YLEISTETYT LINEAARISET MALLIT

Jukka Nyblom

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
2015

Esipuhe

Tämä moniste on syntynyt luennoista, joita olen vuosien varrella pitänyt Jyväskylän yliopistossa. Moniste jakautuu kahteen osaan, jotka yhdessä koostuvat kahdesta lukuvuoden neljänneksen mittaisesta kurssista. Ensimmäisellä 4 opintopisteen aineopintojen kurssilla käsitellään Osa I ja toisella 5 opintopisteen syventävien opintojen kurssilla Osa II.

Ensimmäisessä osassa tutustumme tavalliseen lineaariseen regressiomalliin, logistiseen regressioon ja Poisson-regressioon, jotka ovat sovellusten kannalta kaikkein tärkeimpiä. Keskitymme regeressiomenetelmien ymmärtämiseen ja mallien tulkintaan. Lisäksi tutustumme tilastolliseen laskenta- ja graafiseen ympäristöön (R-ohjelmointiympäristö), R Core Team (2014), josta on tullut vakiotyökalu tilastotieteilijöiden ja monien muidenkin tilastomenetelmiä käyttävien keskuudessa. Ensimmäisen osan ymmärtäminen edellyttää lukijalta tilastotieteen alkeiskurssin tiedot ja halua ymmärtää yksinkertaisia matemaattisia merkintöjä. Lukion matematiikan lyhyt oppimäärä riittää jälkimmäisten taitojen hankkimiseksi.

Monisteen toinen osa on vaativampi. Siinä paneudumme syvemmälle yleistettyjen lineaaristen mallien teoriaan. Yleistä teoriaa valaistaan erityisesti ensimmäisessä osassa käytettyjen mallien osalta. Lisäksi tutustumme mallien laajennuksiin. Toisen osan ymmärtäminen edellyttää huomattavasti enemmän matemaattisen tilastotieteen tietoja ja matemaattisia taitoja; tarpeen on ennen kaikkea harjaantuminen matriisialgebran, differentiaali- ja integraalilaskennan soveltamiseen.

Tämä moniste on johdatus yleistettyihin lineaarisiin malleihin. Se ei käsittele niitä kaikessa laajuudessaan. Ulkopuolelle jäävät esimerkiksi multinomivasteiset mallit, log-lineaariset mallit, elinaikamallit ja pitkittäisaineistohin liittyvät mallit. Satunnaiskomponentti- ja aikasarjamalleista käsittelemme vain pari erikoistapausta. Laajempaa ja syvempää tietämystä janoavia kehotan tutustumaan seuraaviin oppikirjoihin: Gelman and Hill (2007) (sisältää paljon käytännön esimerkkejä ja aineistoja), Dobson (2002) (tiivis johdatus teoriaan) ja Fahrmeir and Tutz (2001) (kattava teoria esimerkkeineen yleistetyistä lineaarisista malleista).

Jyväskylä, elokuussa 2015
Jukka Nyblom

Sisältö

Osa I

Käsitteet ja tulkinnat

Luku 1

Lineaarinen regressio: alkeet

Regressioanalyysillä voidaan kuvata numeerisen *vastemuuttujan* keskiarvojen vaihtelua osapopulaatioissa, jotka voidaan määritellä joidenkin muiden, *prediktorimuuttujien,* lineaaristen funktioiden avulla. Toinen tapa luonnehtia regressioanalyysiä on sanoa sen kuvaavan vastemuuttujan riippuvuutta prediktoreista. Mutta kuten Gelman ja Hill toteavat (Gelman and Hill, 2007, s. 31), edellinen tapa toteaa selvemmin sen tosiasian, etteivät regressiorelaatiot välttämättä kerro muuttujien välisestä kausaalisuhteesta (syy-seuraus -yhteydestä). *Regressiomallilla voidaan ennustaa vasteen arvo annetulla prediktoreiden arvoilla, ja regressiokertoimet voidaan tulkita ennusteiden eroiksi tai tiettyjen keskiarvojen eroiksi aineistossa.*

1.1 Yksi prediktori

Ymmärtääksemme regressiokertoimen merkityksen aloitamme yhden prediktorin tilanteesta pohtimatta estimointiin ja kertoimien epävarmuuteen liittyviä asioita. Tarkastelemme aineistoa, joka on kerätty eräistä Suomen lukioista ja joka sisältää oppilaiden tietoja menestymisestä peruskoulussa ja vuoden 1994 ylioppilaskirjoituksissa.

Dikotominen prediktori

Valitsemme osa-aineistoksi koulut A ja B ja vastemuuttujaksi ruotsin kielen ylioppilaskokeen pistemäärän `ruotsi.pist`. Prediktori `kouluA` saa arvon 1, kun oppilas on koulusta A, ja arvon 0, kun oppilas on koulusta B. Regressiomallimme on

$$\texttt{ruotsi.pist} = \beta_0 + \beta_1 \texttt{kouluA} + \varepsilon,$$

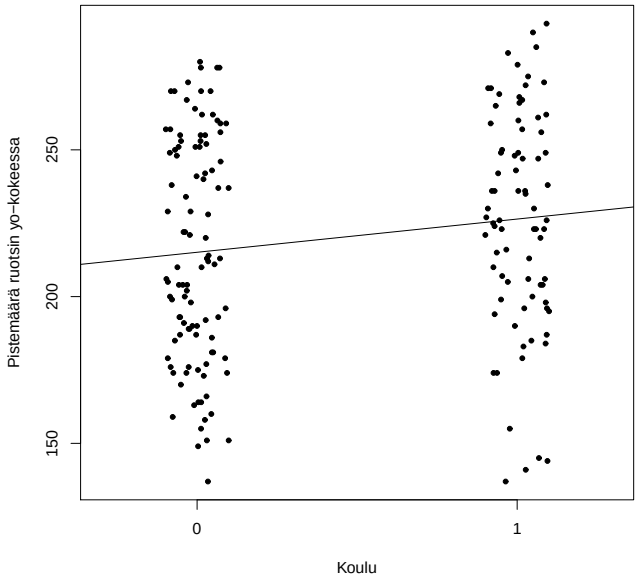
missä vakiot β_0 ja β_1 ovat tuntemattomia vakioita, jotka estimoidaan aineistosta, ja ε tarkoittaa poikkeamaa arvosta $\beta_0 + \beta_1 \texttt{kouluA}$. Termiä ε sanotaan jatkossa myös virheeksi. Aineistosta laskettu *sovite* on

$$\widehat{\texttt{ruotsi.pist}} = 215.1 + 11.3\,\texttt{kouluA}$$

Kun asetamme `kouluA` $= 1$, saamme koulun A keskiarvon $215.1 + 11.3 = 226.4$. Koulun B keskiarvon 215.1 saamme, kun `kouluA` $= 0$. Koulujen keskiarvojen erotus

(koulu A - koulu B) on regressiokerroin 11.3. Siis koulun A kokelaat selviytyivät keskimäärin 11.3 pistettä paremmin kuin koulun B kokelaat.

Jos prediktori on dikotominen, sen regressiokerroin on kahden ryhmän keskiarvojen erotus.



Kuva 1.1: *Ruotsin kielen ylioppilaskokeen pistemäärä on piirretty koulun määrittävän indikaattorin suhteen. Kuvaan on lisätty regressiosuora, joka kulkee ryhmien keskiarvojen kautta. Indikaattori on täristetty niin, etteivät pisteet osu toistensa päälle.*

Jatkuva prediktori

Sovitamme seuraavaksi mallin, jossa prediktorina on peruskoulun päättötodistuksen lukuaineiden keskiarvo `lka`

$$\texttt{ruotsi.pist} = \beta_0 + \beta_1 \texttt{lka} + \varepsilon.$$

Laskettu sovite on

$$\widehat{\texttt{ruotsi.pist}} = -156.1 + 43.7\,\texttt{lka}$$

Kun verrataan kahta osapopulaatiota, jotka poikkeavat lukuaineiden keskiarvojen suhteen yhdellä numerolla, so. verrataan populaatiota, jossa kun $\texttt{lka} = x + 1$ sellaiseen, missä $\texttt{lka} = x$ niin näiden osapopulaatioiden keskiarvojen erotus on

$$(-156.1 + 43.7(x + 1)) - (-156.1 + 43.7x) = 43.7.$$

Samaan tapaan saamme tuloksen, että jos verrataan osapopulaatioita, joissa poikkeama on a:n suuruinen, niin yo-pistemäärien odotettu ero on $43.7\,a$.

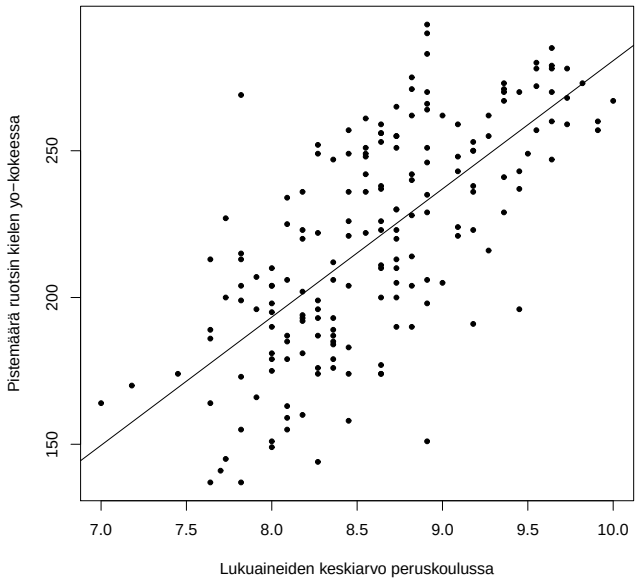
Regressiosuoran vakion ymmärtämiseksi on ajateltava tapausta, jossa prediktorin arvo on 0. Tässä esimerkissä sillä ei ole tulkintaa, koska peruskoulun keskiarvo ei voi olla nolla!

Voimme tehdä yksinkertaisen muunnoksen prediktoriin, joka antaa yhtäpitävän mallin, mutta jonka vakio on helpommin tulkittavissa. Prediktori `lka` keskistetään (siitä vähennetään sen keskiarvo 8.6) ja kirjoitetaan sovite uudelleen

$$\begin{aligned}\widehat{\texttt{ruotsi.pist}} &= -156.1 + 43.7\,\texttt{lka} \\ &= -156.1 + 43.7 \cdot 8.6 + 43.7\,(\texttt{lka} - 8.6) \\ &= 220.0 + 43.7\,(\texttt{lka} - 8.6).\end{aligned}$$

Vakio 220.0 antaa ennusteen kokelaalle jonka lukuaineiden keskiarvo on 8.6. Voi tietysti käyttää muutakin sopivaa referenssiarvoa keskiarvon sijasta.

Jos keskistää dikotomisen muuttujan, mikä tulkinta vakiolla silloin on?



Kuva 1.2: *Ruotsin kielen ylioppilaskokeen pistemäärä on piirretty peruskoulun lukuaineiden keskiarvon suhteen. Kuvaan on lisätty regressiosuora. Jokaisen suoran pisteen voi tulkita sellaisen yo-kokelaan ruotsin kokeen pistemäärän ennusteeksi, jolla on vastaava lukuaineiden keskiarvo. Toisen tulkinnan mukaan suoran piste on sellaisen osapopulaation yo-kokeen pistemäärien keskiarvon estimaatti, joilla on vastaava peruskoulun lukuaineiden keskiarvo.*

1.2 Useita prediktoreita

Regressiokertoimien tulkinta tulee monimutkaisemmaksi, kun mukana on useita prediktoreita. Silloin kunkin kertoimen tulkinta riippuu siitä, mitä muita prediktoreita on mallissa mukana. Yksinkertaistettu neuvo on tulkita tietty kerroin "pitämällä muut prediktorit vakioina". Tämä ei kuitenkaan ole aina mahdollista kuten myöhemmin nähdään.

Tarkastellaan asiaa esimerkin avulla. Sovitetaan nyt malli

$$\texttt{ruotsi.pist} = \beta_0 + \beta_1 \texttt{kouluA} + \beta_2 \texttt{lka} + \varepsilon,$$

jossa on mukana molemmat prediktorit: koulu-indikaattori ja lukuaineiden keskiarvo. Sovitteeksi saadaan

$$\widehat{\texttt{ruotsi.pist}} = -156.9 + 8.8\,\texttt{kouluA} + 43.3\,\texttt{lka}. \tag{1.1}$$

Kertoimet tulkitaan seuraavasti:

1. *Vakio.* Jos yo-kokelas on koulusta B ja hänen lukuaineidensa keskiarvo on 0, niin ennuste ruotsin yo-kokeen pistemäärälle on -156.9. Tämä ennuste ei ole hyödyllinen, sillä kenenkään lukuaineiden keskiarvo ei voi olla nolla! Samaan tapaan kuin edellä prediktorin keskistäminen tuottaa tulkinnallisemman vakion arvon.

2. *Koulun kerroin.* Kun verrataan koulujen A ja B yo-kokelaita, joilla on *sama lukuaineiden keskiarvo* peruskoulussa, niin mallin mukainen ennuste heidän odotetulle erotukselleen (koulu A - koulu B) on 8.8 pistettä ruotsin kielen yo-kokeessa.

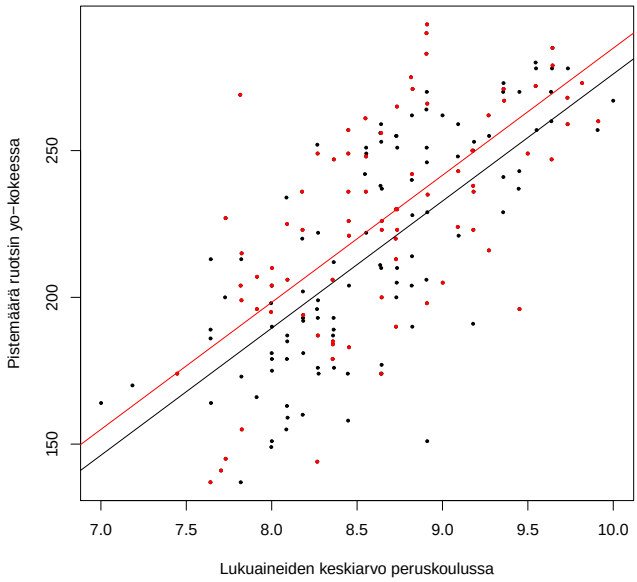
3. *Lukuaineiden keskiarvon kerroin.* Kun verrataan *saman koulun kokelaita*, joiden peruskoulun lukuaineiden keskiarvo eroaa yhdellä numerolla, heidän ennustettu eronsa ruotsin kielen yo-kokeessa on 43.3 pistettä. Malli sisältää oletamuksen, että *suoran kulmakerroin on sama molemmissa kouluissa.* Myöhemmin nähdään, miten voidaan sovittaa malli, joka sallii kouluille eri kulmakertoimet.

Huomaa tulkintojen ero yhden prediktorin ja kahden prediktorin malleissa. Kun ennustetaan koulujen A ja B keskimääräistä eroa ottamatta huomioon peruskoulun lukuaineiden keskiarvoa, ennuste on 11.3. Jos taas ennustetaan koulujen A ja B keskimääräistä eroa niiden osalta, joilla on sama lukuaineiden keskiarvo, ennuste on pienempi 8.8. Vastaava periaatteellinen ero on myös lukuaineiden keskiarvon kertoimien kohdalla. (Tässä esimerkissä kertoimien numeeriset arvot ovat melkein samat, mutta se on eri asia.)

