3 Regressioanalyysi

Regressioanalyysi on käytetyimpiä tilastotieteellisiä menetelmiä. Ei liene olemassa ainakaan kaupallista tilasto-ohjelmistoa, joka ei sisältäisi regressioanalyysiä. Yksi syy lienee, että se mahdollistaa muuttujan vaikutuksen suuruuden toiseen muuttujaan tai vaikutuksen olemassaolon ylipäätään arvioinnin ja testaamisen. Ne ovat polttavia kysymyksiä monen tutkijan mielessä. Regressioanalyysi on tässä mielessä usein hyvin antoisaa ja tuloksellista.

3.1 Yhden selittäjän lineaarinen regressiomalli

Tarkastellaan kahta muuttujaa y ja x. Edellisen pitää olla välimatka-asteikollinen; jälkimmäinen voi olla myös luokitteluasteikollinen. Muuttuja y määräytyy lineaarisen regressiomallin

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

mukaisesti x:n arvoista. Muuttujaa y kutsutaan selitettäväksi muuttujaksi ja muuttujaa x selittäväksi muuttujaksi, selittäjäksi tai regressoriksi. Viimeinen termi ε ("epsilon") on mallin jäännös eli satunnaistermi, jonka odotusarvo on 0 ja varianssi on σ^2 ("sigma toiseen"). Kreikkalaisilla kirjaimilla — esimerkiksi β_0:lla ja β_1:llä ("beeta-nollalla" ja "beeta-yhdellä") — tavataan merkitä tilastollisten mallien parametreja ja niin edelläkin. Mallin parametrit ovat kiinteitä lukuja (esim. $\beta_0 = 90$ ja $\beta_1 = 0,5$), joiden suuruudet (tyypillisesti) ovat tuntemattomia ja joiden selvittämiseen regressioanalyysillä pyritään. Parametria β_0 kutsutaan usein mallin vakioksi ja parametria β_1 (regressio)kertoimeksi.

Luvussa 1 viitattu systemaattinen komponentti on yhden selittäjän regressiossa selitettävän (selittäjän x arvolle ehdollinen) odotusarvo

$$\mathsf{E}(y) = \mathsf{E}(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x.$$

Yllä $\mathsf{E}(\cdot)$ on odotusarvon symboli ja on käytetty oletusta $\mathsf{E}(\varepsilon) = 0$.

Tavalliseen x, y-koordinaatistoon funktio $y = \beta_0 + \beta_1 x$ piirtyy suorana, jonka kulmakerroin on β_1 ja joka leikkaa y-akselin kohdassa β_0. Periaatteessa β_0 kertoo siten selitettävän odotusarvon, kun selitettävän arvo on 0:

$$\mathsf{E}(y) = \mathsf{E}(\beta_0 + \beta_1 \times 0 + \varepsilon) = \beta_0.$$

Empiirisessä analyysissa tämä tulkinta ei ole aina järkevä, mihin palataan myöhemmin (jaksot 3.1.1 ja 3.3).

Mallin mukaan y:n suuruus riippuu x:n suuruudesta lineaarisesti parametrin β_1 välityksellä: Jos x muuttuu yksikön verran, niin y muuttuu β_1:n verran. Esimerkiksi jos y on lapsen pituus, x on isän pituus ja $\beta_1 = 0,5$, niin mallin mukaan lapsen pituus tapaa olla $0,5$ senttimetriä pidempi, jos isä on senttimetrin pidempi. Vakio β_0 asettaa mallin kuvaaman suoran sopivalle korkeudelle. Joissain tilanteissa vakiolla on selkeä tutkittavaan ilmiöön liittyvä tulkinta mutta monesti ei ole (jaksot 3.1.1 ja 3.3). Jäännös ε kuvaa y:n vaihtelua, joka ei selity x:n vaihtelulla. Esimerkiksi lapsen pituuteen vaikuttaa muitakin tekijöitä kuin isän pituus (luku 1). Ne jäävät mallissa huomioimatta ja puristetaan ε:iin.

Erikoistapaus on $\beta_1 = 0$. Tällöin malli (1) typistyy niin, että y on satunnaisesti jakautunut vakion β_0 ympärillä:

$$y = \beta_0 + 0 \times x + \varepsilon = \beta_0 + \varepsilon. \quad (2)$$

Kiinnostavin asia mallissa (1) onkin tyypillisimmin parametrin β_1 suuruus — esimerkiksi poikkeaako se nollasta eli päteekö malli (1) vai (2).

