
Aki Taanila

YHDEN SELITTÄJÄN REGRESSIO

26.4.2011

SISÄLLYS

JOHDANTO.....	1
LINEAARINEN MALLI	1
Selityskerroin.....	3
Excelin funktioita.....	4
EKSPONENTIAALINEN MALLI	4
MALLIN KÄYTTÄMINEN ENNUSTAMISEEN	5
Mallin sopivuus	5
Poikkeavat havainnot.....	5
Mallin käyttöalue.....	5

JOHDANTO

Kahden määrällisen muuttujan riippuvuutta voit tarkastella hajontakuvion avulla. Lisäksi voit laskea lineaarisen riippuvuuden voimakkuutta mittaavan korrelaatiokertoimen. Jos haluat selvittää tarkemmin riippuvuuden luonnetta tai hyödyntää riippuvuutta ennustamistarcoituksiin, niin voit mallintaa riippuvuutta regressiomallin avulla.

Tässä monisteessa käsitellään lineaarista ja eksponentiaalista regressiomallia. Mukana ovat tarvittavat Excel-ohjeet. Viimeisin versio tästä monisteesta ja siihen liittyvästä materiaalista löytyy osoitteesta

<http://myy.haaga-helia.fi/~taaak/m>

Yllä mainitusta osoitteesta löytyvät myös tähän monisteeseen liittyvät Excel-esimerkit ja pidemmälle menevä useamman selittävän muuttujan malleja käsittelevä moniste.

Tutustu myös muihin oppimateriaaleihini <http://myy.haaga-helia.fi/~taaak/>

LINEAARINEN MALLI

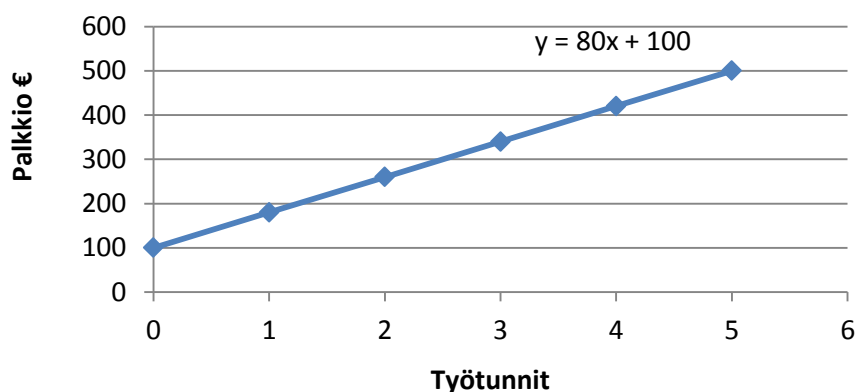
Riippuvuudesta voidaan rakentaa matemaattinen malli. Kahden muuttujan riippuvuutta kuvaava matemaattinen malli on lauseke, jonka avulla voidaan laskea toisen muuttujan arvoja ensimmäisen muuttujan arvojen perusteella. Jos muuttujien välinen riippuvuus on lineaarinen, niin mallina käytetään suoraa.

Suoraa voidaan kuvata lausekkeella $y = b_1x + b_0$. Lauseke kertoo miten y saadaan lasketua, kun x :n arvo tunnetaan.

- Termiä b_0 kutsutaan vakiotermissi. Vakiotermi kertoo, missä kohdassa suora leikkaa y -akselia (tämä nähdään asettamalla x :lle arvo 0, jolloin lausekkeesta jää jäljelle $y=b_0$).
- Termiä b_1 kutsutaan kulmakertoimeksi. Kulmakerroin ilmoittaa minkä verran y muuttuu, kun x kasvaa yhdellä yksiköllä. Laskevaan suoraan liittyy negatiivinen kulmakerroin ja nousevaan suoraan positiivinen kulmakerroin.

