m).$$

Funkcija φ određuje se iz uvjeta da suma kvadrata razlika između funkcije i aproksimacijske funkcije u čvorovima bude minimalna, tj. moramo minimizirati S,

$$S = \sum_{k=0}^{n} (f(x_k) - \varphi(x_k))^2 \to \min .$$

Ovu funkciju S interpretiramo kao funkciju nepoznatih parametara

$$S = S(a_0, \ldots, a_m).$$

Očito je uvijek $S \geq 0$, bez obzira kakvi su parametri. Dakle, zadatak je minimizirati funkciju S kao funkciju više varijabli $a_0, \ldots, a_m$. Nužni uvjet ekstrema je

$$\frac{\partial S}{\partial a_k} = 0, \quad k = 0, \ldots, m.$$

Dobili smo tzv. sustav normalnih jednadžbi.

Ilustrirajmo to sada na slučaju kada je aproksimacijska funkcija pravac:

Zadane su točke $(x_0, f(x_0)), \ldots, (x_n, f(x_n))$ koje po diskretnoj metodi najmanjih kvadrata aproksimiramo pravcem

$$\varphi(x) = a_0 + a_1 x.$$

Greška aproksimacije u čvorovima koju minimiziramo je

$$S = S(a_0, a_1) = \sum_{k=0}^{n} (f(x_k) - \varphi(x_k))^2 = \sum_{k=0}^{n} (f(x_k) - a_0 - a_1 x_k)^2 \to \min .$$

Nađimo parcijalne derivacije po prametrima a_0 i a_1:

$$\begin{aligned} 0 &= \frac{\partial S}{\partial a_0} = -2 \sum_{k=0}^{n} (f(x_k) - a_0 - a_1 x_k), \\ 0 &= \frac{\partial S}{\partial a_1} = -2 \sum_{k=0}^{n} (f(x_k) - a_0 - a_1 x_k) x_k. \end{aligned}$$

Dijeljenjem s -2 i sređivanjem po nepoznanicama $a_0,\ a_1$, dobivamo linearni sustav

$$\begin{aligned} a_0(n+1) + a_1 \sum_{k=0}^{n} x_k &= \sum_{k=0}^{n} f(x_k) \\ a_0 \sum_{k=0}^{n} x_k + a_1 \sum_{k=0}^{n} x_k^2 &= \sum_{k=0}^{n} f(x_k) x_k, \end{aligned}$$

pa je

$$\begin{aligned} a_0 &= \frac{\sum_{k=0}^{n} x_k^2 \sum_{k=0}^{n} f(x_k) - \sum_{k=0}^{n} x_k \sum_{k=0}^{n} f(x_k) x_k}{(n+1)\sum_{k=0}^{n} x_k^2 - \left(\sum_{k=0}^{n} x_k\right)^2}, \\ a_1 &= \frac{(n+1)\sum_{k=0}^{n} x_k f(x_k) - \sum_{k=0}^{n} x_k \sum_{k=0}^{n} f(x_k)}{(n+1)\sum_{k=0}^{n} x_k^2 - \left(\sum_{k=0}^{n} x_k\right)^2}. \end{aligned} \tag{3.1}$$

Za funkciju φ možemo uzeti i funkciju s općenitim baznim funkcijama $\varphi_0, \ldots . \varphi_m$

$$\varphi(x) = a_0 \varphi_0(x) + a_1 \varphi_1(x) + \cdots + a_m \varphi_m(x).$$

U tom slučaju

$$S = \sum_{k=0}^{n} \left(f(x_k) - \sum_{i=0}^{m} a_i \varphi_i(x_k) \right)^2 \ \Rightarrow\ \frac{\partial S}{\partial a_j} = -2 \sum_{k=0}^{n} \left(f(x_k) - \sum_{i=0}^{m} a_i \varphi_i(x_k) \right) (\varphi_j(x_k)) = 0,$$

pa dobijemo sustav

$$\begin{array}{ccccccccc} a_0\langle \varphi_0, \varphi_0 \rangle & + & a_1 \langle \varphi_1, \varphi_0 \rangle & + & \cdots & + & a_m \langle \varphi_m, \varphi_0 \rangle & = & \langle f, \varphi_0 \rangle \\ a_0\langle \varphi_0, \varphi_1 \rangle & + & a_1 \langle \varphi_1, \varphi_1 \rangle & + & \cdots & + & a_m \langle \varphi_m, \varphi_1 \rangle & = & \langle f, \varphi_1 \rangle \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_0\langle \varphi_0, \varphi_m \rangle & + & a_1 \langle \varphi_1, \varphi_m \rangle & + & \cdots & + & a_m \langle \varphi_m, \varphi_m \rangle & = & \langle f, \varphi_m \rangle \end{array}, \tag{3.2}$$

gdje je za funkcije g i h, $\langle g, h \rangle = \sum_{k=0}^{n} g(x_k) h(x_k)$.

Primjer 3.1. *Metodom najmanjih kvadrata odredite polinome prvog i drugog stupnja koji aproksimiraju podatke*

x_k	0	0.4	0.8	1.2	1.6	2.0
$f(x_k)$	0.21	1.25	2.31	2.70	2.65	3.20

.

Rješenje. 1) $P_1 : \varphi(x) = a_0 + a_1 x$

$$n = 5,\ \sum_{k=0}^{5} x_k = 6,\ \sum_{k=0}^{5} x_k^2 = 8.8,\ \sum_{k=0}^{5} x_k f(x_k) = 16.228,\ \sum_{k=0}^{5} f(x_k) = 12.32.$$

Iz (3.1) imamo

$$a_0 = \frac{8.8 \cdot 12.32 - 6 \cdot 16.228}{6 \cdot 8.8 - 36} = 0.6576,\quad a_1 = \frac{6 \cdot 16.228 - 6 \cdot 12.32}{6 \cdot 8.8 - 36} = 1.3957,$$

pa je $\varphi(x) = 0.6576 + 1.3957x$.

2) $P_2 : \varphi(x) = a_0 + a_1 x + a_2 x^2$

Ako u sustav (3.2) stavimo $m = 2$ i $\varphi_0(x) = 1,\ \varphi_1(x) = x,\ \varphi_2(x) = x^2$ imamo

$$\begin{array}{rcrcrcl} a_0(n+1) & + & a_1 \sum_{k=0}^5 x_k & + & a_2 \sum_{k=0}^5 x_k^2 & = & \sum_{k=0}^5 f(x_k) \\ a_0 \sum_{k=0}^5 x_k & + & a_1 \sum_{k=0}^5 x_k^2 & + & a_2 \sum_{k=0}^5 x_k^3 & = & \sum_{k=0}^5 f(x_k) x_k\ , \\ a_0 \sum_{k=0}^5 x_k^2 & + & a_1 \sum_{k=0}^5 x_k^3 & + & a_2 \sum_{k=0}^5 x_k^4 & = & \sum_{k=0}^5 f(x_k) x_k^2 \end{array}$$

pa iz $\sum_{k=0}^5 x_k^3 = 14.4,\ \sum_{k=0}^5 x_k^4 = 25.0624,\ \sum_{k=0}^5 f(x_k) x_k^2 = 25.1504$ imamo

$$\begin{array}{rcrcrcl} 6a_0 & + & 6a_1 & + & 8.8a_2 & = & 12.32 \\ 6a_0 & + & 8.8a_1 & + & 14.4a_2 & = & 16.228\ . \\ 8.8a_0 & + & 14.4a_1 & + & 25.0624a_2 & = & 25.1504 \end{array}$$

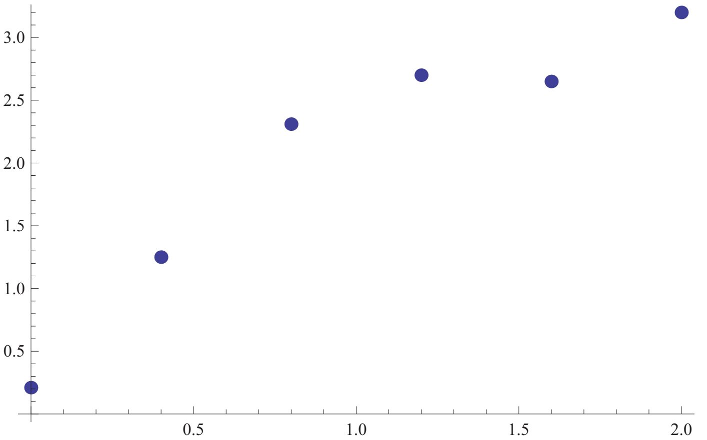
Rješenje sustava je $a_0 = 0.2475,\ a_1 = 2.9337,\ a_2 = -0.769$ pa je polinom $\varphi(x) = 0.2475 + 2.9337x - 0.769x^2$.

Programska realizacija

```
podaci = {{0, 0.21}, {0.4, 1.25}, {0.8, 2.31}, {1.2, 2.7}, {1.6, 2.65}, {2, 3.2}}
{{0, 0.21}, {0.4, 1.25}, {0.8, 2.31}, {1.2, 2.7}, {1.6, 2.65}, {2, 3.2}}
```

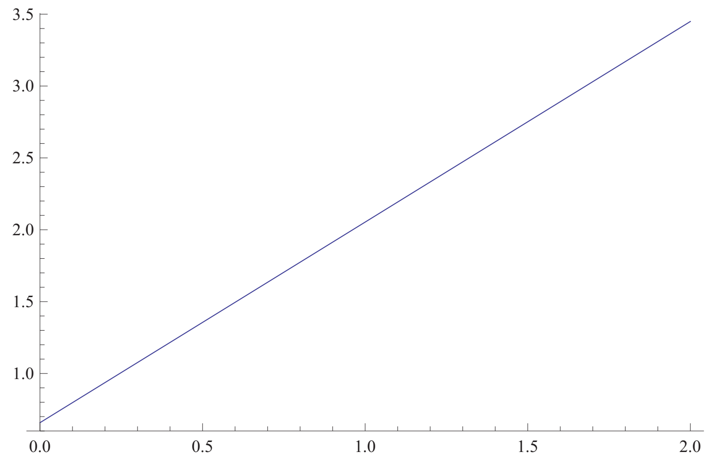
Slika 3.1.

```
p1 = ListPlot[podaci, PlotStyle → PointSize[0.02]]
```



Slika 3.2.

```
Fit[podaci, {1, x}, x]
0.657619 + 1.39571 x
p2 = Plot[%, {x, 0, 2}]
```

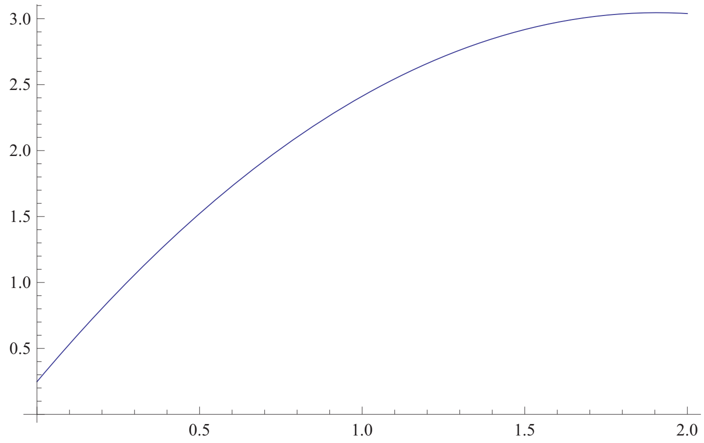


Slika 3.3.

```
Fit[podaci, {1, x, x^2}, x]
```

$0.2475 + 2.93366\,x - 0.768973\,x^2$

```
p3 = Plot[%, {x, 0, 2}]
```

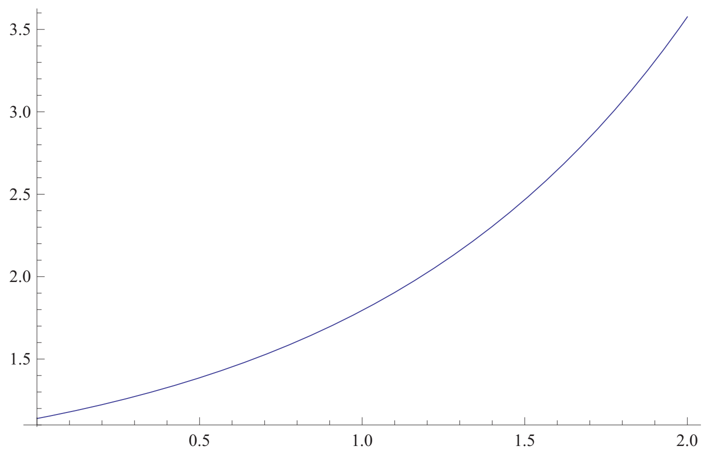


Slika 3.4.

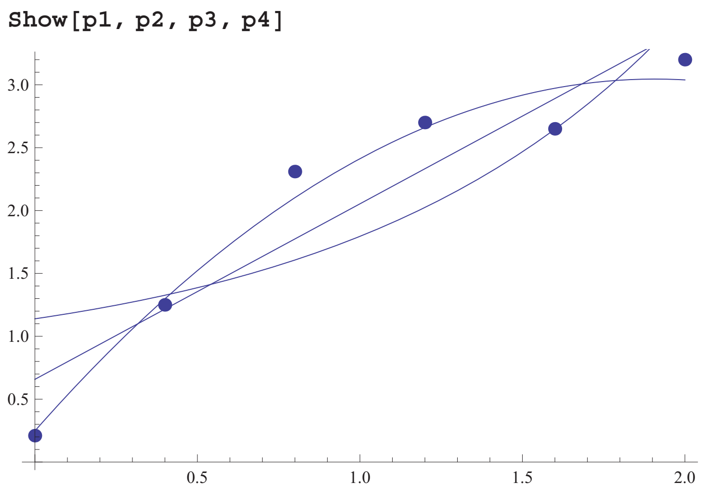
```
Fit[podaci, {1, Exp[x]}, x]
```

$0.757695 + 0.381452\,e^{x}$

```
p4 = Plot[%, {x, 0, 2}]
```



Slika 3.5.



Slika 3.6.