Regressioanalyysi on keino arvioida parametrien suuruutta ja systemaattista komponenttia, kun mallin kuvaamasta ilmiöstä on havaintoaineisto. Mallin (1) kohdalla aineisto koostuisi havaintopareista $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Tässä $n \geq 2$ on havaintojen lukumäärä.

Galtonin herneen siemen -vanhemmat ja niiden jälkipolvi sekä vanhempien ja lasten pituudet -esimerkit (luku 1) havainnollistavat regressiota odotusarvoa kohti mallin (1) mukaisesti, kun $0 < \beta_1 < 1$. Kuvitteellinen sosiaalitukien saajat -esimerkki (luku 1) vastaa mallia (2): Sosiaalitukien saajien lukumäärä edellisenä vuonna (x) ei auta ennustamaan heidän lukumääräänsä kuluvana vuonna (y): Kerroin $\beta_1 = 0$ ja lukumäärät pyrkivät palautumaan kohti odotusarvoaan β_0.

Oheiseen hajontakuvioon on piirretty keinotekoinen aineisto ($n = 50$) — vaikkapa helsinkiläisten isien (x) ja heidän lastensa (y) pituuksista aikuisina. Kukin piste vastaa yhtä havaintoparia (x_i, y_i). Mitä pidempi isä on, sitä pidempi vaikuttaa lapsi olevan. Mutta kuinka paljon? Voitaisiinko havaintopisteiden yhteys tiivistää suoraksi, jonka parametreista voitaisiin päätellä vaikutuksen keskimääräinen suuruus?

3.1.1 Yhden selittäjän lineaarisen regressiomallin estimointi

Yhden selittäjän regressiossa aineistoon sovitetaan regressiosuora, joka summeeraa muuttujien välisen riippuvuuden eli systemaattisen osan. Regressiosuoran parametriarvot ovat vastaus edellä esitetyn tapaisiin kysymyksiin.

Sovittaminen voidaan tehdä periaatteessa monella tavalla. Ylivoimaisesti käytetyin tapa on pienimmän neliösumman (PNS) menetelmä. Siinä parametrit β_0 ja β_1 valitaan niin, että y_i-havaintojen poikkeamat sovitettavasta suorasta neliöidään ja neliöiden summa minimoidaan:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Yllä merkintä "min" tarkoittaa, että sen oikealla puolella oleva lauseke minimoidaan min-merkinnän alapuolelle merkittyjen suureiden suhteen. Minimoinnin voi ajatella tapahtuvan ikään kuin kokeilemalla eri lukuarvoja β_0:lle ja β_1:lle ja valitsemalla sellainen β_0, β_1-pari, että lauseke $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ ei voi saada pienempiä arvoja. (Todellisuudessa tilasto-ohjelmisto ratkaisee minimointitehtävän yhdellä laskutoimituksella eikä kokeile eri arvoja.) Poikkeamien suoralta $y_i - \beta_0 - \beta_1 x_i$ kasvaessa (itseisarvoltaan) kasvaa neliösumma $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ nopeasti. PNS-menetelmä pyrkii siten tuottamaan regressiosuoran, joka ei koskaan sijoittuisi kovin kauas yhdestäkään havaintopisteestä. Termien

$(y_i - \beta_0 - \beta_1 x_i)$ neliöinnin takia minimoinnin kannalta ei ole väliä, onko y_i suurempi tai pienempi kuin mallin mukainen arvo $\beta_0 - \beta_1 x_i$. Kaikkia poikkeamia kohdellaan tässä mielessä samanarvoisesti.

Neliösumman minimoivia parametriarvoja kutsutaan PNS-estimaateiksi ja niitä merkitään $\hat{\beta}_0$:lla ja $\hat{\beta}_1$:lla ("$\wedge$" luetaan "hattu"). Suureita

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

kutsutaan sovitteiksi ja suureita

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

residuaaleiksi ($i = 1, \ldots, n$). Regressiosuora $(\hat{\beta}_0 + \hat{\beta}_1 x_i)$ saadaan piirtämällä hajontakuvioon suora sovitteiden $(\hat{y}_i)$ kautta. Residuaalit $(\hat{\varepsilon}_i)$ ovat jäännösten (ε) estimaatteja.

