5. REGRESSIOANALYYSI

5.1 YHDEN SELITTÄJÄN REGRESSIOMALLIT

Yhden selittäjän regressiomallia on käsitelty Kirjassa 1. Tähän malliin liittyvät kaavat on esitetty myös tämän luvun lopussa.

5.2 USEAN SELITTÄJÄN REGRESSIOMALLIT

5.2.1 JOHDANTO

Riippuvuuksien tutkiminen on tieteen tärkeimpiä tehtäviä. Tilastotieteessä riippuvuuksia analysoidaan mm. *regressiomallien* avulla. Regressioanalyysi käsittää laajan joukon menetelmiä, joissa pyritään mallittamaan jonkin muuttujan, ns. *selitettävän muuttujan* eli *vastemuuttujan* riippuvuutta toisista muuttujista, ns. *selittävistä muuttujista* eli *selittäjistä* sopivilla selittäjien funktioilla. On syytä huomata, että selitettävän ja selittävän muuttujan välinen ero on regressiomalleja sovellettaessa perustavaa laatua. Regressioanalyysin päämääränä on selvittää riippuvuuden muoto ja voimakkuus muuttujista kerätyn havaintoaineiston perusteella. Analyysin eräs osatehtävistä on selvittää riippuuko selitettävä muuttuja todellakin niistä selittäjistä, joista sen on ajeteltu riippuvan.

Seuraavassa tarkastellaan sellaisia tilanteita, joissa selitettävän muuttujan ja selittäjien välistä riippuvuutta kuvataan *lineaarisella funktiolla*.

Oletetaan, että tehtävänä on selvittää miten selitettävän muuttujan Y (satunnaismuuttuja) keskimääräinen arvo riippuu selittäjien $x_1, x_2, \ldots, x_p$ arvoista. Oletamme, että selitettävän muuttujan Y *odotusarvo* $E(Y) = \mu_Y$ on selittävien muuttujien $x_1, x_2, \ldots, x_p$ arvojen *lineaarinen funktio*:

$$E(Y) = \mu_Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

jossa kertoimia $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ sanotaan *regressiokertoimiksi*. Koska regressiokertoimien arvoja ei tavallisesti tunneta, regressioanalyysin eräs osatehtävistä on kertoimien estimointi eli arviointi havaintojen perusteella.

Jos yo. yhtälö pätee, vastemuuttujan arvon eli *vasteen* Y ja selittäjien arvojen $x_1, x_2, \ldots, x_p$ välillä vallitsee *lineaarinen tilastollinen riippuvuus*. Nimitys lineaarinen johtuu siitä, että yo. lausekkeen oikea puoli määrittelee *tason* yhtälön p-ulotteisessa avaruudessa ja taso on esimerkki lineaarisesta funktiosta.[1]

---

[1] *Lineaarinen* = suoraviivainen tai viivallinen.

Yllä esitetty yhtälö on vastemuuttujan ja selittäjien välinen *regressioyhtälö* perusjoukossa. Yhtälöä ei voida havaita suoraan, koska Y-muuttujan havaitut arvot eli vasteet vaihtelevat niiden odotusarvon μ_Y ympärillä. On syytä huomata, että vastemuuttujan Y odotusarvoa tarkastellaan sillä ehdolla, että selittäjät ovat saaneet arvot $x_1, x_2, \ldots, x_p$. Jokainen selittäjien arvojen yhdistelmä määrittelee Y:n mahdollisten arvojen muodostaman perusjoukon osajoukon. Vaste Y vaihtelee näin määritellyissä osajoukoissa siten, että Y:n odotusarvona on yllä määritelty selittäjien arvojen lineaarinen funktio.

Lineaarisen mallin formulaatioossa on mahdollista ottaa huomioon myös sellainen tilanne, jossa selittäjät ovat *satunnaismuuttujia*, mikä onkin tavanomainen tilanne. Selitettävän muuttujan odotusarvoa tarkastellaan tällöin *ehdollisesti* selittäjien havaittujen arvojen suhteen, mikä merkitsee selittäjien arvojen *kiinnittämistä* havaittuihin arvoihinsa. Kiinnittäminen johtaa sihen, että selittäjien arvot voidaan tulkita *ei-satunnaisiksi*. Joissakin tilanteissa selittäjien arvot ovat aidosti ei-satunnaisia. Tämä merkitsee yleensä sitä, että selittäjien arvot on voitu *kiinnittää* tai *valita* halutuiksi. Tämä on tavallisesti mahdollista vain *koeasetelmien* yhteydessä ja tällöinkään ei ole yleensä mahdollista valita kaikkia selitettävään muuttujaan vaikuttavien tekijöiden arvoja.

Kertoimia $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ kutsutaan siis *regressiokertoimiksi*. Erityisesti kerrointa β_0 kutsutaan regressioyhtälön *vakioksi*. Vakio kuvaa selitettävän muuttujan odotettavissa olevaa arvoa, jos kaikilla muilla selittäjillä on arvo 0. Kertoimilla β_i, $i = 1,2,\ldots,p$ on seuraava tulkinta: Tarkastellaan esimerkiksi kerrointa β_1. Oletetaan, että selittäjillä on kiinteät arvot $x_1, x_2, \ldots, x_p$. Oletetaan, että kerrointa β_1 vastaavan selittäjän arvo x_1 kasvaa yhdellä yksiköllä, ts.

$$x_1 \rightarrow x_1 + 1$$

kaikkien muiden selittäjien arvojen pysyessä ennallaan. Tällöin selitettävän muuttujan Y odotusarvo muuttuu β_1 yksiköllä, koska

$$\begin{aligned}\mu_Y &\rightarrow \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \ldots + \beta_p x_p \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \beta_1 \\ &= \mu_Y + \beta_1.\end{aligned}$$

Regressiokerroin kuvaa siis sitä, miten selitettävän muuttujan odotettavissa oleva arvo muuttuu, jos kerrointa vastaavan selittäjän arvo kasvaa yhdellä yksiköllä ja kaikkien muiden selittäjien arvot pysyvät samaan aikaan muuttumattomina. Jos regressiokerroin $\beta_i = 0$, vaste ei riipu vastaavasta selittäjästä.

Seuraavassa tarkastellaan miten tämä matemaattinen malli voidaan liittää *havaintoihin* tilastolliseksi malliksi.

