LINEAARNE REGRESSIOONANALÜÜS

Korrelatsioonanalüüs - seose suund ja rangus
Regressioonanalüüs - seose kuju
Kasutatakse eelkõige prognoosimiseks.

<u>Seose kuju</u> - tendents, mis avaldub uuritava e. sõltuva tunnuse Y väärtuse muutumises seoses tegur- e. sõltumatu tunnuse X väärtuse muutumisega;

Lihtne regressioonanalüüs

- **sirgjooneline (lineaarne),**
- **kõverjooneline.**

Tendentsi iseloomustavat joont nimetatakse *regressioonijooneks.*
Regressioonijoon väljendab uuritava tunnuse Y sellist muutumist, mis esineks siis, kui teiste tegurite X mõju oleks kinnistatud ühele ja samale keskmisele tasemele kõigi vaatluste jaoks ning uuritav tunnus Y oleks tegurtunnuse X funktsioon.

Regressioonijoont kirjeldatakse võrrandiga

$$Y'' = B_0 + B_1X, \text{ kus}$$

B_0, B_1- võrrandi parameetrid,
B_1 - X proportsionaalsuse koefitsient. Näitab, mitme ühiku võrra muutub Y väärtus, kui X väärtus muutub ühe ühiku võrra;
B_0 - regressioonijoone algpunkt koordinaatide süsteemis.
X ja Y empiiriliste väärtuste alusel otsitakse Y teoreetilisi väärtusi sõltuvalt X väärtustest.

Näide.

5 maakonna haiglate kohta teada andmed:

Y - haiglas viibimise kestus (päeva),

X1 - olemasolevaid voodikohti (tuhat),

X2 - hõivatud voodikohti (tuhat).

Maakond	Y	X1	X2
1	9,0	26	21
2	9,3	28	24
3	8,2	14	11
4	9,8	30	25
5	10,7	31	26

Prognoosida haiglas viibimise kestus, kui

X_1 = 24 tuhat, X_2 = 20 tuhat; p = 0,05

1. Regressioonivõrrandi leidmine

$Y'' = B_0 + B_1X_1 + B_2X_2$

B_0 = 6,406; B_1 = 0,086; B_2 = 0,036

$Y'' = 6,406 + 0,086*X_1 + 0,036*X_2$

2. Prognoosimine

Y” = 6,406 + 0,086*24 + 0,036*20 = 9,2 päeva

3. Prognoosi usalduspiiride arvutamine

$SE_{Y.X1X2} = \sigma_Y \sqrt{1- R^2_{Y.X1X2}}$

σ_Y = 0,93; $R^2_{Y.X1X2}$ = 0,543

$SE_{Y.X1X2}$ = 0,93 $\sqrt{}$ 1- 0,543 = 0,629

$Y'' \pm SE* t_{0,05}$

68,26% 9,2 ± 0,629*1,0 = 9,2 ± 0,629

95% 9,2 ± 0,629*1,96 = 9,2 ± 1,23

99% 9,2 ± 0,629*2,54 = 9,2 ± 1,60

$B_1 = r_{YX} * \sigma_Y / \sigma_X$

$B_0 = M_Y - B_1 M_X$

⇒***Milline sirge on parim?***

Vähimruutude meetod - vertikaalhälvete ruutude summa $\Sigma(Y_i - Y_i'')^2$ -> min

⇒***Kui hästi regressioonivõrrand kirjeldab populatsiooni?***

Eeldused:

1) Y jaotuse normaalsus, dispersioonide võrdsus;

2) Y vaatlusandmed üksteisest sõltumatud;

3) Lineaarsus. Populatsiooni keskväärtused $\mu_{Y/Xi}$ asuvad populatsiooni regressioonisirgel.

$Y_i = \beta_0 + \beta_1 X_i + e_i$, kus

$e_i = Y_i - \mu_{Y/Xi}$,

β_0, β_1 - populatsiooni parameetrid.

B_0, B_1 - arvutatakse vähimruutude meetodil, eri valimites väärtused erinevad.

Kui eeldused täidetud, B_0 ja B_1 väärtuste jaotus normaalne keskväärtustega β_0 ja β_1.

$SE_{B0} = \sigma_e \sqrt{1/N + (M_X)^2/[(N-1)\sigma_X^2]}$, kus

σ_e - populatsiooni hälvete e_i standardhälve,

σ_X^2 - tunnuse X dispersioon.

$SE_{B1} = \sigma_e / \sqrt{(N-1)\sigma_X^2}$

σ_e^2 = ?

$S^2 = \Sigma_i [Y_i - (B_0 + B_1 X_i)]^2/(N-2)$, kus

S^2- hälvete e_i dispersioon valimi andmete alusel.