Funkcija φ može i nelinearno ovisiti o parametrima. U tom slučaju često možemo jednostavnim transformacijama problem transformirati u linearni problem najmanjih kvadrata:

(a) Funkcija
$$\varphi(x) = a_0 e^{a_1 x}$$
linearizira se logaritmiranjem
$$\psi(x) = \ln \varphi(x) = \ln a_0 + a_1 x, \quad y_k = \ln f(x_k), \quad k = 0, \dots, n,$$
pa dobivamo linearni problem najmanjih kvadrata
$$\tilde{S} = \tilde{S}(b_0, b_1) = \sum_{k=0}^{n} (y_k - \psi(x_k))^2 = \sum_{k=0}^{n} (y_k - b_0 - b_1 x_k)^2 \to \min,$$
gdje je
$$b_0 = \ln a_0, \quad b_1 = a_1.$$

(b) Funkcija
$$\varphi(x) = a_0 x^{a_1}$$
linearizira se logaritmiranjem
$$\psi(x) = \log \varphi(x) = \log a_0 + a_1 \log x, \quad y_k = \log f(x_k), \quad k = 0, \dots, n.$$
Drugim riječima, dobili smo linearni problem najmanjih kvadrata
$$\tilde{S} = \tilde{S}(b_0, b_1) = \sum_{k=0}^{n} (y_k - b_0 - b_1 \log x_k)^2 \to \min,$$

gdje je

$$b_0 = \log a_0, \quad b_1 = a_1.$$

(c) Funkcija

$$\varphi(x) = \frac{1}{a_0 + a_1 x}$$

linearizira se na sljedeći način

$$\psi(x) = \frac{1}{\varphi(x)} = a_0 + a_1 x, \quad y_k = \frac{1}{f(x_k)}, \quad k = 0, \dots, n.$$

Pripadni linearni problem najmanjih kvadrata je

$$\tilde{S} = \tilde{S}(a_0, a_1) = \sum_{k=0}^{n} (y_k - a_0 - a_1 x_k)^2 \to \min.$$

(d) Funkciju

$$\varphi(x) = \frac{x}{a_0 + a_1 x}$$

možemo linearizirati na više načina. Prvo možemo staviti

$$\psi(x) = \frac{1}{\varphi(x)} = a_0 \frac{1}{x} + a_1, \quad y_k = \frac{1}{f(x_k)}, \quad k = 0, \dots, n.$$

Pripadni linearni problem najmanjih kvadrata je

$$\tilde{S} = \tilde{S}(a_0, a_1) = \sum_{k=0}^{n} (y_k - a_0 \frac{1}{x_k} - a_1)^2 \to \min.$$

Može se koristiti i sljedeći način

$$\psi(x) = \frac{x}{\varphi(x)} = a_0 + a_1 x, \quad y_k = \frac{x_k}{f(x_k)}, \quad k = 0, \dots, n.$$

Pripadni linearni problem najmanjih kvadrata je

$$\tilde{S} = \tilde{S}(a_0, a_1) = \sum_{k=0}^{n} (y_k - a_0 - a_1 x_k)^2 \to \min.$$

(e) Funkcija

$$\varphi(x) = \frac{1}{a_0 + a_1 e^{-x}}$$

linearizira se stavljanjem

$$\psi(x) = \frac{1}{\varphi(x)} = a_0 + a_1 e^{-x}, \quad y_k = \frac{1}{f(x_k)}, \quad k = 0, \dots, n.$$

Pripadni linearni problem najmanjih kvadrata je

$$\tilde{S} = \tilde{S}(a_0, a_1) = \sum_{k=0}^{n} (y_k - a_0 - a_1 e^{-x_k})^2 \to \min.$$

Primjer 3.2. *Odredite vezu oblika* $(x-1)^a = 2(3y+5)^b$ *za podatke*

x_k	1.30	1.35	1.40	1.50	1.58
y_k	5.10	4.00	2.60	1.00	0.33

Rješenje. Linearizacijom dobijemo

$$a\ln(x-1) = \ln 2 + b\ln(3y+5) \;\Rightarrow\; \ln(3y+5) = -\frac{\ln 2}{b} + \frac{a}{b}\ln(x-1),$$

pa ako stavimo

$$\hat{y}_k = \ln(3y_k+5), \quad \hat{x}_k = \ln(x_k-1) \text{ i } a_0 = -\frac{\ln 2}{b},\; a_1 = \frac{a}{b}$$

imamo linerani problem najmanjih kvadrata

$$\tilde{S} = \tilde{S}(a_0, a_1) = \sum_{k=0}^{4}(\hat{y}_k - a_0 - a_1\hat{x}_k)^2 \to \min.$$

Pripadna tablica za $\hat{x}_k$ i $\hat{y}_k$ je

$\hat{x}_k$	−1.204	−1.0498	−0.9163	−0.6931	−0.5447
$\hat{y}_k$	3.0106	2.8332	2.5494	2.0794	1.7901

pa je

$$n = 4, \quad \sum_{k=0}^{4}\hat{x}_k = -4.4079, \quad \sum_{k=0}^{4}\hat{x}_k^2 = 4.1684,$$

$$\sum_{k=0}^{4}\hat{x}_k\hat{y}_k = -11.3514, \quad \sum_{k=0}^{4}\hat{y}_k = 12.2627.$$

Iz (3.1) $a_0 = 0.7647,\; a_1 = -1.9146$ pa je tražena veza

$$(x-1)^{1.7354} = 2(3y+5)^{-0.9064}.$$

3.1.2. Analiza varijance za model regresijskog polinoma drugog stupnja

Oblik:

$$\hat{y} = \varphi(x) = a_0 + a_1x + a_2x^2.$$

Računa se:

$$SP = \sum_{i=0}^{n}(\hat{y}_i - \bar{y})^2 = a_0\sum_{i=0}^{n}y_i + a_1\sum_{i=0}^{n}x_i\cdot y_i + a_2\sum_{i=0}^{n}x_i^2\cdot y_i - (n+1)\bar{y}^2,$$

$$SR = \sum_{i=0}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=0}^{n} y_i^2 - a_0 \sum_{i=0}^{n} y_i - a_1 \sum_{i=0}^{n} x_i \cdot y_i - a_2 \sum_{i=0}^{n} x_i^2 \cdot y_i,$$

$$ST = \sum_{i=0}^{n}(y_i - \bar{y})^2 = \sum_{i=0}^{n} y_i^2 - (n+1)\bar{y}^2.$$

Vrijedi da je $SP+SR = ST$, što predstavlja jednadžbu analize varijance i predstavlja temelj analize reprezentativnosti regresijskog modela.

Standardna greška regresije je apsolutni pokazatelj reprezentativnosti regresijskog modela, a pokazuje prosječni stupanj varijacije stvarnih vrijednosti ovisne varijable u odnosu na očekivane regresijske vrijednosti:

$$\hat{\sigma}_{\hat{y}} = \sqrt{\frac{SR}{n-1}}.$$

Ovaj izraz je standardna greška regresije jednostrukog modela. Izražen je u originalnim jedinicama mjere ovisne varijable y. Stoga je na temelju standardne greške regresije teško uspoređivati reprezentativnost modela s različitim mjernim jedinicama.

Taj problem eliminira relativni pokazatelj koeficijent varijacije regresije, koji predstavlja postotak standardne greške regresije od aritmetičke sredine varijable y:

$$\hat{V}_{\hat{y}} = \frac{\hat{\sigma}_{\hat{y}}}{\bar{y}} \cdot 100.$$

Najmanja vrijednost koeficijenta varijacije je 0%, a najveća nije definirana. Što je koeficijent varijacije regresijskog modela bliži nuli, to je model reprezentativniji. Često se uzima dogovorena granica reprezentativnosti od 10%. Dakle ako je koeficijent varijacije manji od 10% kaže se da je model dobar.

Koeficijent determinacije za model regresijskog polinoma drugog stupnja:

$$r^2 = \frac{SP}{ST}.$$

Korigirani koeficijent determinacije:

$$\bar{r}^2 = 1 - \frac{n}{n-2}(1 - r^2),$$

je asimptotski nepristrana ocjena koeficijenta determinacije.

3.2. Višestruka (multipla) regresija

Kod modela višestruke ili multiple regresije jedna zavisna (ili regresand) varijabla ovisi o $k \geq 2$ nezavisnih (ili regresorskih) varijabli. Grafičko prikazivanje ovdje nije jednostavno jer se radi o tzv. "hiperravninama". Općeniti oblik modela:

$$\hat{y} = \varphi(x) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_kx_k.$$

Ocjenom parametara metodom najmanjih kvadrata traži se minimum zbroja kvadrata empirijskih odstupanja u odnosu na regresijske vrijednosti:

$$S(a_0, a_1, \ldots, a_k) = \sum_{i=0}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=0}^{n}[y_i - (a_0 + a_1x_1 + \cdots + a_kx_k)]^2 = SR,$$

gdje za minimum funkcije f vrijedi da je:

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = \cdots = \frac{\partial S}{\partial a_k} = 0.$$

U matričnom obliku ocjena parametara je $a = (X^TX)^{-1} \cdot (X^Ty)$, gdje je:

$$a = \begin{bmatrix} a_0 \\ a_1 \\ .. \\ a_k \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{10} & x_{20} & .. & x_{k0} \\ 1 & x_{11} & x_{21} & .. & x_{k1} \\ .. & .. & .. & .. & .. \\ 1 & x_{1n} & x_{2n} & .. & x_{kn} \end{bmatrix}, \quad y = \begin{bmatrix} y_0 \\ y_1 \\ .. \\ y_n \end{bmatrix}.$$

Jednadžbe analize varijance višestruke (multiple) regresije:

$$\sum_{i=0}^{n}(\hat{y}_i - \bar{y})^2 \Rightarrow SP = a^T(X^Ty) - (n+1)\bar{y}^2.$$

$$\sum_{i=0}^{n}(y_i - \hat{y}_i)^2 \Rightarrow SR = y^Ty - a^T(X^Ty),$$

$$\sum_{i=0}^{n}(y_i - \bar{y})^2 \Rightarrow ST = y^Ty - (n+1)\bar{y}^2.$$

Interpretacija parametara modela višestruke linearne regresije je: konstantni član, tj. očekivana vrijednost zavisne varijable u originalnim jedinicama mjere kada su sve nezavisne varijable jednake nuli: ($a_0 = \hat{y}$ kada je $x_j = 0$).

Regresijski koeficijenti pokazuju prosječnu promjenu zavisne varijable u originalnim jedinicama mjere kada odgovarajuća nezavisna varijabla poraste za jednu jedinicu, uz uvjet da su sve ostale neovisne varijable neizmijenjene.

Za model s dvije regresorske varijable, tj. za $k = 2$:

$$\hat{y} = \varphi(x) = a_0 + a_1x_1 + a_2x_2,$$

ocjena parametara metodom najmanjih kvadrata računa se sustavom od tri jednadžbe s tri nepoznata parametra: a_0, a_1 i a_2:

$$\begin{aligned}
(n+1)a_0 + a_1\sum_{i=0}^{n} x_{1i} + a_2\sum_{i=0}^{n} x_{2i} &= \sum_{i=0}^{n} y_i \\
a_0\sum_{i=0}^{n} x_{1i} + a_1\sum_{i=0}^{n} x_{1i}^2 + a_2\sum_{i=0}^{n} x_{1i}x_{2i} &= \sum_{i=0}^{n} x_{1i}y_i \\
a_0\sum_{i=0}^{n} x_{2i} + a_1\sum_{i=0}^{n} x_{1i}x_{2i} + a_2\sum_{i=0}^{n} x_{2i}^2 &= \sum_{i=0}^{n} x_{2i}y_i
\end{aligned}$$

Rješenje ovog sustava po parametrima a_j daje rješenje ocjena parametara regresijskog modela s dvije regresorske varijable.

Jednadžbe analize varijance za model s dvije regresorske varijable, tj. za $k = 2$:

$$SP = \sum_{i=0}^{n}(\hat{y}_i - \bar{y})^2 = a_0\sum_{i=0}^{n} y_i + a_1\sum_{i=0}^{n} x_{1i}\cdot y_i + a_2\sum_{i=0}^{n} x_{2i}\cdot y_i - (n+1)\bar{y}^2,$$

$$SR = \sum_{i=0}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=0}^{n} y_i^2 - a_0\sum_{i=0}^{n} y_i - a_1\sum_{i=0}^{n} x_{1i}\cdot y_i - a_2\sum_{i=0}^{n} x_{2i}\cdot y_i,$$

$$ST = \sum_{i=0}^{n}(y_i - \bar{y})^2 = \sum_{i=0}^{n} y_i^2 - (n+1)\bar{y}^2.$$

Vrijedi da je $SP+SR = ST$, što predstavlja jednadžbu analize varijance i predstavlja temelj analize reprezentativnosti regresijskog modela.

Standardna greška regresije je apsolutni pokazatelj reprezentativnosti regresijskog modela, a pokazuje prosječni stupanj varijacije stvarnih vrijednosti ovisne varijable u odnosu na očekivane regresijske vrijednosti:

$$\hat{\sigma}_{\hat{y}} = \sqrt{\frac{SR}{n-1}}.$$

Ovaj izraz je standardna greška regresije modela. Izražen je u originalnim jedinicama mjere ovisne varijable y. Stoga je na temelju standardne greške regresije teško uspoređivati reprezentativnost modela s različitim mjernim jedinicama.

Taj problem eliminira relativni pokazatelj koeficijent varijacije regresije, koji predstavlja postotak standardne greške regresije od aritmetičke sredine varijable y:

$$\hat{V}_{\hat{y}} = \frac{\hat{\sigma}_{\hat{y}}}{\bar{y}}\cdot 100.$$

Najmanja vrijednost koeficijenta varijacije je 0%, a najveća nije definirana. Što je koeficijent varijacije regresijskog modela bliži nuli, to je model reprezentativniji.

Često se uzima dogovorena granica reprezentativnosti od 10%. Dakle ako je koeficijent varijacije manji od 10% kaže se da je model dobar.

Koeficijent multiple determinacije za model višestruke regresije:

$$r^2 = \frac{SP}{ST}.$$

Korigirani koeficijent determinacije:

$$\bar{r}^2 = 1 - \frac{n}{n-k}(1-r^2),$$

je asimptotski nepristrana ocjena koeficijenta determinacije.

