Matemaatiline statistika

Kirjeldav ja järeldav statistika

Statistika on kõik see, mis põhineb andmetel ja tegeleb andmetega.

Enamasti eristatakse kahte statistika valdkonda:

Kirjeldav statistika tegeleb valimi resümeerimise ja kirjeldamisega (sagedustabel, erinevad diagrammid, keskmise taseme näitajad, hajuvuse näitajad jne).

Järeldava statistika ülesanne on üldistuste tegemine laiema objektide hulga – üldkogumi kohta (uurime väikest hiirte kogumit järeldused teeme kõigi hiirte kohta).

Punkthinnangud

Valimi põhjal arvutatud arvkarakteristikud (näit aritmeetiline keskmine, standardhälve jne) on hinnanguteks vastavatele üldkogumi parameetritele. Neid kutsutakse punkthinnanguteks.

Üldkogumi parameetri **punkthinnang** on valimi vastav parameeter.

jaotusparameeter ⟷ punkthinnang

keskväärtus EX ⟷ valimi aritm. keskmine $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

dispersioon DX ⟷ valimi dispersioon s^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

standardhälve σ ⟷ valimi standardhälve s

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Punkthinnangute puudused

Punkthinnangud on juhuslikud suurused, sest nad muutuvad ühelt valimilt teisele ülemineku korral.

Punkthinnangud ei anna informatsiooni hinnangu täpsuse ja usaldatavuse kohta.

Valimi väikeses mahu korral (kui *n* on väike) võib punkthinnang oluliselt erineda hinnatava parameetri tegelikust väärtusest.

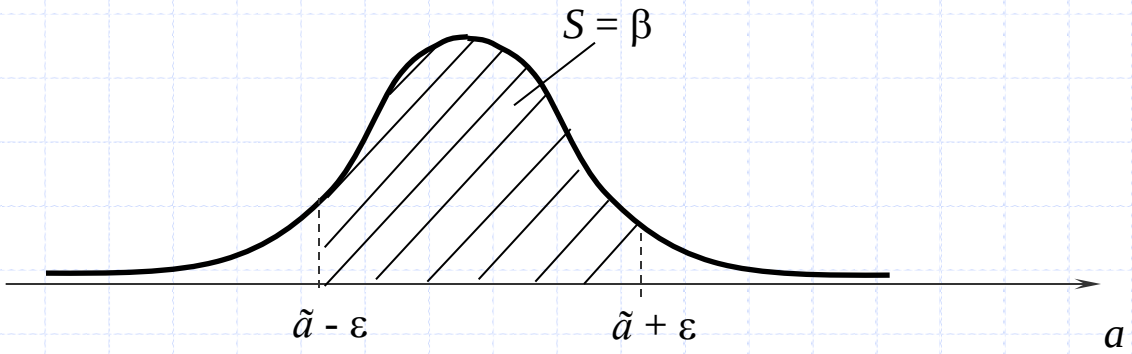
Vahemikhinnangud

Vahemikhinnangu puhul määratakse vahemik, millesse otsitav parameeter etteantud tõenäosusega kuulub. Seda tõenäosust nimetatakse **usaldusnivooks** ja tähistatakse sümboliga β.

Parameetri a **sümmeetriliseks usalduspiirkonnaks** vastavalt **usaldusnivoole** β nimetatakse juhuslikku vahemikku $(\tilde{a} - \varepsilon, \tilde{a} + \varepsilon)$, mis katab hinnatava parameetri a tõenäosusega β:

$$P(|\tilde{a} - a| < \varepsilon) = \beta$$

Arv $\varepsilon > 0$ iseloomustab hinnangu täpsust.