Tärkeä käsite on residuaalineliösumma

$$\mathsf{RNS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Nimityksensä mukaisesti se on summa residuaalien neliöistä. Se on sitä suurempi, mitä enemmän y_i-havainnot poikkeavat sovitteista $\hat{y}_i$. Sen avulla lasketaan estimaatti jäännöksen varianssille:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Jäännösvarianssin estimaatti $\hat{\sigma}^2$ mittaa residuaalin neliön keskimääräistä suuruutta eli vaihtelevuutta aineistossa.[10] Yleensä toivotaan, että $\hat{\sigma}^2$ olisi pieni, koska silloin malli selittää hyvin y:n vaihtelun.

Määritellään vastaavasti kokonaisneliösumma

$$\mathsf{KNS} = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{3}$$

Siinä $\bar{y} = \sum_{i=1}^{n} y_i / n$ eli y_i-havaintojen keskiarvo. Kokonaisneliösumma kuvaa, kuinka suurta on y_i-havaintojen vaihtelu keskiarvonsa ympärillä.

Neliösummista saadaan mallin selityskyvylle mittari selitysosuus

$$\mathsf{R}^2 = \frac{\mathsf{KNS} - \mathsf{RNS}}{\mathsf{KNS}} = 1 - \frac{\mathsf{RNS}}{\mathsf{KNS}}. \tag{4}$$

[10] Syy jakaa RNS $n-2$:lla eikä n:llä liittyy näin saadun σ^2:n estimaatin teoreettisiin ominaisuuksiin ja jaksossa 3.1.2 käsiteltävään jakaumateoriaan. Muistisääntö on, että PNS-estimoinnissa havaintoja "menetetään" yksi kutakin estimoitua parametria kohti. Vrt. RNS:n kaava monen selittäjän regressiomallissa (7).

Selitysosuus saa lähellä yhtä olevia arvoja, mikäli residuaalineliösumma on pieni suhteessa selitettävän kokonaisneliösummaan ($\mathsf{RNS}/\mathsf{KNS} \approx 0$). Tällöin y selittyy hyvin x:llä. Mikäli x:llä ei ole selityskykyä, residuaalineliösumma ei eroa paljoa kokonaisneliösummasta ($\mathsf{RNS}/\mathsf{KNS} \approx 1$). Tällöin selitysosuus on lähellä nollaa.

Selitysosuus on siten hyvin intuitiivinen mittari mallin hyvyydelle. Nyt esillä olevassa yhden selittäjän regression tilanteessa se liittyykin aiemmista opinnoista toivottavasti tuttuun otoskorrelaatiokertoimeen (r) yksinkertaisella tavalla:

$$\mathsf{R}^2 = r^2. \tag{5}$$

Oheinen kuvio havainnollistaa käsitteitä isä-lapsi -aineiston avulla. Regressiosuora määrittää sovitteen kunkin x_i-havainnon kohdalla. Residuaalit ovat x_i, y_i-havainnoista regressiosuoraan pystysuorasti meneviä viivoja. Toisenlainen suora tuottaisi toisenlaiset residuaalit; kuviossa esitettyjen residuaalien neliöiden summa on pienin mahdollinen.

Mallin (1) PNS-estimointi tuotti tästä aineistosta tulokset

$$y = 93,853 + 0,478x + \hat{\varepsilon}, \quad \hat{\sigma}^2 = 3,725, \; \mathsf{R}^2 = 0,323.$$

Malli ennustaa lapselle lisää pituutta $0,478$ eli noin $0,5$ senttimetriä isän pituuden kasvaessa senttimetrillä ja selittää noin 32 % lasten pituuden vaihtelusta aineistossa. Kaavasta (5) seuraa, että otoskorrelaatio on selitysosuuden neliöjuuri: $r = \sqrt{\mathsf{R}^2}$. Pituuksien otoskorrelaatio on siten $\sqrt{0,323} \approx 0,569$.

Koska aineisto oli keinotekoinen, estimointituloksia voidaan verrata aineiston tuottaneeseen todelliseen malliin. Suureiden todelliset arvot ovat $\beta_0 = 90$, $\beta_1 = 0,5$, $\sigma^2 \approx 4,5$, korrelaatio populaatiossa $\rho = 0,5$ ja selitysosuus populaatiossa $\mathsf{R}^2 = (0,5)^2 = 0,25$. Kaikki suureet tulivat estimoiduksi varsin hyvin.