3.2.1. Procjena parametara i testiranje hipoteze o značajnosti regresijskog modela

- $(1-\alpha)\cdot 100\%$ pouzdan interval za a_j:

$$\hat{a}_j - t_{\frac{\alpha}{2}}(n-k)\cdot\hat{\sigma}_{\hat{y}}\cdot\sqrt{s_{jj}} \le a_j \le \hat{a}_j + t_{\frac{\alpha}{2}}(n-k)\cdot\hat{\sigma}_{\hat{y}}\cdot\sqrt{s_{jj}},$$

  gdje s_{jj} predstavlja vrijednost odgovarajućeg dijagonalnog elementa u matrici $(X^TX)^{-1}$.
- test-statistike za testiranje sljedećih nul-hipoteza:
  1. $H_0 : a_1 = a_2 = \cdots = a_k = 0,\ H_1 : \exists a_j \neq 0,$

$$F = \frac{SP(n-k)}{SR\cdot k},$$

     a nultu hipotezu odbacujemo ako

$$F \ge f_\alpha(k, n-k).$$

  2. $H_0 : a_j = 0,\ H_1 : a_j \neq 0,$

$$T = \frac{\hat{a}_j}{\hat{\sigma}_{\hat{y}}\cdot\sqrt{s_{jj}}},$$

     a nultu hipotezu odbacujemo ako

$$T \ge t_{\frac{\alpha}{2}}(n-k).$$

- procjena vrijednosti y za zadane vrijednosti neovisnih varijabli:

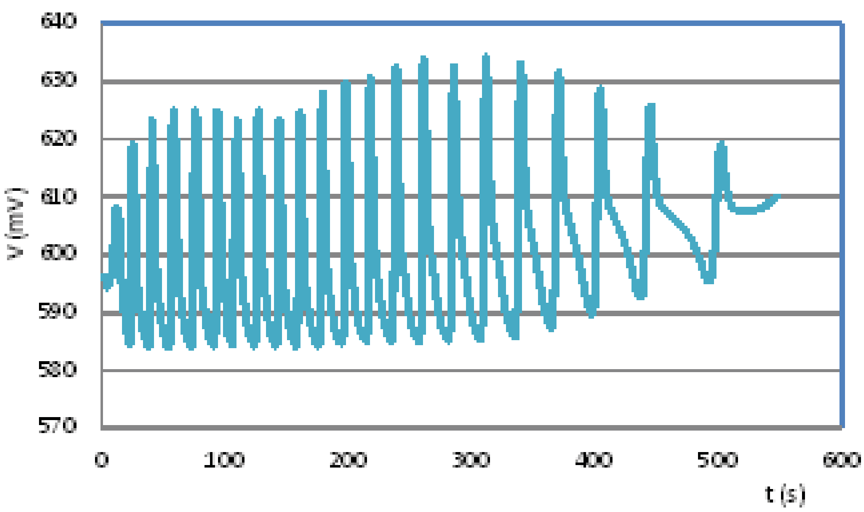
$$\hat{y}_0 = a_0 + a_1x_1^0 + a_2x_2^0 + \cdots + a_kx_k^0.$$

- $(1-\alpha)\cdot 100\%$ pouzdan interval za $\hat{y}_0$:

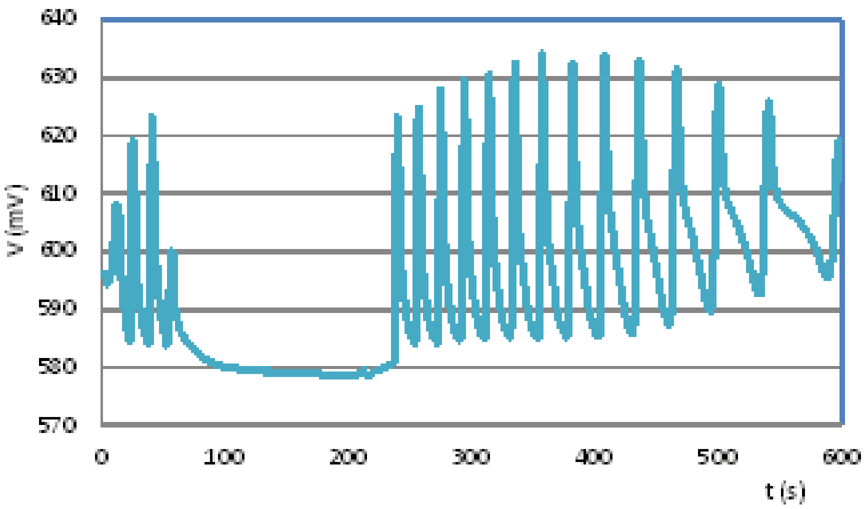
$$\Big[\hat{y}_0 - t_{\frac{\alpha}{2}}(n-k)\cdot\hat{\sigma}_{\hat{y}}\cdot\sqrt{1+x_0(X^TX)^{-1}x_0^T},$$
$$\hat{y}_0 + t_{\frac{\alpha}{2}}(n-k)\cdot\hat{\sigma}_{\hat{y}}\cdot\sqrt{1+x_0(X^TX)^{-1}x_0^T}\Big].$$

3.3. Primjena u struci

Za različita vina (crna i bijela) praćena je antioksidacijska aktivnost Briggs-Rauscher-ovom reakcijom. Svojstvo te reakcije je oscilatornost. Dodatkom uzorka (kao što je u ovom slučaju vino) oscilacije reakcije se prekidaju. Vrijeme prekida oscilacija (inhibition time, IT) reakcija je proporcionalno antioksidacijskom kapacitetu uzorka, odnosno ukupnim fenolima. Ideja je pomoću regresijskog modela, na osnovu podataka o vremenu inhibicije (IT) predvidjeti očekivani sadržaj ukupnih fenola u vinu. (Gajdoš Kljusurić J., Djaković S., Kruhak I., Kovačević Ganić K., Komes D. and Kurtanjek Ž.(2005) Application of Briggs-Rauscher reaction for measurement of antioxidant capacity of Croatian Wines. Acta Alimentaria. 34 (4) 483-492.)



Slika 3.7.Oscilacije Briggs-Rauscherove reakcije



Slika 3.8.Prekid oscilacija uslijed djelovanja antioksidanasa iz bijelog vina, Graševine

Ulazni podaci modela su vrijeme prekida oscilacija BR reakcije (IT) u sekundama i ukupni fenoli izraženi u miligramima po litri kao ekvivalent galne kiseline, te su prikazani u tablici koja sijedi:

IT (s)	Ukupni fenoli (GAE mg/L)
1425.2	2163
1420.2	2148
1428.7	2158
541.2	726
541.8	714
545.1	715
1137.5	1531
1140.1	1527
1143.4	1537
122.1	251
122.9	249
125.1	258
98.9	239
101.0	245
101.2	240
109.0	246
109.2	240
109.9	237

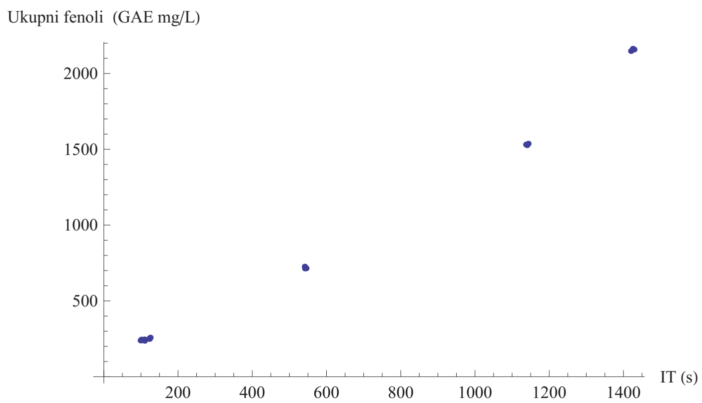
U programskom paketu *Mathematica* postupak unosa podataka iz datoteke "PAP_primjer (radna stranica: primjer 3_3)" koji sadrži podatke iz prethodno navedene tablice je kako slijedi:

```
tablica = Import["F:\\PAP_primjeri.xlsx", {"Sheets", "primjer3_3"}]

{{IT (s), Ukupni fenoli  (GAE mg/L)}, {1425.2, 2163.},
 {1420.2, 2148.}, {1428.7, 2158.}, {541.2, 726.}, {541.8, 714.},
 {545.1, 715.}, {1137.5, 1531.}, {1140.1, 1527.}, {1143.4, 1537.},
 {122.1, 251.}, {122.9, 249.}, {125.1, 258.}, {98.9, 239.},
 {101., 245.}, {101.2, 240.}, {109., 246.}, {109.2, 240.}, {109.9, 237.}}
```

Slika 3.9.

Slijedi grafički prikaz, gdje se na x osi nalaze podaci o vremenu inhibicije IT, a na y-osi vrijednosti izmjerenog sadržaja ukupnih fenola za uzorke vina:

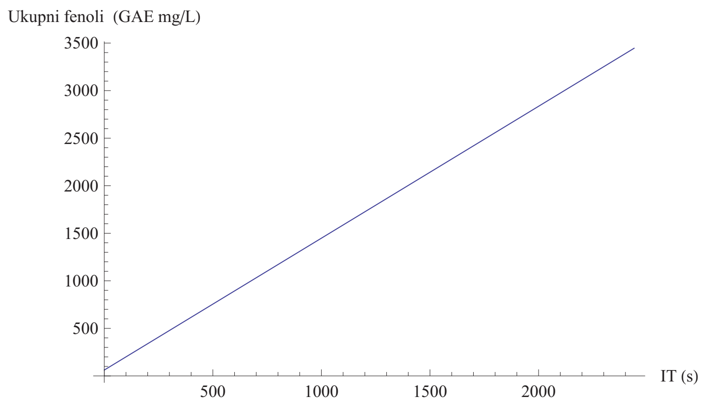
```
g1 = ListPlot[tablica, AxesLabel → tablica[[1]]]
```



Slika 3.10.

Kada se pogledaju podaci na grafu, može se pretpostaviti kako kroz točke možemo provući pravac ili možda polinom. Ako se radi o pravcu imamo:

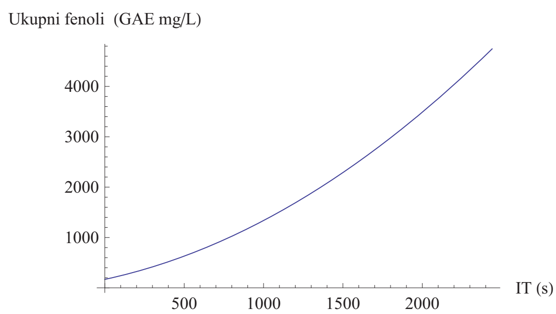
```
podaci = tablica[[2 ;; 19]]
{{1425.2, 2163.}, {1420.2, 2148.}, {1428.7, 2158.}, {541.2, 726.},
 {541.8, 714.}, {545.1, 715.}, {1137.5, 1531.}, {1140.1, 1527.},
 {1143.4, 1537.}, {122.1, 251.}, {122.9, 249.}, {125.1, 258.}, {98.9, 239.},
 {101., 245.}, {101.2, 240.}, {109., 246.}, {109.2, 240.}, {109.9, 237.}}

Fit[podaci, {1, x}, x]
61.5852 + 1.38682 x

g2 = Plot[%, {x, 0, 2440}, AxesLabel → tablica[[1]]]
```



Slika 3.11.

Ako se radi o polinomu (ovdje je izabran polinom drugog stupnja) slijedi:

```
Fit[podaci, {1, x, x^2}, x]
168.508 + 0.678053 x + 0.000490785 x^2

g3 = Plot[%, {x, 0, 2440}, AxesLabel → tablica[[1]]]
```



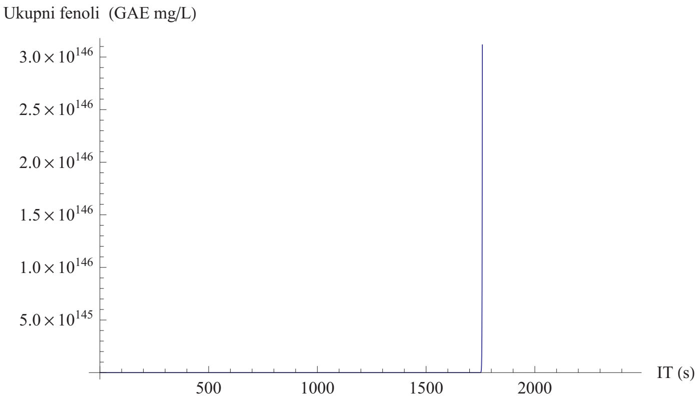
Slika 3.12.

Ako se radi o eksponencijalnoj funkciji slijedi:

```
Fit[podaci, {1, Exp[x]}, x]
```

$775.402 + 4.751439820130531 \times 10^{-618}\, e^{x}$

```
g4 = Plot[%, {x, 0, 2440}, AxesLabel → tablica[[1]]]
```



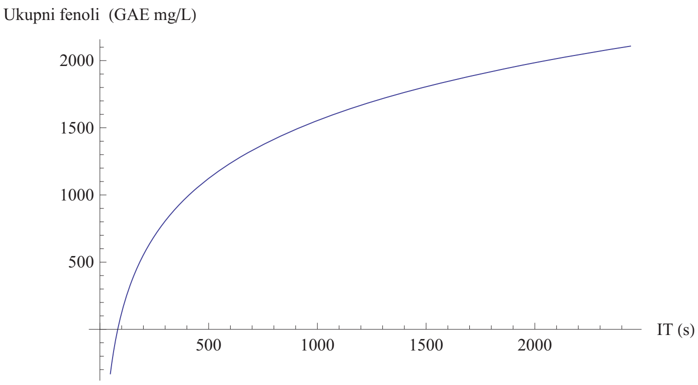
Slika 3.13.

Ako se radi o logaritamskoj funkciji slijedi:

```
Fit[podaci, {1, Log[x]}, x]
```

$-2736.83 + 621.108\, \mathrm{Log}[x]$

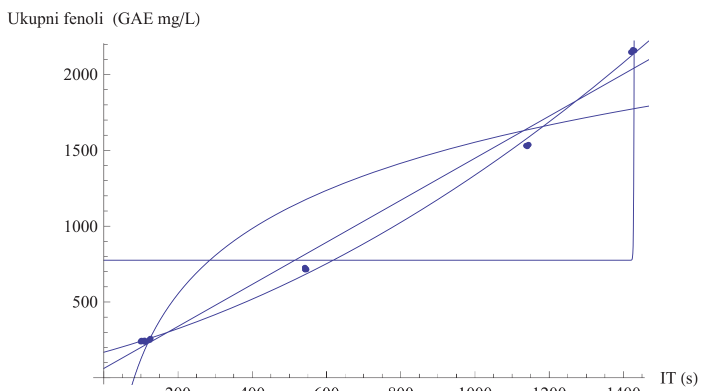
```
g5 = Plot[%, {x, 0, 2440}, AxesLabel → tablica[[1]]]
```



Slika 3.14.