Normaaljaotuse keskväärtuse usalduspiirkond

Usalduspiirkonna leidmine:

$$P(|\overline{X} - m| < \varepsilon) = \beta$$

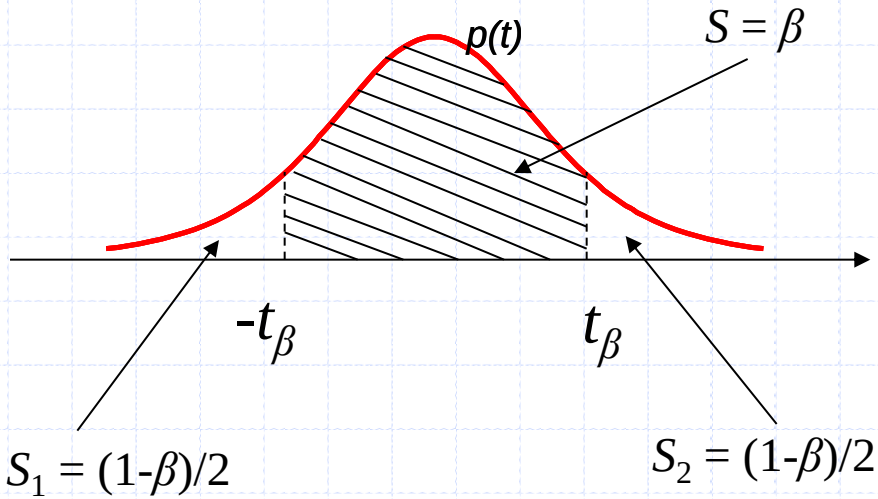
Usalduspiirkonnaks saame:

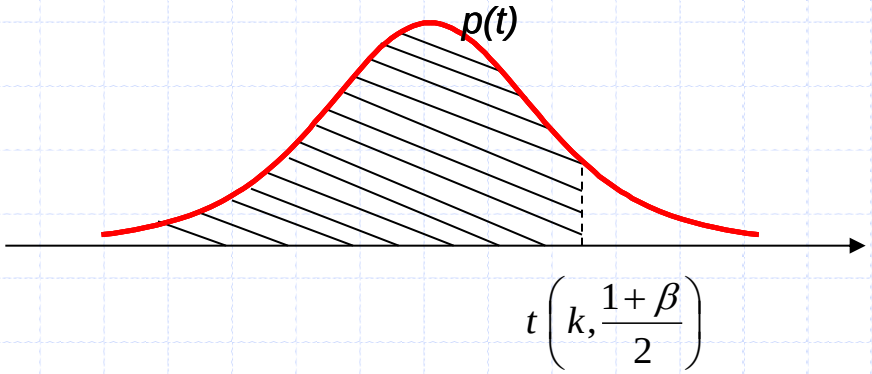
$$\bar{x} - \frac{s \cdot t\left(k, \frac{1+\beta}{2}\right)}{\sqrt{n}} < m < \bar{x} + \frac{s \cdot t\left(k, \frac{1+\beta}{2}\right)}{\sqrt{n}}$$