Kuten usein on, estimoidulla vakiolla ei ole järkevää tulkintaa. Estimoidun mallin ja vakion mukaan lapsen pituus olisi noin 94 senttimetriä, jos isän pituus olisi 0 senttimetriä (kuvio), mikä on järjetön ajatus. Malli ei välttämättä antaisi luotettavaa ennustetta edes periaatteessa mahdollisen mutta poikkeuksellisen lyhyen isän (esim. $x = 155$) lapsen pituudelle. Regressiomalleja ei ylipäänsä kannata yrittää soveltaa aineiston vaihteluvälin ulkopuolella.

Edellä implisiittisesti oletettiin, että kaikki x_i-havainnot eivät ole yhtäsuuria. Jos ne olisivat, β_0- ja β_1-parametreja ei voisi estimoida (oheinen kuvio havainnollistaa β_1:n estimoinnin mahdottomuutta). Seuraavassa jaksossa tarvitaan muitakin oletuksia.

3.1.2 Yhden selittäjän lineaarisen regressiomallin testaus

Ei ole harvinaista, että tutkijan päämielenkiinto on testauksessa. Vaikka se ei olisi, pääsääntöisesti ei koskaan tulisi rajoittaa regressiomallin tarkastelua vain estimointituloksiin. Mallia tulisi aina testata. Filosofia on sama kuin muutenkin tilastotieteessä: Pelkkä estimaatin tai yleisemmin tilastollisen tunnusluvun subjektiivinen arviointi ei ole riittävää; tulee myös testata, poikkeaako tunnusluku

nollasta tai muusta oleelliseksi katsotusta arvosta tilastollisesti merkitsevästi. Hedelmällisen tilastotieteen soveltamisen tunnusmerkkejä on, että on arvioitu sekä laskettujen tunnuslukujen merkittävyyttä sovellusalan kannalta että niiden tilastollista merkitsevyyttä.

Regressioanalyysillä voidaan testata parametreihin liittyviä nollahypoteeseja (H_0). Sellaisia voisivat olla esimerkiksi H_0: $\beta_1 = 0$ tai H_0: $\beta_1 = 1$. Vakion suuruutta testataan harvoin muun muassa, koska sillä ei ole usein selkeää sovellukseen liittyvää merkityksellistä tulkintaa (vrt. isä-lapsi -mallitus edellä).

Hypoteesien testaukseen tarvitaan lisäoletuksia:

- Selittävä muuttuja x on kiinteä (siinä ei ole satunnaisuutta).
- Jäännös noudattaa normaalijakaumaa odotusarvolla 0 ja varianssilla σ^2: $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$, $i = 1, \ldots, n$.
- Jäännökset ε_i eivät korreloi keskenään eli ne ovat riippumattomia toisistaan (normaalijakauman tilanteessa korreloimattomuudesta seuraa riippumattomuus).

Oleellista on ymmärtää, että $\hat{\beta}_1$ on satunnaismuuttuja. Se saa yhden tietyn arvon tutkittavana olevassa aineistossa. Jos tutkittavana olisi toinen — esimerkiksi espoolainen 50 havainnon aineisto isien ja lapsien pituuksista — saataisiin toisensuuruinen $\hat{\beta}_1$. Samoin estimaatti muuttuisi, jos tutkittaisiin 50 vantaalaisen, 50 kaunialaisen jne. aineistoa isien ja lasten pituuksista. Koska $\hat{\beta}_1$ on satunnaismuuttuja, on sillä (ilmeisesti) myös keskihajonta, jota kutsutaan tässä yhteydessä keskivirheeksi. ("Virhe", koska tavoitteena on estimoida β_1, missä ei täysin onnistuta.)

Edellä lueteltujen oletuksien pätiessä estimaatteihin liittyvä jakaumateoria tunnetaan. Tavalla, jota tässä ei selitetä, voidaan laskea (otos)keskivirhe $\hat{\beta}_1$:lle ($\mathsf{SD}(\hat{\beta}_1)$) ja muodostaa niin sanottu t-testisuure eli t-arvo

$$t_{\beta_1=\beta_1^0} = \frac{\hat{\beta}_1 - \beta_1^0}{\mathsf{SD}(\hat{\beta}_1)} \sim \mathsf{t}_{n-2}.$$