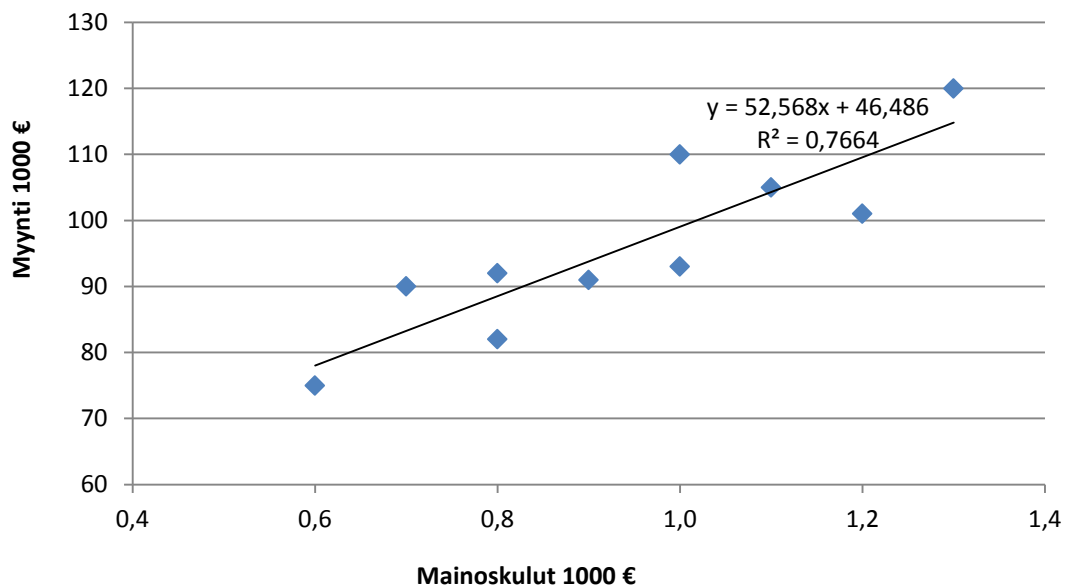
Oletetaan, että konsultti perii palkkiota paikalle saapumisesta 100 euroa ja jokaiselta tehdyltä työtunnilta 80 euroa. Tällöin konsultin kokonaispalkkiota voidaan kuvata lausekkeella $y=80x+100$, missä x on työtuntien määrä. Kyseisessä suoran yhtälössä

- vakioterminä on 100 ja se ilmoittaa y :n arvon, kun $x=0$ (eli esimerkissämme palkkio ilman varsinaisia työtunteja)
- kulmakerroin 80 ilmoittaa palkkion muutoksen, kun työtunnit lisääntyvät yhdellä.



Voit lisätä Excelin hajontakuviioon riippuvuutta kuvaavan mallin kuvaajan, lausekkeen ja selityskertoimen:

1. Napsauta hiiren oikeaa painiketta jonkin hajontakuviion pisteen päällä.
2. Valitse esiin tulevasta valikosta **Insert Trendline** (Lisää trendiviiva).
3. Valitse haluamasi malli, esimerkiksi **Linear** (Lineaarinen).
4. Valitse tulostettavaksi mallin kaava **Display Equation on Chart** (Näytä kaava kaaviossa).
5. Valitse tulostettavaksi mallin selityskerroin kohdasta **Display R-squared Value on Chart** (Näytä korrelaatiokertoimen arvo kaaviossa). Huomaa, että Excelin suomenkielissä versioissa puhutaan virheellisesti korrelaatiokertoimesta vaikka kyseessä on korrelaatiokertoimen neliö eli selityskerroin.



Yläolevaan kuvioon on lisätty malli mainoskulujen ja myynnin väliseen hajontakuviioon. Mallia voidaan tulkita seuraavasti:

- Kulmakertoimesta 52,568 voidaan päätellä, että tuhat euroa mainoskuluissa merkitsee keskimäärin 52568 euroa myynnissä.
- Vakiotermi 46,486 taas ilmoittaa myynnin olevan 46486 euroa, jos mainoskuluja ei ole lainkaan. Tässä tapauksessa vakiotermin antama tieto ei ole käyttökelpoinen eikä luotettava, koska mainoskulujen arvo 0 sijaitsee selvästi havaintoalueen ulkopuolella. Yleensäkin mallin käyttöaluetta ei voi laajentaa kovin paljon havaintoalueen ulkopuolelle.