$t\left(k, \frac{1+\beta}{2}\right)$ on Studenti jaotuse kvantiil

$k = n - 1$

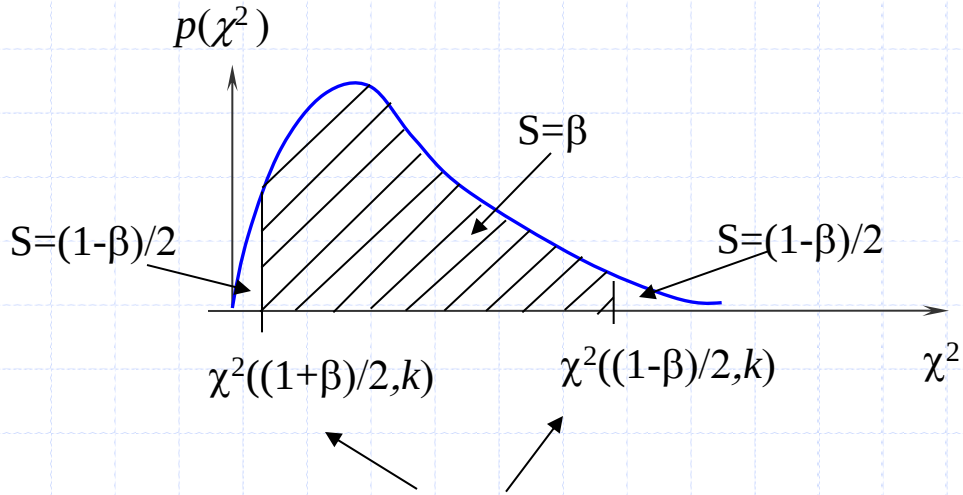
Kvantiilid





$$\frac{1-\beta}{2} + \beta = \frac{1+\beta}{2}$$

χ^2 – jaotuse täiendkvantiilid



χ^2-jaotusega juh. suuruse täiendkvantiilid

$$P\left[\chi^2\left(\frac{1+\beta}{2},k\right) < \chi^2 < \chi^2\left(\frac{1-\beta}{2},k\right)\right] = \beta$$

Dispersiooni ja standardhälbe usalduspiirkond

Tõkked, mille vahel suurus σ^2 tõenäosusega β asub ehk dispersiooni usaldusvahemik usaldusnivooga β:

$$\left(\frac{k \cdot s^2}{\chi^2\left(\frac{1-\beta}{2}, k\right)}, \frac{k \cdot s^2}{\chi^2\left(\frac{1+\beta}{2}, k\right)} \right)$$

Standardhälbe usalduspiirkonna määramisel võetakse dispersiooni kummastki usalduspiirist ruutjuur:

$$\left(\sqrt{\frac{k \cdot s^2}{\chi^2\left(\frac{1-\beta}{2}, k\right)}}, \sqrt{\frac{k \cdot s^2}{\chi^2\left(\frac{1+\beta}{2}, k\right)}} \right)$$

Statistilised hüpoteesid

Statistiliseks hüpoteesiks nimetatakse teatud teineteist välistavate väidete paari üldkogumi(te) või tema parameetrite kohta.

Statistilisteks hüpoteesideks võivad olla näiteks oletused:

- jaotusseaduse tüübi kohta;
- kahe jaotuse parameetrite võrdsusest või olulisest erinevusest;
- juhuslike suuruste vahelise seose olemasolust või puudumisest.

H_1 – sisukas e. **alternatiivne** e. **konkureeriv hüpotees**, mida uurija soovib tõestada (tavaliselt mingi erinevuse, mõju või seose olemasolu).

H_0 – nullhüpotees, mis tavaliselt väljendab uurijat mittehuvitavat juhtu (üldkogumi vastamine teatud standardile).

H_0: sild on nõrk
H_1: sild on piisavalt tugev

H_0: kohtualune ei ole süüdi
H_1: kohtualune on süüdi

Vead hüpoteeside kontrollimisel

Kuna statistiliste hüpoteeside kontrollimisel tehakse valimi põhjal järeldusi üldkogumi kohta, on võimatu vältida vigu. Vigu saab olla kaht liiki.

- **Esimest liiki viga** tekib siis, kui võetakse vastu sisukas hüpotees, aga tegelikult on õige nullhüpotees.

  See on raske viga, mis tähendab, et uurija "tõestas" erinevuse, mõju või seose mida tegelikult ei ole, vaid mis juhuslikult ilmnes mõõdetud valimis.

- **Teist liiki viga** tekib siis, kui jäädakse nullhüpoteesi juurde, ehkki tegelikult on õige sisukas hüpotees.

  See on kergem viga, mis enamasti tähendab, et soovitu tõestamiseks tuleb mõõtmisandmeid juurde koguda.

Vea tekkimise suurust mõõdetakse tõenäosusega.