1.3 Interaktio

Edellä pakotettiin lukuaineiden keskiarvon kulmakerroin samaksi molemmissa kouluissa. Malli, joka sallii eri kulmakertoimet, on

$$\texttt{ruotsi.pist} = \beta_0 + \beta_1 \texttt{kouluA} + \beta_2 \texttt{lka} + \beta_3 \texttt{kouluA} \cdot \texttt{lka} + \varepsilon,$$



Kuva 1.3: *Ruotsin kielen ylioppilaskokeen pistemäärä on piirretty peruskoulun lukuaineiden keskiarvon suhteen. Punaiset pisteet liittyvät kouluun A ja mustat kouluun B. Kuvaan on lisätty molempien koulujen regressiosuorat. Punainen suora liittyy kouluun A ja musta kouluun B. Suorien vertikaalinen etäisyys on ennuste koulujen A ja B piste-erolle ruotsinkielen yo-kokeessa niiden yo-kokelaiden osalta, joilla on saman lukuaineiden keskiarvo.*

missä tulotermiä $\texttt{kouluA} \cdot \texttt{lka}$ sanotaan *interaktioksi*. Sovitteeksi saadaan

$$\widehat{\texttt{ruotsi.pist}} = -186.7 + 78.7 \cdot \texttt{kouluA} + 46.8 \cdot \texttt{lka} - 8.1 \cdot \texttt{kouluA} \cdot \texttt{lka}.$$

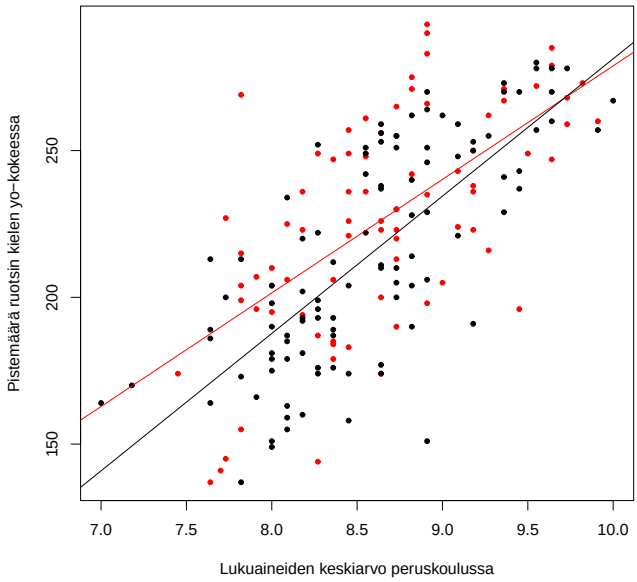
Kertoimien tulkinnassa on syytä olla varovainen. Kaikilla kertoimilla ei tässä esimerkissä ole mielekästä tulkintaa.

1. Vakio antaa ennusteen yo-kokeen pistemäärälle, kun kokelas on koulusta B ja hänen lukuaineidensa keskiarvo peruskoulussa on 0. Koska keskiarvo 0 on mahdoton, tulkinta on hyödytön.

2. Prediktorin `kouluA` kerroin antaa ennusteen koulun A ja koulun B kokelaiden pistemäärien erotukselle siinä tapauksessa, että heillä molemmilla peruskoulun lukuaineiden keskiarvot ovat nollia! Jälleen hyödytön tulkinta.

3. Prediktorin `lka` kerroin antaa ennusteen koulun B kokelaiden pistemäärien erotukselle silloin, kun kokelaiden lukuaineiden keskiarvot poikkeavat yhdellä numerolla.

4. Interaktiotermin kerroin kertoo prediktorin `lka` kulmakertoimien erotuksen kouluissa A ja B (koulu A - koulu B).



Kuva 1.4: *Ruotsin kielen ylioppilaskokeen pistemäärä on piirretty peruskoulun lukuaineiden keskiarvon suhteen. Punaiset pisteet liittyvät kouluun A ja mustat kouluun B. Kuvaan on lisätty molempien koulujen regressiosuorat. Punainen suora liittyy kouluun A ja musta kouluun B. Interaktio näkyy suorien kulmakertoimien erisuuruutena.*

Keskistämällä prediktorin `lka` saamme jälleen helpommin tulkittavan vakion ja kertoimen prediktorille `kouluA`:

$$\begin{aligned}\widehat{\texttt{ruotsi.pist}} &= 216.3 + 8.9 \cdot \texttt{kouluA} + 46.8 \cdot (\texttt{lka} - 8.6) \\ &\quad -8.1 \cdot \texttt{kouluA} \cdot (\texttt{lka} - 8.6).\end{aligned}$$

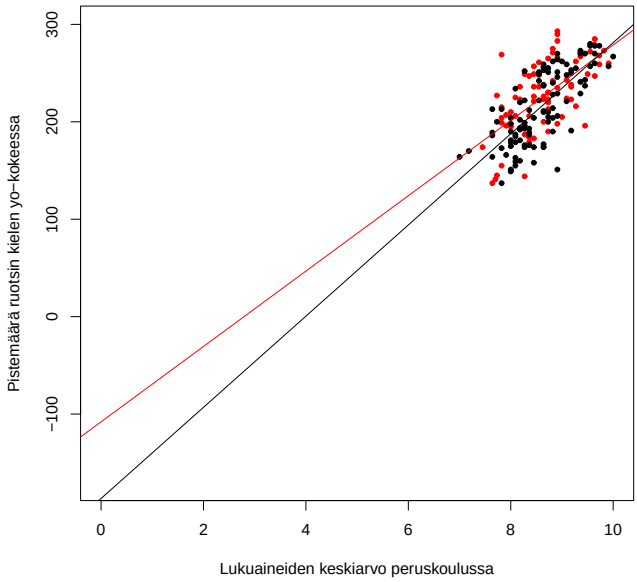
Vakio 216.3 on ennuste sellaisten koulun B oppilaiden pistemäärälle, joiden lukuaineiden keskiarvo peruskoulussa on ollut 8.6. Prediktorin `kouluA` kerroin 8.9 on koulujen A ja B keskiarvojen erotus niiden kokelaiden osalta, joiden peruskoulun lukuaineiden keskiarvo on ollut 8.6.

Sovitteen voi esittää myös ao. muodossa, joka on usein hyödyllinen

$$\begin{aligned}\text{koulu A:} \quad \widehat{\texttt{ruotsi.pist}} &= 225.1 + 38.7 \cdot (\texttt{lka} - 8.6) \\ \text{koulu B:} \quad \widehat{\texttt{ruotsi.pist}} &= 216.3 + 46.8 \cdot (\texttt{lka} - 8.6).\end{aligned}$$

Molemmat vakiot kertovat sellaisten kokelaiden keskiarvon, joilla oli peruskoulussa lukuaineiden keskiarvo 8.6. Kulmakerroin koulussa A on 38.7 ja koulussa B 46.8. Jos kaksi kokelasta poikkeaa 1 numerolla keskiarvon suhteen, heidän ennustettu eronsa ruotsin yo-kokeen pistemäärissä on suurempi koulussa B kuin koulussa A.

Interaktiot voivat olla hyvinkin tärkeitä, kuten jatkossa tulemme huomaamaan.



Kuva 1.5: *Tässä kuvassa on sama tilanne kuin kuvassa 1.4, mutta tässä näkyy myös vakion arvo joka on mustan suoran arvo origossa ja* `koulu`*:n kerroin, joka on suorien erotus origossa.*

1.4 Tilastollinen päättely

Tilastoyksiköt, prediktorit ja syötemuuttujat

Yksittäisiä tilastoyksiköitä sanotaan joskus myös tapauksiksi. Ne ovat yleensä ihmisiä, kouluja, kuntia, onnettomuuksia yms. Regressiomallin ns. X-muuttujia sanotaan prediktoreiksi ja Y-muuttujaa vastemuuttujaksi. Syötemuuttujat ovat sellaisia muuttujia, joiden antaman informaation avulla prediktorit muodostetaan. Syötemuuttujat ja prediktorit ovat eri asioita. Esim. mallissa

$$\widehat{\texttt{ruotsi.pist}} = -186.7 - 78.7 \cdot \texttt{kouluA} + 46.8 \cdot \texttt{lka} - 8.1 \cdot \texttt{kouluA} \cdot \texttt{lka}$$

on kolme prediktoria: `kouluA`, lukuaineiden keskiarvo `lka`, interaktio $\texttt{kouluA} \cdot \texttt{lka}$ ja vakiotermi, mutta vain kaksi syötemuuttujaa: `kouluA` ja `lka`.

Regressiomallin matriisiesitys

Seuraavaksi siirrymme matemaattisempaan esitysmuotoon. Merkitsemme vastemuuttujaa Y:llä ja prediktoreita X_j:llä, $j = 1, \ldots, p$. Tärkein käsite on Y:n ehdollinen odotusarvo

$$E[Y|X_1 = x_1, \ldots, X_p = x_p],$$

missä prediktoreille on annettu arvot $X_1 = x_1, \ldots, X_p = x_p$. Mallin virhetermi (teoreettinen jäännös) on erotus

$$Y - E[Y|X_1 = x_1, \ldots, X_p = x_p] = \varepsilon.$$

Termi ε on satunnaismuuttuja, jonka odotusarvo on $E(\varepsilon) = 0$. Tämä seuraa ε:n määritelmästä. Lisäksi tehdään olettamus että $\varepsilon \sim N(0, \sigma^2)$ ja että se on tilastollisesti riippumaton selittäjistä.

Lineaarisessa regressiossa oletetaan, että

$$E[Y|X_1 = x_1, \ldots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{1.2}$$

Silloin virhetermiä koskevista olettamuksista seuraa, että voidaan kirjoittaa myös

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \\ Y \,|\, X_1 = x_1, \ldots, X_p = x_p &\quad \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2). \end{aligned}$$

Oletetaan, että vasteen Y arvot ovat $y_1, \ldots, y_n$ ja että selittäjän X_j:n arvot ovat $x_{1j}, \ldots, x_{nj}$, $j = 1, \ldots, p$. Tapaus i voidaan kirjoittaa muotoon

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \tag{1.3}$$

virheet ε_i ovat satunnaismuuttujia. Ne ovat keskenään riippumattomia sekä riippumattomia kaikista selittäjistä. Kukin noudattaa normaalijakaumaa $N(0, \sigma^2)$. Kaavaan (1.3) liittyen otamme käyttöön merkinnät

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Silloin voimme kirjoittaa lineaarisen regressiomallin lyhyesti matriisisalgebran merkinnöin seuraavasti

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

Tapauskohtaisesti voimme kirjoittaa

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{NID}(0, \sigma^2),$$

missä $\boldsymbol{x}_i' = (1, x_{i1}, \ldots, x_{ip})$, ja NID on lyhennys fraasista "Normally and Independently Distributed".

Regressiomallin sovitus `R`:n avulla

Lineaarisen regressiomallin voimme sovittaa funktion `lm()` avulla. Saamme koko joukon hyödyllisiä funktioita ottamalla käyttöömme `R`-paketin `arm`. Sovitamme aluksi mallin, jossa vasteena on ruotsin kielen yo-kirjoitusten pistemäärä ja prediktoreina koulu (koulut A ja B) ja peruskoulun lukuaineiden keskiarvo. Annetaan mallille nimeksi `lukiot.3`. Ladataan ensin `arm` ja ajetaan regressio:

```
library(arm)
lukiot.3 <- lm(ruotsi.pist ~ kouluA + lka)
display(lukiot.3)
```

Tuloksena on

```
            coef.est coef.se
(Intercept) -156.90    29.27
kouluA         8.79     4.09
lka           43.33     3.39
---
n = 182, k = 3
residual sd = 27.28, R-Squared = 0.49
```

Funktio `display()`, joka on tuotti eo. tulostuksen tulee paketin `arm` mukana. Se antaa enemmän tietoja kuin `print()` mutta vähemmän kuin `summary()`.

Pienimmän neliösumman estimaatit, jäännökset ja sovitteet

Mallin (1.3) pienimmän neliösumman (p.n.s.) estimaatit saadaan minimoimalla neliösumma

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

kertoimien $\beta_0, \beta_1, \ldots, \beta_p$ suhteen. Menetelmä on vaikuttaa järkevältä, sillä se etsii sellaiset kertoimet, jotka minimoivat ennustevirheiden neliösumman aineistossa. Ratkaisu $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)'$ voidaan kirjoittaa matriisien avulla muodossa

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Tässä vaiheessa on hyödyllistä huomata, että estimaatti $\widehat{\boldsymbol{\beta}}$ saadaan lineaarisesti, (matriisikertolaskulla) vasteesta $\boldsymbol{y}$. Käytännössä numeeriset laskut hoituvat tehokkaiden algoritmien avulla laskematta itse käänteismatriisia.

Mallin sovitteet ja jäännökset ovat vastaavasti

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, \\ e_i &= y_i - \hat{y}_i, \quad i = 1, \ldots, n, \end{aligned}$$

Varianssin σ^2 estimaatti saadaan jäännöksistä:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - p - 1}.$$

Kertoimien epävarmuus: keskivirheet

Tulostuksessa on mukana kertoimien keskivirheet. Karkeasti ajatellen korkeintaan kahden keskivirheen poikkeamat lasketusta estimaatista eivät ole ristiriidassa aineiston kanssa. Esim. `kouluA`:n kertoimen estimaatti on 8.79, ja sen keskivirhe on 4.09. Siis aineisto on sopusoinnussa β_1:n arvojen kanssa, jotka ovat välillä $[8.79 \pm 2 \cdot 4.09] = [0.61, 16.97]$. Mikäli mallin kaikki olettamuksen ovat voimassa, saatu väli on n. 95 %:n *luottamusväli*. Tämä tarkoittaa, että estimaatti $\hat{\beta}_1$ poikkeaa oikeasta arvosta β_1 vain 5 %:ssa tapauksista enemmän kuin kahden keskivirheen verran. Jos laskettaisiin kahden keskivirheen sijasta yhden keskivirheen poikkeamia, saataisiin n. 68%:n ($\approx 2/3$) luottamusväli.

Täsmällisempi tapa laskea tämä luottamusväli on käyttää t-jakaumaa, jonka *vapausasteet* saadaan vähentämällä tapausten lukumäärästä n estimoitujen kertoimien lukumäärä $p + 1$ (tulostuksessa $k = p + 1$). Tässä esimerkissä $n = 182$, $p = 2$ ja t-jakaumasta saatu arvo on 1.97.

Kertoimien epävarmuudesta kertoo niiden estimoitu kovarianssimatriisi $\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Merkitään $\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, ja indeksoidaan $\boldsymbol{C}$:n rivit ja sarakkeet poikkeuksellisesti 0:sta p:hen, so. $\boldsymbol{C} = (c_{ij}), i, j = 0, 1, \ldots, p$. Silloin $\hat{\beta}_1$:n estimoitu varianssi on $\hat{\sigma}^2 c_{11}$. Estimaattien $\hat{\beta}_1$ ja $\hat{\beta}_2$ estimoitu kovarianssi on $\hat{\sigma}^2 c_{12}$, ja niiden korrelaatiokerroin on $c_{12}/\sqrt{c_{11}c_{22}}$.

Jäännösten keskihajonta ja selitysaste

Esimerkissämme jäännösten keskihajonta on $\hat{\sigma} = 27.28 \approx 27$, minkä voi tulkita niin, että yksittäisen kokelaan pistemäärä poikkeaa ennustetusta vähemmän kuin 27 pistettä todennäköisyydellä 2/3.

Sovitteen hyvyyttä voidaan mitata $\hat{\sigma}$:n avulla (mitä pienempi $\hat{\sigma}$ sen parempi yhteensopivuus) tai selitysasteen avulla. Se kertoo, kuinka suuren osan vasteen vaihtelusta malli selittää. Selitysasteen saa kaavasta

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}. \tag{1.4}$$

Esimerkissämme koulujen välinen ero ja kokelaiden erot peruskoulun lukuaineiden keskiarvoissa selittävät ruotsin kielen pistemäärien vaihtelusta 49 %. Pistemäärää voidaan siis jossain määrin ennustaa peruskoulumenestyksen avulla. Toisaalta kovin suuri ennustetarkkuus ei ehkä olisi toivottavaakaan.