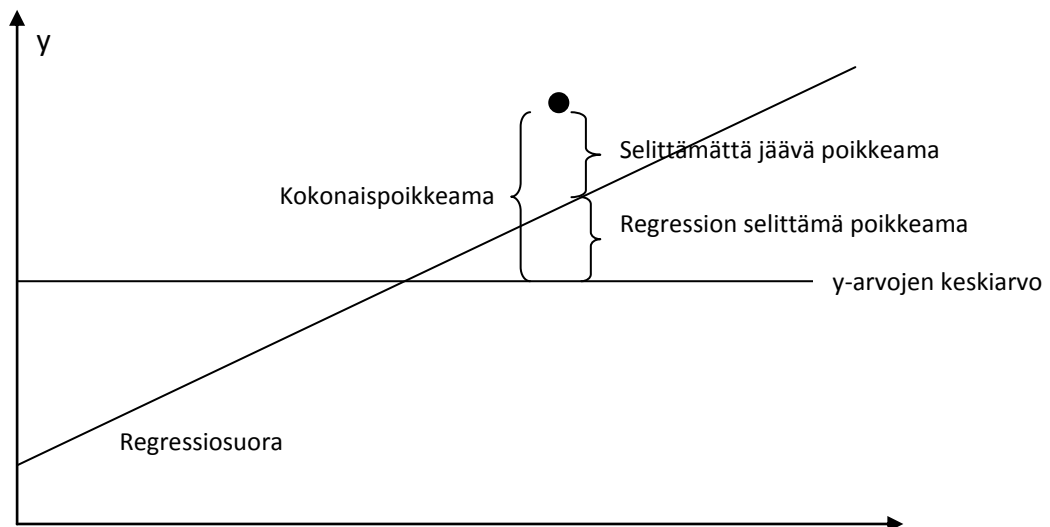
Mallin avulla voidaan laskea esimerkiksi seuraavat ennusteet:

- Jos mainontaan aiotaan käyttää 900 euroa, niin mallin mukainen myyntiennuste saadaan laskemalla $52,568 \cdot 0,9 + 46,486 \approx 93,8$ eli 93 800 euroa.
- Jos tavoitteena on 90 000 euron myynti, niin mallin mukaan mainontaan pitäisi käyttää $(90 - 46,486) / 52,568 \approx 0,83$ eli 830 euroa.

Selityskerroin

Äskeisessä esimerkissä selityskerroin on 0,7664 eli 76,64%. Tämä tulkitaan seuraavasti: 76,64% myynnin vaihtelusta voidaan selittää mainoskulojen vaihtelulla. Regression tarkoituksena on selittää y:n arvojen vaihtelua x:n arvojen vaihtelulla. Selityskertoimella mitataan kuinka hyvin tässä on onnistuttu.

Tarkastellaan seuraavaksi, mihin selityskertoimen laskenta perustuu. Kunkin havainnon y-arvon kokonaispoikkeama y-arvojen keskiarvosta koostuu kahdesta osasta: regression selittämästä poikkeamasta ja poikkeamasta, jota regressio ei selitä. Seuraavassa kuviossa havaintopisteen kokonaispoikkeama on jaettu regression selittämään poikkeamaan ja selittämättä jäävään poikkeamaan.



Jos merkitään regression selittämien poikkeamien neliöiden summaa SSR (sum of squares due to **regression**) ja selittämättömien poikkeamien neliöiden summaa SSE (sum of squares due to **error**), niin kokonaispoikkeamien neliöiden summa SST (**total** sum of squares) jakaantuu kahteen komponenttiin

$$SST = SSR + SSE$$

Selityskerroin R^2 on regression selittämän vaihtelun osuus kokonaisvaihtelusta eli

$$R^2 = \frac{SSR}{SST}$$

Jos käytetään lineaarista mallia, niin selityskerroin voidaan laskea myös korrelaatiokertoimen neliönä.

Regressiosuoran laskentamenetelmä liittyy sekin neliösummiin. Regressiosuora lasketaan pienimmän neliösumman menetelmää käyttäen. Kaikkien mahdollisten pistejoukon läpi kulkevien suorien joukosta valitaan se, jonka kohdalla neliösumma SSE (vaihtelu, jota regressio ei selitä) saa pienimmän mahdollisen arvon.

Excelin funktioita

=**FORECAST(x;tunnetut y;tunnetut x)**-funktioilla (ENNUSTE) voit kätevästi laskea lineaariseen malliin liittyviä ennusteita. Funktio laskee x-arvoon liittyvän y-arvon regressiosuoran yhtälöä käyttäen (taustalla Excel laskee tunnettujen y:n arvojen ja tunnettujen x-arvojen perusteella regressiosuoran yhtälön).

=**INTERCEPT(tunnetut y;tunnetut x)**-funktioilla (LEIKKAUSPISTE) voit laskea regressiosuoran vakiotermin.