Olulisuse nivoo e. riskiprotsent

Esimest liiki vea tegemise suurimat lubatavat tõenäosust nimetatakse **olulisuse nivooks e. riskiprotsendiks**.

Olulisuse nivood tähistatakse tähega α.

Olulisuse nivooks valitakse mingi väike arv, sageli

0,1
0,05
0,01

(sõltuvalt selles, kui rasketele tagajärgedele võib 1. liiki vea tegemine viia).

Olulisuse nivoo määrab uurija, selle suuruses lepitakse tavaliselt kokku enne uuringu algust.

Statistilise hüpoteesi kontrollimine

Otsuse langetamiseks valitakse teatav juhuslik suurus (teststatistik), mille teoreetiline jaotus nullhüpoteesi kehtivuse korral on teada.

Valimi(te) andmetel arvutatakse teststatistiku empiiriline väärtus, kui see on võrreldes tema teoreetilise jaotusega on ebatõenäoline, siis võetakse vastu sisukas hüpotees, vastasel juhul jäädakse nullhüpoteesi juurde.

Matemaatilise statistika meetoditega **saab üldiselt tõestada ainult sisukat hüpoteesi H_1.**

See, kui arutluskäigu tulemusena võetakse vastu nullhüpotees, ei ole selle hüpoteesi tõestus. Nullhüpoteesi vastuvõtmine tähendab, et kui uurija tahab mingit erinevust, mõju või seose olemasolu tõestada, siis tuleb tal mõõtmisi jätkata.

Kahe normaaljaotuse keskväärtuse võrdlemine

X, Y - normaaljaotusega juhuslikud suurused

Valimid : $x_1, \ldots, x_n$, $n < 30$,
$y_1, \ldots, y_m$, $m < 30$,

Hüpoteesid: $H_0 : EX = EY$

$H_1 : EX \neq EY$

Lahendus:

1) Teststatistikuna kasutatakse statistikut

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

Kahe normaaljaotuse keskväärtuse võrdlemine

kui valimite mahud on väikesed siis see statistik on ligikaudselt **Studenti jaotusega** vabadusastmete arvuga

$$k = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{\left(\frac{s_x^2}{n}\right)^2}{n-1} + \frac{\left(\frac{s_y^2}{m}\right)^2}{m-1}}$$

ümardatakse lähima vähima täisarvuni

2) Etteantud olulisuse nivoo α korral **leitakse kriitiline punkt t_{kr}** Studenti jaotuse kvantiilide tabelist.

$$t_{kr} = t_{k;1-\frac{\alpha}{2}}$$

3) Juhul kui $|t_{emp}| > t_{kr}$, siis lükatakse nullhüpotees tagasi, vastupidisel juhul jäädakse nullhüpoteesi juurde.

Kahe normaaljaotuse dispersioonide võrdlemine

Hüpoteesid: $H_0 : DX = DY$

$H_1 : DX \neq DY$

Nende hüpoteeside kehtivuse kontrollimisel olulisuse nivool α

1) Leitakse kaks kriitilist punkti:

$$F_p = F(k_1, k_2, \alpha/2) \quad ja \quad F_v = \frac{1}{F(k_2, k_1, \alpha/2)}$$

F-jaotuse täiendkvantiilid

seejuures k_1 on suurema dispersioonihinnanguga valimi vabadusastmete arv.