Tilastollinen merkitsevyys

Suurin piirtein pätee, että *jos regressiokerroin poikkeaa enemmän kuin 2 keskivirheen verran nollasta, niin kerroin on* ***tilastollisesti merkitsevä***. Silloin voimme olla melko varmoja, että kertoimen etumerkki on oikein.

```
lka <- lka - mean(lka)
lm(formula = ruotsi.pist ~ kouluA + clka + kouluA:clka)
            coef.est coef.se
(Intercept) 213.88     2.69
kouluA        9.27     4.10
clka         46.79     4.48
kouluA:clka  -8.11     6.85
---
n = 182, k = 4
residual sd = 27.25, R-Squared = 0.49
```

Yo. analyysissa `kouluA`:n kerroin 9.27 on merkitsevä. Siis koulun A oppilaat, joiden lukuaineiden keskiarvo on 8.6 (= lukuaineiden keskiarvojen keskiarvo) menestyvät keskimäärin merkitsevästi paremmin kuin vastaavat koulun B oppilaat. Sen sijaan interaktio ei ole merkitsevä, ts. prediktorin `lka` kulmakertoimet kouluissa A ja B eivät eroa merkitsevästi toisistaan. Juuri koskaan *emme* ole kiinnostuneita siitä, poikkeaako vakio tilastollisesti merkitsevästi nollasta.

Tarkkaan ottaen regressiokertoimen merkitsevyys perustuu siihen, että

$$T = \frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)} \sim t(n-p-1),$$

missä β_j on oikea (tuntematon) arvo. Kun vapausasteet eivät ole kovin pieniä pätee $P(|T| > 2) \approx 0.05$. Näin ollen likimääräinen 95%:n luottamusväli saadaan, kun etsitään ne β_j:n arvot, jotka toteuttavat epäyhtälön $|T| \leq 2$. Epäyhtälön ratkaisu on selvästikin väli, jonka päätepisteet ovat $\hat{\beta}_j \pm 2\mathrm{se}(\hat{\beta}_j)$. Kertoimen merkitsevyys siis tarkoittaa, että arvo $\beta_j = 0$ ei kuulu luottamusvälille. Kuten aikaisemmin on todettu, arvon 2 käyttö on nopea likimääräinen menettely. Tarkemman arvon saa t-jakaumasta.

Jotkut ovat sitä mieltä, että jos kerroin ei ole merkitsevä, niin vastaavaa prediktori on jätettävä mallista pois. Näin yksioikoisesti ei kuitenkaan ole syytä ajatella. Palaamme asiaan myöhemmin.

Jäännösvarianssin epävarmuus

Mallin olettamusten nojalla jäännösvarianssin $\hat{\sigma}^2$ otosjakauma on verrannollinen $\chi^2(n-p-1)$-jakaumaan. Täsmällisesti pätee

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1).$$

1.5 Olettamukset ja diagnostiikka

Seuraavat olettamukset ovat *laskevassa* tärkeysjärjestyksessä.

1. *Mallin pätevyys.* Mallin pitää antaa vastaus asetettuun tutkimusongelmaan. Vastemuuttujan pitää mitata sitä ilmiötä tai ominaisuutta, josta ollaan kiinnostuneita, ja mallin pitää sisältää kaikki relevantit prediktorit. Mallin pitää päteä niihin tapauksiin, joihin sitä tullaan soveltamaan.

   Syötemuuttujien valinta on usein analyysin hankalin vaihe. Pitäisi ottaa kaikki relevantit muuttujat, mutta käytännössä on voi olla vaikeata päättää, mitkä ovat välttämättömiä. Lisäksi on vaikeata tulkita sellaisia kertoimia, joiden keskivirheet ovat suuria.

2. *Lineaarisuus.* Tärkein matemaattinen olettamus on mallin lineaarisuus. Deterministinen komponentti on lineaarinen funktio prediktoreista: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p$. Jos lineaarisuus ei päde, voi yrittää muunnosta, esim. voi ottaa vasteeksi $\log y$ tai $1/y$ tms. Joskus saattaa olla tarpeen ottaa prediktoriksi x:n lisäksi myös sen neliö x^2. Tavoitteena on lineaarinen malli, joka on riittävä approksimaatio tutkittavalle ilmiölle.

3. *Virheiden riippumattomuus.* Tavallisessa regressiomallissa virheet oletetaan riippumattomiksi. Myöhemmin käsittelemme tilanteita, joissa tämä olettamus ei päde.

4. *Virhevarianssien vakioisuus.* Jos virhevarianssit $\text{var}(\varepsilon_i) = \sigma_i^2$ vaihtelevat, tehokkaampi estimointi edellyttää painotettua pienimmän neliösumman menetelmää, jossa painot ovat käänteisesti verrannollisia virhevariansseihin. Palaamme tähänkin myöhemmin. Useimmissa tapauksissa virhevarianssin poikkeama vakioisuudesta on vaikutuksiltaan vähäinen. Tavallisin poikkeama virhevarianssin vakioisuudesta on sellainen, että jäännösten vaihtelu on suurempaa silloin, kun sovitteen arvot ovat suuria. Tämä näkyy ns. "megafonikuviona", kun piirretään jäännökset sovitteen suhteen.

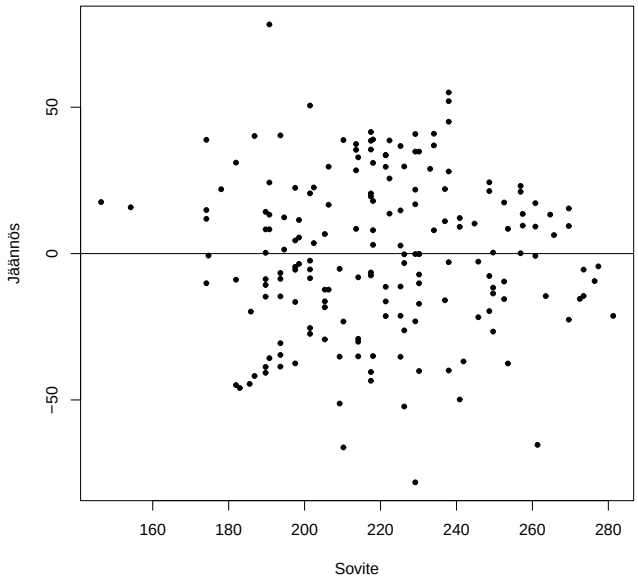
5. *Virheiden normaalisuus.* Tämä olettamus on yleensä vähiten tärkeä. Useimmiten kertoimen otosjakaumat ovat keskeisen raja-arvolauseen nojalla melko normaalisia riippumatta virheen jakaumasta. Virheiden normaalisuudella on merkitystä, kun mallin avulla lasketaan vasteen tulevien arvojen ennustevälejä. Jäännöstarkasteluissa kannattaa kiinnittää huomiota poikkeuksellisen suuriin (positiivisiin) ja pieniin (negatiivisiin) arvoihin (outliereihin). Niitä vastaavilla tapauksilla saattaa olla merkittävä vaikutus kertoimien arvoihin ja niiden keskivirheisiin. Jäännösten huomattava vinous saattaa kertoa tarpeesta tehdä muunnos vasteeseen.

Jäännöskuviot

Hyvä tapa tutkia poikkeamia lineaarisuudesta on piirtää jäännökset sovitteen suhteen, so. jäännökset pystyakselille ja sovitteet vaaka-akselille. Lisäksi jäännökset voi

piirtää jokaisen prediktorin suhteen. *Mikäli mallin olettamukset ovat voimassa, mitään systemaattista kuviota ei pitäisi näkyä.* Diktomisen prediktorin kohdalla kannattaa useinkin täristää kuten kuvassa 1.1. Voi myös piirtää *vasteen arvot sovitteen suhteen*, jolloin olettamusten vallitessa pisteparven pitäisi keskittyä suoran ympärille, jonka kulmakerroin on yksi. Kuva 1.6 näyttää aika hyvältä. Tavanomaisesta poikkeavasti jäännösten varianssi ehkä hieman vähenee vasteen kasvaessa.

Vaikka normaalisuus ei kovin tärkeä tässä esimerkissä olekaan, saattaa olla hyödyllistä piirtää otoskvantiilit teoreettisten suhteen ("Normal QQ-plot"), ks. kuva 1.7. Kuva ei paljasta mitään dramaattista.



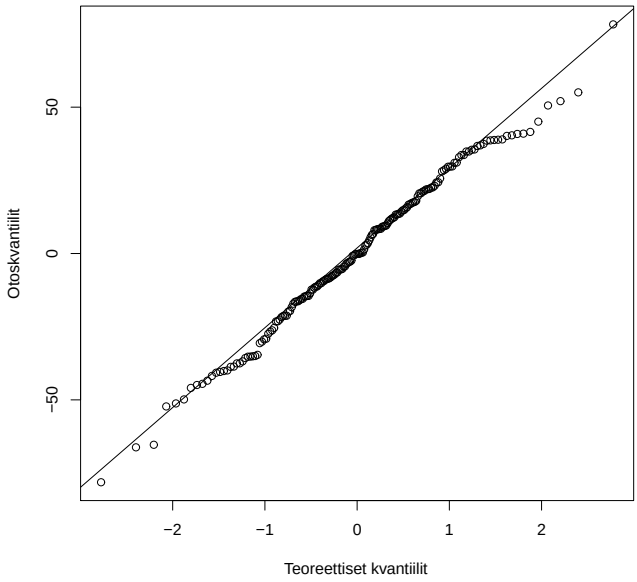
Kuva 1.6: *Mallin* (1.1) *jäännökset sovitteen suhteen.*

1.6 Ennustaminen ja validointi

Joskus regressiomallin tarkoituksen on ennustaa tulevia havaintoja. Joskus jo olemassa olevia otoksen ulkopuolisia havaintoja ennustetaan otoksen perusteella, jolloin tarkoituksena on mallin hyvyyden arviointi.

Ennustaminen

Mallin (1.1) $\widehat{\texttt{ruotsi.pist}} = -156.9 + 8.8\,\texttt{kouluA} + 43.3\,\texttt{lka}$ mukaisesti, jos koulun A oppilaan lukuaineiden keskiarvo on 7, ennustamme hänen ruotsin ylioppilaskokeen pistemääräkseen $-156.9 + 8.8 + 43.3 \cdot 7 = 155.2 \approx 155$. Jos estimoitu malli olisi täsmälleen oikein, niin ennusteen keskivirhe olisi $\hat{\sigma} = 27.3$ pistettä. Itse asiassa



Kuva 1.7: *Mallin* (1.1) *jäännösten normaalisuus.*

keskivirhe on suurempi, kun otetaan huomioon regressiokertoimien estimointivirhe. Matriisimerkinnöin ennusteen keskivirheen kaava on

$$\hat{\sigma}_{\text{pred}} = \hat{\sigma}\sqrt{1 + \boldsymbol{x}_a'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_a},$$

missä $\boldsymbol{x}_a$ sisältää ennustettavaan arvoon liittyvien prediktorien arvot. Meidän tapauksessamme $\boldsymbol{x}_a = (1, 1, 7)$, joka tuottaa keskivirheen 28.0. Vastaava ennuste koulun B oppilaalle saadaan, kun valitaan $\boldsymbol{x}_a = (1, 0, 7)$. Ennuste on silloin 146, jonka keskivirhe on 27.9.

Jos meillä on otoksen ulkopuolisia havaintoja, voi mallin ennustekykyä voi validoida niiden suhteen. Useinkaan tätä mahdollisuutta ei ole.

Luku 2

Lineaarinen regressio: mallin rakentaminen

Aina ei ole syytä sovittaa regressiomallia aineistoon sen alkuperäisessä muodossa. Tässä luvussa tarkastelemme erilaisia muunnoksia ja syötemuuttujien yhdistelyjä sopivien prediktoreiden valitsemiseksi.

2.1 Lineaarinen muunnos

Lineaariset muunnokset eivät vaikuta sovitteen eivätkä ennusteiden hyvyyteen, mutta niiden avulla voi vaikuttaa mallin tulkittavuuteen. Olemme jo huomanneet, että keskistämällä jatkuvat prediktorit saamme tulkinnan regressioyhtälön vakiolle.

Prediktorin skaalaus

Tarkastellaan erästä amerikkalaista aineistoa, jossa vasteena asunnon hinta (dollareina) ja prediktorina pinta-ala neliöjalkoina. Estimoitu regressioyhtälö on

$$\widehat{\texttt{hinta}} = -3721 + 54.8\,\texttt{ala} \text{ (neliöjalkaa)}.$$

Kertoimen tulkinta on, että jos asuntojen pinta-alat poikkeavat yhden neliöjalan verran, niin hintaeron ennuste on 54.8 \$. Koska neliöjalka on jokseenkin epähavainnollinen pinta-alamitta mannereurooppalaiselle, se kannattaa muuntaa neliömetreiksi. Kun vielä keskistää prediktorin, saa regressioyhtälön

$$\widehat{\texttt{hinta}} = 78800 + 590\,(\texttt{ala} - 140) \text{ (neliömetriä)}.$$

Siis neliömetrin ero pinta-aloissa ennustaa hinnan eroksi 590 \$. Vakio ennustaa keskikokoisen asunnon (140 m^2) hinnaksi 78800 \$.

Standardointi

Joskus prediktori standardoidaan l. siitä vähennetään keskiarvo ja se jaetaan keskihajonnalla. Sitä voi käyttää silloin, kun prediktorilla ei ole kunnon mittayksikköä. Yleensä mittayksikölliset prediktorit kannattaa jättää silleen tai skaalata tilanteen mukaan sopivasti, esim. tulot tuhansina euroina, ikä kymmeninä vuosina tms. Jos mallissa on interaktioita, kannattaa tulkinnallisista syistä usein keskistää jatkuvat prediktorit. Sama pätee silloin, kun haluaa vakiolle tulkinnan.