Nollahypoteesin H_0: $\beta_1 = \beta_1^0$ pätiessä se noudattaa ("$\sim$"-merkki yllä) t-jakaumaa vapausasteilla $n - 2$. Huomionarvoista on, että jakauma riippuu havaintojen lukumäärästä mutta tunnetaan kaikilla havaintomäärillä. Testisuure on hyvin intuitiivinen. Estimaatin $\hat{\beta}_1$ poikkeama nollahypoteesin mukaisesta arvosta β_1^0 suhteutetaan estimaatin keskivirheeseen. Suurikaan poikkeama ei ole tilastollisesti merkitsevä, jos $\hat{\beta}_1$:n keskivirhe on suuri. Toisaalta pienikin poikkeama on tilastollisesti merkitsevä, jos $\hat{\beta}_1$:n keskivirhe on hyvin pieni. Keskivirhe pienenee havaintojen lukumäärän kasvaessa. (Muutkin tekijät vaikuttavat keskivirheen suuruuteen.)

Tyypillisimmin testataan nollahypoteesia $\beta_1 = 0$. Tällöin testisuure on yksinkertaisesti β_1:n estimaatti jaettuna keskivirheellään:

$$t_{\beta_1=0} = \frac{\hat{\beta}_1}{\mathsf{SD}(\hat{\beta}_1)} \sim \mathsf{t}_{n-2}. \tag{6}$$

Monet tilasto-ohjelmistot tulostavat tämän testisuureen regressoitaessa selitettävää yhdellä selittävällä muuttujalla. Toiset ohjelmistot raportoivat PNS-estimaatin ja sen keskivirheen, jolloin käyttäjän tehtävä on muodostaa osamäärä $\hat{\beta}_1/\mathsf{SD}(\hat{\beta}_1)$. Tieteellisissä artikkeleissa käytäntö vaihtelee: Joissain raportoidaan estimaatti ja t-arvo ja toisissa estimaatti ja sen keskivirhe. Jälkimmäisessä tilanteessa lukijan tulee osata itse muodostaa t-arvo, jos haluaa tietää sen suuruuden.

Testaaminen etenee tämän jälkeen tavanomaiseen tapaan eli valitaan sopivaksi katsottu riskitaso (esim. 5, 1 tai 0,1 %), ja katsotaan, onko testisuureen itseisarvo suurempi kuin riskitasoon liittyvä kriittinen arvo (kaksisuuntainen testaus). Esimerkiksi 5 %:n riskitasoa käytettäessä kriittiset arvot olisivat isä-lapsi esimerkissä t-jakauman $50 - 2 = 48$:lla vapausasteella 2,5. tai 97,5. persentiilit.

Tilasto-ohjelmistot usein raportoivat testisuureeseen liittyvän p-arvon, joka on todennäköisyys saada havaittu tai vielä poikkeavampi testisuureen arvo nollahypoteesin pätiessä. Jos ohjelmiston raportoima p-arvo on esimerkiksi alle 0,05, niin nollahypoteesi hylätään 5 %:n riskitasolla. Jos ohjelmisto ei raportoi p-arvoa, voi kriittiset arvot silti usein laskea tilasto-ohjelmiston avulla. Vaihtoehtoisesti voidaan kriittisiä arvoja katsoa t-jakaumataulukoista. Niistä ei ole taulukoitu kriittisiä arvoja kaikille mahdollisille havaintojen lukumäärille mutta on tietenkin käytännön kannalta riittävälle määrälle.

Isä-lapsi -esimerkissä $\hat{\beta}_1$:n keskivirhe on 0,099, joten t-arvo on $0,478/0,099 \approx 4,788$. Esimerkin laskussa käytetty R-ohjelmisto[11] raportoi sekä keskivirheen että t-arvon, joka täsmää juuri lasketun kanssa. Ohjelmiston mukaan p-arvo on pienempi kuin 0,0001, joten selvästikin nollahypoteesi $\beta_1 = 0$ hylätään 5 %:n riskitasolla ja paljon pienemmilläkin riskitasoilla. Samaan tulokseen päädytään vertaamalla t-arvoa 4,788 t-jakauman 50. vapausasteella 97,5. persentiiliin 2,009 (oheisesta taulukosta). Taulukosta ei löydy persentiilejä t-jakaumalle 48. vapausasteella, joten vertailujakaumana käytettiin t-jakaumaa 50. vapausasteella.