Oletetaan, että selitettävästä muuttujasta Y ja selittäjistä $x_1, x_2, \ldots, x_p$ on käytettävissä havaintoyksiköitä $i = 1,2,\ldots,n$ koskevat havaintoarvot:

Havaintoyksikkö 1: $Y_1, x_{11}, x_{12}, \ldots, x_{1p}$,

Havaintoyksikkö 2: $Y_2, x_{21}, x_{22}, \ldots, x_{2p}$,

...

Havaintoyksikkö n: $Y_n, x_{n1}, x_{n2}, \ldots, x_{np}$.

Siten

Y_j = selitettävän muuttujan Y arvo havaintoyksikössä j,

x_{ji} = selittäjän x_i arvo havaintoyksikössä j,

n = havaintojen lukumäärä,

p = selittäjien lukumäärä.

Havainnot toteuttavat vain poikkeuksellisesti yllä esitetyn perusjoukon regressioyhtälön. Poikkeamat tästä yhtälöstä vaihtelevat satunnaisesti havainnosta toiseen. Tekemällä poikkeamista sopivia oletuksia voidaan havainnoille muodostaa tilastollinen malli, jota kutsutaan *lineaariseksi regressiomalliksi*.

LINEAARISTA REGRESSIOMALLIA KOSKEVAT STANDARDIOLETUKSET

Oletetaan, että havaintojen välillä vallitsee yhtälö

$$Y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_p x_{jp} + \varepsilon_j, \qquad j = 1,2,\ldots,n,$$

jossa

ε_j = *poikkeama* eli *virhe-* eli *jäännöstermi* havainnossa j.

Vastemuuttujan Y arvot Y_j voidaan jakaa tämän yhtälön mukaan *rakenneosaan* ja *jäännökseen* niin, että

$$Y_j = \textit{rakenne} + \textit{jäännös},$$

jossa

$$\textit{rakenne} = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_p x_{jp}$$

on selittäjien arvon lineaarinen funktio ja kuvaa *systemaattista osaa* havaintoarvossa Y_j ja

$$\textit{jäännös} = \varepsilon_j$$

kuvaa *satunnaista osaa* havaintoarvossa Y_j.

Selittäjistä tehdään seuraavat oletukset:

1. Selittäjien arvot ovat kiinteitä eli ei-satunnaisia.

   Tästä oletuksesta on mahdollista luopua, kunhan jäännöstermin ominaisuuksia koskevat oletukset 3 ja 4 muunnetaan sopivaan muotoon. Näitä muunnettuja oletuksia tarvitaan erityisesti aikasarja-aineistojen yhteydessä, mutta myös monissa muissa sovelluksissa selittäjien arvot ovat satunnaisia. Sivuutamme tässä näiden muunnettujen oletusten tarkemman käsittelyn.

2. Minkään selittäjän arvot eivät riipu lineaarisesti toisten selittäjien arvoista.

   Oletus 2 on tekninen ehto, joka takaa sen, että regressiokertoimille voidaan määrätä pienimmän neliösumman estimaattorit tavanomaiseen tapaan. Ehdosta seuraa mm. se, että sama selittäjä ei saa esiintyä kahdesti selittäjien joukossa ja

se, että minkään selittäjän arvoja ei voida lausua kahden muun selittäjän arvojen (painotettuna) summana. Jos ehto ei jostakin syystä päde, joutuvat kertoimien estimoinnissa käytetyt tilasto-ohjelmat tavallisesti vaikeuksiin. Missään tavanomaisessa tapauksessa ei ole pelkoa siitä, ettei ehto pätisi.

Virhetermistä ε_j tehdään seuraavat oletukset:

3. Virhetermin ε_j odotusarvo on nolla:

$$E(\varepsilon_j) = 0 \text{ kaikille } j = 1,2,\ldots,n.$$

Ehto takaa sen, että selitettävän muuttujan arvojen eli vasteiden odotusarvot ovat muotoa

$$E(Y_j) = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_p x_{jp} \text{ kaikille } j = 1,2,\ldots,n.$$

Tämä merkitsee sitä, että havaintoarvon Y_j odotusarvo yhtyy mallin *systemaattiseen osaan.*

4. Virhetermin ε_j varianssi eli *jäännösvarianssi* on vakio:

$$D^2(\varepsilon_j) = \sigma^2 \text{ kaikille } j = 1,2,\ldots,n.$$

Ehdosta seuraa se, että

$$D^2(Y_j) = \sigma^2.$$

Tämä merkitsee sitä, että havaintoarvojen Y_j keskittyneisyys oman odotusarvonsa $E(Y_j)$ ympärille on kaikille havainnoille j sama.

Oletusta 4 kutsutaan usein *homoskedastisuusoletukseksi.* Jos se ei päde sanotaan, että virhetermit ovat *heteroskedastisia.* Homoskedastisuusoletuksen mukaan virhetermien jakauma keskittyy

Koska σ^2 on tavallisesti tuntematon, regressioanalyysin eräs osatehtävistä on jäännösvarianssin estimointi havaintojen perusteella.

5. Virhetermit ε_j ovat keskenään korreloimattomia.

Oletus 5 on merkityksellinen lähinnä aikasarja-aineistojen yhteydessä. Sivuutamme ehdon tarkemman käsittelyn tässä yhteydessä.

Oletuksia 1.—5. kutsutaan lineaariseen regressiomalliin liittyviksi *standardioletuksiksi.*

Oletuksiin 3 ja 4 liitetään usein normaalisuusoletus:

6. Virhetermi ε_j on jakautunut normaalijakauman mukaan:

$$\varepsilon_j \sim N(0, \sigma^2).$$

On syytä huomata, että jos oletukset 1.—6. eivät päde saattaa olla syytä käyttää toisenlaisia tilastollisia menetelmiä, kuin niitä, joita käsitellään alla.

Lineaarisen regressiomallin *parametreja* ovat regressiokertoimet $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ ja jäännösvarianssi σ^2. Regressioanalyysin päätehtäviä on näiden parametrien estimointi ja niitä koskevien hypoteesien testaaminen. Lisäksi analyysiin pitää aina liittää mallista tehtyjen oletusten 1.—6. tarkistaminen. On hyvä tietää, että standardioletuksia voidaan tutkia graafisesti tai tekemällä sopivia tilastollisia testejä. Standardioletusten tarkistamista kutsutaan usein *diagnostiikaksi.* Sivuutamme

diagnostisten menetelmien kuvailun tässä esityksessä. Jos malli toteuttaa siitä tehdyt oletukset ja malli on muutenkin järkevä, estimoitua mallia voidaan käyttää paitsi selitettävän muuttujan ja selittäjien välisen riippuvuuden *kuvaamiseen*, selitettävän muuttujan arvojen *ennustamiseen* ja jopa selitettävän muuttujan arvojen *kontrollointiin*.