Prediktoreiden lineaarinen yhdistely

Seuraava esimerkki osoittaa, että joskus kertoimien tulkinta vaatii hiukan ponnistelua. Sovitetaan asuntoaineistoon malli, jossa vasteena on edelleen hinta (nyt tuhansia dollareita), mutta prediktoreina ovat huoneiden lukumäärä `hluku` ja makuuhuoneiden lukumäärä `mhluku`. estimoitu yhtälö on

$$\widehat{\texttt{hinta}} = -14.9 + 9.8\,\texttt{hluku} + 10.9\,\texttt{mhluku}.$$

Kun verrataan kahta asuntoryhmää, joista toisessa on yksi huone enemmän, mutta molemmissa ryhmissä on yhtä monta makuuhuonetta, niin ne asunnot, joissa on yksi huone enemmän, ovat keskimäärin 9800 $ kalliimpia. Tässä täytyy huomata, että tämä yksi *lisähuone on jokin muu kuin makuuhuone.* Kun taas verrataan kahta ryhmää, joissa on yhtä monta huonetta, niin se ryhmä, jossa on yksi makuuhuone enemmän, on keskimäärin 10900 $ kalliimpi. Yksinkertaisempi tulkinta saadaan, kun valitaan prediktoreiksi `mhluku` ja erotus `hluku` - `mhluku` = `muuhuone`. Silloin

$$\widehat{\texttt{hinta}} = -14.9 + 9.8\,\texttt{muuhuone} + 20.8\,\texttt{mhluku}.$$

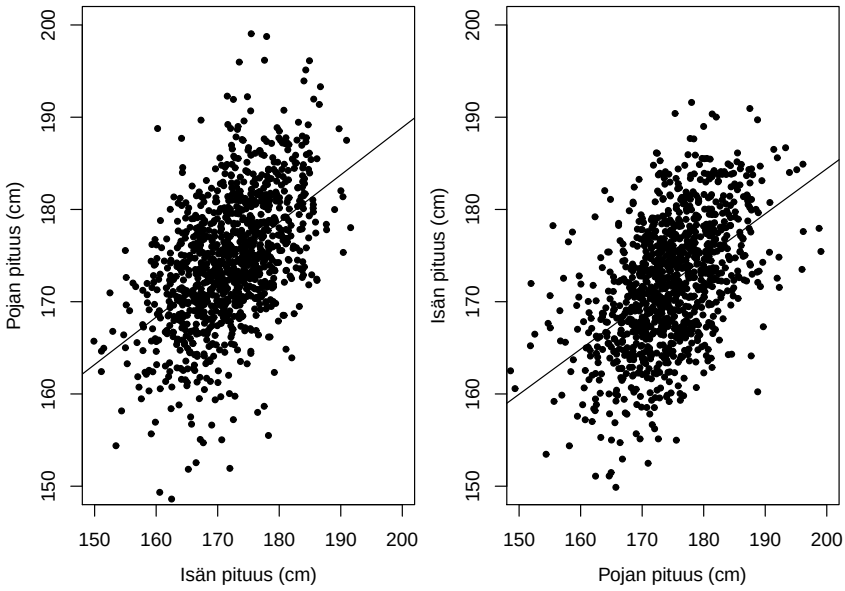
Siis kun verrataan kahta ryhmää joista toisessa on yksi makuuhuone enemmän, mutta muita huoneita on sama määrä, niin hintaero on 20800 $. Jälkimmäinen malli on helpommin tulkittavissa, vaikka molemmissa malleissa on sama informaatio. Huomaa myös, että edellisen mallin prediktorin `hluku` kerroin on sama kuin jälkimäisen mallin `muuhuone`.

Korrelaatio ja "regressio kohti keskiarvoa"

Tarkastellaan yhden prediktorin mallia $y = \beta_0 + \beta_1 x + \varepsilon$. Voidaan osoittaa, että kertoimen β_1 p.n.s estimaatti on

$$\hat{\beta}_1 = r\frac{s_y}{s_x}, \tag{2.1}$$

missä r on muuttujien x ja y korrelaatiokerroin ja s_x ja s_y ovat niiden keskihajonnat vastaavasti. Jos sattuu olemaan $s_x = s_y$, niin regressiokerroin on sama kuin korrelaatiokerroin, jonka tiedetään olevan välillä $[-1, 1]$. Tämä tilanne saadaan aikaiseksi jos sekä vaste että prediktori standardoidaan. Voimme lisäksi päätellä yleisessä tilanteessa kaavasta (2.1), että jos $\hat{\beta}_1 > 1$ tai $\hat{\beta}_1 < -1$ niin vasteen keskihajonta on suurempi kuin prediktorin keskihajonta.



Kuva 2.1: *Pearsonin aineisto isistä ja pojista. Suora on molemmissa kuvissa pienimmän neliösumman suora.*

Tarkastellaan nyt Karl Pearsonin aineistoa isien ja poikien pituuksista. Kuvassa 2.1 on piirretty sekä poikien pituudet isien pituuksien suhteen että päinvastoin. Huomaamme vasemmanpuoleisesta kuviosta, että lyhyillä isillä on lyhyitä poikia, mutta pojat ovat kuitenkin keskimäärin pidempiä kuin isänsä. Pitkillä isillä on pitkiä poikia, mutta pojat ovat keskimäärin lyhyempiä kuin isänsä. Kun vaihdetaan isien ja poikien roolit, näemme oikeanpuoleisesta kuvasta, että lyhyillä pojilla on lyhyitä isiä, mutta isät ovat keskimäärin pidempiä kuin poikansa ja että pitkillä pojilla on pitkiä isiä, mutta nämä ovat keskimäärin lyhyempiä kuin poikansa. Tällaista ilmiötä sanotaan regressioksi kohti keskiarvoa (tässä regressio tarkoittaa taantumista tai paluuta). Tätä ei pidä tulkita niin, että sukupolvien kuluessa kaikista tulisi samanmittaisia. Pitää muistaa, että ennusteeseen liittyy virhe, joka pitää vaihtelua yllä.

Kun ennustetaan isän pituudella pojan pituutta saamme yhtälön

$$\widehat{\texttt{pojan pituus}} = 86.1 + 0.51\,\texttt{isän pituus}.$$

Ennuste 155 senttisen isän pojalle on 166 cm ja 185 senttisen isän pojalle 181 cm. Vastaavat ennusteet isälle kun poika on 155 tai 185 cm ovat 162 cm ja 177 cm. Jälkimmäiset saamme kaavasta

$$\widehat{\texttt{isän pituus}} = 86.6 + 0.49\,\texttt{pojan pituus}.$$

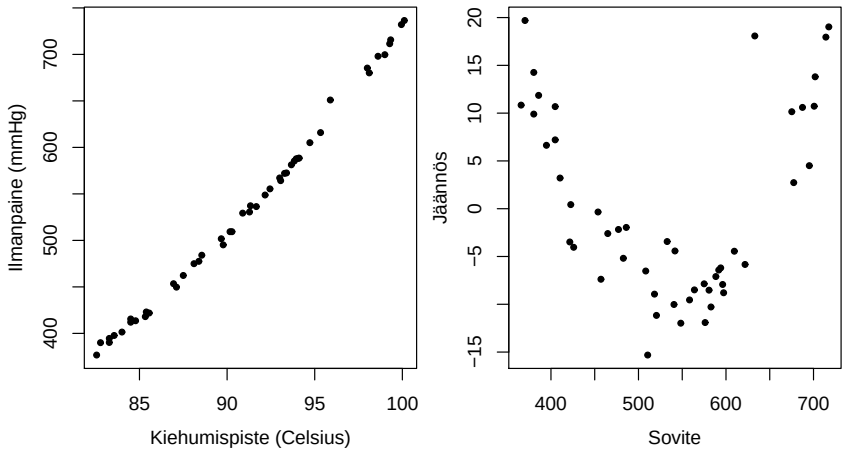
2.2 Logaritmimuunnos

Logaritmi vasteesta

Aineistoissa `forbes` ja `hooker` on mittauksia veden kiehumispisteestä (`Temp`, Fahrenheit) ja ilmanpaineesta (`Pressure`, tuumaa elohopeaa). Forbes keräsi aineistonsa Alpeilla ja Skotlannissa ja Hooker Himalajalla (Weisberg, 2005, s. 5 ja 40). Tarkoituksena on ennustaa ilmanpaine veden kiehumispisteen avulla. Ilmanpaineesta voidaan sitten päätellä miten korkealla vuoristossa ollaan. Alkuperäiset mittayksiköt (paine = inchHg, lämpötila = Fahrenheit) on muutettu tutummiksi mmHg (millimetriä elohopeaa) ja Celsius-asteiksi. Estimoitu regressioyhtälö on

$$\widehat{\texttt{paine}} = -1287 + 20.03\,\texttt{lämpötila}.$$

Regressiokertoimen perusteella yhden Celsius-asteen ero kiehumispisteissä ennustaa 20 mmHg:n eroa paineessa. Vakio antaa paineen, kun kiehumispiste on 0° C. Ennuste on tietysti mieletön, koska paine ei voi olla negatiivinen. Tämä on yksi argumentti muunnoksen puolesta.



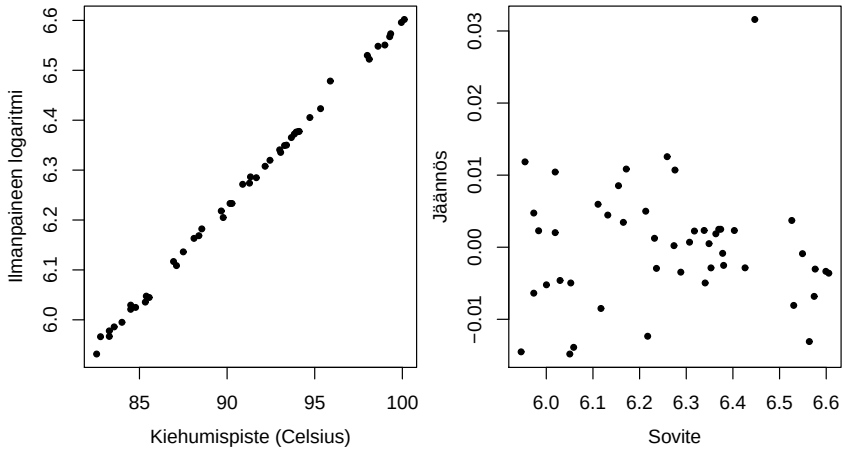
Kuva 2.2: *Vasemmalla ilmanpaine kiehumispisteen suhteen Forbesin ja Hookerin mittauksista. Oikealla additiivisen mallin jäännökset sovitteen suhteen.*

Kuva 2.2, että mallissa on systemaattista virhettä. Vasemmanpuoleisessa kuvassa näemme hienoisen poikkeaman lineaarisuudesta. Oikea puoli paljastaa selvemmin, että ilmanpaineen ja kiehumispisteen yhteys on epälineaarinen. Tällaisessa tilanteessa kannattaa kokeilla logaritmimuunnosta vasteeseen. Yleensäkin kannattaa kokeilla logaritmimuunnosta sellaisissa tilanteissa, joissa vasteen arvot ovat positiivisia, jos sovite alkuperäisissä yksiköissä näyttää huonolta. Molemmat kuvat 2.3 ovat selvästi tyydyttävämpiä. Jälkimmäisessä näkyy yksi selvä poikkeava jäännös. Palaamme tähän myöhemmin.

Kuvan 2.3 malli on

$$\widehat{\log \texttt{paine}} = 2.84 + 0.0376\,\texttt{lämpötila}.$$

Vakiota ei kannata ehkä tässäkään ottaa vakavasti, koska mitatut lämpötilat ovat hyvin kaukana nolla-asteesta. Yhden Celsius-asteen ero ennustaa paineen logaritmissa



Kuva 2.3: *Vasemmalla ilmanpaineen logaritmi kiehumispisteen suhteen Forbesin ja Hookerin mittauksista. Oikealla logaritmisen mallin jäännökset sovitteen suhteen.*

eroa 0.0376. Soveltamalla eksponenttimuunnosta puolittain saadaan ennuste paineelle

$$\widehat{\texttt{paine}} = e^{2.84+0.0376\,\texttt{lämpötila}} = e^{2.84}e^{0.0376\texttt{lämpötila}}.$$

Kun tarkastellaan kahta lämpötilaa x ja $x+1$, niin ennusteiden osamäärä on

$$\frac{e^{2.84}e^{0.0376(x+1)}}{e^{2.84}e^{0.0376x}} = e^{0.0376} = 1.0383.$$

Siis paineen ennuste on 3.8 % suurempi korkeammassa lämpötilassa. Jos regressiokerroin pyöristetään kahteen merkitsevään numeroon, saamme 0.038, joka on sadalla kerrottuna sama kuin aikaisempi muutosprosentti.

Tarkastellaan useamman prediktorin mallia

$$\log y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

Oletetaan, $i = 1, 2$ ja että tilanteessa $i = 2$ $x_{2j} = x_{1j} + 1$ mutta muut prediktorit ovat ennallaan. Silloin

$$\frac{E(y_2)}{E(y_1)} = \frac{e^{\beta_0}e^{\beta_1 x_{11}} \cdots e^{\beta_j(x_{1j}+1)} \cdots e^{\beta_p x_{1p}} E(e^{\varepsilon_1})}{e^{\beta_0}e^{\beta_1 x_{11}} \cdots e^{\beta_j x_{1j}} \cdots e^{\beta_p x_{1p}} E(e^{\varepsilon_2})} = e^{\beta_j}. \tag{2.2}$$

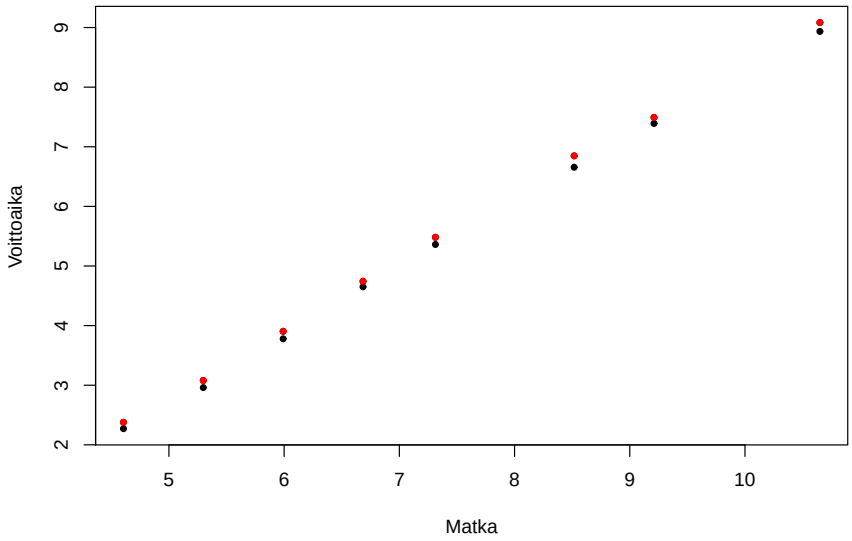
Siis kerroin e^{β_j} antaa suhteellisen eron ennusteissa. Jos kerroin β_j on pieni pätee

$$e^{\beta_j} \approx 1 + \beta_j,$$

joten kerroin β_j sellaisenaan antaa likiarvon suhteelliselle erolle. Kertomalla 100:lla saadaan prosenttiluku. Tämän yksinkertaisen likimääräisen tulkinnan takia suosimme yleensä luonnollista logaritmia 10-kantaisen sijaan. Jos edellisessä esimerkissä käytetään 10-kantaista logaritmia, saadaan yhtälö

$$\widehat{\log_{10}(\texttt{paine})} = 1.24 + 0.0163\,\texttt{lämpötila}.$$

Kun verrataan paineiden suhteellisia eroja aikaisemman tapaan, saamme tuloksen, että yhden asteen ero lämpötilassa ennustaa paineiden suhteellista eroa $10^{0.0163} = 1.0383$, joka on täsmälleen sama kuin aikaisemmin. Tässä mielessä kantaluvun valinnalla ei ole väliä, mutta luonnollisen logaritmin käyttö antaa likiarvon nopeasti.



Kuva 2.4: *Pekingin olympialaisten "sileiden"juoksumatkojen voittoajat matkan pituuden suhteen. Molemmissa on logaritmiasteikko. Naiset on merkitty punaisella ja miehet mustalla.*

Logaritmi vasteesta ja prediktorista

Joskus on syytä ottaa logaritmi sekä vasteesta että prediktorista. Kuvassa 2.4 on Pekingin olympialaisten juoksumatkojen voittoajat matkan suhteen. Molemmat on muunnettu logaritmiasteikolle. Estimoitu regressioyhtälö on

$$\widehat{\log \texttt{aika}} = -2.88 + 0.126\,\texttt{nainen} + 1.12\,\texttt{log(matka)},$$

joka alkuperäiseen skaalaan muunnettuna on

$$\widehat{\texttt{aika}} = e^{-2.88} e^{0.126\,\texttt{nainen}} \texttt{matka}^{1.12}.$$

Vakiolla ei tässä ole kovin hyvää tulkintaa, mutta nopea likiarvo naisten ja miesten erolle saadaan: Kun verrataan miehiä ja naisia heidän juostessaan saman matkan, niin regressiokertoimesta saadaan suoraan, että naiset ovat keskimäärin $12.6 \approx 13$ % hitaampia; tässä myös $e^{0.126} = 1.134 \approx 1.13$. Kun verrataan miehiä keskenään ja naisia keskenään, niin 1 %:n ero matkassa ennustaa aikojen suhteellista eroa

$$\frac{e^{-2.88} e^{0.126\,\texttt{nainen}} (\texttt{1.01 matka})^{1.12}}{e^{-2.88} e^{0.126\,\texttt{nainen}} \texttt{matka}^{1.12}} = 1.01^{1.12} = 1.0112.$$

Suhteellinen ero on siis 1.12 %, joka on sama kuin regressiokerroin sellaisenaan. Tulos ei ole sattuma. Samanlaisella päättelyllä kuin kaavassa (2.2) saadaan nimittäin tulos, että 1 %:n ero j. prediktorissa, joka on mallisssa logaritmoituna vasteenkin ollessa logaritmoitu, ennustaa suhteellista eroa

$$\left(1 + \frac{1}{100}\right)^{\beta_j} \approx 1 + \frac{\beta_j}{100}.$$

Siis β_j % on suoraan likimääräinen prosentuaalinen ero, kun verrataan prediktorin arvoja, jotka poikkeavat 1 %:lla.