⇒ ***Olulisuse kontroll***

H_0 : Y ja X vahel pole olulist lineaarset seost.
H_1 : Y ja X vahel on oluline lineaarne seos.

Kontroll:

$B_1 = 0$ $t = B_1/S_{B1}$,
$B_0 = 0$ $t = B_0/S_{B0}$; df = N - 2; (p = 0,05)

kui t_{ARV} olulisus $\leq$ p, siis hülgame H_0 ja võtame vastu H_1.

- populatsiooni parameetrite β_0, β_1 usalduspiirid

β_1 - $B_1 \pm SE_{B1}*t_{0,05}$
β_0 - $B_0 \pm SE_{B2}*t_{0,05}$
df = N-2

Näide.

	B	SE_B	95%
B_1	1,909	0,047	1,816...2,003
B_0	771,282	355,472	72,779...1469,785

⇒ ***Mudeli headuse hindamine***

R^2 - determinatsioonikoefitsient, näitab, kui suur osa (%) Y varieeruvusest on kirjeldatud mudeliga Y" (R^2 = 0.775 => 77,5%).

R^2 = 1, kui tegelikud väärtused regressiooni-sirgel.

Tolerantsus (= 1 - R^2) näitab, kui suur osa Y varieeruvusest jääb kirjeldamata.

H_0: R^2=0, s.t. sõltuva tunnuse Y ja sõltumatu(te) tunnus(t)e X vahel pole olulist lineaarset seost (kui hästi mudel kirjeldab populatsiooni)

$R_a^2 = R^2 - P(1- R^2)/(N-P-1)$, kus
R_a^2 - korrigeeritud R^2 populatsiooni jaoks,
P - sõltumatute tunnuste arv.

Y varieeruvus: $Y_i - Y'' = (Y_i - Y_i'') + (Y_i'' - M_Y)$
1.Jääk 2.Regressioon
$F = {}_i\Sigma(Y_i'' - M_Y)^2 / {}_i\Sigma(Y_i - Y_i'')^2$, i = 1,...,N, kus
Y_i - Y väärtus i-ndas objektis,
Y_i'' - Y ooteväärtus i-ndas objektis,
M_Y - Y keskväärtus.
$df_1 = N-P-1$; $df_2 = P$; ($\alpha = 0,05$)
Kui olulisus $F_{ARV} \leq \alpha$, siis hüljatakse H_0.

$\Rightarrow$*Millised andmed võtta analüüsi?*
Iga objekti korral:
1) X ja Y teisendada Z-tulemuseks
$Z_X = (X_i - M_X)/ \sigma_X$, $-3,0 \leq Z_X \leq 3,0$
$Z_Y = (Y_i - M_Y)/ \sigma_Y$, $-3,0 \leq Z_Y \leq 3,0$

2) Hälvete $e_i = Y_i - Y_i''$ analüüs
a) $Z_{ei} = (e_i - M_e)/ \sigma_X$, $-3,0 \leq Z_e \leq 3,0$
b) kas hälvete jaotus normaalne?
c) Cook'i kaugus objektile j, j=1,...,N
$C_j = {}_i\Sigma({}^jY_i'' - Y_i'')^2 / (p+1)S^2$, i=1,...,N, kus

${}^jY_i''$ - i-nda objekti korral Y oodatav väärtus, kui objekt j on analüüsist välja lülitatud,
S - esindusviga.

Mitmene regressioonanalüüs

$Y'' = B_0 + B_1X_1 + B_2X_2 + ... + B_PX_P$, kus

$\Rightarrow$ ***Kuidas hinnata tunnuste X_K tähtsust?***

1) kuivõrd iga X_K üksinda mõjutab Y-it?

2) kuivõrd kõik X_K üheskoos mõjutavad Y-it?

1) Mida suurem r_{XY}, seda tugevam lineaarne seos;

2) a) B_K , X_K erinevad mõõtühikud;

b) $BETA_K = B_K * S_K / S_Y$, kus

S_K , S_Y - tunnuste X_K ja Y standardhälbed,

c) $_jH_{R2} = R^2 - {}_jR^2$, kus

$_jR^2$ - determinatsioonikordaja, mille arvutamisel tunnus X_j välja jäetud.

Mida suurem H_{R2} , seda tähtsam tunnus.

H_0: $_jH_{R2} = 0$;

H_1: $_jH_{R2} \neq 0$; (p=0,05)

Kui F_{ARV} olulisus $\leq$ p => hülgame H_0.

Mittelineaarne regressioonanalüüs

Enamlevinud funktsioonid:

$Y = a + b/X$ $Y = a + bX + cX$ $Y = ab^X$

$Y = a + b\log X$ $\log Y = a + b\log X$