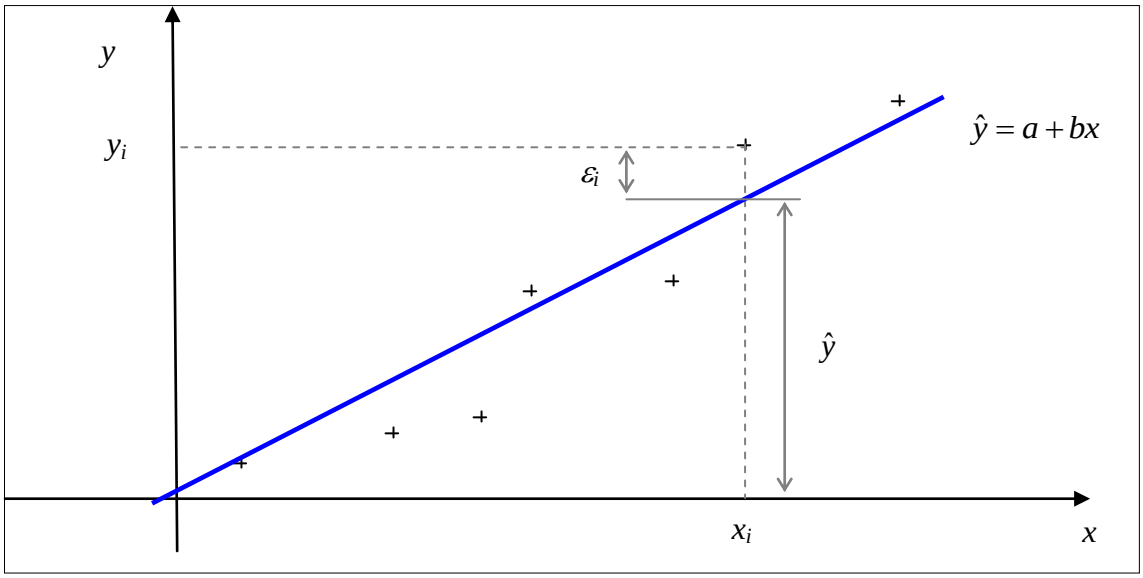
Regressioonanalüüs

Tunnustevahelise seose suuruse ja suuna hindamiseks kasutatakse korrelatsioonikordajaid. Sageli on lisaks sõltuvuse tugevuse hinnangule võimalik väljendada tunnuste *X* ja *Y* vahelist sõltuvust funktsioonina. (Uuritavatelt tunnustelt *X* ja *Y* nõutakse, et nad mõlemad oleksid arvulised ja sisust lähtudes määratakse kindlaks kumb on sõltumatu - *X* ja kumb sõltuv - *Y*.) Sellisel juhul on võimalik ühe tunnuse väärtuse abil prognoosida teise tunnuse väärtust ja me räägime, et meil on antud tunnuste matemaatiline mudel. Üldjuhul võib $Y = f(X)$ olla suvaline matemaatiline funktsioon. Lihtsaimal juhul vaadeldakse lineaarfunktsiooni: $Y = bX + a$. Vastavat sirget nimetatakse regressioonisirgeks, kusjuures *b* kannab nime regressioonikordaja ning *a* – vabaliige.

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

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

ruutude summa

$$s_r^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (a + bx_i - y_i)^2 .$$

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:

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

Regressioonisirge võrrand on avaldatav ka kujul $\hat{y} = \bar{y} - b(x - \bar{x})$, millest on näha, et regressioonisirge läbib punkti $(\bar{x}, \bar{y})$, mida nimetatakse **korrelatsiooni keskpunktiks**.

Näide. Järgnevas tabelis on toodud aparaatide testimiseks kulutatud aeg:

Seadmete arv	4	6	2	5	7	6	3	8	5	3	1	5
Aeg (min.)	197	272	100	228	327	279	148	377	238	142	66	239

Uurida aparaatide arvu ja nende testimiseks kulunud aja vahelist sõltuvust. Arvutada regressioonisirge parameetrid *a* ja *b*. Joonestada graafik, millele on kantud antud punktid $(x_i; y_i)$ ning regressioonisirge. Lahendus on toodud lisana Exceli failis.