Kuvion mukaan mitään ilmeistä interaktiota ei ole. Ennuste pätee näin ollen sekä miehille että naisille. Tarkemmassa analyysissä huomaa kyllä, että mallissa on systemaattista virhettä, mutta karkeana arviona tulos kuitenkin pätee.

Kun sekä vaste että prediktori ovat logaritmoituja, niin kantaluvulla ei ole väliä: regressiokerroin pysyy samana. On kuitenkin muistettava, että jos mallissa on mukana myös prediktoreita, joita ei ole logaritmoitu, niin niiden kertoimien nopea likimääräinen tulkinta menetetään, kun käytetään jotakin muuta kuin luonnollista logaritmia.

2.3 Muita muunnoksia

Neliöjuurimuunnos

Neliöjuurimuunnos on joskus hyödyllinen, kun halutaan suurien arvojen vaikutusta vähentää. Se on lievempi kuin logaritmimuunnos. Esim. vuosiansioiden ero 0 euron ja 10000 euron välillä tuntuu paljon suuremmalta kuin ero 80000 euron ja 90000 euron välillä. Logaritmiasteikolla 5000 euron ja 10000 euron ero on sama kuin 40000 euron ja 80000 euron välillä. Neliöjuuriasteikolla 0 euron ja 10000 euron ero on sama kuin 10000 euron ja 40000 euron välillä ja 40000 euron ja 90000 euron välillä. Neliöjuurimuunnoksen ikävä puoli on se, että regressiokertoimilla ei ole yhtä selkeää tulkintaa kuin aritmeettisella ja logaritmiasteikolla. Neliöjuurimuunnosta kannattaa ehkä käyttää vain silloin, kun mallilla ennustetaan, mutta kertoimien tulkinnalla ei ole suurta merkitystä.

Käänteislukumuunnos

Käänteismuunnos on myös joskus hyödyllinen. Polttoaineen kulutus on tapana ilmoittaa Suomessa yksikössä litra/100 km. Jos kulutus on 7 l/100 km, niin litralla voi ajaa $100/7 = 14.3$ kilometriä. Samanlainen muuttuja on kuukausipalkka. Käänteismuunnos vähentää suurten lukujen eroa vielä enemmän kuin logaritmimuunnos.

Box-Cox -muunnos

Kaikki nämä muunnokset ovat muotoa $u = y^\alpha$, kun sovitaan, että $\alpha = 0$ vastaa logaritmimuunnosta. Sopimus on järkevä, sillä

$$\lim_{\alpha \to 0} \frac{y^\alpha - 1}{\alpha} = \log y.$$

Näitä muunnoksia sanotaan Box-Cox -muunnoksiksi. Ne sopivat sellaisille muuttujille, jotka saavat vain positiivisia arvoja. Kun $\alpha > 0$, niin periaatteessa nollakaan ei tuota ongelmia.

Jatkuvien prediktoreiden luokittelu

Yleensä jatkuvat muuttujat kannattaa pitää jatkuvina ennemmin kuin tehdä mitään luokitteluja. Luokittelussa aina häviää informaatiota. Ikä on kuitenkin usein sellainen muuttuja, joka usein luokitellaan, koska sukupolvien väliset erot ovat usein vaihtelevia. Poliittisia asenteita tutkittaessa yksi luokittelu voisi olla 18–29, 30–44, 45–64 ja 65+.

2.4 Luokittelevat prediktorit

Kategoriset l. luokittelevat syötemuuttujat ovat tavallisia. Aikaisemmin vertailimme kahden koulun eroja ruotsin kielen ylioppilaskokeessa. Aineistossa `yoruotsi` on tulokset neljästä eri koulusta: A, B, C ja D. Voimme tehdä neljä indikaattorimuuttujaa

- `kouluA` = 1, kun kokelas on koulusta A, 0 muulloin,
- `kouluB` = 1, kun kokelas on koulusta B, 0 muulloin,
- `kouluC` = 1, kun kokelas on koulusta C, 0 muulloin,
- `kouluD` = 1, kun kokelas on koulusta D, 0 muulloin.

Sovitetaan malli, jossa prediktoreina ovat indikaattorit lukuun ottamatta indikaattoria `kouluD`.

```
lm.koulu <- lm(ruotsi.pist ~ kouluA + kouluB + kouluC)
display(lm.koulu)
lm(formula = ruotsi.pist ~ kouluA + kouluB + kouluC)
            coef.est coef.se
(Intercept) 224.00     3.60
kouluA        2.41     5.38
kouluB       -8.91     5.02
kouluC        2.23     5.11
---
n = 375, k = 4
residual sd = 35.50, R-Squared = 0.02
```

Kertoimien tulkinta: Vakio ilmoittaa koulun D keskiarvon. Prediktorin `kouluA` kerroin tarkoittaa, että koulun A keskiarvo on 2.41 pistettä suurempi kuin koulun D kerroin. Muut prediktorit tulkitaan analogisesti. Huomaamme, että kertoimet ovat erotuksia pois jätetyn koulun suhteen ja että vakio on pois jätetyn koulun keskiarvo. Pois jätetty koulu D on *perustaso*, johon muita verrataan.

Tulkinnan helpottamiseksi on joskus mukava jättää pois se indikaattori, jolla on pienin keskiarvo, jolloin kaikki kertoimet ovat positiivisia. Tehdään uusi ajo, mistä `kouluB` on jätetty pois

```
lm(formula = ruotsi.pist ~ kouluA + kouluC + kouluD)
            coef.est coef.se
(Intercept) 215.09     3.50
kouluA       11.32     5.31
kouluC       11.14     5.04
kouluD        8.91     5.02
---
n = 375, k = 4
residual sd = 35.50, R-Squared = 0.02
```

Prediktoreiden kertoimet ilmoittavat siis eron kouluun B nähden. Muiden koulujen erot saadaan vastaavien kertoimien erotuksena. Esim. Koulujen A ja C ero on 0.18 pistettä. Vakio ilmoittaa nyt koulun B keskiarvon.

Koska syötemuuttuja `koulu` ei ole numeerinen muuttuja, vaan sen arvot ovat kirjaimia A,B,C, ja D, niin `R` tulkitsee sen automaattisesti *faktoriksi*, jonka *tasoja* nämä kirjaimet ovat. `R`:n funktiot osaavat käsitellä faktoreita automaattisesti. Voimme tehdä ajon

```
lm(formula = ruotsi.pist ~ koulu)
            coef.est coef.se
(Intercept) 226.41     3.99
kouluB      -11.32     5.31
kouluC       -0.18     5.39
kouluD       -2.41     5.38
---
n = 375, k = 4
residual sd = 35.50, R-Squared = 0.02
```

Huomaamme, että perustasoksi on nyt valikoitunut koulu A, joka on aakkosjärjestyksessä ensimmäinen. Jos `koulu` olisi koodattu luvuilla 1,2,3,4, ja haluttaisiin käyttää sitä faktorina, se täytyy muuttaa faktoriksi funktiolla `factor()`. Tasojen nimet voivat olla usean merkin mittaisia, esim. sukupuoli voidaan koodata merkkijonoilla "nainen" ja "mies".

Luokittelevan syötemuuttujan merkitsevyys

Edellä olevista tulosteista näemme, että saamme helposti estimaatit ryhmien välisille eroille perustasoon verrattuna. Lisäksi saamme näiden estimaattien keskivirheet

ja niiden kautta erojen merkitsevyydet. Edellä olevista kolmesta sovituksesta saamme kyllä kaikki parittaiset erot ja niiden merkitsevyydet. Huomaamme, että lukiot A ja C eroavat merkitsevästi koulusta B ja että muut erot eivät ole merkitseviä.

Yhden kertoimen merkitsevyys näkyy osamäärästä $|\hat{\beta}_j/\mathrm{se}(\hat{\beta}_j)|$. Jos se ylittää arvon 2 (tai täsmällisemmin t-jakaumasta vapausastein $n-p-1$ saadun todennäköisyyttä 0.975 vastaavan kvanttiilin arvon, joka on likimain 2), niin sanomme kerrointa merkitseväksi. Kolmen tai useamman tason faktori tuottaa prediktoreita, joita on yksi vähemmän kuin tasoja. Miten regressiossa saadaan tällaisen faktorin merkitsevyys? Oletetaan nyt, että tutkittavien prediktoreiden kertoimet ovat vektorissa $\boldsymbol{b}$ ja että niiden estimoitu kovarianssimatriisi on $\boldsymbol{S}$. Jos vektorissa $\boldsymbol{b}$ on r komponenttia, niin sopiva testisuure on

$$F = \boldsymbol{b}'\boldsymbol{S}^{-1}\boldsymbol{b}/r,$$

jota verrataan F jakaumaan vapausastein $(r, n-p-1)$. Katsotaan esimerkkiä.

```
lm.1 <- lm(ruotsi.pist ~ koulu)
b <- coef(lm.1)
S <- vcov(lm.1)
F <- b[2:4]%*%solve(S[2:4, 2:4])%*%b[2:4]/3

F
[1] 2.224669                              # testisuureen arvo

qf(.95, df1=3, df2=lm.1$df.residual)
[1] 2.628968                              # kriittinen arvo

1 - pf(F, df1=3, df2=lm.1$df.residual)
[1,] 0.08491422                           # p-arvo
```

Funktio `vcov()` poimii estimaattien kovarianssimatriisin. Testisuureen arvo on tässä 2.22, joka jää 5 %:n kriittisen arvon 2.63 alle. Toinen tapa on laskea ns. p-arvo. Ne p-arvot, jotka jäävät alle 5 %:n ovat merkitseviä. Tässä p-arvo ylittää 0.05:n Saamme johtopäätöksen, että koulujen välinen ero ei ole merkitsevä. Tämä näyttäisi olevan ristiriidassa aikaisemman havainnon kanssa. Silloinhan saimme tuloksen, että koulu B poikkeaa merkitsevästi kouluista A ja C. Mistä tämä johtuu?

Parittaisissa vertailuissa teemme itse asiassa $g(g-1)/2$ vertailua, kun ryhmiä on g kpl. Kun $g = 4$, vertailuja on 6. Kun tehdään 6 testiä 5 %:n merkitsevyystasolla, niin tulee keskimäärin $6 \cdot 0.05 = 0.3$ merkitsevää eroa, vaikka todellisuudessa ryhmien välillä ei olisi eroja ollenkaan. Tästä syystä faktorin merkitsevyyden arvioinnissa käytetään eo. F-testiä, jolla on se ominaisuus, että jos käytetään merkitsevyys tasoa α ($= 0.05$ tavallisesti), niin faktori tulee merkitseväksi todennäköisyydellä α, kun kaikkien faktoriin liittyvien prediktoreiden kertoimet ovat nollia (tässä koulujen välillä ei ole eroa). Esimerkin F-testin perusteella toteamme, että aivan riittävää näyttöä koulujen eroista ei ole.

2.5 Ennustemallin rakentaminen

On yleensä monia tapoja rakentaa järkevä regressiomalli. Siihen vaikuttavat mallin käyttötarkoitus ja se miten aineisto on kerätty. Tärkeitä valintoja tehdään, kun pohditaan, mitkä ovat mahdollisia syötemuuttujia, miten niitä mitataan ja koodataan ja miten niitä yhdistellään. Nämä ovat isoja kysymyksiä. Jos malliin valitaan liian paljon syötemuuttujia, regressiokertoimien estimaatit tulevat niin epätarkoiksi, että niistä tulee käytännössä hyödyttömiä. Seuraavassa on lueteltu eräitä yleisiä periaatteita (Gelman and Hill, 2007, s. 69):

1. Ota tarkasteluun kaikki sellaiset muuttujat, joiden voi asiaperustein odottaa olevan tärkeitä vasteen ennustamisessa.

2. Ei ole aina välttämätöntä sisällyttää kaikkia potentiaalisia syötemuuttujia sellaisenaan malliin. Esim. niitä voi yhdistää laskemalla summia tai keskiarvoja ja käyttää niitä prediktoreina.

3. Jos joillakin syötemuuttujilla on suuret vaikutukset, kannattaa harkita myös niiden interaktioiden mukaan ottamista.

4. Gelman ja Hill suosittelevat prediktoreiden mukaan ottamisesta seuraavaa strategiaa, joka perustuu regressiokertoimen estimaatin etumerkkiin ja merkitsevyyteen (tyypillisesti 5%:n tasolla):

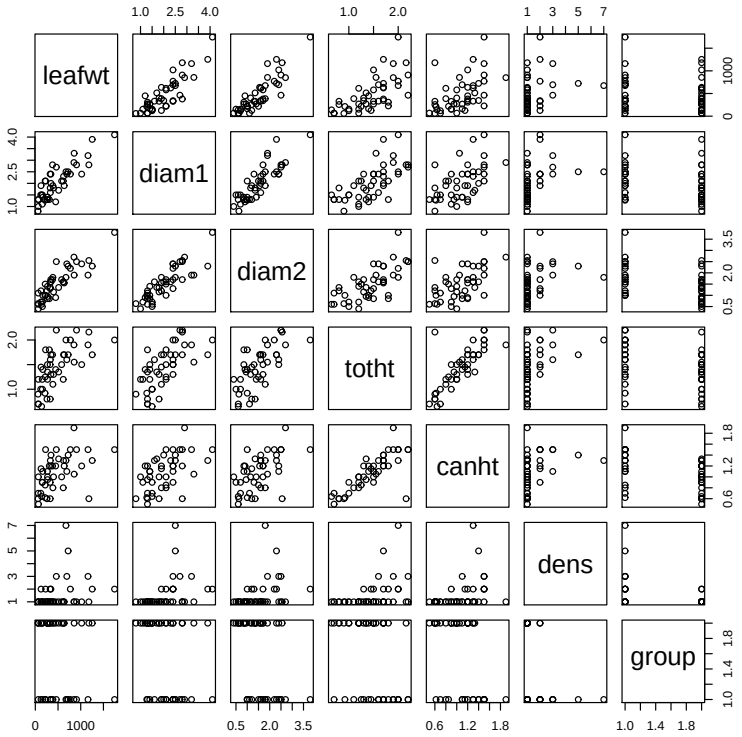
    - Jos prediktorin kerroin on oikean merkkinen, muttei merkitsevä, sen voi pitää mallissa, jos on asiaperusteita sen mukaan ottamiseksi. Se ei ehkä paranna ennustamista kovin paljon, mutta se ei todennäköisesti tuota vahinkoakaan.
    - Jos prediktorin kerroin ei ole merkitsevä, ja sen etumerkki on odotusten vastainen, harkitse sen poistamista mallista.
    - Jos prediktorin kerroin on merkitsevä, mutta sen etumerkki on odotusten vastainen, mieti perusteellisesti, onko siinä järkeä. Sulje pois mahdollinen virhe mallissa, koodauksessa tms. Jos kysymyksessä ei ole virhe, yritä kerätä aineistoa jostakin puuttuvasta muuttujasta, jota mallissa ei vielä ole.
    - Jos prediktorin kerroin on merkitsevä ja oikean merkkinen pidä se ilman muuta mallissa.