Cilj je izabrati onu krivulju koja je "bliža" eksperimentalnim podacima, odnosno razlike između krivulje i mjerenih podataka su manji, pa je najbolje pogledati obje krivulje i same eksperimentalne podatke na istom grafu:

```
Show[g1, g2, g3, g4, g5]
```



Slika 3.15.

U opisanom primjeru, bolje slaganje eksperimentalnih podataka je sa regresijskom krivuljom polinoma drugog stupnja.

Pogledajmo drugi primjer koji ima jednu zavisnu, a više nezavisnih varijabli: za desertno vino Prošek analitičkim metodama su prikupljeni podaci o masenim koncentracijama flavonoida, neflavonoida i ukupnih fenola te je Briggs-Rauscherovom reakcijom praćena antioksidacijska aktivnost. Cilj je bio utvrditi odnos antioksidacijske aktivnosti i koncentracija flavonoida, neflavonoida i ukupnih fenola za desertno vino Prošek pomoću regresijskih modela parcijalne linearne regresije i viestruke linearne regresije. Utvrđenim odnosom antioksidacijske aktivnosti sa ostalim navedenim parametrima omogućilo bi predikciju očekivane antioksidacijske aktivnosti, ukoliko su poznati podaci o masenim koncentracijama flavonoida, neflavonoida i ukupnih fenola za desertno vino Prošek. (Gajdoš Kljusurić J., Budić-Leto I., Kurtanjek Ž., Lovrić T.(2009) Regresijski modeli predviđanja antioksidacijske aktivnosti desertnog vina Prošek. XXI. Hrvatski skup kemičara i kemijskih inženjera,. str. 124.)

U tablici su navedeni eksperimentalni podaci za desertno vino Prošek:

Ukupni fenoli (UF) GAE mg/L	Flavonoidi (fl) GAE mg/L	Neflavonoidi (nfl) GAE mg/L	Vrijeme inhibicije (IT) s
390	600	790	21.69
414	756	950	235.05
2440	874	1566	861.82
3433	1364	2069	1287.97
3160	1013	2146	1255.92
2914	1272	1642	1033.8
2907	1478	1428	1438.46
3157	1327	1829	692.97
2686	1178	1508	1125.17
2856	1249	1607	978.28
3136	1386	1750	912.64
2185	1156	1600	422.23
2907	1478	1428	2260.75

Ako navedenu tablicu zamislimo kao matricu A, potrebno ju je transponirati kako bi podaci jednog stupca bili zapisani u redovima (kao liste sa kojim se pojednostavljuje daljnji rad u programu *Mathematica*), pa imamo:

```
SetDirectory["C:\\Documents and Settings\\pbf\\Desktop\PAP"];
A = Import["model2a.dat"];
MatrixForm[A];
B = Transpose[A]; MatrixForm[B]
```

$$\begin{pmatrix} 390 & 414 & 2440 & 3433 & 3160 & 2914 & 2907 & 3157 & 2686 & 2856 & 3136 & 2185 \\ 600 & 790 & 874 & 1364 & 1013 & 1272 & 1478 & 1327 & 1178 & 1249 & 1386 & 1156 \\ 756 & 950 & 1566 & 2069 & 2146 & 1642 & 1428 & 1829 & 1508 & 1607 & 1750 & 1600 \\ 21.6938 & 235.05 & 861.825 & 1287.98 & 1255.92 & 1033.8 & 1438.46 & 692.975 & 1125.18 & 978.28 & 912.64 & 422.233 \end{pmatrix}$$

Slika 3.16.

U cilju izdvajanja naslova stupaca od brojeva, tablica se definira kao podaci te se svakoj od 4 liste pridružuje ime varijable; ukupni fenoli (UF); flavonoidi (fl); neflavonoidi (nfl) i vrijeme inhibicije (IT).

Retke matrice B izdvajamo na sljedeći način:

```
uF = B[[1]]; fl = B[[2]]; nfl = B[[3]]; IT = B[[4]];
```

Slika 3.17.

U parcijalnoj regresiji promatraju se zasebno sljedeći parovi:

1. vrijeme inhibicije, IT i ukupni fenoli, uF
2. vrijeme inhibicije, IT i flavonoidi, fl
3. vrijeme inhibicije, IT i neflavonoidi, nfl

Treba naći pripadajući regresijski model za svaki od parova.

```
tablica = Import["F:\\PAP_primjeri.xlsx", {"Sheets", "primjer"}]
{{Ukupni fenoli (GAE mg/L), Flavonoidi (GAE mg/L),
  Neflavonoidi (GAE mg/L), Vrijeme inhibicije (s)}, {390., 600., 790., 21.69},
 {414., 756., 950., 235.05}, {2440., 874., 1566., 861.82},
 {3433., 1364., 2069., 1287.97}, {3160., 1013., 2146., 1255.92},
 {2914., 1272., 1642., 1033.8}, {2907., 1478., 1428., 1438.46},
 {3157., 1327., 1829., 692.97}, {2686., 1178., 1508., 1125.17},
 {2856., 1249., 1607., 978.28}, {3136., 1386., 1750., 912.64},
 {2185., 1156., 1600., 422.23}, {2907., 1478., 1428., 2260.75}}

podaci = tablica[[2 ;; 14]]
{{390., 600., 790., 21.69},
 {414., 756., 950., 235.05}, {2440., 874., 1566., 861.82},
 {3433., 1364., 2069., 1287.97}, {3160., 1013., 2146., 1255.92},
 {2914., 1272., 1642., 1033.8}, {2907., 1478., 1428., 1438.46},
 {3157., 1327., 1829., 692.97}, {2686., 1178., 1508., 1125.17},
 {2856., 1249., 1607., 978.28}, {3136., 1386., 1750., 912.64},
 {2185., 1156., 1600., 422.23}, {2907., 1478., 1428., 2260.75}}

UF = Transpose[podaci][[1]]
{390., 414., 2440., 3433., 3160., 2914.,
 2907., 3157., 2686., 2856., 3136., 2185., 2907.}

fl = Transpose[podaci][[2]]
{600., 756., 874., 1364., 1013., 1272.,
 1478., 1327., 1178., 1249., 1386., 1156., 1478.}
```

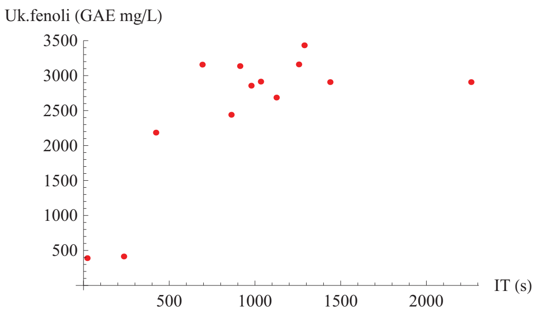
Slika 3.18.

```
nfl = Transpose[podaci][[3]]
{790., 950., 1566., 2069., 2146., 1642.,
 1428., 1829., 1508., 1607., 1750., 1600., 1428.}

IT = Transpose[podaci][[4]]
{21.69, 235.05, 861.82, 1287.97, 1255.92, 1033.8,
 1438.46, 692.97, 1125.17, 978.28, 912.64, 422.23, 2260.75}
```

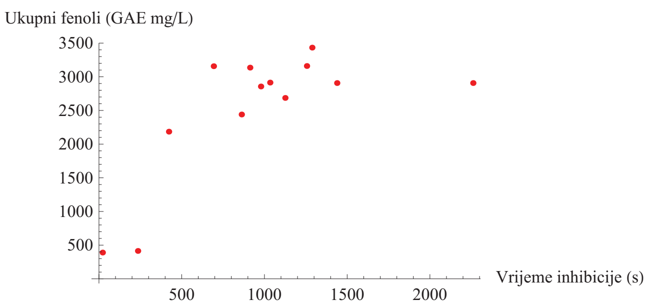
Slika 3.19.

Model I: vrijeme inhibicije, IT i ukupni fenoli, uF:

```
m1 = Table[{IT[[i]], UF[[i]]}, {i, 1, 13}]
{{21.69, 390.}, {235.05, 414.}, {861.82, 2440.},
 {1287.97, 3433.}, {1255.92, 3160.}, {1033.8, 2914.},
 {1438.46, 2907.}, {692.97, 3157.}, {1125.17, 2686.},
 {978.28, 2856.}, {912.64, 3136.}, {422.23, 2185.}, {2260.75, 2907.}}

ListPlot[m1, AxesLabel → {"IT (s)", "Uk.fenoli (GAE mg/L)"}, PlotStyle -> {Red}]
```




```
model1 =
 ListPlot[m1, AxesLabel → {tablica[[1, 4]], tablica[[1, 1]]}, PlotStyle -> {Red}]
```

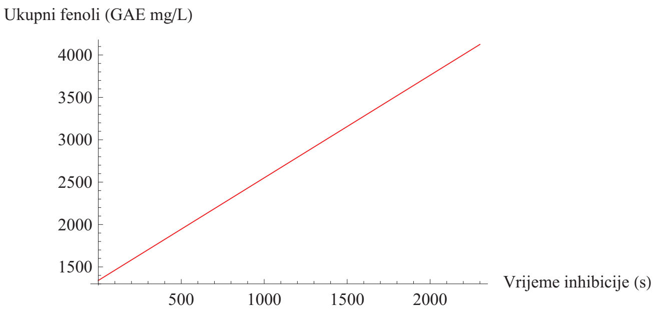


Slika 3.20.

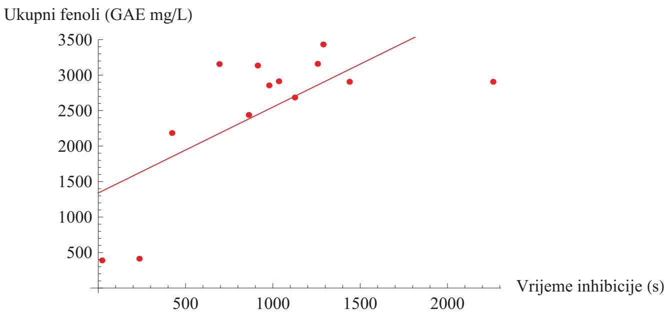
```
Fit[m1, {1, x}, x]
```

```
1339.92 + 1.2107 x
```

```
g1 = Plot[%, {x, 0, 2300},
  AxesLabel → {tablica[[1, 4]], tablica[[1, 1]]}, PlotStyle -> {Red}]
```



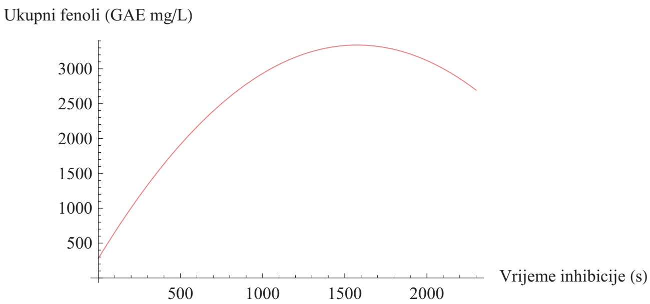

```
Show[model1, g1]
```




```
Fit[m1, {1, x, x^2}, x]
```

$279.373 + 3.88651\,x - 0.00123341\,x^2$

```
g2 = Plot[%, {x, 0, 2300},
  AxesLabel → {tablica[[1, 4]], tablica[[1, 1]]}, PlotStyle → {Pink}]
```

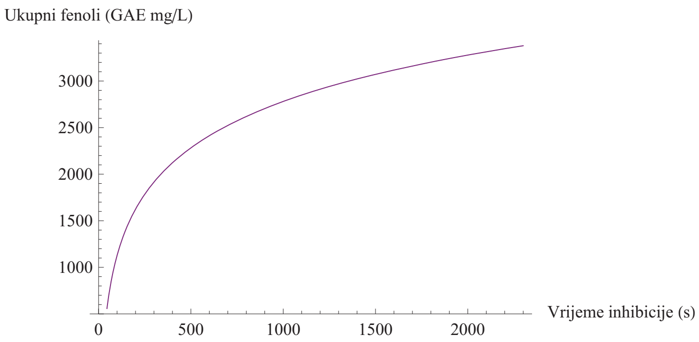


Slika 3.21.

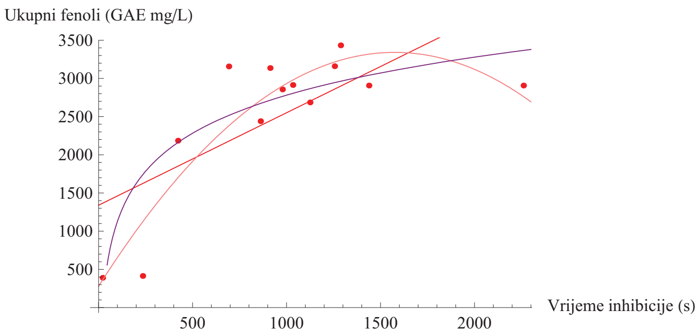
```
Fit[m1, {1, Log[x]}, x]
```

```
-2182.47 + 718.476 Log[x]
```

```
g3 = Plot[%, {x, 0, 2300},
  AxesLabel → {tablica[[1, 4]], tablica[[1, 1]]}, PlotStyle -> {Purple}]
```




```
Show[model1, g1, g2, g3]
```



Slika 3.22.