ESIMERKKI 1.

Alkoholin kulutusta tutkittaessa formuloidaan usein regressiomalli, jolla pyritään selittämään alkoholin vuosikulutusta absoluuttialkoholilitroina *per capita*[1], kun selittäjinä ovat alkoholin *reaalihinta* ja *reaalinen vuositulo per capita*.[2]

Eräs sellainen malli on muodoltaan seuraava:

$$Y_j = \beta_0 + \beta_1 p_j + \beta_2 Q_j + \varepsilon_j, \qquad j = 1,2,\ldots,n,$$

jossa

Y_j = alkoholin kulutus absoluuttialkoholina (l) per capita vuonna j,

p_j = alkoholin reaalihinta (mk) vuonna j,

Q_j = reaalitulo per capita (mk) vuonna j.

Oletetaan, että malli toteuttaa standardioletukset sopivasti muunnettuina: Alkoholin *nimellishinta* on ALKO:n hallintoneuvoston valittavissa ja on siten ei-satunnainen. Sen sijaan reaalihinta on satunnaismuuttuja, koska elinkustannusindeksi on satunnaismuuttuja. Samoin reaalitulo on satunnaismuuttuja.

Tällaista mallia voidaan käyttää sekä alkoholin kulutuksen *ennustamiseen* että *kontrollointiin*: Jos reaalitulojen ennustetaan pienenevän seuraavana vuonna, voidaan mallin perusteella ennustaa paljonko kulutus tulee muuttumaan. Jos alkoholin kulutusta halutaan vähentää tietyn määrän, voidaan mallin perusteella määrätä paljonko alkoholin hintaa on muutettava, jotta tavoite saavutetaan.

Regressiokertoimilla on seuraavat tulkinnat:

β_0: Kerroin ilmaisee paljonko alkoholia kulutettaisiin, jos alkoholin reaalihinta ja reaalitulot saisivat arvon 0.

β_1: Kerroin ilmaisee paljonko alkoholin kulutus muuttuu, jos alkoholin reaalihinta kasvaa yhdellä markalla.

β_2: Kerroin ilmaisee paljonko alkoholin kulutus muuttuu, jos reaalitulot kasvavat yhdellä markalla.

---

[1] *Per capita* tarkoittaa henkilöä kohti.

[2] *Reaalihinnalla* tarkoitetaan hintaa, jossa on otettu huomioon inflaatio eli rahan arvon huononeminen. Reaalihinta saadaan *nimellishinnan* ja elinkustannusindeksin suhteena. Vastaavasti reaalitulolla tarkoitetaan tuloa, jossa on otettu huomioon inflaatio eli rahan arvon huononeminen. Reaalitulo saadaan *nimellistulon* ja elinkustannusindeksin suhteena.

Huomaa, että regressiokertoimien merkeistä on käytettävissä *priori-* eli *ennakkotietoa*:

$\beta_1 \leq 0$:

> Kerroin β_1 on *negatiivinen*, koska minkä tahansa tuotteen reaalihinnan nousu vähentää tuotteen kulutusta.

$\beta_2 \geq 0$:

> Kerroin β_2 on luultavasti *positiivinen*, koska reaalitulojen nousu lisää sellaisten tuotteiden kulutusta kuin alkoholi.

Regressioanalyysin tarkoituksena on mm. estimoida regressiokertoimien arvot ja tarkistaa ovatko kertoimien merkit ennakkokäsitysten mukaisia. Tällainen järkevyystarkistus on keskeisessä asemassa regressioanalyysissa. Jos hinta-muuttujaa vastaava regeressiokertoimen β_1 estimaatti saa positiivisen merkin, ollaan pahoissa vaikeuksissa, vaikka malli olisi *tilastollisilta ominaisuuksiltaan* kuinka hyväksyttävä tahansa. ●

5.2.2 LINEAARISEN REGRESSIOMALLIN ESTIMOINTI

Lineaarisen regressiomallin

$$Y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_p x_{jp} + \varepsilon_j, \qquad j = 1,2,\ldots,n,$$

parametrit estimoidaan tavallisesti *pienimmän neliösumman keinolla*, lyh. PNS-keinolla. Regressiokertoimien $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ PNS-estimaattorit saadaan minimoimalla neliösumma

$$\sum_{j=1}^{n} \varepsilon_j^2 = \sum_{j=1}^{n} (Y_j - \beta_0 - \beta_1 x_{j1} - \beta_j x_{j2} - \cdots - \beta_p x_{jp})^2$$

parametrien $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ suhteen. Jos virhetermit ε_j ovat normaalijakautuneita, *suurimman uskottavuuden* menetelmä tuottaa parametreille $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ samat estimaattorit kuin PNS-menetelmä. Sivuutamme tässä kokonaan regressiokertoimien estimaattien laskukaavat. Alla tosin esitetään yhden selittäjän malliin liittyvät kaavat.

Oletetaan, että PNS-estimaattorit kertoimille $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ ovat $b_0, b_1, b_2, \ldots, b_p$. Niiden avulla voidaan määritellä seuraavat käsitteet:

Sovite:

$$\hat{Y}_j = b_0 + b_1 x_{j1} + b_2 x_{j2} + \ldots + b_p x_{jp}, \qquad j = 1,2,\ldots,n.$$

Sovite on estimoidun mallin selitettävälle muuttujalle Y antama, selittäjien arvoja $x_{j1}, x_{j2}, \ldots, x_{jp}$ vastaava arvo. $\hat{Y}_j$ luetaan y-j-hattu.

Residuaali:

$$e_j = Y_j - \hat{Y}_j = Y_j - b_0 - b_1 x_{j1} - b_2 x_{j2} - \ldots - b_p x_{jp}, \qquad j = 1,2,\ldots,n.$$

Residuaali on selitettävän muuttujan havaitun arvon Y_j ja estimoidun mallin selitettävälle muuttujalle antaman arvon $\hat{Y}_j$ erotus. Residuaalia voidaan pitää jäännöstermin ε_j havaittuna vastineena, jos malli on oikein muodostettu.

Lienee ilmeistä, että malli *selittää* selitettävän muuttujan arvojen vaihtelun sitä paremmin, mitä lähempänä sovite on selitettävän muuttujan havaittuja arvoja eli mitä pienempiä ovat residuaalit.