Esimerkki 2.1. Tässä esimerkissä on tarkoitus kehittää malli, jonka avulla voidaan ennustaa mesquite-pensaan lehtien biomassa (`leafwt`, grammoina) käyttämällä joitakin tunnuslukuja, jotka ovat helposti mitattavissa yksittäisistä pensaista, ks. Freund et al. (2006, s. 209) ja Gelman and Hill (2007, s. 70–73):

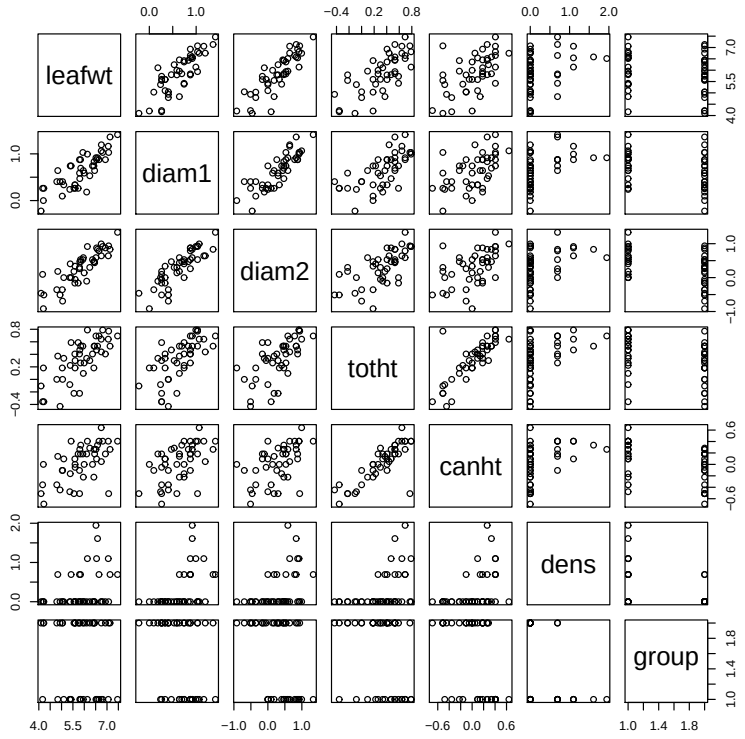
`diam1`	latvuston pidempi halkaisija , metriä,
`diam2`	latvuston lyhyempi halkaisija , metriä,
`canht`	latvuston korkeus, metriä,
`totht`	pensaan koko korkeus, metriä,
`dens`	pensaassa olevien runkojen lukumäärä,
`group`	kaksi pensasryhmää, MCD (26 kpl) ja ALS (19 kpl).

Kuvasta 2.5 näemme, että pensaiden kokoa mittaavat muuttujat ovat jokseenkin lineaarisesti yhteydessä biomassan määrään. Poikkeuksen tekee muuttuja `dens`, joka ei näytä olevan suoraan yhteydessä biomassan määrään. Voisimme siis yksinkertaisesti sovittaa mallin

```
lm(formula = leafwt ~ diam1 + diam2 + canht + totht + dens +
    group)
            coef.est coef.se
(Intercept) -597.59   112.61
diam1        159.20    63.57
diam2        331.02    70.35
canht         30.03   123.77
totht         65.31   107.10
dens           8.28    23.54
```



Kuva 2.5: *Parittaiset pisteparvet* `mesquite`*-aineistosta.*



Kuva 2.6: *Parittaiset pisteparvet* `mesquite`*-aineistosta logaritmiskaalassa.*

```
groupMCD     197.96    59.39
---
n = 45, k = 7
residual sd = 151.41, R-Squared = 0.86
```

Tarkemmin katsottaessa näyttää vasteen varianssi kuitenkin olevan jonkin verran suurempi kookkaammilla pensailla. Lisäksi on luontevaa ajatella, että biomassan määrä riippuu paremminkin multiplikatiivisesti kuin additiivisesti kokoa mittaavista muuttujista. Myös kuvan 2.6 perusteella vasteen varianssi logaritmiskaalassa näyttää vakioiselta. Sovitetaan malli:

```
lm(formula = log(leafwt) ~ log(diam1) + log(diam2) + log(totht) +
    log(canht) + log(dens) + group)
             coef.est coef.se
(Intercept) 4.79     0.16
log(diam1)  0.38     0.28
log(diam2)  1.14     0.21
log(totht)  0.43     0.32
log(canht)  0.32     0.29
log(dens)   0.05     0.13
```

```
groupMCD    0.56     0.13
---
n = 45, k = 7
residual sd = 0.33, R-Squared = 0.87
```

Sen sijaan, että ajattelisimme yhden metrin eron latvuksen korkeudessa liittyvän 30.03 gramman eroon lehtien biomassassa, ajattelemme, että 1 %:n ero latvusten korkeudessa liittyy 0.32 %:n eroon biomassassa (molemmat laskettuna niin, että muut prediktorit pysyvät ennallaan).

Huomaamme, että selitysaste on 87 %, mikä on varsin hyvä ennustuskykyä silmällä pitäen. Kertoimet ovat periaatteessa suoraviivaisesti tulkittavissa, esim. kun verrataan saman ryhmän sellaisia pensaita, jotka poikkeavat vain latvuksen lyhyemmän halkaisijan suhteen 1 %, niin biomassat poikkeavat keskimäärin 1.14 %. Lisäksi näemme, että samankokoisten pensaiden keskimääräinen ero ryhmien välillä on merkitsevä. Ongelmaksi jää se, että muut kertoimet eivät ole merkitseviä. Kokeillaan aivan yksinkertaista mallia, jossa latvuksen tilavuudella ennustetaan latvuksen biomassaa. Tilavuus saadaan likimäärin kertolaskulla $\texttt{diam1} \cdot \texttt{diam2} \cdot \texttt{canht} = \texttt{can.vol}$. Pidetään tietysti ryhmät erotteleva muuttuja myös mukana.

```
lm(formula = log(leafwt) ~ log(can.vol) + group)
             coef.est coef.se
(Intercept)  4.75     0.11
log(can.vol) 0.82     0.05
groupMCD     0.54     0.12
---
n = 45, k = 3
residual sd = 0.34, R-Squared = 0.85
```

Tämän mallin selitysaste (85 %) on lähes yhtä hyvä kuin sen, missä on kaikki selittäjät mukana (87 %). Saadun relaation voi kirjoittaa myös muotoon

$$\begin{aligned} \text{leafwt} &\approx e^{4.75}(\text{diam1} \cdot \text{diam2} \cdot \text{canht})^{0.82}, \quad \text{ALS-ryhmälle,} \\ &\approx e^{4.75+0.54}(\text{diam1} \cdot \text{diam2} \cdot \text{canht})^{0.82}, \quad \text{MCD-ryhmälle,} \end{aligned}$$

Koska lehtimassa on enemmänkin latvuksen reunoilla, niin toinen vaihtoehto on estimoida kaikkien kolmen dimension eksponentit erikseen:

```
# Malli A
lm(formula = log(leafwt) ~ log(diam1) + log(diam2) + log(canht) +
             group)
            coef.est coef.se
(Intercept) 4.86     0.15
log(diam1)  0.46     0.27
log(diam2)  1.20     0.20
log(canht)  0.62     0.19
groupMCD    0.57     0.12
```

```
---
n = 45, k = 5
residual sd = 0.33, R-Squared = 0.86
```

Kertoimien summa on $2.27 \approx 2$, joka sopii yhteen sen ajatuksen kanssa, että latvuksen pinta-ala paremmin kuin kuin tilavuus liittyy lehtimassan määrään. Ennustekaava on nyt

$$\begin{aligned} \text{leafwt} &\approx e^{4.86}\text{diam1}^{0.46} \cdot \text{diam2}^{1.20} \cdot \text{canht}^{0.62}, \quad \text{ALS-ryhmälle,} \\ &\approx e^{4.86+0.57}\text{diam1}^{0.46} \cdot \text{diam2}^{1.20} \cdot \text{canht}^{0.62}, \quad \text{MCD-ryhmälle,} \end{aligned}$$

Toinen näkökulma saadaan mallin kun määritellään prediktorit hieman toisin: `can.area` = `diam1` $\cdot$ `diam2` ja `can.shape` = `diam1`/`diam2`. Silloin

```
# Malli B
lm(formula = log(leafwt) ~ log(can.area) + log(can.shape) +
             log(canht) + group)
               coef.est coef.se
(Intercept)     4.86     0.15
log(can.area)   0.83     0.08
log(can.shape) -0.37     0.22
log(canht)      0.62     0.19
groupMCD        0.57     0.12
---
n = 45, k = 5
residual sd = 0.33, R-Squared = 0.86
```

Selitysaste on tietysti sama kuin mallissa A, ja ennusteetkin ovat samat, mutta kertoimien tulkinta on ehkä helpompi:

- Latvuksen poikkileikkausalan kerroin positiivinen, joten odotusten mukaisesti suurempi latvus liittyy suurempaan biomassaan. Kun latvuksen muoto ja korkeus pidetään vakiona, niin ne pensaat, joiden latvuksen poikkileikkausalat poikkeavat 1 %:n, poikkeavat biomassaltaan keskimäärin 0.83 %. 95 %:n luottamusväli on 0.67 % – 0.99 %.

- Latvuksen korkeuden kerroin positiivinen, joten odotusten mukaisesti korkeampi latvus liittyy suurempaan biomassaan. Kun latvuksen muoto ja koko pidetään vakiona, niin pensaat, joiden latvuksen korkeudet poikkeavat 1 %:n, poikkeavat biomassaltaan keskimäärin 0.62 %. 95 %:n luottamusväli on 0.22 % – 1.01 %.

- Latvuksen muodon kerroin on negatiivinen, joten mitä enemmän poikkileikkauksen muoto poikkeaa ympyrästä, sitä pienempi biomassa. Kerroin ei ole aivan merkitsevä, sillä p-arvo on 0.10. Ei ole aivan selvää, että meidän pitäisi "uskoa" tätä; 95 %:n luottamusväli on -0.82 % – 0.07 %. Tämän muuttujan voi pitää mallissa tai jättää pois.

- Ryhmien ero on merkitsevä ja muuttuja on syytä pitää mallissa. Koska meillä ei ole tarkempaa tietoa aineistosta, emme voi sanoa, miksi MCD-ryhmässä saman kokoisten ja muotoisten pensaiden biomassa on $100 \times (e^{0.57} - 1) = 76$ % suurempi; luottamusväli on 40 % – 123 %. Mahdollisesti on olemassa jokin relevantti kvantitatiivinen selittäjä, joka puuttuu mallista ja selittäisi ryhmien välistä eroa.

- Olisi tietysti perusteltua tutkia ryhmien ja muiden selittäjien interaktioita, mutta aineiston pienuuden takia se ei tuota tässä tulosta.

Voimme ennustamisessa käyttää joko mallia B tai ao. mallia C.

```
# Malli C
 lm(formula = log(leafwt) ~ log(can.area) + log(canht) + group)
                coef.est coef.se
 (Intercept)    4.71     0.12
 log(can.area)  0.89     0.07
 log(canht)     0.57     0.20
 groupMCD       0.52     0.12
 ---
 n = 45, k = 4
 residual sd = 0.34, R-Squared = 0.85
```

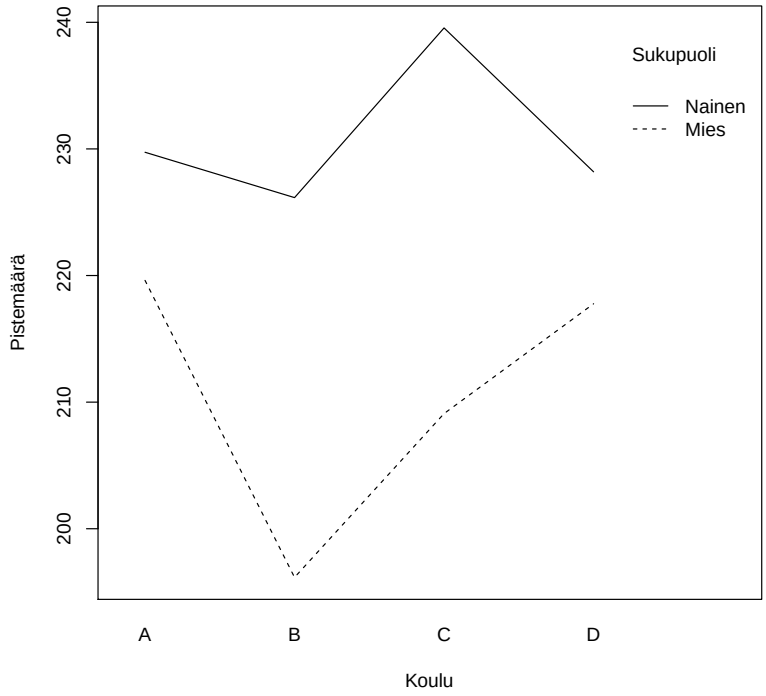
2.6 Ryhmien vertailu

Tässä palaamme ruotsin kielen ylioppilaskirjoituksiin. Aikaisemmin vertailimme kouluja vain niiden keskiarvojen kautta ja saimme tuloksen, etteivät koulujen keskiarvot poikkea toisistaan merkitsevästi. Toisaalta tiedämme, että sukupuolten välillä on yleensä eroja ehkä erityisesti juuri ruotsin kielessä. Tilanne saattaa näyttää erilaiselta, kun vertailemme kouluja vain samaa sukupuolta olevien kesken.

Additiivinen malli

Syötemuuttuja `koulu` on sama kuin edellä. Prediktori `mies` saa arvon 1, kun kokelas on mies ja arvon 0, kun kokelas on nainen. Sovitetaan ensiksi malli, jossa ei ole interaktiota, ja valitaan perustasoksi koulu B.

```
 lm(formula = ruotsi.pist ~ mies + koulu)
             coef.est coef.se
 (Intercept) 222.85     3.61
 mies        -21.03     3.62
 kouluA       10.48     5.09
 kouluC       12.58     4.83
 kouluD        9.61     4.82
```



Kuva 2.7: *Keskiarvot koulun ja sukupuolen mukaan.*

```
---
n = 375, k = 5
residual sd = 34.03, R-Squared = 0.10
```

Kertoimien tulkinta on varsin yksinkertainen. Vakio antaa ennusteen naisten keskiarvolle (≈ 223) koulussa B. Kun verrataan *samassa koulussa* olevien miesten ja naisten eroa, niin huomaamme, että miehet saavat n. 21 pistettä vähemmän verrattuna naisiin. Kun verrataan sama sukupuolta olevia kokelaita kouluissa A ja B, niin koulussa A he saavat n. 10.5 pistettä enemmän. Samoin tehdään koulujen C ja D vertailut B:hen nähden. Koulujen A ja C vertailu saadaan erotuksesta $10.48 - 12.58 = -2.10$, so. samaa sukupuolta olevien kokelaiden keskimääräinen ero on 2.1 pistettä koulun C hyväksi. Muut vertailut tehdään samaan tapaan.

