Sisällysluettelo

Joissain tapauksissa havaitsemme, että residuaalit eivät olekaan homoskedastisia selitettävän muuttujan suhteen. Huomaamme, että myös **selitettävää muuttujaa saatetaan joutua korjaamaan ja muuntamaan**. Mainitut nyrkkisäännönomaiset muunnokset soveltuvat myös tähän tapaukseen. Mainittujen muunnosten lisäksi on olemassa myös sofistikoidumpi tapa etsiä oikeaa muunnosta. Box-Cox -muunnoksella voidaan **etsiä sitä Y-muuttujan potenssia, joka toisi parhaan mahdollisen korjauksen heteroskedastiseen tilanteeseen**. Mikäli nimittäin Y:n potenssi on 1, ei muunnosta tarvita, mutta jos Y:n potenssi olisi esimerkiksi ½, tarvittaisiin neliöjuurimuunnos. Box ja Cox esittivät (1964), että parasta potentiaalista potenssia voidaan etsiä mallittamalla Y:n potenssin $Y^{(1-b)}$ parametri *b* (*Slope*). Jo aiemmin mainittuja muunnoksia vastaavat eri *b*:n arvot seuraavasti:

b:n arvo	Kaavan potenssi	Suositeltava muunnos
-1	2	neliöön korottaminen (Y^2)
0	1	ei muunnosta (Y)
½	½	neliöjuuri ($\sqrt{Y}$)
1	0	logaritmi (LOG(Y) tai LN(Y))
1½	-½	käänteisluku ja neliöjuuri ($1/Y^2$)
2	-1	käänteisluku (1/Y)

SPSS-ohjelmistossa ei pystytä suoraan laskemaan tätä teknistä muunnosehdotusta, mutta samaa ideaa käyttäen pystytään muunnosta arvioimaan ns. *Spead-versus-level* -kuvalla. Tämä tapahtuu *Analyze*-valikon *Descriptives*-alavalikon *Explore*-valinnan avulla. Riippuviksi muuttujiksi (*Dependent List*) valitaan malliin mukaan tulleet selittävät muuttujat. Nämä tekijät asetetaan selitettävää muuttujaa (*Factor List*) vastaan. *Plots*-valinnasta valitaan *Spread vs. Level with Levene Test* -kuva ja siellä erityisesti *Power estimation*. Kuvan alareunaan tulee aineiston pohjalta laskettu Y:n potenssin *b*-arvo (*Slope*) sekä ehdotettu muunnos (*Power for transformation*). Harvoin ehdotettu muunnos on kuitenkaan juuri täsmälleen edeltäneen taulukon mukainen. Muunnosta kokeillaan sillä potenssilla, joka lähinnä vastaa taulukon arvoa. Jos siis ehdotettu muunnoksen potenssi on 0.189, voidaan kokeilla logaritmimuunnosta, sillä 0.189 on lähempänä nollaa kuin puolta. Uudella muunnetulla muuttujalla tehdään joko edellä kuvattu *Spread-versus-level* -kuva tai tavallinen histogrammi ja tarkistetaan, olisiko nyt ehdotettu muunnoksen potenssi lähellä arvoa 1 eli ettei tarvittaisi muunnosta.

Ennustearvo, residuaalit ja ristiinvalidointi

Lisäanalyysina voidaan laskea ennustearvo (*Predicted value*). Kyseistä arvoa voidaan hyödyntää esimerkiksi ennustettaessa muiden kuin otoksessa tai laskelmissa mukana olleiden havaintojen arvoja. Lisäanalyysina voidaan tehdä **analyyseja myös residuaaleille**. Voimme pyrkiä selittämään sitä, miksi osa havainnoista ei esimerkiksi saavuta ennustearvoa (residuaalit ovat negatiivisia) ja osa saa ennustetta paremman arvon (residuaalit positiivisia). **Ristiinvalidoinnissa** (*Cross-validation*) enkä kolmasosa tai jopa puolet havainnoista otetaan mallin rakentamiseen ja malli testataan lopuilla. On nimittäin niin, että tekninen mallittaminen pystyy kyllä löytämään hyvän mallin – jopa liian hyvän mallin – siihen aineistoon, joka oli mukana analyysissa. Emme vain tiedä, miten luotu malli ennustaa niitä havaintoja, jotka eivät ole mukana itse analyysissa. **Yleinen tapa tutkia mallin hyvyyttä tai toimivuutta onkin ottaa mukaan analyysiin vain osa havainnoista ja tutkia mallin osuvuutta lopuilla havainnoilla**. Tämä vaatii kuitenkin huomattavan määrän havaintoja.