Kahe normaaljaotuse dispersioonide võrdlemine

2) Arvutatakse välja teststatistiku väärtus

$$F = \frac{s_1^2}{s_2^2}, \quad kus \quad s_1^2 > s_2^2.$$

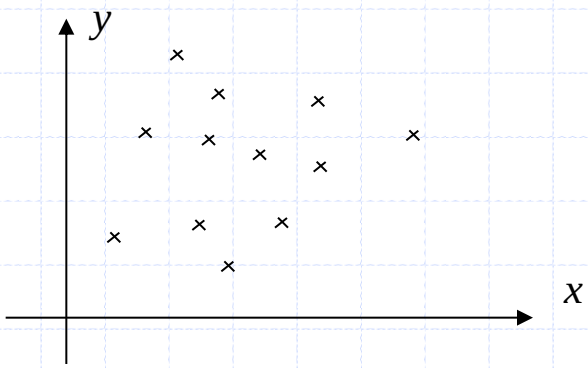
3) Tehakse järeldus. Kui

$$F_v < F_{emp} < F_p,$$

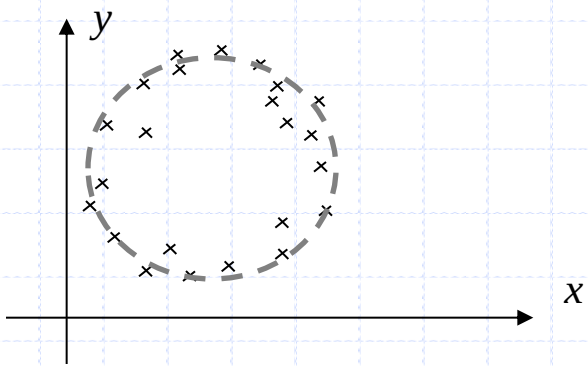
siis pole alust nullhüpoteesi tagasi lükata.

Kui üks võrratustest on rikutud, võetakse vastu alternatiivne hüpotees.

Statistiline sõltuvus



Sõltuvus puudub



Statistiline sõltuvus olemas

Kui üht suurust pole võimalik teise kaudu täpselt arvutada, vaid selle asemel on ühe muutuja **tendents muutuda kindlas suunas** teise muutuja muutumisel, siis on tegemist **statistilise e. stohhastilise sõltuvusega**.

Korrelatsioonikordajad

Lineaarne korrelatsioonikordaja iseloomustab tunnustevahelise sõltuvuse lähedust lineaarsele seosele

$$r = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{\left\{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]\cdot\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]\right\}^{1/2}}$$

Spearmani korrelatsioonikordaja iseloomustab tunnustevahelise sõltuvuse lähedust monotoonsele seosele.

$$r = 1 - \frac{6\cdot\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Korrelatsioonikordajad on defineeritud nii, et

$$-1 \le r \le 1$$

Korrelatsioonikordaja tõlgendamine

Kui $r > 0$, siis räägitakse **positiivsest korrelatsioonist**, see tähendab, et juhuslikel suurustel X ja Y on tendents muutuda samas suunas.

Negatiivse korrelatsiooni korral ($r < 0$) on ühe suuruse kasvamisel teisel suurusel tendents kahaneda.

Korrelatsiooni tugevuse kohta võib öelda, et

- korrelatsioon on tugev, kui $|r| \geq 0{,}8$
- korrelatsioon on märgatav, kui $0{,}6 \leq |r| < 0{,}8$
- korrelatsioon on nõrk, kui $0{,}3 \leq |r| < 0{,}6$
- korrelatsioon on väga nõrk, kui $|r| < 0{,}3$

Spearmani korrelatsioonikordaja

Spearmani korrelatsioonikordaja arvutamiseks tuleb:

1) korrastada ühe tunnuse väärtused x_i järjestatud hulka, alustades vähimast ja lõpetades suurimaga;

2) nummerdada saadud järjestatud hulga elemendid, alustades ühest (öeldakse elementidele x_i omistatakse astakud r_i), kui X väärtuste seas on korduvaid, siis võetakse neile vastavateks astakuteks esialgsete astakute aritmeetiline keskmine;

3) omistada teise tunnuse Y väärtustele astakud q_i;

4) arvutada summa $\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (r_i - q_i)^2$

5) leida Spearmani korrelatsioonikordaja $r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$

Korrelatiivse sõltuvuse olemasolu kontroll

$H_0: r_{xy} = 0$ (üldkogumis tunnuste vahel pole seost)

$H_1: r_{xy} \neq 0$ (üldkogumis tunnuste vahel on seos)

Nullhüpoteesi kontrollimiseks leitakse suurus $t_{emp} = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}}$

Nullhüpoteesi kehtivuse korral on see suurus ligikaudselt Studenti jaotusega vabadusastmete arvuga $k = n - 2$.