Luokittelevien muuttujien interaktio

Kun malliin lisätään interaktio tulkinnat monimutkaistuvat.

```
lm(formula = ruotsi.pist ~ mies + koulu + mies:koulu)
            coef.est coef.se
(Intercept) 226.15     4.19
mies        -30.00     6.91
```

```
kouluA        3.58     6.26
kouluC       13.40     6.23
kouluD        2.04     6.11
mies:kouluA  19.88    10.64
mies:kouluC  -0.46     9.80
mies:kouluD  19.58     9.84
---
n = 375, k = 8
residual sd = 33.82, R-Squared = 0.12
```

Kertoimien tulkinta (vrt. kuvaan 2.7):

1. Vakio ilmoittaa naisten keskiarvon koulussa B.
2. Prediktorin `mies` kerroin kertoo, että koulussa B miehet saavat keskimäärin 30 pistettä vähemmän kuin naiset.
3. Faktoriin `koulu` liittyvät kertoimet ilmoittavat ko. koulun naisten keskiarvon eron koulun B naisten keskiarvoon verrattuna. Esim. koulussa C ero on 13.4 pistettä.
4. Interaktiot ovat jokseenkin monimutkaisia. Esim. interaktio `mies:kouluA` tarkoittaa, että miesten ja naisten ero koulussa A verrattuna miesten ja naisten eroon koulussa B on $-30+19.88 = -10.12$ pistettä, so. 19.88 suurempi. Vastaavasti tulkitaan muutkin interaktiot. Koska interaktio `mies:kouluC` on melkein 0, niin kouluissa B ja C miesten ja naisten ero on suunnilleen sama ts. miehet saavat keskimäärin 30 vähemmän. Samoin koska interaktiot `mies:kouluA` ja `mies:kouluD` ovat lähes yhtä suuria ja n. 20 pistettä, niin miesten naisten ero näissä kouluissa on suunnilleen yhtä suuri ja miehet saavat n. 10 pistettä vähemmän ($-30 + 20 = -10$).

Kuten edellä olemme huomanneet, tulkinnat ovat vertailuja perustasoon nähden. Jos haluamme vertailuja muiden tasojen välillä, se vaatii lisää laskutoimituksia. esim. koulujen C ja A naisten ero on $13.40 - 3.58 = 9.82$ pistettä. Mallin rakentamisessa edetään yleensä niin, että kaikki interaktiot pidetään mallissa tai ne kaikki jätetään pois.

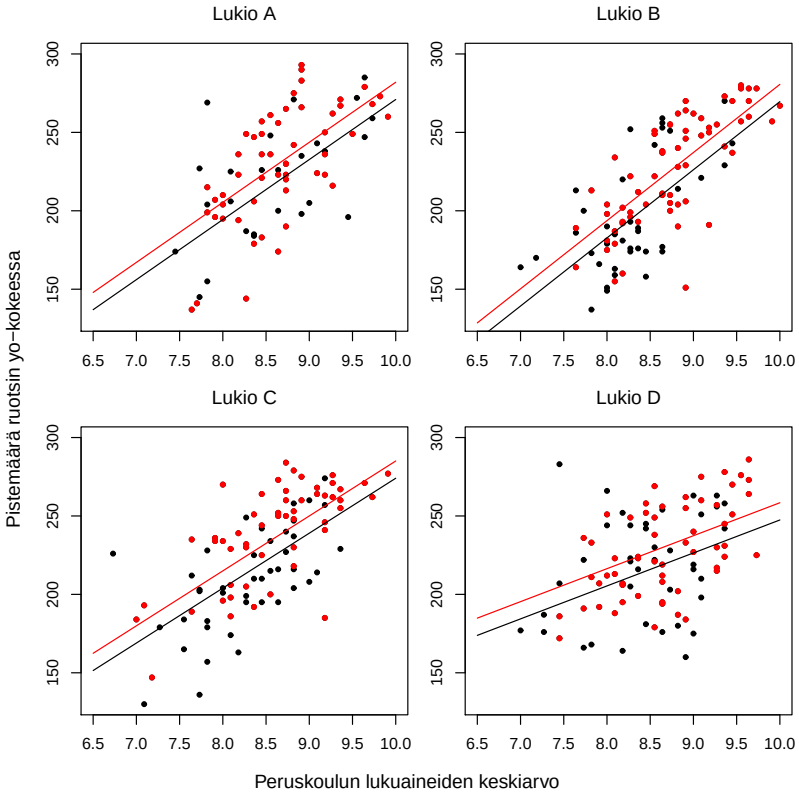
Miten saadaan interaktioiden merkitsevyys selville? Menettelemme samaan tapaan kuin faktorin merkitsevyyttä testattaessa.

```
b <- coef(lm.int)
S <- vcov(lm.int)
F <- b[6:8]%*%solve(S[6:8,6:8])%*%b[6:8]/3
F
         [,1]
[1,] 2.544175

qf(.95, df1=3, df2=lm.int$df.residual)
```

```
[1] 2.629231

1 - pf(F,df1=3, df2=lm.int$df.residual)
          [,1]
[1,] 0.05592405
```

Havaitsemme, että likimain 5 %:n tasolla interaktiot ovat merkitseviä, joten voinemme uskoa niiden tulkintoihin.

Tiedämme, että oppilasaineksen erot vaikuttavat huomattavasti koulujen välisiin eroihin ylioppilaskirjoituksissa. Useimmiten haluamme tehdä vertailun ottamalla näitä tekijöitä huomioon mahdollisuuksien mukaan. Käytössämme on tiedot kokelaiden menestyksestä peruskoulussa, joka luultavasti ennustaa myös menestystä yokirjoituksissa. Edelleen tiedämme, että sukupuolten välillä on eroja aineesta riippuen, joten pidetään sekin mallissa mukana. Lisäksi on yleinen käsitys, että oppilaiden sosioekonominen tausta ja vanhempien koulutustausta tuo eroja oppilaiden välille. Ikävä kyllä näitä tietoja ei aineistossamme ole, mikä aiheuttaa sen, että vertailumme ei koske puhtaasti koulusta aiheutuvia eroja. Tyydymme prediktoreihin `koulu`, `mies` ja `lka`, jota keskistettynä merkitään `clka`:lla. Valitaan perustasoksi koulu B.



Kuva 2.8: *Regressiosuorat kouluittain; miehet mustalla ja naiset punaisella.*

Kun sovitetaan malli, jossa on mukana kaikki parittaiset interaktiot `mies*koulu`,

`mies*clka` ja `koulu*clka`, paljastuu, että ainoastaan interaktio `koulu*clka` on merkitsevä. Alla on malli, jossa on interaktioista on mukana vain merkitsevä `koulu*clka`.

```
lm(formula = ruotsi.pist ~ mies + koulu * clka)
            coef.est coef.se
(Intercept) 218.12     2.86
mies        -11.00     2.94
kouluA        8.77     4.03
kouluC       16.54     3.83
kouluD       10.08     3.79
clka         43.45     4.50
kouluA:clka  -5.16     6.78
kouluC:clka  -8.42     6.15
kouluD:clka -22.43     6.25
---
n = 375, k = 9
residual sd = 26.79, R-Squared = 0.45
```

Tehtävä: Miten miehet ja naiset poikkeavat toisistaan? Kirjoita näkyviin kolukohtaiset mallit ja tulkitse `clka`:n kertoimet, vrt. 2.8. Pohdi, miten kouluja voisi vertailla.

Luku 3

Logistinen regressio

Logistinen regressio on nykyään tavallisin tapa mallittaa dikotomisia vasteita. Ne ovat vasteita, jotka saavat vain kaksi eri arvoa 0 tai 1. Tutustumme aluksi terminologiaan. Oletetaan nyt, että y on dikotominen ja että $P(y = 1) = \pi$, joten $P(y = 0) = 1 - \pi$. *Tässä olemme merkinneet kirjaimella π todennäköisyyttä emmekä matemaattista vakiota. Pidämme tämän sopimuksen voimassa myös jatkossa, ellei toisin mainita.*

Tarkastelemme nyt peliä, jossa pelaaja A voittaa todennäköisyydellä π ja pelaaja B todennäköisyydellä $1 - \pi$. Osamäärää

$$\frac{\pi}{1-\pi}$$

sanotaan vedonlyöntisuhteeksi (tai lyhyesti vetosuhteeksi, engl. odds). Nimi tulee reilun pelin käsitteestä, mikä tarkoittaa sitä, että molempien pelaajien voiton odotusarvo on 0. Silloin A:n ja B:n panosten suhde pitää olla sama kuin vedonlyöntisuhde. Siis B laittaa reilussa pelissä panokseksi x euroa ja A $x\pi/(1-\pi)$ euroa. Jos $\pi = 1/2$, molemmat laittavat x euroa. Jos taas $\pi = 2/3$, niin A laittaa $2x$ euroa ja B x euroa.

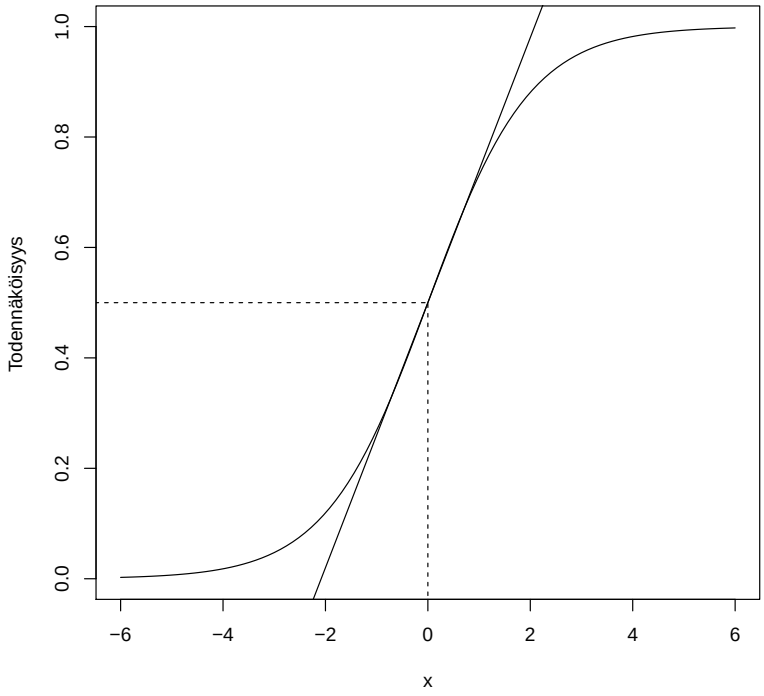
Logistisessa regressiossa tulee tärkeäksi kahteen todennäköisyyteen, π_1 ja π_2, liittyvä vedonlyöntisuhteiden osamäärä

$$\frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)} = \frac{\pi_2(1-\pi_1)}{\pi_1(1-\pi_2)}. \tag{3.1}$$

Suomeksi sitä useimmiten sanotaan ristitulosuhteeksi (engl. odds ratio). Tässä monisteessa käytetään usein ristitulosuhteesta englannista johtuvaa lyhennettä OR. OR:n tulkinnan kannalta seuraava yksinkertainen huomio saattaa auttaa. Kun $\pi_1 = 1/3$ ja $\pi_2 = 1/2$, niin OR on 2. Samoin OR on 2, kun $\pi_1 = 1/2$ ja $\pi_2 = 2/3$. Siis todennäköisyyksien muutokset $1/3 \to 1/2$ ja $1/2 \to 2/3$ vastaavat OR:ssä 2 yksikön muutoksia.

Otamme käyttöön myös merkinnät

$$\begin{aligned} \text{logit}(\pi) &= \log\frac{\pi}{1-\pi}, \quad 0 < \pi < 1, \\ \text{logit}^{-1}(x) &= \frac{e^x}{1+e^x}, \qquad -\infty < x < \infty. \end{aligned}$$



Kuva 3.1: *Logistinen funktio* logit^{-1} *ja sen lineaarinen approksimaatio origon ympäristössä.*

Huomaamme, että logit-funktion arvot kattavat koko reaaliakselin. Kuvassa 3.1 on logit^{-1}-funktio ja sen lineaarinen approksimaatio. Käytämme jatkossa tämän funktiolla paria hyödyllistä ominaisuutta:

$$\text{logit}^{-1}(x) \approx \frac{1}{4}x, \text{ kun } |x| \text{ on pieni,} \tag{3.2}$$

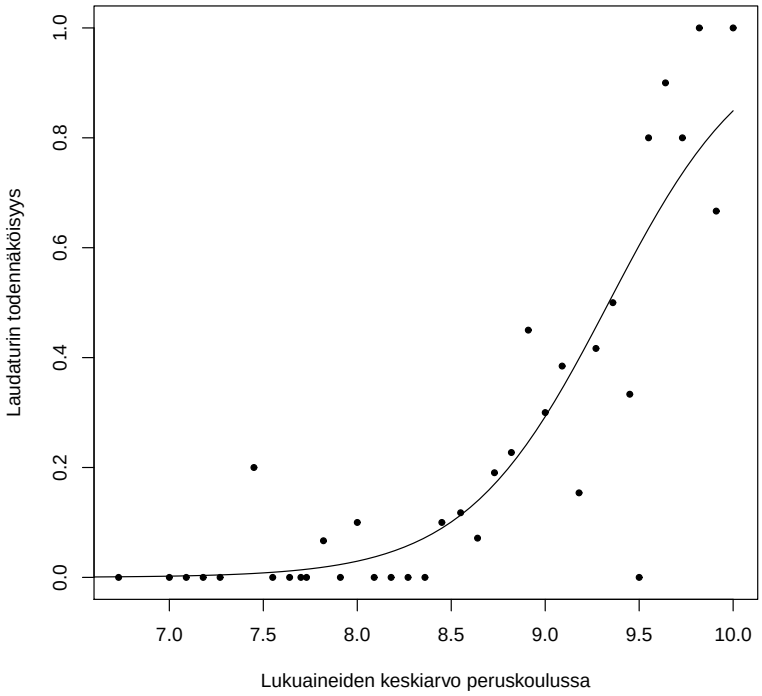
$$\text{logit}^{-1}(x_2) - \text{logit}^{-1}(x_1) \leq \frac{x_2 - x_1}{4}, \quad x_1 \leq x_2. \tag{3.3}$$

Kuvan 3.1 approksimoiva suora on piirretty kohtaan, missä funktio kasvaa jyrkimmin, ts. sen derivaatta on maksimissaan. On helppo nähdä, että se tapahtuu, kun $x = 0$. Derivaatan arvo tässä pisteessä on $1/4$.

3.1 Yhden prediktorin logistinen regressio

Oletetaan nyt, että dikotomisen vasteen arvoon y_i liittyy prediktorin arvo x_i. Merkitään

$$\begin{aligned} P(y_i = 1 \,|\, x_i) &= \pi_i, \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 x_i. \end{aligned}$$



Kuva 3.2: *Pisteet ovat laudaturin suhteellisia frekvenssejä eri lukuaineiden keskiarvon suhteen. Käyrä on logistisen regression sovitus.*

Tarkastellaan kahta arvoa $i = 1, 2$. Silloin

$$\text{logit}(\pi_2) - \text{logit}(\pi_1) = \beta_1(x_2 - x_1)$$

Kun sovellamme eksponenttifunktiota puolittain, saamme

$$\left(\frac{\pi_2}{1-\pi_2}\right) \Big/ \left(\frac{\pi_1}{1-\pi_1}\right) = e^{\beta_1(x_2-x_1)}.$$

Kun $x_2 = x_1 + 1$, niin e^{β_1} on prediktorin arvoihin x_1 ja $x_1 + 1$ liittyvistä todennäköisyyksistä laskettu OR. Siis OR liittyy luonnollisella tavalla logistiseen regressioon ja antaa tulkinnan kertoimelle β_1. Tulkinta ei kuitenkaan ole kovin havainnollinen.