Testin mukaan isien pituus vaikuttaa lasten pituuteen ($\beta_1 \neq 0$). Tulos on tietenkin odotettu. Mikäli tutkittava ilmiö olisi tuntemattomampi. keskeinen osa regressioanalyysia olisi testata, poikkeaako parametri β_1 nollasta. Mikäli nollahypoteesia ei hylättäisi (t-arvo olisi itseisarvoltaan pienempi kuin kriittiset arvot), pääteltäisiin, että muuttujien välillä ei ole yhteyttä tai että aineisto ei ainakaan anna tukea yhteyden olemassaololle. Mallin selitysosuus olisi tällöin lähellä nollaa (mieti miksi!), ja kaavan (5) perusteella muuttujien välinen otoskorrelaatio olisi samoin lähellä nollaa.

[11] R-versio 2.15.0 (2012-03-30). Copyright (C) 2012 The R Foundation for Statistical Computing.

3.2 Monen selittäjän lineaarinen regressiomalli

Selitettävä muuttuja y määräytyy nyt monen selittävän muuttujan x_i ($i = 1, \ldots, k$) lineaarisesta regressiomallista

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon. \tag{7}$$

Mallin tulkinta on samantapainen kuin mallin (1). Selitettävä y on välimatka-asteikollinen, selittäjät x_i voivat olla myös luokitteluasteikollisia ja jäännös ε on satunnaistermi odotusarvolla 0 ja varianssilla σ^2. Siihen tiivistyy y:n vaihtelu, joka ei selity x_i:den vaihtelulla. Parametrit $\beta_0, \ldots, \beta_k$ ovat kiinteitä yleensä tuntemattomia lukuja, joiden suuruudet pyritään selvittämään regressioanalyysilla (eritoten β_1:stä β_k:hon). Parametria β_0 kutsutaan vakioksi ja parametreja $\beta_1, \ldots, \beta_k$ (regressio)kertoimiksi.

Mallin (7) systemaattinen komponentti on selitettävän (selittäjien x_i arvolle ehdollinen) odotusarvo

$$\mathsf{E}(y) = \mathsf{E}(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{8}$$

Vakio kuvaa nyt selitettävän odotusarvoa, kun kaikki selittäjät saavat arvon 0:

$$\mathsf{E}(y) = \mathsf{E}(\beta_0 + \beta_1 \times 0 + \cdots + \beta_k \times 0 + \varepsilon) = \beta_0.$$

Kuten yhden selittäjän regressiossa (jakso 3.1.1), tämä tulkinta ei ole aina järkevä.

Kerroin β_i kuvaa x_i:n yksikön suuruisen muutoksen vaikutuksen y:hyn, kun muut selittäjät eivät muutu. Monesti mielenkiintoisin kysymys on, ovatko β_i-kertoimet nollia eli selittääkö x_i:den vaihtelu lainkaan y:n vaihtelua.

Monen selittäjän regressiomallin (7) systemaattisen komponentin ja parametrien selvittäminen edellyttää n:stä havaintovektorista $[x_{11} \ \ldots \ x_{1k} \ y_1], \ldots, [x_{n1} \ \ldots \ x_{nk} \ y_n]$ koostuvaa aineistoa ($n \geq k$). Muuttujien ensimmäinen indeksi on havainnon numero ($i = 1, \ldots, n$) ja jälkimmäinen indeksi kertoo, mistä selittäjästä havaintoarvo x_{ij} on ($j = 1, \ldots, k$). Jos selittäjiä on kaksi, niin aineisto $[x_{11} \ \ x_{12} \ y_1], \ldots, [x_{n1} \ x_{n2} \ y_n]$ on esitettävissä kolmiulotteisessa koordinaatistossa vaikkapa tietokoneen näyttöruudulla. Jos selittäjiä on enemmän, vastaava visualisointi ei ole mahdollista.

3.2.1 Monen selittäjän lineaarisen regressiomallin estimointi

Monen selittäjän regressiossa aineistoon sovitetaan selitettävän ja selittäjien välisen riippuvuuden summeeraava lineaarinen funktio eli mallin (7) systemaattinen osa (8). Systemaattisen osan geometrinen tulkinta ei ole yhtä helppo kuin yhden selittäjän regressiossa, jossa aineistoon sovitettiin suora. Mikäli selittäjiä on kaksi, sovitetaan aineistoon kaksiulotteinen taso (kuvio)[12].

[12] Kuvio on kirjasta B. Burt Gertsman (2008): *Basic Biostatistics — Statistics for Public Health Practice.* Jones and Bartlett.