Virhetermin varianssi eli jäännösvarianssi σ^2 voidaan estimoida residuaalien avulla: Voidaan osoittaa, että

$$s^2 = \frac{1}{n-p-1} \sum_{j=1}^{n} e_j^2$$

on jäännösvarianssin σ^2 harhaton estimaattori. s^2 on residuaalien varianssi ja kuvaa siten residuaalien vaihtelua niiden keskiarvon ympärillä. Tulkinnan kannalta on tärkeätä huomata se, että s^2 kuvaa selitettävän muuttujan havaittujen arvojen vaihtelua *regressiotason*

$$y = b_0 - b_1 x_1 - b_2 x_2 - \ldots - b_p x_p$$

ympärillä.

Se, että edellä esitetty kaava s^2:lle on todellakin residuaalien *varianssi*, seuraa siitä, että residuaalien aritmeettinen keskiarvo

$$\bar{e} = \frac{1}{n}\sum_{j=1}^{n} e_j = 0$$

aina, kun mallissa on vakio.

Kuten tunnettua selitettävän muuttujan arvojen vaihtelun mittaaminen voidaan perustaa neliösummaan

$$SST = \sum_{j=1}^{n} (Y_j - \bar{Y})^2,$$

jossa

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

on selitettävän muuttujan arvojen otoskeskiarvo. Neliösummaa *SST* kutsutaan *kokonaisneliösummaksi*, koska se kuvaa vasteiden vaihtelua niiden keskiarvon ympärillä. Selitettävän muuttujan Y havaittujen arvojen $Y_1, Y_2, \ldots, Y_n$ otosvarianssi s_Y^2 voidaan määritellä kokonaisneliösumman *SST* avulla seuraavalla kaavalla:

$$s_Y^2 = \frac{SST}{n-1}.$$

Residuaalien vaihtelun mittaaminen taas voidaan perustaa neliösummaan

$$SSE = \sum_{j=1}^{n} e_j^2.$$

Neliösummaa *SSE* kutsutaan *jäännösneliösummaksi*. Jäännösvarianssin estimaattori s^2 voidaan määritellä jäännösneliösumman *SSE* avulla seuraavalla kaavalla:

$$s^2 = \frac{SSE}{n-p-1}.$$

Regressioanalyysin keskeisiä tuloksia on se, että jäännösneliösumma on aina korkeintaan yhtä suuri kuin kokonaisneliösumma:

$$SSE \leq SST.$$

Erotusta

$$SSM = SST - SSE$$

kutsutaan *regressioneliösummaksi* tai *mallineliösummaksi*. Voidaan osoittaa, että

$$SSM = \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2.$$

Jäännösneliösumman ja kokonaisneliösumman välinen epäyhtälö mahdollistaa sen, että suuretta

$$R^2 = 1 - \frac{SSE}{SST}$$

voidaan käyttää mallin *selitysasteen* mittaamiseen. Selitysasteella R^2 on seuraavat ominaisuudet:

1. $0 \leq R^2 \leq 1.$

2. Jos kaikki residuaalit ovat nollia, jolloin malli sopii havaintoihin täydellisesti,

   $R^2 = 1.$

3. Jos kaikki residuaalit ovat muotoa $Y_j - \overline{Y}$, jäännösneliösumma yhtyy kokonaisneliösummaan ja

   $R^2 = 0.$

   Tällöin malli ei selitä ollenkaan selitettävän muuttujan arvojen vaihtelua.

Selitysaste kuvaa siis sitä, miten hyvin malli selittää selitettävän muuttujan vaihtelua.

REGRESSIOKERTOIMIEN OTOSJAKAUMA

Regressiokertoimien PNS-estimaattorit ovat regressiomallin standardioletuksien 1—5 pätiessä *harhattomia*:

$$E(b_i) = \beta_i.$$

Voidaan osoittaa, että standardioletuksien 1—5 pätiessä regressiokertoimien estimaattorit ovat suurissa otoksissa approksimatiivisesti normaalisia $N(\beta_i, \sigma^2_{b_i})$, jossa $\sigma^2_{b_i}$ on kerroinestimaattorin b_i varianssi:

$$b_i \sim_a N(\beta_i, \sigma^2_{\beta_i}).$$

Jos jäännöstermit ovat normaalisia eli, jos standardioletukset 1—6 pätevät, regressiokertoimen estimaattorin jakaumaa koskeva tulos on eksakti eli tarkka.

REGRESSIOKERTOIMIEN LUOTTAMUSVÄLI

Rergressiokertoimien estimaattoreiden varianssit voidaan estimoida havainnoista. Sivuutamme tässä varianssien laskukaavat. Merkitään kertoimen b_i estimoitua varianssia symbolilla $s_{b_i}^2$. Kertoimen β_i luottamusväli luottamustasolla $1-\alpha$ on standardioletusten 1—5 pätiessä muotoa

$$b_i \pm z s_{b_i},$$

jossa z on luottamustasoa $1-\alpha$ vastaava *luottamuskerroin*. Luottamuskerroin z saadaan normaalijakaumasta, jos regressiokertoimen otosjakauma on suurissa otoksissa approksimatiivisesti normaalinen. Jos regressiokertoimen otosjakauma on eksaktisti normaalinen, z saadaan t-jakaumasta, jossa vapausasteiden lukumäärä on $n-p-1$.

5.2.3 REGRESSIOMALLIIN LIITTYVÄT TESTIT

Käsittelemme seuraavassa regressiomallin kertoimille tehtäviä testejä. Regressiokertoimille on aina tapana tehdä seuraavat testit:

1. *Yleistestinä* kertoimille käytetään testiä, jolla tutkitaan auttavatko malliin valitut *selittävät muuttujat yhdessä* selittämään selitettävän muuttujan arvojen vaihtelua. Tällä testillä tutkitaan *onko mallilla ilmaistua regressiota olemassa.*
2. Jokaiselle kertoimelle on tapana tehdä *kerroinkohtainen testi*, jolla tutkitaan auttaako *se selittävä muuttuja*, jonka kerrointa testataan, selittämään selitettävän muuttujan arvojen vaihtelua.

ONKO REGRESSIOTA?

Yleistestissä regressiokertoimille testataan nollahypoteesia

$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_p = 0.$$

Jos H_0 pätee, selitettävä muuttuja Y ei riipu lineaarisesti *yhdestäkään* selittäjästä $x_1, x_2, \ldots, x_p$. Tässä on esimerkki tilanteesta, jossa tavallisesti toivotaan nollahypoteesin kumoutumista.