2.1.5 Tekninen suoritus SPSS-ohjelmistolla ja tulkinta

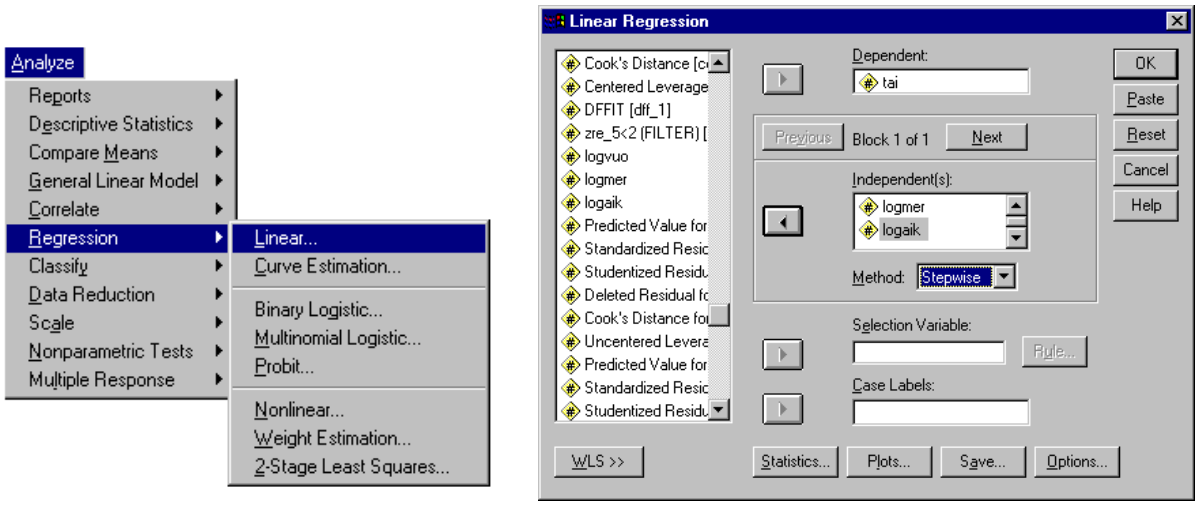
Tarkoituksemme on regressioanalyysin avulla selvittää, **millä sitoutumiseen liittyvillä tekijöillä voidaan selittää sitä, että kyselyyn vastaaja kokee olevansa taitava oman alansa harrastajana**. Korrelaatiomatriisin perusteella mukaan on valittu vain sellaisia tekijöitä, jotka korreloivat taitotason kanssa (Taulukko 21). Yleisesti ottaen korrelaatiot eivät ole korkeita (vaihtelevat välillä 0.29 - 0.50), mikä indikoi sitä, että mallin selitysaste ei tule olemaan korkea. Koska otoskoko on kohtuullinen (n. 700), korrelaatiot eroavat nollasta tilastollisesti merkitsevästi. Ennen regressioanalyysia on tutkittu muuttujien jakaumia ja havaittu, että erityisesti kaikki selittävät muuttujat ovat voimakkaasti vinoja. Kaikille selittäville muuttujille on tehty logaritmimuunnokset. Muuttujat AIK ja VUO ovat voimakkaasti painottuneita pieniin arvoihin, joten niille on tehty suora luonnollinen logaritmointi. Logaritmointia varten muuttuja pitää olla positiivinen; koska jotkut vastaajista olivat harrastaneet vain vähän aikaa (vuosina 0) ja toisaalta jotkut ilmoittivat käyttävänsä 0 tuntia viikossa harrastukseensa, on muutujiin ensin lisätty 1. Muuttuja MER puolestaan on painottuu suuriin arvoihin, joten se on ensin käännetty ja sitten otettu luonnollinen logaritmi. Näillä muunnoksilla pyritään siihen, että jakaumat olisivat normaalimmat, eikä jälkikäteen tarvitsisi tehdä muunnoksia residuaalien heteroskedastisuuden tai poikkeavien havaintojen vuoksi.

Taulukko 2.1 Analyysiin mukaan tulevat muuttujat

Mja	selite	Merkitys mallissa	Korrelaatio Y:n kanssa
TAI	Itse koettu taitotaso	Y	
LOGMER	Harrastuksen koettu merkityksellisyys (logaritmimuunnos [LN(6-MER)])	X	-.29
LOGAIK	Harrastukseen käytetty aika kuukaudessa (logaritmimuunnos [LN(AIK+1)])	X	.29
LOGVUO	Harrastuksen kesto vuosina (logaritmimuunnos [LN(AIK+1)])	X	.50

Perusnäkymä

SPSS-ympäristössä regressioanalyysi alkaa valinnoilla *Analyze* – *Regression* – *Linear*… Regressioanalyysin päävalikko näyttää seuraavalta:



Valinnat

Muistamme, että regressioanalyysi voidaan tehdä usealla eri menettelyllä. Oletuksena SPSS-ohjelmistossa on, että halutaan tehdä pakotettu mallitus (*Method*-valinnan vaihtoehto *Enter*). Nut valitaan kuitenkin askeltava eli *Stepwise*-menettely, sillä käytössä ei ole teoriaa, joka väittäisi, juuri kyseiset valitut selittävät muuttujat muodostaisivat taitotason. Aiemmin jo todettiin, että on varmempaa kokeilla erilaisia regressiomenettelyjä luotettavimman tuloksen löytämiseksi. Päävalikossa on muuttujien ja regressiomenettelyn lisäksi valittavana neljä erilaista seikkaa, joilla voidaan vaikuttaa analyysin sisältöön ja laajuuteen***:*** *Statistics, Plots, Save* ja *Options* . Tehdään oletusten lisäksi seuraavat valinnat:

- **Statistics**

  ☒ R squared change, ☒ Collinearity diagnostics

  Residuals: ☒ Casewise Diagnostics

  [Haluamme tutkia multikollineaarisuutta ja pyrimme löytämään outliereita.]