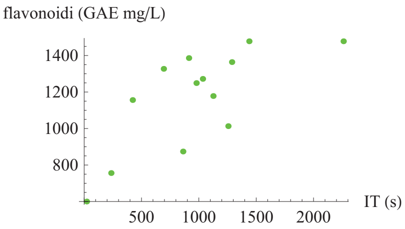
Model II: vrijeme inhibicije, IT i flavonoidi, fl:

```
m2 = Table[{IT[[i]], fl[[i]]}, {i, 1, 13}]
```

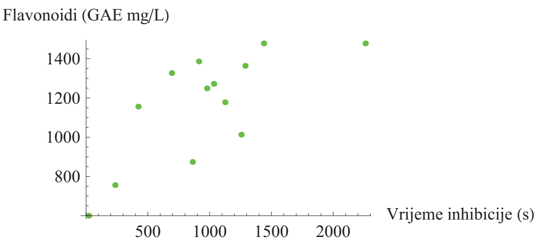
```
{{21.69, 600.}, {235.05, 756.}, {861.82, 874.}, {1287.97, 1364.}, {1255.92, 1013.},
 {1033.8, 1272.}, {1438.46, 1478.}, {692.97, 1327.}, {1125.17, 1178.},
 {978.28, 1249.}, {912.64, 1386.}, {422.23, 1156.}, {2260.75, 1478.}}
```

Slika 3.23.

```
ListPlot[m2,
 AxesLabel → {"IT (s)", "flavonoidi (GAE mg/L)"}, PlotStyle -> {Green}]
```



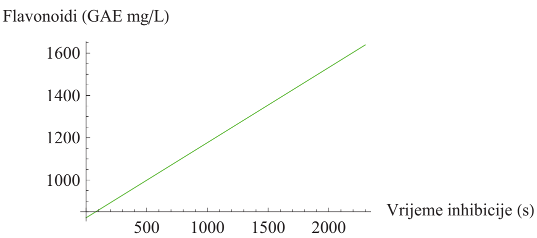

```
model2 =
 ListPlot[m2, AxesLabel → {tablica[[1, 4]], tablica[[1, 2]]}, PlotStyle -> {Green}]
```



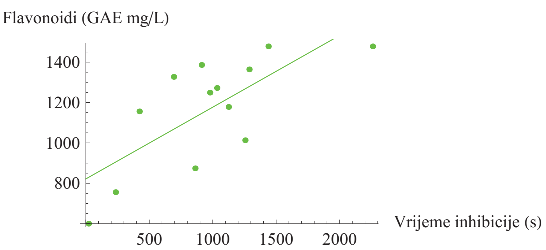

```
Fit[m2, {1, x}, x]
```

```
821.61 + 0.355246 x
```

```
g4 = Plot[%, {x, 0, 2300},
  AxesLabel → {tablica[[1, 4]], tablica[[1, 2]]}, PlotStyle -> {Green}]
```



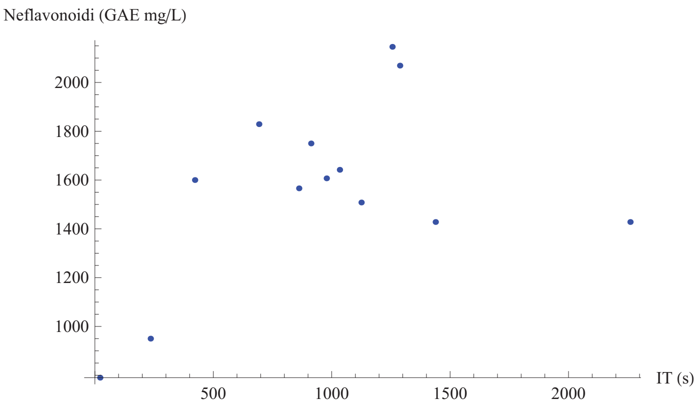

```
Show[model2, g4]
```



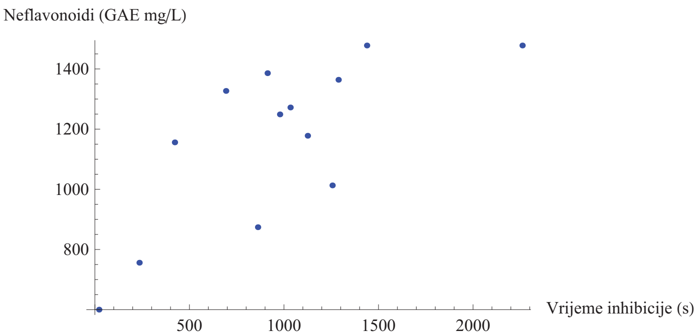
Slika 3.24.

Model III: vrijeme inhibicije, IT i neflavonoidi, nfl

```
m3 = Table[{IT[[i]], nfl[[i]]}, {i, 1, 13}]
{{21.69, 790.}, {235.05, 950.}, {861.82, 1566.},
 {1287.97, 2069.}, {1255.92, 2146.}, {1033.8, 1642.},
 {1438.46, 1428.}, {692.97, 1829.}, {1125.17, 1508.},
 {978.28, 1607.}, {912.64, 1750.}, {422.23, 1600.}, {2260.75, 1428.}}

ListPlot[m3,
 AxesLabel → {"IT (s)", "Neflavonoidi (GAE mg/L)"}, PlotStyle -> {Blue}]
```




```
model3 =
 ListPlot[m2, AxesLabel → {tablica[[1, 4]], tablica[[1, 3]]}, PlotStyle -> {Blue}]
```

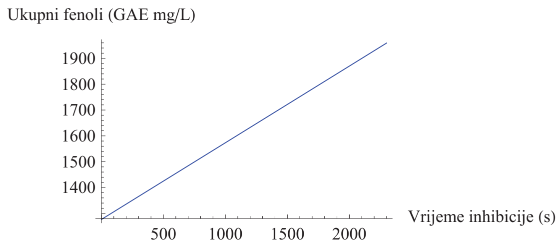


Slika 3.25.

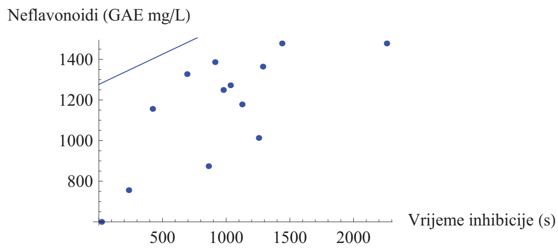
```
Fit[m3, {1, x}, x]
```

1276.57 + 0.296767 x

```
g5 = Plot[%, {x, 0, 2300},
  AxesLabel → {tablica[[1, 4]], tablica[[1, 1]]}, PlotStyle -> {Blue}]
```




```
Show[model3, g5]
```



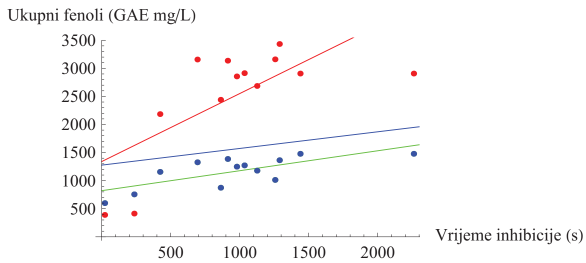
Slika 3.26.

Za svaki promatrani model, može se izračunati regresijska jednadžba:

	Model I	Model II	Model III
Naredba	**Fit**[p1,{1,x},x]	**Fit**[p2,{1,x},x]	**Fit**[p3,{1,x},x]
Regresijski model	1339.92 + 1.2107 x	821.61 + 0.355246 x	1276.57 + 0.296767 x

Jednostavnije je promotrati regresijske modele i podatke na zajedničkom grafu:

```
Show[model1, g1, model2, g4, model3, g5]
```



Slika 3.27.

Ukoliko se želi ispitati kako bi se procijenila antioksidacijska aktivnost (odnosno u ovom slučaju vrijeme inhibicije, IT, koje je proporcionalno antioksidacijskoj aktivnosti) ukoliko se kao nezavisne varijable promatraju sve ostale varijable, tj. masene koncentracije ukupnih fenola, flavonoida i neflavonoida, postavka bi bila kako slijedi:

```
podaci = Table[{uF[[i]], fl[[i]], nfl[[i]], IT[[i]]}, {i, 1, 12}]

{{390, 600, 756, 21.6938}, {414, 790, 950, 235.05}, {2440, 874, 1566, 861.825},
 {3433, 1364, 2069, 1287.98}, {3160, 1013, 2146, 1255.92}, {2914, 1272, 1642, 1033.8},
 {2907, 1478, 1428, 1438.46}, {3157, 1327, 1829, 692.975}, {2686, 1178, 1508, 1125.18},
 {2856, 1249, 1607, 978.28}, {3136, 1386, 1750, 912.64}, {2185, 1156, 1600, 422.233}}
```

Slika 3.28.

Navedena relacija podrazumijeva da će se kao izlazna veličina računati vrijeme zadrške oscilacijske reakcije (inhibicijsko vrijeme, IT), a x = uk.fenoli, $x1$ = flavonoidi, $x2$ = neflavonoidi.

```
tablica = Import["F:\\PAP_primjeri.xlsx", {"Sheets", "primjer"}]

{{Ukupni fenoli (GAE mg/L), Flavonoidi (GAE mg/L),
  Neflavonoidi (GAE mg/L), Vrijeme inhibicije (s)}, {390., 600., 790., 21.69},
 {414., 756., 950., 235.05}, {2440., 874., 1566., 861.82},
 {3433., 1364., 2069., 1287.97}, {3160., 1013., 2146., 1255.92},
 {2914., 1272., 1642., 1033.8}, {2907., 1478., 1428., 1438.46},
 {3157., 1327., 1829., 692.97}, {2686., 1178., 1508., 1125.17},
 {2856., 1249., 1607., 978.28}, {3136., 1386., 1750., 912.64},
 {2185., 1156., 1600., 422.23}, {2907., 1478., 1428., 2260.75}}

podaci = tablica[[2 ;; 14]]

{{390., 600., 790., 21.69},
 {414., 756., 950., 235.05}, {2440., 874., 1566., 861.82},
 {3433., 1364., 2069., 1287.97}, {3160., 1013., 2146., 1255.92},
 {2914., 1272., 1642., 1033.8}, {2907., 1478., 1428., 1438.46},
 {3157., 1327., 1829., 692.97}, {2686., 1178., 1508., 1125.17},
 {2856., 1249., 1607., 978.28}, {3136., 1386., 1750., 912.64},
 {2185., 1156., 1600., 422.23}, {2907., 1478., 1428., 2260.75}}
```

Slika 3.29.

```
UF = Transpose[podaci][[1]]
{390., 414., 2440., 3433., 3160., 2914.,
 2907., 3157., 2686., 2856., 3136., 2185., 2907.}

fl = Transpose[podaci][[2]]
{600., 756., 874., 1364., 1013., 1272.,
 1478., 1327., 1178., 1249., 1386., 1156., 1478.}

nfl = Transpose[podaci][[3]]
{790., 950., 1566., 2069., 2146., 1642.,
 1428., 1829., 1508., 1607., 1750., 1600., 1428.}

IT = Transpose[podaci][[4]]
{21.69, 235.05, 861.82, 1287.97, 1255.92, 1033.8,
 1438.46, 692.97, 1125.17, 978.28, 912.64, 422.23, 2260.75}

m1 = Table[{IT[[i]], UF[[i]]}, {i, 1, 13}]
{{21.69, 390.}, {235.05, 414.}, {861.82, 2440.},
 {1287.97, 3433.}, {1255.92, 3160.}, {1033.8, 2914.},
 {1438.46, 2907.}, {692.97, 3157.}, {1125.17, 2686.},
 {978.28, 2856.}, {912.64, 3136.}, {422.23, 2185.}, {2260.75, 2907.}}

m2 = Table[{IT[[i]], fl[[i]]}, {i, 1, 13}]
{{21.69, 600.}, {235.05, 756.}, {861.82, 874.}, {1287.97, 1364.}, {1255.92, 1013.},
 {1033.8, 1272.}, {1438.46, 1478.}, {692.97, 1327.}, {1125.17, 1178.},
 {978.28, 1249.}, {912.64, 1386.}, {422.23, 1156.}, {2260.75, 1478.}}

m3 = Table[{IT[[i]], nfl[[i]]}, {i, 1, 13}]
{{21.69, 790.}, {235.05, 950.}, {861.82, 1566.},
 {1287.97, 2069.}, {1255.92, 2146.}, {1033.8, 1642.},
 {1438.46, 1428.}, {692.97, 1829.}, {1125.17, 1508.},
 {978.28, 1607.}, {912.64, 1750.}, {422.23, 1600.}, {2260.75, 1428.}}

model = LinearModelFit[m1, x, x]
FittedModel[ 1339.92+1.2107 x ]

model["ParameterTable"]
```

	Estimate	Standard Error	t-Statistic	P-Value
1	1339.92	411.887	3.25311	0.0076925[illegible]
x	1.2107	0.371294	3.26076	0.0075888

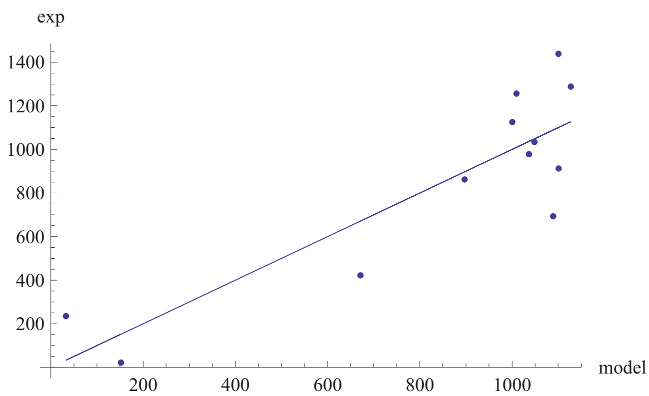
Slika 3.30.

```
model = LinearModelFit[podaci, {x1, x2, x3}, {x1, x2, x3}]

FittedModel[ 639.498 + 0.74403 x1 + 0.162436 x2 - 1.10711 x3 ]

model["ParameterTable"]
```

	Estimate	Standard Error	t-Statistic	P-Value
1	639.498	1142.09	0.559934	0.589187
x1	0.74403	0.498839	1.49152	0.170018
x2	0.162436	1.00015	0.162412	0.874569
x3	−1.10711	0.877375	−1.26185	0.238724

Slika 3.31.

Kako bi se provjerila potencijalna primjerenost modela u predikciji antioksidacijske aktivnosti koja je u ovom primjeru izražena kako vrijeme inhibicije, mogu se na istom grafu opisati eksperimentalno dobivene vrijednosti i vrijednosti dobivene za IT računom iz modela. Te bi tako slijedilo:



Slika 3.32.