Nollahypoteesia H_0 voidaan testata testisuureella

$$F = \frac{n-p-1}{p} \cdot \frac{SSM}{SSE}$$

$$= \frac{n-p-1}{p} \cdot \frac{R^2}{1-R^2},$$

jossa SSM on *mallineliösumma*, SSE on *jäännösneliösumma* ja R^2 on *selitysaste* (kts. edellinen kappale). Testisuure F on jakautunut kuten F-jakauma vapausastein p ja $n-p-1$, *jos nollahypoteesi* H_0 *pätee* ja, jos standardioletukset 1—6 pätevät. Vaikka normaalisuusoletus 6 ei olisikaan tosi, tulos pätee approksimatiivisesti suurissa otoksissa, jos regressiokertoimien jakauma on approksimatiivisesti normaalinen.

Nollahypoteesin pätiessä testisuureen odotusarvo on

$$E(F) = \frac{n-p-1}{n-p-3} \approx 1.$$

Jos testisuure saa odotettavissa olevaa arvoaan kyllin paljon suurempia arvoja, nollahypoteesi joudutaan hylkäämään.

Testiä voidaan pitää *varianssien vertailutestinä*, joka perustuu *jäännösvarianssin*

$$s^2 = \frac{1}{n-p-1} SSE$$

ja *mallivarianssin*

$$\frac{1}{p} SSM$$

vertailuun.

YKSITTÄISEN SELITTÄJÄN MERKITYS

Edellisessä kappaleessa esitetyn yleistestin *jälkeen* on tapana tutkia jokaisen selittäjän merkitystä mallissa. Testattavat nollahypoteesit ovat muotoa

$$H_{0i}: \quad \beta_i = 0, \qquad i = 1,2,\ldots p.$$

Jos nollahypoteesi H_{0i} *pätee*, selitettävä muuttuja Y ei riipu lineaarisesti selittäjästä x_i. Tällöin selittäjä x_i on mallissa turha ja se voidaan poistaa mallista.

Muodostetaan testisuure

$$t_i = \frac{b_i}{s_{b_i}},$$

jossa b_i on kertoimen β_i PNS-estimaattori ja s_{b_i} on estimaattorin b_i hajonnan estimaattori. Testisuuretta t_i kutsutaan tavallisesti kertoimen β_i *t-arvoksi*.

Testisuure t_i on jakautunut suurissa otoksissa approksimatiivisesti kuten normaalijakauma, *jos nollahypoteesi pätee* ja, jos estimaattorin b_i otosjakauma on suurissa otoksissa approksimatiivisesti normaalinen. Testisuure t_i on jakautunut kuten t-jakauma $(n-p-1)$:llä vapausasteella, *jos nollahypoteesi pätee* ja, jos estimaattorin b_i otosjakauma on eksaktisti normaalinen. Nollahypoteesin pätiessä testisuureen odotusarvo on

$$E(t_i) = 0.$$

Jos kertoimen β_i t-arvo poikkeaa nollasta kyllin paljon, joudutaan nollahypoteesi hylkäämään.

Selittäjät voidaan jakaa niiden regressiokertoimien t-arvojen avulla *tilastollisesti merkityksellisiin ja merkityksettömiin*. Valitaan sopiva, pieni todennäköisyys α.[1] Jos regressiokertoimen β_i t-arvoa t_i vastaava P-arvo on *suurempi* kuin

[1] Yleensä $\alpha = 0.05$ tai 0.01.

α, sanotaan, että kerrointa β_i vastaavaa selittäjä x_i on tilastollisesti *merkityksetön*. Jos taas regressiokertoimen β_i t-arvoa t_i vastaava *P*-arvo on *pienempi* kuin α, sanotaan, että kerrointa β_i vastaava selittäjä x_i on tilastollisesti *merkitsellinen*.

Tavallisesti regrssiomalleja käytetään sellaisissa tilanteissa, joissa selittäjiksi on tarjolla useita *selittäjäkandidaatteja* ja tutkimus kohdistuu siihen, mistä kandidaateista selitettävä muuttuja Y todellakin riippuu.[1] Malliin pyritään valitsemaan ne selittäjät, jotka auttavat selittämään vastemuuttujan Y arvojen vaihtelua kerroinkohtaisten t-testien avulla. Selittäjien valinta voidaan tehdä *askeltaen*. Käytössä on useita erilaisia *askellusstrategioita*. Seuraavassa kuvataan kahta yleisimmin käytettyä strategiaa, joita kutsutaan suuntansa takia *alaspäiseksi* ja *ylöspäiseksi askellukseksi*. Alaspäisessä askelluksessa mallista poistetaan merkityksettömiä selittäjiä yksi kerrallaan, kunnes saadaan malli, jossa kaikki selittäjät ovat merkityksellisiä. Ylöspäisessä askelluksessa malliin lisätään merkityksellisiä selittäjiä yksi kerrallaan, kunnes saadaan malli, jonka ulkopuolelle jätettyjen selittäjäkandidaattien joukossa ei ole enää merkityksellisiä selittäjiä.

ALASPÄINEN ASKELLUS

Lähdetään liikkeelle mallista, jossa ovat mukana kaikki selittäjäkandidaatit ja estimoidaan mallin kertoimet. Poistetaan mallista turhat selittäjät vaiheittain siten, että jokaisessa vaiheessa poistetaan tilastollisesti *merkityksettömistä* selittäjistä se, jota vastaava t-arvo on itseisarvoltaan *pienin*. Jokaisen poiston jälkeen malli estimoidaan uudelleen. Tätä jatketaan kunnes on saatu malli, jossa kaikki selittäjät ovat tilastollisesti merkitseviä. Näin saadulle mallille tehdään lopuksi kertoimien merkin ja suuruuden järkevyystarkastelu.

YLÖSPÄINEN ASKELLUS

Lähdetään liikkeelle mallista, jossa on mukana pelkkä vakiotermi. Estimoidaan kaikki mahdolliset *yhden* selittäjän mallit ja valitaan malliin ensimmäisenä *merkityksellisistä* selittäjäkandidaateista se, jota vastaava t-arvo on itseisarvoltaan *suurin*. Estimoidaan tämän jälkeen kaikki mahdolliset *kahden* selittäjän mallit, joissa on mukana ensimmäisenä valittu selittäjä ja valitaan malliin toisena merkityksellisistä selittäjäkandidaateista se, jota vastaava t-arvo on itseisarvoltaan suurin. Tehdään sama kaikille mahdollisille kolmen selittäjän malleille, joissa on mukana kaksi ensimmäisenä valittua selittäjää. Tätä jatketaan kunnes on saatu malli, jonka ulkopuolelle jätetyistä selittäjistä yksikään ei ole merkityksellinen. Näin saadulle mallille tehdään lopuksi kertoimien merkin ja suuruuden järkevyystarkastelu.