Sovittaminen tehdään yleisimmin PNS-menetelmällä jaksossa 3.1.1 esitettyyn tapaan. Parametrien $\beta_0,\ldots,\beta_k$ lukuarvot valitaan minimoimaan y_i-havaintojen poikkeamien systemaattisesta komponentista neliöiden summa:

$$\min_{\beta_0,\ldots,\beta_k} \sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)]^2 = \min_{\beta_0,\ldots,\beta_k} \sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_1-\cdots-\beta_k x_k)^2.$$

Neliösumman minimoivat parametriarvot ovat PNS-estimaateja $\hat{\beta}_0,\ldots,\hat{\beta}_k$. Jaksossa 3.1.1 selitetyt käsitteet yleistyvät muutenkin suoraviivaisesti k:n selittäjän tilanteeseen. Sovitteet $(\hat{y}_i)$, residuaalit (ε_i), residuaalineliösumma ja jäännöksen varianssin estimaatti ovat nyt

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik},$$

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik},$$

$$\mathsf{RNS} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2$$

ja

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2.$$

Kokonaisneliösumman ja selitysosuuden kaavat (3) ja (4) eivät muutu. Merkittävä ero on, että selitysosuuden ja otoskorrelaation neliöt sitova kaava (5) pätee nyt, kun otoskorrelaatiokerroin r on laskettu selitettävän muuttujan y_i ja sen sovitteen $\hat{y}_i$välille. (Tämä tulkinta on mahdollinen myös yhden selittäjän regression mallin kohdalla.)

3.2.2 Monen selittäjän lineaarisen regressiomallin testaus

Testaamista varten jaksossa 3.1.2 tehtyjä oletuksia pitää täydentää olettamalla nyt, että kaikki selittävät muuttujat x_i ovat kiinteitä (niissä ei ole satunnaisuutta). Lisäksi yhdenkään selittäjän x_i arvot eivät saa riippua täydellisesti lineaarisesti muiden selittäjien x_j, $j \neq i$, arvoista.[13]

Ehkä tärkein ja useimmin testattu monen selittäjän regressiomallin (7) β_i-kertoimia koskeva nollahypoteesi on, että ne ovat kaikki nollia (H_0: $\beta_1 = \cdots = \beta_k = 0$). Nollahypoteesin mukaan selittäjillä x_i ei ole tällöin lainkaan selityskykyä selitettävän muuttujan y suhteen. Tämän hypoteesin päteminen tai

[13] Ei saa olla olemassa γ_k-kertoimia niin, että pätisi $x_i = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_{i-1}x_{i-1} + \gamma_{i+1}x_{i+1} + \cdots + \gamma_k x_k$ kaikkien havaintojen kohdalla. Yhden selittäjän tilanteessa jälkimmäinen ehto merkitsee, että kaikki ainoan selittäjän havainnot eivät saa olla samoja. Tätä tilannetta sivuttiin jakson 3.1.1 lopussa. Asia on matemaattinen ongelma, johon harvemmin törmää empiirisessä työssä ja se sivuutetaan siksi tässä enemmittä pohdinnoitta.

pätemättömyys on tutkijalle usein keskeisimpiä kysymyksiä. Nollahypoteesia testaava F-testisuure on hyvin yksinkertainen:

$$F = \frac{\mathsf{R}^2/k}{(1-\mathsf{R}^2)/(n-k-1)} \sim \mathsf{F}_{k,n-k-1} \tag{9}$$

Nollahypoteesin H_0: $\beta_1 = \cdots = \beta_k = 0$ pätiessä se noudattaa F-jakaumaa k:lla ja $n-k-1$:llä vapausasteella. Jälleen (vrt. jakso 3.1.2) jakauma riippuu havaintojen lukumäärästä ja on tunnettu kaikilla havaintomäärillä.

Useimmat tilasto-ohjelmistot laskevat F-testisuureen automaattisesti regression yhteydessä. Testisuureen suuret arvot ovat testin kannalta hälyttäviä. Mikäli tilasto-ohjelmisto ei ilmoita F-testisuureen p-arvoa, voidaan testisuureen arvon tilastollinen merkitsevyys arvioida esimerkiksi F-jakauman taulukoitujen kriittisten arvojen avulla.