- **Plots**

  ☒ Produce all Partial plot

  Valitaan kuvat: Y:SDRESID [Standardoitut residuaalit]
  X:ZPRED [Ennustearvo]

  ja

  Y:ZPRED [Ennustearvo]
  X:DEPENDNT [Selitettävä muuttuja Y]

  Standardized Residual plots: ☒ Normal probability plot

  [Voimme graafisesti tarkastella, ovatko residuaalit normaalisia.]

- **Save**

  Predicted: ☒ Standardized

  Residuals: ☒ Studentized, ☒ Studentized deleted

  Distance: ☒ Mahalanobis, ☒ Cook's, ☒ Leverage value

  Influence statistics: ☒ DfFit

  [Saamme hyvän käsityksen mallin oletusten toteutumisesta.]

- **Options**

  [Annamme oletusten olla voimassa. Täällä määrätään millä sisäänotto- ja ulosheitto-kriteereillä muuttujia käsitellään askeltavassa valinnassa.]

Tulokset ja niiden tulkinta

Seuraavassa käymme läpi regressioanalyysin tulokset tilanteessa, että meillä on kolme selittävää muuttujaa ja yksi selitettävä muuttuja. Regressiomenetelmänä käytetään tässä askeltavaa menettelyä (*Stepwise*). Logistisen regression yhteydessä luvussa 3.2 teemme pakotetun mallituksen. Taulujen selitykset eivät oleellisesti poikkea erilaisia regressiomenettelyjä käytettäessä. Halukas lukija voi konsultoida esimerkiksi SPSS-manuaalia (1999a, 205 - 230) saadakseen laajemman kuvan erilaisista tulkintavaihtoehdoista.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	LOGVUO	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	LOGMER	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
3	LOGAIK	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: TAI

Ensimmäinen taulu kertoo, missä **järjestyksessä muuttujat on otettu mukaan malliin** (*Variables Entered/Removed*). Osoittautuu, että kaikki mukaan valitut muuttujat ovat tilastollisessa mielessä hyviä selittäjiä – tätähän olimme tietenkin varmistaneet valitsemalla mukaan muuttujia, jotka korreloivat muuttujan TAI kanssa. SPSS-ohjelmisto käsittelee kaikkia löytämiään erilaisia regressiomalleja erillisinä malleina (*Model*); viimeinen malli numero 3 on siis $TAI = C(vakio) + \beta_1 LOGVUO + \beta_2 LOGMER + \beta_3 LOGAIK + \varepsilon$. *Method* -sarakkeessa kerrataan, millä sisäänottokriteerillä (F-testin p-arvo ≤ 0.05) ja poisheittokriteerillä ($p \geq 0.1$) menettely on toiminut.

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,616[a]	,379	,378	,701	,379	440,547	1	722	,000
2	,637[b]	,406	,405	,685	,027	33,317	1	721	,000
3	,643[c]	,414	,411	,682	,007	8,913	1	720	,003

a. Predictors: (Constant), LOGVUO

b. Predictors: (Constant), LOGVUO, LOGMER

c. Predictors: (Constant), LOGVUO, LOGMER, LOGAIK

d. Dependent Variable: TAI

Mallien hyvyyttä kuvataan **mallien yhteenveto** -taulukossa (*Model Summary*). Jokaisessa mallissa on mukana vakiotermi (*Constant*) sekä erinäinen joukko muita muuttujia. R-sarake (multippelikorrelaatiokerroin) kertoo selittävien muuttujien muodostaman muuttujajoukon ja selitettävän muuttujan välisen yhteiskorrelaation. Tätä tärkeämpi on kuitenkin multippelikorrelaatiokertoimen neliö R^2 (*R Square*), joka kertoo mallin selitysasteen eli sen, kuinka monta prosenttia muuttujat yhdessä selittävät TAI-muuttujasta. Nyt kaikki kolme tekijää selittävät taitotasosta yhteensä vain n. 40 % (tarkalleen 41.4 %), mitä ei voi pitää kovin suurena osuutena. Koetusta taitotasosta jää siis selittymättä 59.6 %. Otoskoolla ja selittäjien määrällä korjattu R^2 (*Adjusted R Square*) on hieman pienempi (koko mallille 0.411) kuin korjaamaton. Olimme lisäksi pyytäneet R:n muutosta koskevia tietoja. Taulun loppuosa kertoo, että muuttujien lisääminen on tilastollisessa mielessä vakuuttavasti R^2:n arvoa kasvattavaa. Vaikka viimeisen AIK-muuttujan lisääminen ei kasvata selitysastetta kuin 0.7 %, on muutos kuitenkin tilastollisesti merkitsevä ($p = 0.003$).