Ylöspäisen askelluksen tuloksena saatavan mallin kaikki selittäjät eivät ole välttämättä tilastollisesti merkityksellisiä. Tämä huomio on johtanut sellaisten

---

[1] Huomaa, että regressiomallin avulla mallitettavan *tilastollisen riippuvuuden* olemassaolosta ei saa tehdä johtopäätöksiä *kausaalisen riippuvuuden* olemassaolosta.

valikointimenetelmien kehittämiseen, joissa sovelletaan jollakin strategialla askellusta molempiin suuntiin. Sivuutamme tällaisten menetelmien kuvaamisen tässä yhteydessä. Todettakoon kuitenkin, että monet tilasto-ohjelmistot sisältävät mahdollisuuden soveltaa *automaattista askellusta*. On kuitenkin syytä tietää, että *automaattisen askelluksen tulos ei ole millään muotoa objektiivinen.*

Todettakoon vielä lopuksi seuraavat seikat: Askelluksen tulos riippuu todennäköisyydestä α. Lisäksi erilaisten askellusstrategioiden käyttö saattaa johtaa erilaisiin malleihin. Askelluksen tuloksena syntyvälle mallille on aina tehtävä kertoimia koskevat järkevyystarkistukset ja standardioletuksien vaatimat diagnostiset tarkistukset. Jos malli havaitaan näissä tarkistuksissa joiltakin osiltaan puutteelliseksi, mallia saatetaan joutua korjaamaan monellakin tavalla. Tämä merkitsee sitä, että hyvän regressiomallin rakentaminen on vaativa tehtävä, johon täytyy tavallisesti uhrata paljon työtä ja aikaa.

ESIMERKKI 1.

Kulutustutkimus on eräs empiirisen taloustieteen osa-alueista. Yksityiset kulutusmenot voidaan jakaa sellaisiin ryhmiin kuten kulutusmenot ruokaan, vaatteisiin, kestokulutushyödykkeisiin, liikennevälineisiin jne. Kulutusmenoja voidaan kussakin ryhmässä selittää esimerkiksi hinnoilla ja käytettävissä olevilla tuloilla.

Seuraavassa tarkastellaan kulutusmenoja kotitalouden tekstiileihin. Kotitalouden kokonaiskulutusmenoista kuluu kotitalouden tekstiilien ostamiseen keskimäärin noin 1%.

Olkoon mallina

$$\log(q_j) = \beta_0 + \beta_1 \log(p_j) + \beta_2 \log(Q_j) + \varepsilon_j,$$

jossa

q_j = kulutusmenot kotitalouden tekstiileihin (kiinteisiin eli reaalisiin hintoihin) per capita (mk),

p_j = kotitalouden tekstiilien reaalihinta (mk),

Q_j = kokonaiskulutusmenot (kiinteisiin eli reaaliisiin hintoihin) per capita (mk).

Oletetaan lisäksi, että jäännnöstermit ε_j ovat riippumattomia ja normaalisti jakautuneita kaikille j:

$$\varepsilon_j \sim N(0, \sigma^2).$$

Kokonaiskulutusmenot edustavat mallissa tuloja. Huomaa, että sekä hintamuuttuja että kokonaiskulutusmenot ovat satunnaismuuttujia. Oletetaan, että jäännöstermistä ε_j voidaan tehdä sellaiset *muunnetut* oletukset, että selitettävän muuttujan *ehdollinen* odotusarvo hintojen ja kokonaiskulutusmenojen suhteen on muotoa

$$E(\log(q_j)) = \beta_0 + \beta_1 \log(p_j) + \beta_2 \log(Q_j).$$

Miksi mallin muuttujat on logaritmoitu? Tämä johtuu siitä, että tällöin regressiokertoimet voidaan tulkita ns. *joustoiksi*:

β_1 on *hintajousto*:

> Kerroin kuvaa *kuinka monta prosenttia* kulutusmenot kotitalouden tekstiileihin muuttuvat, kun niiden hinta kasvaa 1%:n (ja kokonaistulot pysyvät ennallaan).

β_2 on *kokonaiskulutusmenojen jousto* (tulojousto):

> Kerroin kuvaa *kuinka monta prosenttia* kulutusmenot kotitalouden tekstiileihin muuttuvat, kun tulot kasvavat 1%:n (ja tekstiilien hinta pysyy ennallaan).

Estimointitulokset perustuvat Tilastokeskuksen tuottamiin aikasarjoihin vuosilta 1950 — 1978 (n =29). Kulutusmenot ja hinnat on määrätty vuoden 1975 rahassa.

Estimointitulokset voidaan esittää seuraavana taulukkona:

Kerroin	Estimaatti	Hajonta
β_0	−10.3	1.72
β_1	0.136	0.190
β_2	1.55	0.0967

Mallin selitysaste on

$$R^2 = 0.980.$$

Kerroinestimaateista voidaan tehdä seuraavat johtopäätökset:

Jos tekstiilien hinta nousee 1%:n, tekstiilien kulutus nousee 0.136%.

Jos kulutusmenot nousevat 1%:n, tekstiilien kulutus nousee 1.55%.

Hinnan vaikutusta koskeva johtopäätös on järjen vastainen, mutta onneksi hinta ei ole tilastollisesti merkitsevä kulutusta määräävä tekijä, kuten alla nähdään.

Tarkastellaan ensin nollahypoteesia

$$H_0: \quad \beta_1 = \beta_2 = 0.$$

Jos nollahypoteesi pätee, vastemuuttuja $\log(q_i)$ ei riipu lineaarisesti kummastakaan selittäjästä.

Testisuure tälle nollahypoteesille on

$$\begin{aligned} F &= \frac{n-p-1}{p} \cdot \frac{R^2}{1-R^2} \\ &= \frac{29-2-1}{2} \cdot \frac{0.980}{1-0.980} \\ &\approx 637. \end{aligned}$$

Jos nollahypoteesi pätee, testisuure F on jakautunut kuten F-jakauma 2:lla ja 26:lla vapausasteella. Testisuureen arvoa vastaava P-arvo on 0 kaikilla tavanomaisilla lukutarkkuuksilla, ts.

$$P(F > 637) \approx 0.$$

Nollahypoteesi siis hylätään: Vaste riippuu molemmista selittäjistä yhdessä.