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

Suurused s_v^2, s_s^2 ja s_r^2 on seotud valemiga $s_v^2 = s_s^2 + s_r^2$. 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.

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

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

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

Näide. Arvutame eelmise näite andmetel determinatsioonikordaja. Lahendus on toodud lisana Exceli failis.

Regressioonisirge usaldatavus

Eespool toodud regressioonisirge parameetrite arvutusvalemitega arvutatavaid suurusi tuleb vaadelda kui hinnanguid regressioonikordajatele. Nende täpsuse hindamiseks leitakse vastavad vahemikhinnangud ehk usalduspiirkonnad. Tutvume siinkohal vaid usalduspiirkondade leidmise algoritmiga.

Regressioonisirge parameetrite *a* ja *b* usalduspiirkondade leidmine:

1) Leiame prognoosijäägi ε_i standardhälbe hinnangu:

$$s_e = \sqrt{\frac{s_r^2}{n-2}}$$

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

2) Leiame parameetrite *a* ja *b* standardhälvete hinnangud:

$$s_a = s_e \sqrt{\frac{\sum x_i^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}},$$

$$s_b = \frac{s_e}{\sqrt{\sum x_i^2 - n\left(\bar{x}\right)^2}}.$$

3) Etteantud usaldusnivoo β puhul on *a* ja *b* usalduspiirkonnad:

$$\left(a - s_a \cdot t\left(k; \frac{1+\beta}{2}\right); \quad a + s_a \cdot t\left(k; \frac{1+\beta}{2}\right)\right),$$

$$\left(b - s_b \cdot t\left(k; \frac{1+\beta}{2}\right); \quad b + s_b \cdot t\left(k; \frac{1+\beta}{2}\right)\right),$$

kus

$k = n - 2$,

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

(*MS Exceli* keskkonnas kasutame Studenti jaotuse kvantiili leidmiseks funktsiooni *TINV*(α; *k*), kus $\alpha = 1 - \beta$ ja $k = n - 2$.)

Näide. Leiame eelmises näites leitud regressioonisirge parameetrite usalduspiirkonnad usaldusnivooga 0,95. Lahendus on toodud lisana Exceli failis.

Statistilised prognoosid

Üks regressioonianalüüsi levinumaid rakendusi on seotud **statistiliste prognoosidega**. Juhusliku suuruse *Y* käitumist on võimalik teatava tõenäosusega prognoosida regressioonivõrrandi abil. Siinjuures tuleb arvestada, et regressioonimudel kehtib vaid vaatlusandmetega kaetud piirkonnas. Ekstrapoleerimist vaadeldavast piirkonnast väljapoole võib lubada vaid vähesel määral ning see on seotud suure riskiga.

Prognoosi punkthinnangu saamiseks tuleb regressioonivõrrandisse asendada vastav x_p ja arvutada $\hat{y}_p = a + b \cdot x_p$. Punkthinnangu $\hat{y}_p$ täpsuse ja usaldatavuse hindamiseks arvutatakse prognoosi punkthinnangu usalduspiirkond.

Prognoosi usalduspiirid usaldusnivooga β arvutatakse järgneva valemi kohaselt:

$$\left(\hat{y}_p - s_u \cdot t\left(k, \frac{1+\beta}{2}\right); \quad \hat{y}_p + s_u \cdot t\left(k, \frac{1+\beta}{2}\right)\right)$$

kus

$s_u = s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x_i^2 - n \cdot (\bar{x})^2}}$ prognoosi punkthinnangu standardhälve,

$k = n - 2$,

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

(*MS Exceli* keskkonnas kasutame Studenti jaotuse kvantiili leidmiseks funktsiooni *TINV*(α; *k*), kus $\alpha = 1 - \beta$ ja $k = n - 2$.)

Näide. Prognoosida eelmise näite põhjal muutuja *Y* väärtust, kui $x_p = 6{,}2$. Leida prognoosi 90%-lised usalduspiirid. Lahendus on toodud lisana Exceli failis.