F-testisuureella on selkeä intuitio. Mikäli selittäjät x_i kykenevät selittämään suuren osan selitettävän y vaihtelusta (KNS), niin jäljelle jäävä selittämätön vaihtelu (RNS) muodostuu pieneksi ja selitysosuus R^2 suureksi (kaava (4)). Kaavasta (9) nähdään, että mitä suurempi R^2 on, sitä suurempi on F-testisuure. F-testi siis hälyttää, kun selittäjillä on aineistossa hyvä selityskyky. Myös havaintojen lukumäärän kasvattaminen pyrkii kasvattamaan F-testisuuretta ja kasvattamaan todennäköisyyttä hylätä nollahypoteesi, kun se ei päde. Mikäli nollahypoteesi pätee, R^2 tapaa jäädä pieneksi ja F-testisuure samoin.

Yleisiä mallin (7) parametreja koskevia nollahypoteeseja ovat, että i:nnen selittäjän kerroin on nolla (H_0: $\beta_i = 0$) tai että se on tietyn suuruinen (H_0: $\beta_i = \beta_i^0$). Edellisessä tilanteessa i:nnettä selittäjää ei tarvittaisi regressiossa (7). Näitä nollahypoteeseja voidaan testata jaksosta 3.1.2 tutuilla t-testisuureilla:

$$t_{\beta_i=0} = \frac{\hat{\beta}_i - \beta_i^0}{\mathsf{SD}(\hat{\beta}_i)} \sim \mathsf{t}_{n-k-1}.$$

ja

$$t_{\beta_i=0} = \frac{\hat{\beta}_i}{\mathsf{SD}(\hat{\beta}_i)} \sim \mathsf{t}_{n-k-1}. \tag{10}$$

Vastaavan nollahypoteesin pätiessä ne noudattavat t-jakaumaa vapausasteilla $n-k-1$. Jakauma riippuu havaintojen lukumäärästä mutta tunnetaan kaikilla havaintomäärillä. Tilasto-ohjelmisto raportoi yleensä automaattisesti jälkimmäisen t-arvon kaikkien selittäjien estimoiduille kertoimille tai niiden keskivirheet $\mathsf{SD}(\hat{\beta}_i)$, $i = 1, \ldots, n$. Testaus tapahtuu käytännössä jaksossa 3.1.2 selitetyllä tavalla. Siellä kuvattiin myös t-testisuureiden intuitio.

Monesti on kiinnostavaa testata, olisikovatko mallin (7) d ($0 < d \leq k$) oikeanpuoleisinta selittäjää tarpeettomia eli päteekö $\beta_{k_r+1} = \cdots = \beta_k = 0$, jossa $k_r = k - d > 0$. (Oletus tarpettomien selittäjien sijoittumisesta mallin oikeanpuoleisimmiksi tehdään merkintöjen yksinkertaistamiseksi.) Edellä "r"

viittaa rajoitettuun. Näin rajoitettu malli olisi

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k_r} x_{k_r} + \varepsilon. \tag{11}$$

Se saadaan mallista (7) erikoistapauksena asettamalla d kappaletta β_i-kertoimia nollaksi $(k_r + 1)$. muuttujasta lähtien.

Nollahypoteesia H_0: $\beta_{k_r+1} = \cdots = \beta_k = 0$ voidaan testata testisuureella

$$\frac{(\mathsf{R}^2 - \mathsf{R}_r^2)/d}{(1 - \mathsf{R}^2)/(n - k - 1)} \sim F_{d,n-k-1}.$$

Testisuure vaatii sekä regression (7) että regression (11) laskemisen. Jälkimmäisen regression selitysastetta on merkitty yllä R_r^2:lla. Nollahypoteesin pätiessä testisuure noudattaa F-jakaumaa d:llä ja $n - k - 1$:llä vapausasteella. Testisuureen suuret arvot ovat hälyttäviä.

Tämänkin testisuureen toimintaperiaate on hyvin ymmärrettävä. Mikäli kertoimet $\beta_{k_r+1}, \ldots, \beta_k$ poikkeavat tai osa niistä poikkeaa nollasta, mallien selitysasteiden tulisi erota selvästi. Tällöin erotus $\mathsf{R}^2 - \mathsf{R}_r^2$ testisuureen osoittajassa muodostuu suureksi ja testisuure samoin. Mikäli d. viimeisellä selittäjällä ei ole selitysvoimaa (nollahypoteesi pätee), erotus ja testisuure jäävät pieniksi.

Muunkinlaisia rajoituksia (esim. $\beta_1 = \beta_2$ tai $\beta_1 + \cdots + \beta_k = 1$) mallin (7) parametreille voidaan testata. Asia jätetään tässä maininnan varaan.