Tutkitaan seuraavaksi vasteen riippuvuutta yksittäisistä selittäjistä.

Tarkastellaan ensin nollahypoteesia

$$H_{01}: \quad \beta_1 = 0.$$

Testisuure tälle nollahypoteesille on

$$t_1 = \frac{b_1}{s_{b_1}}$$
$$= \frac{0.136}{0.190}$$
$$\approx 0.716.$$

Jos nollahypoteesi pätee, testisuure t_1 on jakautunut kuten t-jakauma 26:lla vapausasteella. Testisuureen arvoa vastaava P-arvo on 0.24. Siten nollahypoteesi voidaan jättää voimaan: Vaste ei riipu hintamuuttujasta.

Tarkastellaan toiseksi nollahypoteesia

$$H_{02}: \quad \beta_2 = 0.$$

Testisuure tälle nollahypoteesille on

$$t_2 = \frac{b_2}{s_{b_2}}$$
$$= \frac{1.55}{0.0967}$$
$$\approx 16.0.$$

Jos nollahypoteesi pätee, testisuure t_2 on jakautunut kuten t-jakauma 26:lla vapausasteella. Testisuureen arvoa vastaava P-arvo on 0 kaikilla tavallisilla lukutarkkuuksilla. Siten nollahypoteesi voidaan hylätä: Vaste riippuu tulomuuttujasta.

Yllä esitetyistä estimointituloksista voidaan siis päätellä, että kulutusmenot kotitalouden tekstiileihin riippuvat pelkästään käytettävissä olevista tuloista. Jos malli estimoidaan uudelleen muodossa

$$\log(q_j) = \beta_0 + \beta_2 \log(Q_j) + \varepsilon_j,$$

saadaan seuraavat estimointitulokset:

Kerroin	Estimaatti	Hajonta
β_0	–9.05	0.364
β_2	1.48	0.0407

Mallin selitysaste on

$$R^2 = 0.980.$$

Huomaa, että uudelleenestimoinnin tuloksena jäljelle jääneiden regressiokertoimien estimaatit muuttuivat. Jäljelle jääneiden regressiokertoimien estimaatit muuttuvatkin tavallisesti aina poistettiinpa mallista tai lisättiinpä malliin selittäjiä.

Tarkastellaan nyt nollahypoteesia

$$H_{02}: \quad \beta_2 = 0.$$

Testisuure tälle nollahypoteesille on

$$t_2 = \frac{b_2}{s_{b_2}}$$
$$= \frac{1.48}{0.0407}$$
$$\approx 36.4.$$

Jos nollahypoteesi pätee, testisuure t_2 on jakautunut kuten t-jakauma 27:llä vapausasteella. Testisuureen arvoa vastaava P-arvo on 0 kaikilla tavallisilla lukutarkkuuksilla. Siten nollahypoteesi voidaan hylätä: Vaste riippuu tulomuuttujasta.

Huomaa, että tässä tapauksessa ei ole tarvetta tehdä F-testiä yleishypoteesille, koska testin tulos on sama kuin ylläkuvatun t-testin. Tämä johtuu siitä, että yhden selittäjän tapauksessa

$$F = t^2.$$

Todettakoon vielä, että regressiomallia koskevat estimointitulokset esitetään usein seuraavassa muodossa:

$$\log(q_j) = -9.05 + 1.48 \cdot \log(Q_j) \qquad R^2 = 0.980$$
$$(0.364) \quad (0.0407)$$

Kerroinestimaattien alapuolella on ilmoitettu vastaavat hajonnat. Toinen paljon käytetty tapa on ilmoittaa kerroinestimaattien alla niitä vastaavat t-arvot.

95%:n luottamusväli kertoimelle β_2 on muotoa

$$1.48 \pm z \cdot 0.0407,$$

jossa z valitaan t-jakauman taulukosta niin, että

$$P(-z \leq t \leq +z) = 0.95,$$

jossa t-jakautuneen satunnaismuuttujan t vapausasteiden luku on 27. t-jakauman taulukoista saadaan

$$z = 1.703.$$

95%:n luottamusväli kertoimelle β_2 saa siis muodon

$$1.48 \pm 0.069$$

eli

$$[1.41, 1.55].$$

Olemme käyttäneet sekä testeissä, että luottamusvälin konstruktiossa t-jakaumaa. t-jakauman käyttö on luvallista, koska jäännöstermistä ε_t tehtiin normaalisuusoletus. Jos normaalisuusoletus ei päde, joudutaan vetoamaan suurten otosten approksimaatiotulokseen ja turvautumaan normaalijakauman

taulukoihin. Näin saatava luottamusväli on vain approksimatiivisesti 95%:n luottamusväli. ●

On syytä huomata, että edellisessä esimerkissä kiinnitettiin huomiota ainoastaan *regressiokertoimia koskeviin testeihin.* Näiden testien käyttö on tarkasti ottaen luvallista vain, jos mallista tehdyt standardioletukset 1—5 (ja 6) pätevät. Siksi regressioanalyysiin pitää aina liittää *diagnostiset testit* oletusten tarkistamiseksi. Sivuutamme kuitenkin sellaiset testit tässä yhteydessä.

5.2.4 YHDEN SELITTÄJÄN MALLI

Käsittelemme seuraavassa *yhden selittäjän lineaarisen regressiomallin* estimointiin liittyviä kaavoja.

Olkoon

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \qquad \varepsilon_j \sim N(0, \sigma^2), \qquad j = 1,2,\ldots,n.$$

Parametrien β_0 ja β_1 PNS-estimaattorit saadaan kaavoista

$$b_0 = \bar{Y} - b_1\bar{x},$$

$$b_1 = r\frac{s_Y}{s_x},$$

jossa

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

on vastemuuttujan havaintoarvojen aritmeettinen keskiarvo,

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

on selittäjän havaintoarvojen aritmeettinen keskiarvo,

$$s_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

on vastemuuttujan havaintoarvojen otosvarianssi,

$$s_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

on selittäjän havaintoarvojen otosvarianssi ja

$$r = \frac{s_{Yx}}{s_Y s_x}$$

on vastemuuttujan ja selittäjän havaintoarvojen otoskorrelaatiokerroin, jossa

$$s_{Yx} = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})(x_j - \bar{x})$$

on vastemuuttujan ja selittäjän havaintoarvojen otoskovarianssi.