Iz navedenog se vidi kako model čije su ulazne veličine sve promatrane masene koncentracije (ukupnih fenola, flavonoida i neflavonoida) nije baš primjeren za predikciju očekivane antioksidacijske aktivnosti pomoću vremena inhibicije. Navode potvrđuje i višekriterijski model sa svojim članovima uz nezavisne varijable, koji su negativni za koncentracije flavonoida i neflavonoida, a pozitivan je jedino za ukupne fenole. Takav rezultat ukazuje da bi parcijalna linearna regresija bila prihvatljivija metoda.

3.4. Faktorska analiza

U znanstvenim disciplinama gdje je dominantan strukturalni pristup organizacije podataka, posebna važnost se pridaje multivarijatnim metodama jer će mjerenjima dati jasniji smisao. Strukturalni ili holistički pristup proučavanju nekog problema očituje se u tome da pojedini proučavani elementi u kombinaciji s nekim drugim mogu imati različito značenje za strukturu (cjelinu) pa tako i različiti doprinos njenom objašnjavanju.

UNIVARIJATNI PRISTUP sastoji se u analizi varijabiliteta u jednoj varijabli (obilježju). U ovom pristupu svaka se varijabla proučava nezavisno od drugih varijabli.

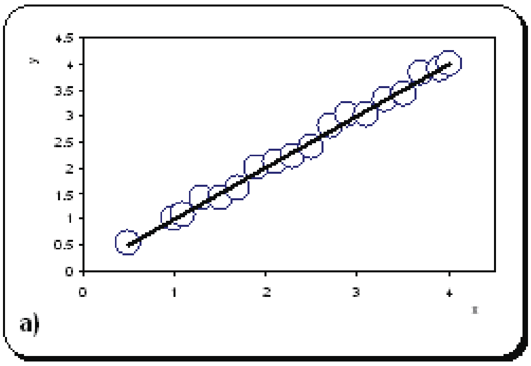
BIVARIJATNI PRISTUP proučava, analizira međusobni odnos dviju varijabli. Za razliku od univarijantnog, ovdje je bit u utvrđivanju kovarijabiliteta (kovarijance) koja je osnovni pojam u znanosti. A razlog leži u činjenici kako je jedan od osnovnih ciljeva znanosti utvrđivanje povezanosti između pojava bilo da se radi o kauzalnom odnosu ili samo korelaciji.

MULTIVARIJATNI PRISTUP proučava, analizira međusobni odnos više od dvije varijable istovremeno. Takav pristup je posebno važan kod problema koji zahtijevaju ispitivanje većeg broja obilježja da bi se moglo odgovoriti na taj problem istraživanja.

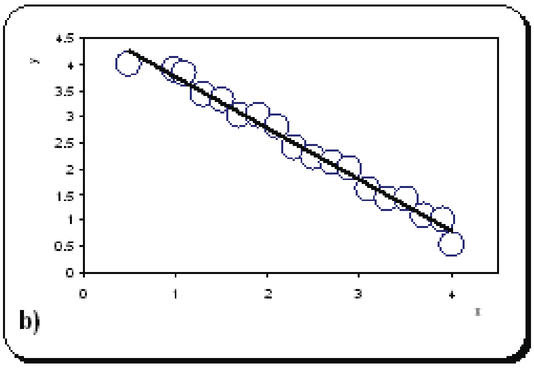
Primjer 3.3. • *U medicini terapija uključuje cijelu osobu, a ne samo organ*

- *U pedagogiji odgoj je integralan usmjeren na cijeli - ukupan razvoj djeteta*
- *U kineziologiji - U zdravom tijelu zdrav duh*
- *U dijetoterapiji - povezivanje načina prehrane i bolesti*
- *U kontroli kvalitete - međpovezivanje većeg broja analiza (npr. senzorike i analitike istog uzorka)*

Međuodnos varijabli utvrđuje se korelacijom sa kojom započinje multivarijantna analiza. Povezanost između dvije varijable može biti različitog stupnja i može biti pozitivna i negativna. Stupanj povezanosti se iskazuje koeficijentom korelacije a koji se kreću u rasponu od -1 preko 0 do $+1$.



Slika 3.33.Primjer za pozitivnu korelaciju



Slika 3.34.Primjer za negativnu korelaciju

Faktorska analiza je jedna od metoda višedimenzionalne analize. Jedna od zadaća faktorske analize jest sažimanje većega broja međusobno povezanih izvornih varijabli u manji broj zajedničkih faktora koji će ih opisivati i objasniti njihovu međusobnu povezanost. Polazi se od pretpostavke da među varijablama postoji linearna korelacija, a svaki izlučeni faktor nije u korelaciji sa drugim faktorima. Tako se umjesto nad velikim brojem koreliranih izvornih varijabli, analiza provodi na nekoreliranim faktorima i na taj se način otklanja problem kolinearnosti varijabli.

Potreba za sažimanjem velikoga broja karakteristika nekog proizvoda na manji broj vidi se pri ispitivanju osnovnih karakteristika. Primjerice, ako se želi saznati koje su osnovne karakteristike nekog proizvoda, dolazi se do saznanja da kod ispitanika postoji velik broj izjava kojima se objašnjava određeni proizvod. Osim toga, u tom velikom broju izjava o karakteristikama proizvoda neke se izjave i preklapaju. Stoga se postavlja pitanje jesu li te izjave doista različite ili bi se mogle grupirati u određene skupine? Iz navedenoga proizlazi primjena faktorske analize koja omogućuje da se taj velik broj karakteristika nekog proizvoda reducira na manji broj osnovnih karakteristika koje će ga opisati, odnosno da se veći broj izvornih varijabli reducira na manji broj faktora.

Jedna je od osnovnih pretpostavki za primjenu faktorske analize ta da su podaci mjereni na intervalnoj skali. No u empirijskim se istraživanjima često javlja problem mjerenja varijabli, na primjer različite izjave ispitanika, njihova mišljenja, stavovi i slično. U tim se slučajvima od ispitanika traži da na skali od recimo 1 do 7 procjene svaku navedenu izjavu, pri čemu bi 1 označivalo "uopće se ne slažem s navedenom izjavom", a 7 "izrazito se slažem s navedenom izjavom". Iako je ta skala ordinalna, ona se može smatrati i intervalnom ako se polazi od pretpostavke da su intervali na skali jednaki. U empirijskim istraživanjima često se primjenjuju ordinalne skale, ali je uobičajeno da se prikupljeni podaci analiziraju kao da su prikupljeni na intervalnoj skali.

Faktorska analiza provodi se u više koraka:

a) procjena prikladnosti podataka za primjenu faktorske analize,

b) utvrđivanje inicijalnih rezultata za izlučivanje faktora,

c) određivanje matrice faktorske strukture i završnih rezultata nakon izlučivanja faktora,

d) provođenje rotacije faktora ako inicijalna matrica faktorske strukture nije interpretabilna, ili ako ne udovoljava postavljenom kriteriju jednostavne strukture,

e) utvrđivanje faktorskih matrica i završnih rezultata nakon rotacije faktora i

f) interpretacija izlučenih faktora nakon rotacije.

Korelacijska matrica koja sadrži koeficijente jednostavne linearne korelacije svakog para varijabli osnovica je za provođenje faktorske analize. Jedan od preduvjeta provođenja faktorske analize jest povezanost između izvornih varijabli, a osnova za uočavanje skupina povezanih varijabli korelacijska je matrica.

Kaiser-Meyer-Olkinova mjera je kriterij kojim se može ispitati prikladnost podataka za primjenu faktorske analize. Vrijednost Kaiser-Meyer-Olkinove mjere (K) izračunava se uz pomoć sljedeće formule:

$$K = \frac{\sum_{i=1,i\neq k}^{p} \sum_{k=1}^{p} r_{ik}^2}{\sum_{i=1,i\neq k}^{p} \sum_{k=1}^{p} r_{ik}^2 + \sum_{i=1,i\neq k}^{p} \sum_{k=1}^{p} q_{ik}^2}, \quad 0 \leq K \leq 1,$$

pri čemu je $r_{ik}^2, (i \neq k)$ kvadrat izvandijagonalnog elementa korelacijske matrice (kvadrat koeficijenta korelacije između i-te i k-te varijable), a $q_{ik}^2, (i \neq k)$ kvadrat izvandijagonalnog elementa antiimidž korelacijske matrice (kvadrat koeficijenta parcijalne korelacije između i-te i k-te varijable). Parcijalna korelacija od varijabli v_1 i v_2 uz isključivanje utjecaja varijable v_3 računa se s formulom $r_{12,3} = \frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$.

Kaiser-Meyer-Olkinova mjera kreće se u zatvorenom intervalu od 0 do 1. Ako je vrijednost te mjere manja od 0.5, korelacijska matrica nije prikladna za faktorsku analizu. Osim što se vrijednost Kaiser-Meyer-Olkinove mjere može izračunati za cijelu matricu, može se izračunati i za pojedine varijable. Kaiser-Meyer-Olkinova mjera za svaku pojedinu varijablu (k_i) računa se uz pomoć sljedećega izraza:

$$k_i = \frac{\sum_{i=1,i\neq k}^{p} r_{ik}^2}{\sum_{i=1,i\neq k}^{p} r_{ik}^2 + \sum_{i=1,i\neq k}^{p} q_{ik}^2}, \quad 0 \leq k_i \leq 1.$$

Na taj se način može ispitati prikladnost svake pojedine varijable u analizi i mogu se isključiti varijable koje nemaju dovoljno veliku vrijednost. Time se povećava vrijednost Kaiser-Meyer-Olkinove mjere cijele matrice.

Pri izlučivanju faktora promatraju se njihove svojstvene vrijednosti. U faktorskoj analizi svojstvena vrijednost određenog faktora jednaka je zbroju kvadrata faktorskih opterećenja po svim varijablama za taj faktor. Za faktorsku analizu glavnih komponenti karakteristično je da se izlučuju faktori kojih su svojstvene vrijednosti veće od jedan.

U primjeni faktorske analize postotak varijance svakog pojedinoga faktora izračunava se na osnovi svojstvene vrijednosti toga faktora, odnosno kao omjer svojstvene vrijednosti i zbroja svojstvenih vrijednosti pomnožen sa sto:

$$p_j = \frac{\lambda_j}{p} \lambda \cdot 100, j = 1, 2, \dots, m.$$

Nakon što je utvrđen broj faktora, potrebno je odrediti matricu faktorske strukture izlučenih faktora. Matrica faktorske strukture sadrži faktorska opterećenja koja predočuju koeficijente korelacije između izlučenih faktora i varijabli. Faktorska opterećenja ukazuju na važnost svake varijable za pojedini faktor.

Budući da inicijalna matrica nema obilježja jednostavne strukture, provodi se rotacija faktora kojom se mijenja odnos između varijabli i faktora. Rotaciju faktora moguće je provesti uz pomoć jedne od metoda ortogonalne rotacije (gdje su faktorske osi zadržane pod pravim kutom) ili jedne od metoda kosokutne rotacije (gdje kutovi među osima mogu biti različiti od 90 stupnjeva.

Matrica faktorske strukture i matrica faktorskog sklopa nakon provedene rotacije osnovica su za interpretaciju faktora. Matrica faktorske strukture i matrica faktorskog sklopa jednake su ako su faktori ortogonalni. Te su matrice različite ako su faktori u međusobnoj korelaciji, odnosno nakon provedene kosokutne rotacije. Kod kosokutne rotacije se još, pored navedenih matrica, javlja i matrica korelacija među faktorima.

Struktura faktorskih opterećenja nakon provedene rotacije omogućuje bolju interpretaciju faktora u odnosu na inicijalnu faktorsku matricu.

Primjer 3.4. *Utvrđivanje osnovnih karakteristika proizvoda "kava" primjenom faktorske analize. Podaci su prikupljeni metodom ispitivanja na uzorku studenta prve godine Ekonomskog fakulteta. Instrument istraživanja - upitnik (na skali od* 1 *do* 7 *ispitanici* *su davali odgovore, pri čemu je jedan označavalo "uopće se ne slažem s navedenom izjavom", a 7 "izrazito se slažem s navedenom izjavom"). Za prikaz metodologije korišteno je 10 rezultata ankete.*

Popis i pojašnjenje promatranih 15 varijabli:

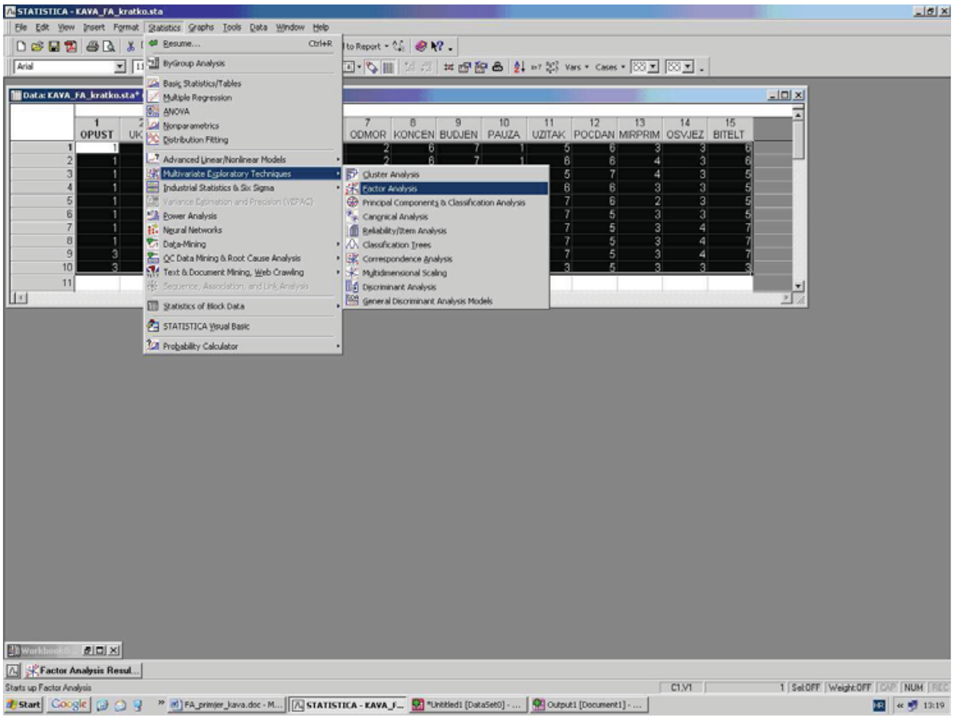
OPUST – *kava je dobra za opuštanje,*
UKUS – *važno je da kava ima dobar ukus,*
NESPA – *kada popijem kavu, ne spava mi se,*
DOBOS – *dobro se osjećam dok pijem kavu,*
NEZIV – *ne mogu živjeti bez kave,*
MIRIS – *važno je da kava ima dobar miris,*
ODMOR – *kava je važna za odmor,*
KONCEN – *kavu pijem kad želim biti koncentriran/a na učenje,*
BUDJEN – *kava je važna za razbuđivanje,*
PAUZA – *kava znači pauzu,*
UZITAK – *kava mi pruža užitak,*
POCDAN – *kava je važna za početak dana,*
MIRPRIM – *miris kave je primamljiv,*
OSVEZ – *kava je osvježenje,*
BITELT – *kava mi je bitan element u životu.*