Leitakse kriitiline punkt t-jaotuse kvantiilide tabelist

$$t_{kr} = t(k, 1-\frac{\alpha}{2})$$

Kui $|t_{emp}| < t_{kr}$ siis jäädakse nullhüpoteesi juurde. Vastupidisel juhul võetakse vastu alternatiivne hüpotees, mis tähendab, et r erineb oluliselt nullist ning X ja Y vahel on korrelatiivne sõltuvus.

Regressioonanalüüs

Tunnustevahelise seose suuruse ja suuna hindamiseks kasutatakse korrelatsioonikordajaid.

Sageli on lisaks sõltuvuse tugevuse hinnangule võimalik väljendada tunnuste X ja Y vahelist sõltuvust funktsioonina $Y = f(X)$.

Üldjuhul võib $Y = f(X)$ olla suvaline matemaatiline funktsioon.

Lihtsaimal juhul vaadeldakse lineaarfunktsiooni:

$$Y = bX + a.$$

Vastavat sirget nimetatakse regressioonisirgeks, kusjuures

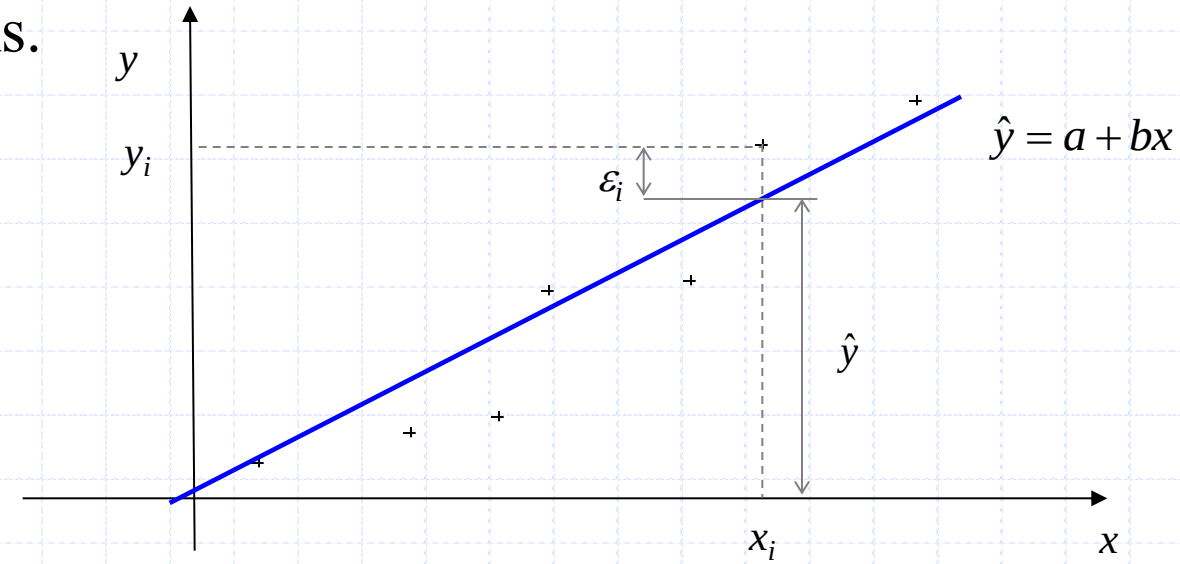
b – regressioonikordaja

a – vabaliige.

Üldiselt: ühe tunnuse (juhusliku suuruse) modelleerimisega teise tunnuse abil tegeleb **regressioonanalüüs**.