Selitysaste R^2 yhtyy yhden selittäjän mallin tapauksessa vastemuuttujan ja selittäjän havaintoarvojen korrelaatiokertoimen neliöön:

$$R^2 = r^2.$$

Jäännöstermin varianssin σ^2 harhaton estimaattori saadaan esimerkiksi kaavasta

$$s^2 = \frac{n-1}{n-2}(1-r^2)s_Y^2.$$

Kertoimen b_1 varianssin harhaton estimaattori saadaan kaavasta

$$s_{b_1}^2 = \frac{s^2}{(n-1)s_x^2}.$$

Olkoon nollahypoteesina

$$H_0: \quad \beta_1 = 0.$$

Tavanomainen t-testisuure nollahypoteesille H_0 on muotoa

$$t = \frac{b_1}{s / \sqrt{(n-1)s_x^2}}.$$

Jos nollahypoteesi pätee ja jäännöstermi ε_i on normaalinen, testisuure t noudattaa t-jakaumaa vapausastein n–2. Jos jäännöstermistä ei voida tehdä normaalisuusoletusta, joudutaan soveltamaan suurten otosten tulosta, jonka mukaan testisuure t noudattaa nollahypoteesin pätiessä suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa N(0,1).

Tässä tapauksessa F-testisuure nollahypoteesille H_0 on muotoa

$$F = \frac{b_1^2}{s^2 / ((n-1)s_x^2)}.$$

Jos jäännöstermi on normaalisti jakautunut, testisuure F on jakautunut kuten F-jakauma vapausastein 1 ja n–2. F-testisuureen käyttö johtaa tässä tapauksessa aivan samaan tulokseen kuin t-testin käyttö normaalijakautuneen jäännöstermin tapauksessa. Tämä nähdään siitä, että tässä tapauksessa

$$F = t^2$$

ja seuraavasta tuloksesta: Jos satunnaismuuttuja noudattaa t-jakaumaa vapausastein f, niin sen neliö noudattaa F-jakaumaa vapausastein 1 ja f.

Yllä esitetty F-testisuure voidaan kirjoittaa myös tutumpaan muotoon

$$F = (n-2)\frac{r^2}{1-r^2},$$

kun käytetään apuna yhtälöä $R^2 = r^2$. On syytä huomata, että tämän testisuureen neliöjuuri on t-testisuure nollahypoteesille

$$H_0: \quad \rho = 0,$$

jossa ρ on selitettävän muuttujan Y ja selittäjän X välinen korrelaatiokerroin perusjoukossa.

Edellä esitetystä nähdään, että korreloimattomuushypoteesi $\rho = 0$ ja hypoteesi $\beta_1 = 0$ ovat yhden selittäjän regressiomallin tapauksessa yhtäpitäviä. Edellä esitetystä nähdään myös se, että testisuureet näille yhtäpitäville hypoteeseille voidaan formuloida monilla yhtäpitävillä tavoilla.

ESIMERKKI 1.

Menestyminen opinnoissa saattaa vaikuttaa vastavalmistuneen alkupalkkaan. Asiaa tutkittiin USA:ssa poimimalla eräästä korkeakoulusta valmistuneiden joukosta yksinkertainen satunnaisotos, jonka koko oli 15.

Otokseen poimituilta kysyttiin heidän arvosanapisteidensä keskiarvoa x ja alkupalkkaansa Y (vuosipalkka tuhansina dollareina). Oletetaan, että regressiomalli alkupalkan riippuvuudelle arvosanapisteiden keskiarvosta on muotoa

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \qquad j = 1,2,\ldots,15,$$

jossa jäännöstermistä ε_j voidaan tehdä tavanomaiset oletukset.

Otosta kuvaavat perustunnusluvut olivat

$$\bar{x} = 3.04, \qquad \bar{Y} = 18.05,$$
$$s_x^2 = 0.063, \qquad s_Y^2 = 5.81,$$
$$r = 0.848.$$

Näistä tunnusluvuista saadaan seuraavat estimaatit regressiokertoimille:

$$b_1 = r\frac{s_Y}{s_x} = 8.12,$$

$$b_0 = \bar{Y} - b_1\bar{x} = -6.63.$$

Selitysasteeksi saadaan

$$R^2 = r^2 = 0.718.$$

Siten malli selittää 71.8% selitettävän muuttujan arvojen vaihtelusta.

Testisuure korreloimattomuushypoteesille

$$H_0: \quad \rho = 0$$

saa arvon

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 5.77.$$

Vastaava P-arvo on 0.000033, joten nollahypoteesi voidaan hylätä: Menestyminen opinnoissa parantaa alkupalkkaa.

Jäännösvarianssin estimaatiksi saadaan

$$s^2 = \frac{n-1}{n-2}(1-r^2)s_Y^2 = 1.76.$$

Estimaattorin b_1 varianssin estimaattori on

$$s_{b_1}^2 = \frac{s^2}{(n-1)s_x^2} = 2.00.$$

Siten testisuureen arvoksi hypoteesille

$$H_0: \quad \beta_1 = 0$$

saadaan

$$t = \frac{b_1}{s_{b_1}} = 5.76.$$

Tämä testisuureen arvo on käytännössä sama kuin ylläesitetyn korreloimattomuushypoteesin testisuureen arvo kuten pitääkin (ero johtuu käytetystä laskutarkkuudesta).

Neliösummista

$$SST = (n-1)s_Y^2 = 81.3,$$

$$SSE = (n-2)s^2 = 22.9,$$

$$SSM = SST - SSE = 58.4,$$

voidaan edellä esitetyn nollahypoteesin $\beta_1 = 0$ t-testisuureelle muodostaa sen kanssa yhtäpitävä F-testisuure

$$F = (n-2)\frac{SSM}{SSE} = 33.2.$$

Huomaa, että (laskutarkkuudesta johtuvaa eroa lukuunottamatta)

$$F = t^2.$$

Regressioanalyysin tulokset voidaan esittää esimerkiksi seuraavassa muodossa:

$$Y_j = -6.63 + 8.12 \cdot x_j \qquad R^2 = 0.718$$

(4.30) (1.41)

Sulkuihin on merkitty kerroinestimaattoreiden hajontojen estimaattorit.

Estimointituloksista voidaan tehdä seuraava johtopäätös: Jos kahdella otokseen poimitun henkilön arvosanapisteiden keskiarvot eroavat yhdellä pisteellä, heidän alkupalkkansa eroavat 8,120 dollaria. ●

Tässäkin tapauksessa analyysiä olisi täydennettävä mallin oletuksia koskevilla *diagnostisilla testeillä*.