Rezultati ankete u tablici programa Statistica v.7.1.:

	1 OPUST	2 UKUS	3 NESPA	4 DOBOSJ	5 NEZIV	6 MIRIS	7 ODMOR	8 KONCEN	9 BUDJEN	10 PAUZA	11 UZITAK	12 POCDAN	13 MIRPRIM	14 OSVJEZ	15 BITELT
1	1	1	6	1	5	3	2	6	7	1	6	6	3	3	6
2	1	1	6	1	6	3	2	6	7	1	6	6	4	3	6
3	1	1	7	1	5	2	2	7	7	1	6	7	4	3	5
4	1	1	6	1	6	3	2	6	6	1	6	6	3	3	5
5	1	1	6	1	7	3	2	6	5	1	6	6	2	3	5
6	1	1	5	1	7	3	1	5	5	1	6	5	3	3	5
7	1	2	5	2	7	3	2	5	5	1	6	5	3	4	7
8	1	2	5	2	7	3	2	5	5	1	7	5	3	4	7
9	3	3	5	2	7	3	3	5	5	3	7	5	3	4	7
10	3	4	5	2	3	3	4	5	6	3	3	5	3	3	3

Slika 3.35.

Sljedeći korak je izbor faktorske analize kao jedne od multivarijantne analize:



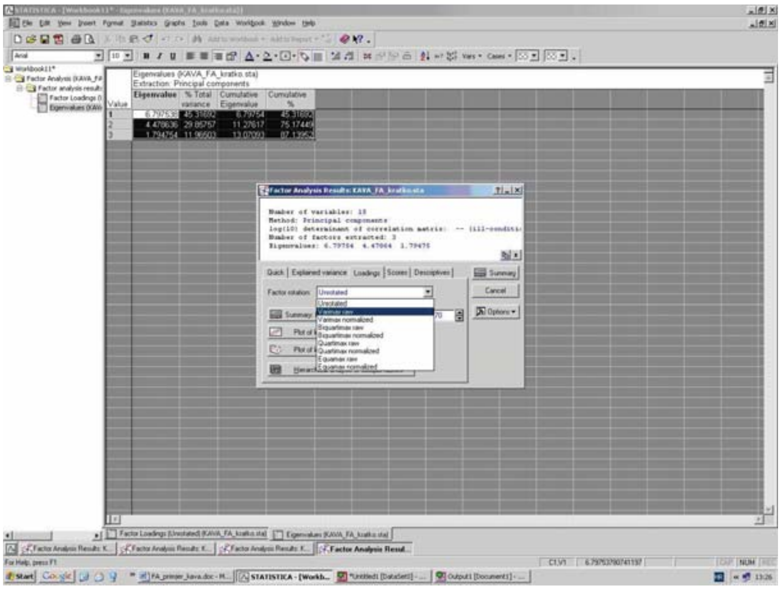
Slika 3.36.

Kao što je pojašnjeno u teorijskom dijelu, prvi korak je izračun korelacija, te dobivamo matricu faktorske strukture:

Variable	Factor Loadings (Unrotated) (KAVA_FA_kratko.sta Extraction: Principal components (Marked loadings are >.700000) Factor 1	Factor 2	Factor 3
OPUST	-0.72455	0.54933	0.18965
UKUS	-0.86042	0.46383	0.18963
NESPA	0.91236	0.24648	0.17393
DOBOSJ	-0.88187	0.00425	0.36448
NEZIV	-0.05298	-0.94737	0.01635
MIRIS	-0.58053	-0.29586	-0.51824
ODMOR	-0.57952	0.72278	0.20108
KONCEN	0.91236	0.24648	0.17393
BUDJEN	0.66368	0.54579	0.24227
PAUZA	-0.72455	0.54933	0.18965
UZITAK	0.20302	-0.83844	0.37606
POCDAN	0.91236	0.24648	0.17393
MIRPRIM	0.40459	0.26068	0.63327
OSVJEZ	-0.63101	-0.52862	0.52896
BITELT	-0.11425	-0.78860	0.52250
Expl.Var	6.79753	4.47863	1.79475
Prp.Totl	0.45316	0.29857	0.11965

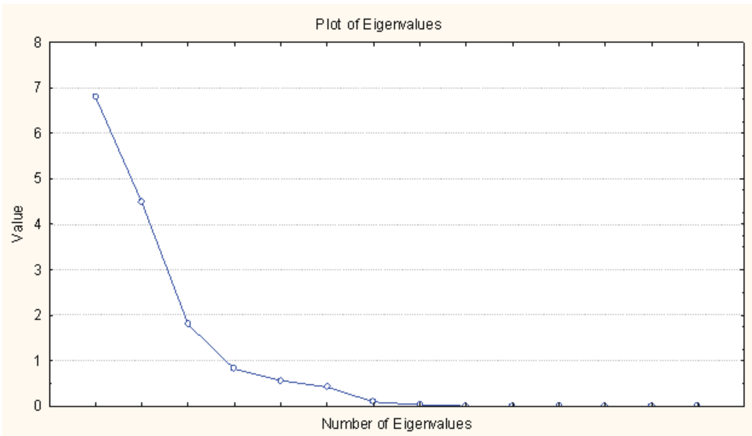
Slika 3.37.

Daljnji postupak nalaže rotaciju matrice (Verimax):



Slika 3.38.

Slijedi pregled uz pomoć Cattellijevog dijagrama (scree plot) koji potvrđuje kako je broj faktora koji nakon rotacije opisuju varijable jednak 3 *(sve točke na grafu čija y vrijednost je veća od* 1*)*



Slika 3.39.

Izgled matrice raspodjele varijabli prema faktorima je sada slijedeći:

Variable	Factor Loadings (Varimax raw) (PAP_primjeri) Extraction: Principal components (Marked loadings are >,700000) Factor 1	Factor 2	Factor 3
opust	0.168662	-0.903383	0.134789
ukus	0.296515	-0.949255	0.049220
nespa	-0.833592	0.422972	0.222806
dobosj	0.387921	-0.773886	-0.401489
neziv	0.402305	0.501072	-0.698338
miris	0.831688	-0.016559	0.033961
odmor	-0.007273	-0.906536	0.277213
koncen	-0.833592	0.422972	0.222806
budjen	-0.821579	0.047801	0.346121
pauza	0.168662	-0.903383	0.134789
uzitak	-0.035142	0.465230	-0.817291
pocdan	-0.833592	0.422972	0.222806
mirprim	-0.765569	-0.135024	-0.168400
osvjez	0.326356	-0.361400	-0.848710
bitel	0.076022	0.152533	-0.937503
Expl.Var	4.608141	5.132887	3.329900
Prp.Totl	0.307209	0.342192	0.221993

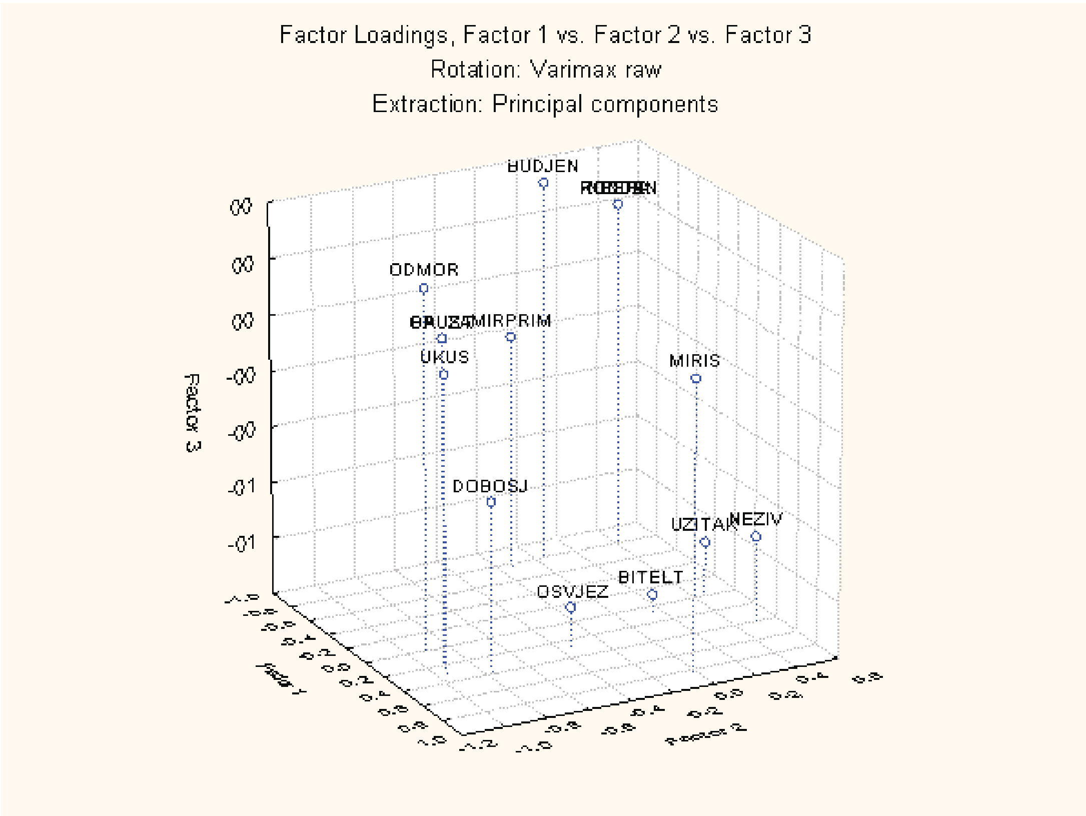
Slika 3.40.

Te se analizom postotka varijance faktora vidi raspodjela varijance prema faktorima:

Value	Eigenvalues (PAP_primjeri) Extraction: Principal components Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	6.797538	45.31692	6.79754	45.31692
2	4.478636	29.85757	11.27617	75.17449
3	1.794754	11.96503	13.07093	87.13952

Slika 3.41.

Prikaže li se navedena analiza grafom, slijedi:



Slika 3.42.

Zaključak: Faktorska analiza je omogućila grupiranje velikog broja praćenih podataka (15 varijabli) u 3 glavne skupine kojima se naziv faktora dodjeljuje prema osnovnim varijablama koje ga čine. Tako prema gore navedenom grafu ili tablici, mogu se grupirati sljedeće varijable:

Faktor 1: NESPA; MIRIS; KONCEN; BUDJEN; POCETDAN; MIRPRIM

Faktor 2: OPUST; UKUS; DOBOSJ; ODMOR; PAUZA

Faktor 3: UZITAK; OSVJEZ; BITELT

Prvo što se mora uočiti je da varijabla "NEZIV" nije našla svoje mjesto niti u jednom faktoru, te se može zaključiti kako tu varijablu nije nužno uzimati u daljnju obradu (dakle ujedno je došlo do ekstrakcije jedne varijable - NEZIV).

Slijedi korak pridruživanja naziva skupovima (tj. faktorima). Tako bi npr. nazivi za navedene faktore mogli biti sljedeći:

Faktor 1: "Koncentracija i ovisnost"

Faktor 2: "Opuštanje"

Faktor 3: "Osvježenje i užitak"

3.5. Složena (dvofaktorska) analiza varijance

Složena analiza varijance je statistički postupak koji se primjenjuje u onim slučajevima kada uspoređujemo rezultate više zavisnih varijabli unutar više nezavisnih varijabli koje imaju različite razine ili kategorije. Za razliku od jednostavne analize varijance, gdje imamo samo jednu zavisnu varijablu, kod složene analize varijance, osim utjecaja većeg broja nezavisnih varijabli na zavisnu, gledamo i međudjelovanje (interakciju) između većeg broja zavisnih varijabli.

U slučaju kada se podaci mogu klasificirati duž dvije različite dimenzije (faktora), složena (dvofaktorska) analiza varijance daje nam odgovor na 3 pitanja:

1. Postoji li statistički značajan utjecaj prvog faktora na zavisnu varijablu (pritom zanemarujući mogući utjecaj drugog faktora)?
2. Postoji li statistički značajan utjevaj drugog faktora na zavisnu varijablu (pritom zanemarujući utjecaj prvog faktora)?
3. Postoji li statistički značajna interakcija između promatrana dva faktora, tj. utječu li "istovremene" promjene u oba faktora statistički značajno na promjene zavisne varijable?

Promatra se utjecaj dvaju faktora, pri čemu prvi faktor ima m_1, a drugi faktor m_2 razina. Dakle, promatramo $m_1 \cdot m_2$ populacija. Pretpostavimo da iz svake populacije uzimamo nezavisne slučajne uzorke jednake duljine l , svaki za obilježje X reprezentirano sa $X_{ij} \sim N(\mu_{ij}, \sigma^2)$ u populaciji ij , za $i = 1, \ldots m_1,\ j = 1, \ldots m_2$.