Lineaarne regressioon

Regressioonisirge leidmisel lähtutakse tunnuste X ja Y väärtustest milleks on paaride hulk $\{(x_i\,;y_i)\}$.
Ülesandeks on leida sirge, mis nende punktidega võimalikult hästi kokku sobiks.



Kui ühe tunnuse väärtus oleks täpselt avaldatav teise tunnuse lineaarfunktsiooniga, siis kehtiks iga i korral $y_i = bx_i + a$.
Tegelikult ei asetse tunnustepaarile vastavad punktid sirgel, vaid on sellest mõnevõrra (ε_i võrra) erinevad: $\varepsilon_i = (bx_i + a) - y_i$

Vähimruutude meetod

Et leitava sirgjoone võrrand kirjeldaks andmeid võimalikult täpselt nõuame, et punktide (x_i ; y_i) kaugused sellest sirgest oleksid võimalikult väikesed.

Niisugust sirgjoone konstrueerimise meetodit nimetatakse **vähimruutude meetodiks**.

Vähimruutude meetodi idee seisneb selles, et minimiseeritakse hälvete

$$\varepsilon_i = (a + bx_i) - y_i$$

ruutude summa

$$s_r^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (a + bx_i - y_i)^2 = G(a,b) \geq 0$$

Vähimruutude meetod

Hälvete ruutude summa on kahe muutuja *a* ja *b* mittenegatiivne funktsioon ning tal on alati olemas miinimum.

Seega tuleb arvutada osatuletised *a* ja *b* järgi, võrdsustada need nulliga ja lahendada saadud võrrandisüsteem.

Tulemusena saame regressioonikordajate vähimruutude hinnangud:

$$a = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2},$$

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}.$$

Determinatsioonikordaja

Determinatsioonikordaja mõõdab, kui hästi regressioonisirge lähendab vaatlusandmeid.

Selleks, et determinatsioonikordajat leida, arvutatakse:

1) mõõdetud väärtuste y_i aritmeetiline keskmine $\bar{y}$

2) mõõdetud väärtuste y_i koguvariatsioon $s_v^2 = \sum (y_i - \bar{y})^2$

3) lineaarse regressioonimudeliga selgitatav variatsioon

$$s_s^2 = \sum (\hat{y}_i - \bar{y})^2,$$

kus $\hat{y}_i = a + b \cdot x_i$ on lineaarse regressioonimudeli kohaselt arvutatud väärtus („silutud väärtus“)

4) variatsioon, mis ei ole selgitatav lineaarse regressiooniga

$$s_r^2 = \sum (\hat{y}_i - y_i)^2.$$

Determinatsioonikordaja

Suurused s_v^2, s_s^2 ja s_r^2 on seotud valemiga

$$s_v^2 = s_s^2 + s_r^2.$$

5) Determinatsioonikordaja on lineaarse regressioonimudeliga selgitatava variatsiooni s_s^2 ja koguvariatsiooni s_v^2 suhe

$$r^2 = \frac{s_s^2}{s_v^2} = \frac{s_v^2 - s_r^2}{s_v^2} = 1 - \frac{s_r^2}{s_v^2}.$$

Determinatsioonikordaja väärtus r^2 rahuldab võrratusi

$$0 \le r^2 \le 1$$

ning ta väljendab, kui suur osa sõltuva muutuja Y kogumuudust on selgitatav sõltumatu muutuja X muuduga.

Determinatsioonikordaja

Kui $r^2 \geq 0{,}9$ siis lähendab lineaarne regressioonimudel vaatlusandmeid väga hästi.

Kui $0{,}8 \leq r^2 \leq 0{,}9$ siis lähendab lineaarne regressioonimudel vaatlusandmeid hästi.

Kui $0{,}6 \leq r^2 \leq 0{,}8$ siis võib mõne rakenduse puhul lugeda tulemust rahuldavaks, kuid statistilisi prognoose tehes peab olema ettevaatlik.