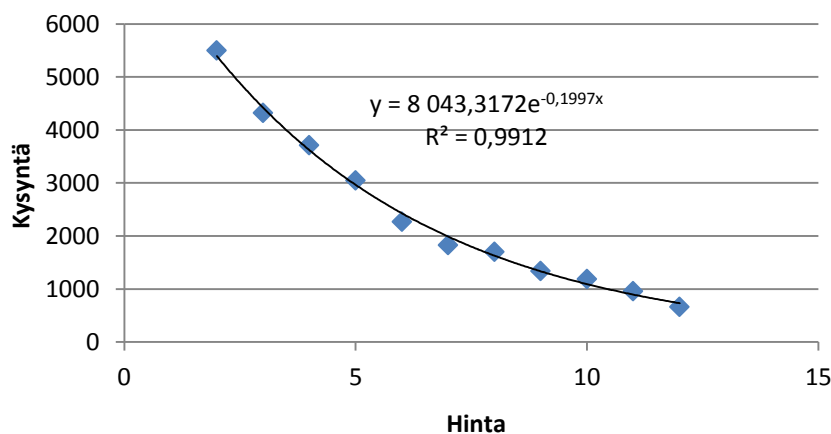
=**SLOPE(tunnetut y;tunnetut x)**-funktioilla (KULMAKERROIN) voit laskea regressiosuoran kulmakertoimen.

EKSPONENTIAALINEN MALLI

Liiketaloudessa esiintyy usein riippuvuus, jonka kuvaamiseen sopii eksponentiaalinen malli. Eksponentiaalinen malli on muotoa $y=b_0e^{bx}$

- e on luonnollisen logaritmijärjestelmän kantaluku, jonka likiarvo on 2,718
- Kerroin b_0 ilmoittaa y:n suuruuden, kun $x=0$.
- Kerroin b ilmoittaa y:n prosentuaalisen muutoksen, kun x kasvaa yhdellä yksiköllä. Huomaa, että lineaarisessa mallissa kulmakerroin ilmoittaa y:n absoluuttisen muutoksen x:n kasvaessa yhdellä yksiköllä, mutta eksponentiaalisessa mallissa y:n muutos on prosentuaalinen.

Seuraavassa kuviossa x:n kasvaessa yhdellä yksiköllä y:n arvo pienenee 19,97 %. Selityskertoimen mukaan 99,12 % kysynnän vaihtelusta voidaan selittää hinnan vaihteluilla.



Kun käytät eksponentiaalista mallia ennusteiden laskemiseen Excelissä, niin tarvitset **EXP** (EKSPONENTTI) -funktioita. Edelliseen hajontakuviioon lasketussa mallissa voit ennustaa 6 euron hintaan liittyvää kysyntää, kirjoittamalla Exceliin kaava **=8043,3172*EXP(-0,1997*6)**

MALLIN KÄYTTÄMINEN ENNUSTAMISEEN

Mallin sopivuus

Mallin avulla voidaan ennustaa y , kun x tunnetaan tai x , kun y tunnetaan. Olipa sitten kyseessä lineaarinen tai eksponentiaalinen malli (tai jokin muu), niin mallin soveltuvuus ennustamiseen riippuu selittämättömän vaihtelun osuudesta. Hajontakuviosta voit arvioida selittämättömän, epäsäännöllisen vaihtelun suuruutta ja yli päättään mallin sopivuutta havaintoaineistoon. Mitä enemmän havainnot "pomppivat" mallin molemmin puolin sitä enemmän ennusteeseen sisältyy epävarmuutta.

Poikkeavat havainnot

Mallit ovat herkkiä poikkeaville arvoille. Jos kuviosta erottuu selvästi muista poikkeavia havaintoja, niin niiden alkuperä on selvitettävä:

- ovatko poikkeavat havainnot virheellisiä tietoja
- ovatko poikkeavat havainnot väärin syötettyjä tietoja
- jos kyse ei ole virheestä, niin löytyykö poikkeaville arvoille luonnollinen selitys?

Kun poikkeavien havaintojen alkuperä selviää, niin seuraavaksi mietitään onko hyvä pitää havainnot mukana vai olisiko perusteltua jättää ne pois tarkasteluista.

- jos virheelliset tai väärin syötetyt tiedot voidaan korjata, niin ne voidaan pitää tarkasteluissa mukana
- jos virheellisille tai väärin syötetyille tiedoille ei syystä tai toisesta saada korjattuja arvoja, niin ne on syytä pudottaa pois tarkasteluista
- jos poikkeavuudelle löytyy luonnollinen selitys, niin asiaa on ajateltava tarkasteltavan ilmiön kannalta.

Mallin käyttöalue

Havaintoaineistoa on käytettävissä vain tietyiltä muuttujan arvoilta ja mallin pätevyyttä voidaan arvioida vain havaintoalueella. Havaintoalueen ulkopuolella olevien muuttujan arvojen kohdalla ei voida tietää, onko malli pätevä. Tämän vuoksi mallia ei ole perusteltua käyttää havaintoalueen ulkopuolella.