Stavimo: $\mu = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mu_{ij}$,

$$\mu_i' = \frac{1}{m_2} \sum_{j=1}^{m_2} \mu_{ij},\ \delta_i' = \mu_i' - \mu,\ i = 1, \ldots m_1,\ \mu_j'' = \frac{1}{m_1} \sum_{i=1}^{m_1} \mu_{ij},\ \delta_j'' = \mu_j'' - \mu,\ j = 1, \ldots m_2,$$

$$\delta_{ij} = \mu_{ij} - \mu - \delta_i' - \delta_j'',\ i = 1, \ldots m_1,\ j = 1, \ldots m_2,$$

gdje je μ opća srednja vrijednost, δ_i' glavni efekt i-te razine prvog faktora, δ_j'' glavni efekt j-te razine drugog faktora i δ_{ij} interakcijski efekt i-te razine prvog i j-te razine drugog faktora.

Testiramo sljedeće hipoteze:

$H_{0,1} : \delta_1' = \delta_2' = \cdots = \delta_{m_1}' = 0$, tj. prvi faktor je beznačajan,

$H_{0,2} : \delta_1'' = \delta_2'' = \cdots = \delta_{m_2}'' = 0$, tj. drugi faktor je beznačajan i

$H_{0,12} : \delta_{ij} = 0$, tj. nema interakcije medu faktorima.

Računamo:

$$\bar{x}_{ij} = \frac{1}{l}\sum_{k=1}^{l} x_{ijk},\ S_{ij}^2 = \frac{1}{l-1}\sum_{k=1}^{l}(x_{ijk} - \bar{x}_{ij})^2,\ i = 1, \ldots m_1,\ j = 1, \ldots m_2,$$

$$\bar{x}'_i = \frac{1}{m_2 l}\sum_{j=1}^{m_2}\sum_{k=1}^{l} x_{ijk} = \frac{1}{m_2}\sum_{j=1}^{m_2} \bar{x}_{ij},\ i = 1, \ldots m_1,$$

$$\bar{x}''_j = \frac{1}{m_1 l}\sum_{i=1}^{m_1}\sum_{k=1}^{l} x_{ijk} = \frac{1}{m_1}\sum_{j=1}^{m_1} \bar{x}_{ij},\ j = 1, \ldots m_2,$$

$$\bar{x} = \frac{1}{m_1 m_2 l}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{l} x_{ijk} = \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \bar{x}_{ij} = \frac{1}{m_1}\sum_{i=1}^{m_1} \bar{x}'_i = \frac{1}{m_2}\sum_{j=1}^{m_2} \bar{x}''_j,$$

$$SST_1 = m_2 l\sum_{i=1}^{m_1}(\bar{x}'_i - \bar{x})^2,\ SST_2 = m_1 l\sum_{j=1}^{m_2}(\bar{x}''_j - \bar{x})^2,\ SST_{12} = l\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(\bar{x}_{ij} - \bar{x}'_i - \bar{x}''_j + \bar{x})^2,$$

$$SSE = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{l}(x_{ijk} - \bar{x}_{ij})^2 = (l-1)\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} S_{ij}^2,$$

$$MST_1 = \frac{SST_1}{m_1 - 1},\ MST_2 = \frac{SST_2}{m_2 - 1},\ MST_{12} = \frac{SST_{12}}{(m_1 - 1)(m_2 - 1)},$$

$$MSE = \frac{SSE}{m_1 m_2 (l-1)}.$$

Testovi i pripadne testne statistike:

$$H_{0,1} : \delta'_i = 0,\ \forall i$$
$$H_{1,1} : \exists i\ \text{ td. }\ \delta'_i \neq 0$$

$$F_1 = \frac{MST_1}{MSE} \overset{H_{0,1}}{\sim} F(df_1, df_2),\ df_1 = m_1 - 1,\ df_2 = m_1 m_2 (l-1),$$

$$H_{0,2} : \delta''_j = 0,\ \forall j$$
$$H_{1,2} : \exists j\ \text{ td. }\ \delta''_j \neq 0$$

$$F_2 = \frac{MST_2}{MSE} \overset{H_{0,2}}{\sim} F(df_1, df_2),\ df_1 = m_2 - 1,\ df_2 = m_1 m_2 (l-1),$$

$$H_{0,12} : \delta_{ij} = 0,\ \forall i, j$$
$$H_{1,12} : \exists i, j\ \text{ td. }\ \delta_{i,j} \neq 0$$

$$F_{12} = \frac{MST_{12}}{MSE} \overset{H_{0,12}}{\sim} F(df_1, df_2),\ df_1 = (m_1 - 1)(m_2 - 1),\ df_2 = m_1 m_2 (l-1).$$

ANOVA tablica:

izvor rasipanja	stupnjevi slobode	suma kvadrata	srednjekvadratno odstupanje	vrijednost test-statistike
prvi faktor	$m_1 - 1$	SST_1	MST_1	F_1
drugi faktor	$m_2 - 1$	SST_2	MST_2	F_2
interakcija	$(m_1 - 1)(m_2 - 1)$	SST_{12}	MST_{12}	F_{12}
zbog greške	$m_1 m_2 (l - 1)$	SSE	MSE	
$\sum$	$m_1 m_2 l - 1$	SS		

gdje je $SS = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \sum_{k=1}^{l} (x_{ijk} - \bar{x})^2$.

Primjer 3.5. *U tablici su zadani podaci o utjecaju konzumiranja čokolade na broj bodova na testu iz matematike* 16 *ispitanika. Ispitanici su na temelju dviju dimenzija (spol: muški i ženski; količina konzumirane čokolade sat vremena prije pisanja testa) podijeljeni u* 4 *jednakobrojne grupe. Odgovorite na pitanja:*

a) Postoji li statistički značajan utjecaj konzumirane čokolade na sposobnost rješavanja matematičkih zadataka (zanemarujući spol)?

b) Postoji li statistički značajan utjecaj spola na sposobnost rješavanja matematičkih zadataka (zanemarujući konzumiranu čokoladu)?

c) Postoji li statistički značajna interakcija između spola i količine konzumirane čokolade pri utjecaju na sposobnost rješavanja matematičkih zadataka? Obrazložite odgovor.

	M	*Ž*
100g čokolade	*40*	*45*
	38	*42*
	35	*47*
	42	*41*
0g čokolade	*39*	*30*
	40	*25*
	38	*24*
	45	*20*

Rješenje. U navedenom se eksperimentu promatra utjecaj dvaju faktora. Prvi je količina konzumirane čokolade, a drugi spol. Dakle, $m_1 = 2$ jer imamo 100g i 0g čokolade, a $m_2 = 2$ (muškarci i žene). Za svaku populaciju $m_1 \cdot m_2 = 4$ uzimamo nezavisni slučajan uzorak duljine $l = 4$. Testiramo sljedeće hipoteze:

$H_{0,1}$: ne postoji statistički značajan utjecaj konzumirane čoklade na sposobnost rješavanja matematičkih zadataka,

$H_{0,2}$: ne postoji statistički značajan utjecaj spola na sposobnost rješavanja matematičkih zadataka i

$H_{0,12}$: ne postoji statistički značajna interakcija izmedu spola i količine konzumirane čokolade pri utjecaju na sposobnost rješavanja matematičkih zadataka.

Prvo računomo srednje vrijednosti:

$$\bar{x}_{1,1} = \bar{x}_{100g,M} = 38.75,\ \ \bar{x}_{1,2} = \bar{x}_{100g,Z} = 43.75,\ \ \bar{x}'_1 = \bar{x}'_{100g} = 41.25,$$

$$\bar{x}_{2,1} = \bar{x}_{0g,M} = 40.5,\ \ \bar{x}_{2,2} = \bar{x}_{0g,Z} = 24.75,\ \ \bar{x}'_2 = \bar{x}'_{0g} = 32.625,$$

$$\bar{x}''_1 = \bar{x}''_M = 39.625,\ \ \bar{x}''_2 = \bar{x}''_Z = 34.25,\ \ \bar{x} = 36.9375.$$

ANOVA tablica sada izgleda ovako:

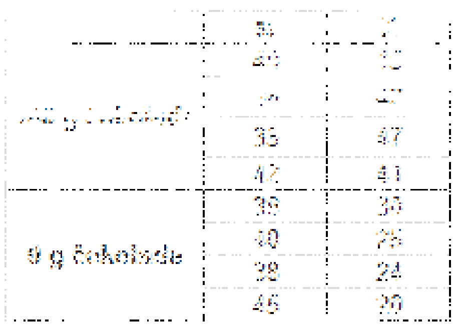
izvor rasipanja	stupnjevi slobode	suma kvadrata	srednjekvadratno odstupanje	vrijednost test-statistike
prvi faktor	1	297.5625	297.5625	27.62669
drugi faktor	1	115.5625	115.5625	10.72921
interakcija	1	430.5625	430.5625	39.97486
zbog greške	12	129.25	10.77083	
$\sum$	15	972.9375		

Za $\alpha = 0.05$ iz tablice za F-razdiobu potrebno je očitati:

$$f_{0.05}(1, 12) = 4,7472.$$

Kako je $F_1 > f_{0.05}(1, 12)$, $F_2 > f_{0.05}(1, 12)$ i $F_{12} > f_{0.05}(1, 12)$ vidimo da su vrijednosti sve tri test-statistike upale u kritično područje što znači da hipoteze $H_{0,1}, H_{0,2}, H_{0,12}$ moramo odbaciti.

Primjer ćemo rješiti i uz pomoć MS Excel-a: Unosimo podatke iz tablice:



Slika 3.43.

Kao i kod jednostavne analize varijance, u složenoj (dvofaktorskoj) analizi varijance koristimo odgovarajuću proceduru iz *Data Analysis*. Kliknemo na ***Tools*** i u padajućem izborniku biramo opciju ***Data Analysis***.

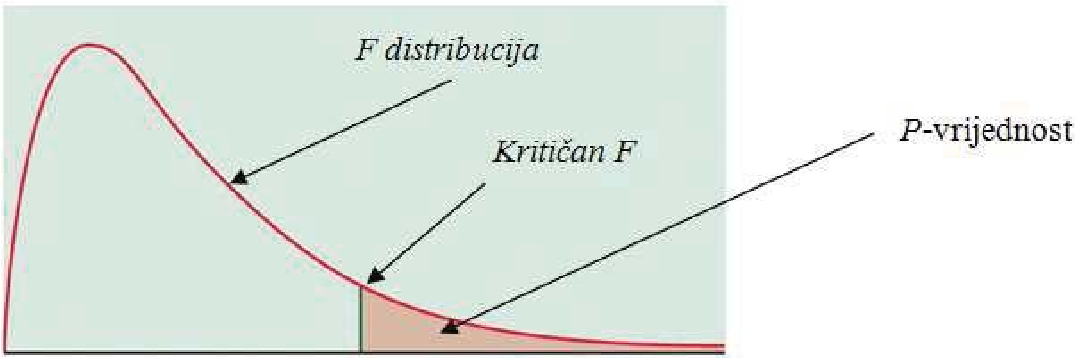
U dobivenom okviru označimo proceduru ***Anova: Two-Factor With Replication*** i kliknemo OK.

Ispunjavamo dijaloški okvir *Anova: Two-Factor With Replication*. U ***Input Range*** upisujemo blok ćelija u kojem se nalaze kvantitativni podaci koje analiziramo zajedno s naslovnim ćelijama u kojima su upisane kategorije (modaliteti) dva promatrana faktora. Pod ***Rows per sample*** upisujemo broj redaka u kojima se nalaze podaci u svakom od uzoraka koje analiziramo. U našem slučaju to je **4**. ***Alpha*** predstavlja razinu značajnosti (rizika) i u ovom ćemo primjeru mi uzeti da je ona **0.05**. U drugom dijelu dijaloškog okvira (*Output options*) određujemo gdje želimo da nam program izbaci konačnu output-tablicu.

Anova: Two-Factor With Replication						
SUMMARY	M	Ž	Total			
Count	4	4	8			
Sum	155	175	330			
Average	38,75	43,75	41,25			
Variance	8,916667	7,583333	14,21429			
Count	4	4	8			
Sum	162	99	261			
Average	40,5	24,75	32,625			
Variance	9,666667	16,91667	82,26786			
Total						
Count	8	8				
Sum	317	274				
Average	39,625	34,25				
Variance	8,839286	113,6429				
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	297,5625	1	297,5625	27,62669	0,000202	4,747225
Columns	115,5625	1	115,5625	10,72921	0,006635	4,747225
Interaction	430,5625	1	430,5625	39,97485	3,82E-05	4,747225
Within	129,25	12	10,77083			
Total	972,9375	15				

Slika 3.44.

Odgovore na postavljena pitanja omogućuju nam dobivene p-vrijednosti u drugom dijelu output-tablice - tablici *ANOVA*.



Slika 3.45.

a) Kao što možemo očitati iz tablice p-vrijednost vezana uz faktor konzumiranja čokolade (pod *Sample*) iznosi 0.000202. Dakle *postoji statistički značajan utjecaj konzumirane čoklade na sposobnost rješavanja matematičkih zadataka (zanemarujući spol).* Iz deskriptivnih podataka vidimo da ispitanici koji su sat vremena prije testa pojeli 100g čokolade ($\bar{x}'_{100g} = 41.25$) postižu bolji rezultat na testu iz matematike od ispitanika koji nisu pojeli čokoladu prije pisanja testa ($\bar{x}'_{0g} = 32.625$).

b) Pripadajuća p-vrijednost faktora spola (pod *Columns*) je 0.006635, *što znači da postoji statistički značajan utjecaj spola na sposobnost rješavanja matematičkih zadataka (zanemarujući konzumiranu čokoladu).* Ispitanici muškog spola ($\bar{x}''_M = 39.625$) postižu bolje rezultate na testu iz matematike od ispitanika ženskog spola ($\bar{x}''_Z = 34.25$).

c) P-vrijednost vezana uz interakciju (pod *Interaction*) iznosi $0.382 \cdot 10^{-4}$. *Zaključujemo da postoji statistički značajna interakcija između spola i količine konzumirane čokolade pri utjecaju na sposobnost rješavanja matematičkih zadataka.* Ispitanici ženskog spola koji su konzumirali 100g čokolade sat vremena prije pisanja testa ($\bar{x}_{100g,Z} = 43.75$) bolje su napisali test od muškaraca koji su pojeli istu količinu čokolade ($\bar{x}_{100g,M} = 38.75$), dok su žene koje nisu konzumirale čokoladu ($\bar{x}_{0g,Z} = 24.75$) znatno gore napisale test iz matematike od muških ispitanika koji isto nisu konzumirali čokoladu sat vremena prije pisanja testa ($\bar{x}_{0g,M} = 40.50$).