Jos molemmat todennäköisyydet π_1 ja π_2 ovat pieniä, pätee likiarvo

$$\left(\frac{\pi_2}{1-\pi_2}\right) \Big/ \left(\frac{\pi_1}{1-\pi_1}\right) \approx \frac{\pi_2}{\pi_1}.$$

Oletetaan, että todennäköisyydet liittyvät jonkin harvinaisen taudin puhkeamiseen: $\pi_1 = 0.01$ on taudin puhkeamisen todennäköisyys, kun henkilöä ei ole rokotettu ja $\pi_2 = 0.001$ todennäköisyys kun henkilöä on rokotettu. Koska $0.01/0.001 = 10$, niin

rokottamattoman riski sairastua tautiin on 10-kertainen rokotettuun nähden. Vastaava OR on 10.1, joka on lähes sama kuin riskisuhde. Osamäärää π_2/π_1 sanotaan tällaisessa tilanteessa riskisuhteeksi. Kun todennäköisyydet ovat pieniä, myös OR voidaan tulkita riskisuhteeksi. Tässä tulkinnassa on syytä kuitenkin olla varovainen. Jos jonkin tapahtuman todennäköisyys on normaalisti 0.3 ja erityistilanteessa 0.6, niin erityistilanteen riski on kaksinkertainen, mutta OR = 3.5. Tässä tilanteessa OR liioittelee riskiä huomattavasti. Yleensäkin pätee, että OR $> \pi_2/\pi_1$, kun $\pi_2 > \pi_1$ ja OR $< \pi_2/\pi_1$, kun $\pi_2 < \pi_1$.

Kaavan (3.3) perusteella saamme arvioita todennäköisyyksien erotuksille eri prediktorin arvoilla. Kuvion perusteella arvio on hyödyllinen erityisesti silloin, kun todennäköisyydet ovat lähellä puolikasta, mikä tapahtuu, kun $x = -\beta_0/\beta_1$. Silloin

$$\text{logit}^{-1}(\beta_0 + \beta_1 x_2) - \text{logit}^{-1}(\beta_0 + \beta_1 x_1) \approx \frac{\beta_1(x_2 - x_1)}{4},$$

mistä saamme arvion, että yksikkömuutos prediktorissa vastaa likimäärin muutosta $\beta_1/4$ todennäköisyydessä. Arvion etumerkki on oikein ja itseisarvo $|\beta_1|/4$ antaa ylärajan muutoksen itseisarvolle.

Aloitamme yhden prediktorin mallilla ja etenemme sitten usean prediktorin malleihin ja interaktiomalleihin.

Esimerkki logistisesta regressiosta

Tarkastellaan laudaturin saamisen todennäköisyyttä ruotsin kielen ylioppilaskokeessa. Otetaan vasteeksi `laudatur`, joka saa arvon 1, kun kokelas saa laudaturin ja arvon 0 muulloin. Prediktoriksi valitaan `clka`, joka on lukuaineiden keskiarvo keskistettynä. Merkitään laudaturin saamisen todennäköisyyttä π_L:llä. Malli on

$$\begin{aligned} \text{logit}(\pi_L) &= \beta_0 + \beta_1 \texttt{clka}, \\ \pi_L &= \text{logit}^{-1}(\beta_0 + \beta_1 \texttt{clka}) \\ &= \frac{e^{\beta_0 + \beta_1 \texttt{clka}}}{1 + e^{\beta_0 + \beta_1 \texttt{clka}}}. \end{aligned} \tag{3.4}$$

Mallin sovitus saadaan seuraavasti

```
glm.1 <- glm(laudatur ~ clka, family=binomial(link="logit"))
display(glm.1)

glm(formula = laudatur ~ clka, family = binomial(link = "logit"))
            coef.est coef.se
(Intercept) -2.03     0.20
clka         2.61     0.33
---
  n = 375, k = 2
  residual deviance = 278.0, null deviance = 372.5 (difference = 94.5)
```

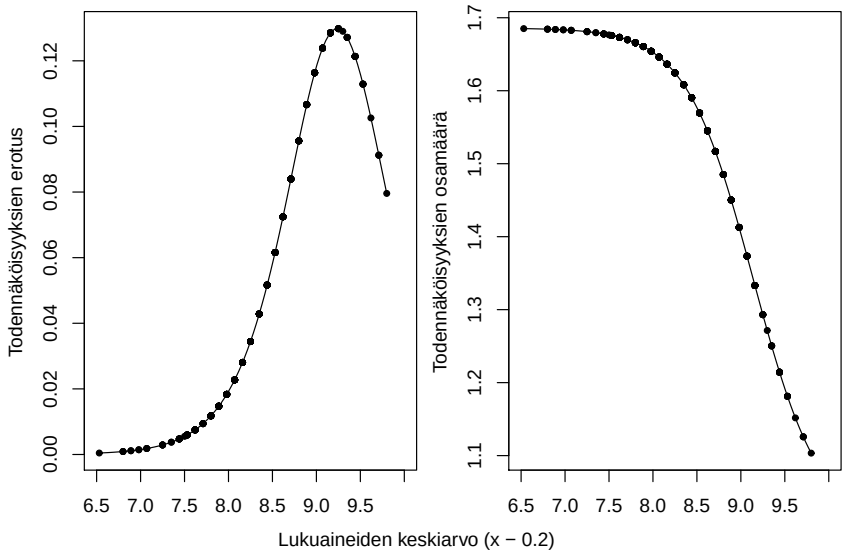
Vakio tulkitaan asettamalla `clka` $= 0$. Silloin vakion avulla saadaan ennuste laudaturin todennäköisyydelle niiden osalta, joiden lukuaineiden keskiarvo on sama kuin aineiston keskiarvo 8.6. Silloin (3.4) antaa laudaturin saamisen todennäköisyydeksi

$$\text{logit}^{-1}(-2.03) = \frac{e^{-2.03}}{1 + e^{-2.03}} = 0.12.$$

Vastaava vedonlyöntisuhde on $e^{-2.03} = 0.13$. Siis laudaturin todennäköisyys on vain 13 % sen saamatta jäämisen todennäköisyydestä.

Regressiokertoimen tulkinta on monimutkaisempi. Koska yhden numeron ero keskiarvossa on kuitenkin aika suuri ero, tarkastellaan kahden kymmenyksen eroa, Silloin logit-skaalalla heidän ennustettu eronsa on $0.2 \cdot 2.61 = 0.522$. Kaavan (3.3) perusteella kahden kymmenyksen ero keskiarvossa vastaa korkeintaan eroa $0.522/4 \approx 0.13$ todennäköisyyksissä. Jos keskiarvot ovat 9.2 ja 9.00, niin tarkka ero todennäköisyyksissä on

```
invlogit(b[1] + b[2]*(9.2 - mean(lka))) -
+ invlogit(b[1] + b[2]*(9 - mean(lka)))

 0.1182378
```



Kuva 3.3: *Laudaturin saamisen todennäköisyyksien vertailu, kun lukuaineiden keskiarvot poikkeavat kahdella kymmenyksellä (x ja $x-0.2$). Vasemmalla todennäköisyyksien erotus ja oikealla todennäköisyyksien osamäärä. Mustat pisteet vastaavat mallin antamia todennäköisyyksien erotuksia ja osamääriä, kun kaikkien kokelaiden lukuaineiden keskiarvoista on vähennetty 0.2.*

Kuvan 3.2 perusteella laudaturin todennäköisyys on kovin pieni (< 0.03), kun keskiarvo on alle 8. Kun keskiarvo ylittää 8.5:n todennäköisyys alkaa kasvaa nopeasti.

Kuvassa 3.3 on todennäköisyyksien vertailua erotuksena ja suhteellisesti. Vertailtavina ovat keskiarvot x ja $x - 0.2$. Kun keskiarvo (x) on välillä 9 – 10, niin erotus on välillä 0.08 – 0.13. Maksimi saavutetaan kohdassa $x - 0.2 = 9.26$. Suhteellisesta vertailusta näemme, että vertailtaessa niitä, joiden keskiarvo on alle 8 niihin joilla se on 2 kymmenystä pienempi on lähes 70 % suurempi todennäköisyys saada laudatur. Tämä suhteellinen ero kuitenkin vähenee melko tasaisesti kohti 10 %, kun lähestytään keskiarvoa 10. Mutta kun verrataan ryhmiä, joiden keskiarvot ovat 9.7 ja 9.5, niin edellisessä ryhmässä laudaturin todennäköisyys on n. 20 % suurempi. Tässä tapauksessa on helppo myös taulukoida todennäköisyyksiä:

Lukuaineiden keskiarvo	8.0	8.5	9.0	9.5	10.0
Laudaturin todennäköisyys	0.03	0.10	0.29	0.60	0.85

Jos haluamme pitäytyä vedonlyöntisuhteissa ja OR:ssä, niin verrattaessa ryhmiä, joiden keskiarvot poikkeavat kahdella kymmenyksellä, saamme OR:ksi $e^{0.2 \cdot 2.61} = 1.69$. Tämä tarkoittaa, että ryhmässä, jossa keskiarvo on kahta kymmenystä suurempi, niin vedonlyöntisuhde on lähes 1.7 kertainen (70 % suurempi). Saman voi myös sanoa niin, että laudaturin saamisen todennäköisyys suhteessa sen saamatta jäämiseen on 1.7 kertainen (70 % suurempi) siinä ryhmässä, jossa keskiarvo on kahta kymmenystä suurempi. Tässä tulkinnassa vain keskiarvojen ero on ratkaisevaa, ei se missä kohtaa ero on. Huomaa kuitenkin, ettei laudaturin saamisen *todennäköisyys* ole 70% suurempi.

3.2 Tilastollinen päättely

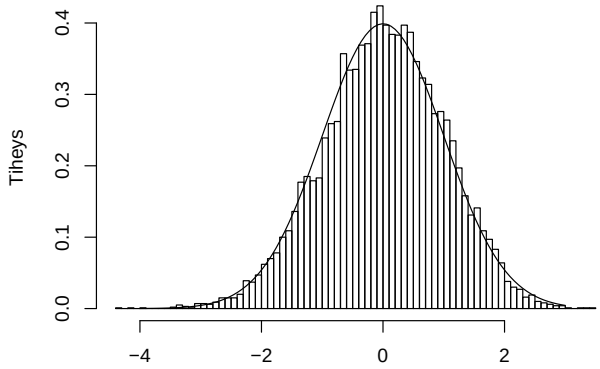
Kertoimet logistisessa regressiossa estimoidaan suurimman uskottavuuden menetelmällä (engl. method of maximum likelihood). Menetelmä yleensä toimii kun tapauksia on riittävästi eikä prediktoreita ole kovin paljon. Joissakin tapauksissa voi kuitenkin tulla ongelmia: Jos esim. esimerkissämme kaikki tapaukset, joissa lukuaineiden keskiarvon ylittää tietyn kynnyksen, saavat laudaturin ja kynnyksen alle jääneet saavat alemman arvosanan, niin kertoimen estimaatiksi tulee $+\infty$. Päinvastainen tilanne johtaisi estimaattiin $-\infty$. Yleisemmin pätee, että jos jokin prediktoreiden lineaarinen kombinaatio jakaa tietyn kynnyksen mukaan tapaukset kahteen ryhmään, joista toisessa vaste on aina 1 ja toisessa 0, niin jotkin estimoidut kertoimet ovat $\pm\infty$.

Kertoimien luottamusvälit voidaan laskea estimaattien ja keskivirheiden avulla samaan tapaan kuin lineaarisessa regressiossa. Esimerkissämme $\hat{\beta}_1 = 2.61$ ja sen keskivirhe 0.33, joten luottamusväliksi tulee $2.61 \pm 2 \cdot 0.33 = [1.95, 3.28]$. Tämä on kuitenkin likimääräinen menettely. Logistisessa regressiossa pätee nimittäin eräin yleisin säännöllisyysehdoin

$$z = \frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)} \sim N(0,1) \quad \text{likimäärin.} \tag{3.5}$$

Sen sijaan lineaarisessa regressiossa luottamusväli on täsmälleen oikein, kun normaalijakauma korvataan t-jakaumalla, mikäli kaikki olettamukset ovat voimassa.

Likimääräisen menettelyn (3.5) voi korvata bootstrap-menetelmällä (ns. bootstrap-t, Efron and Tibshirani, 1993, Sec. 12.5):



Kuva 3.4: *Suureen* (3.5) *bootstrap-tiheys verrattuna N(0,1)-tiheyteen.*

1. Sovita logistinen regressio ja laske estimoidut todennäköisyydet $\hat{p}_i = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x)$, $i = 1, \ldots, n$.

2. Generoi estimoidusta mallista uudet "vasteet" $y_i^* \sim \text{Ber}(\hat{\pi}_i)$, $i = 1, \ldots, n$.

3. Estimoi uudet kertoimet β_0^* ja β_1^* käyttämällä vasteita y_i^* ja laske "z-arvot"

$$z_j^* = \frac{\beta_j^* - \hat{\beta}_j}{\text{se}(\beta_j^*)}, \quad j = 0, 1.$$

4. Toista N kertaa kohdat 2 ja 3 ja järjestä $z_{(j,1)}^* < \cdots < z_{(j,N)}^*$, $j = 0, 1$.

5. Valitse α ja sen perusteella $k_\alpha = N\alpha$, $k_{1-\alpha} = N(1-\alpha)$. Luottamusvälit kertoimella $1 - 2\alpha$ ovat

$$[\hat{\beta}_j - z_{(j,k_{1-\alpha})}^* \text{se}(\hat{\beta}_j), \hat{\beta}_j - z_{(j,k_\alpha)}^* \text{se}(\hat{\beta}_j).$$

Luottamusväli saadaan samalla logiikalla kuin tavanomainen luottamusväli

$$z_\alpha < \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} < z_{1-\alpha} \Leftrightarrow \hat{\beta}_j - z_{1-\alpha}\text{se}(\hat{\beta}_j) < \beta_j < \hat{\beta}_j - z_\alpha \text{se}(\hat{\beta}_j).$$

Tavanomaisessa tilanteessa arvot $z_\alpha, z_{1-\alpha}$ saadaan normaalijakaumasta, bootstrap-tilanteessa ne otetaan bootstrap-jakaumasta. Lisäksi on huomattava, että normaalijakauman tilanteessa $z_\alpha = -z_{1-\alpha}$ symmetriasyistä, mutta logistisessa regressiossa jakauma voi olla jonkin verran vino.

```
probs <- fitted(glm.1)

sim.logit <- function(probs, x){
y <- rbinom(length(probs), probs, size=1)
out <- glm(y ~ x, family=binomial)
cbind(coef(out), se.coef(out))
}

b.sim <- replicate(10000, sim.logit(probs, clka))

t.boot <- quantile((b.sim[2,1,] - b[2])/b.sim[2,2,],c(.025,.975))
b[2] - t.boot[2:1]*se.coef(glm.1)[2]

   97.5%     2.5%
 1.986340 3.264887
```

Ehkä hiukan yllättäen luottamusväli on hiukan lyhyempi kuin standardi menetelmällä saatu. Jos jälkimmäisessä käytettäisiin arvoa 1.96 2:n sijasta, välit olisivat lähempänä toisiaan. Kuva 3.4 kertoo, että bootstrap-jakauma on aavistuksen vino vasemmalle mutta varsin lähellä normaalijakaumaa. Standarditeorian mukaiset johtopäätökset lienevät varsin luotettavia tässä tapauksessa.

3.3 Useita prediktoreita

Jos mallissa on useita prediktoreita

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

niin $\text{logit}^{-1}(\beta_0)$ antaa todennäköisyyden tapahtumalle $y = 1$, kun kaikki prediktorit ovat nollia. Regressiokertoimien eksponenttimuunnokset e^{β_j} ovat OR:iä, kun verrataan j. prediktorin arvoja x_j ja $x_j + 1$ muiden pysyessä muuttumattomina. Tässä täytyy olettaa, että muut prediktorit voidaan todella pitää muuttumattomina. Katsotaan esimerkkiä.

```
koulu <- relevel(koulu, ref="B")
glm.2 <- glm(laudatur ~ mies + koulu + clka, family=binomial)
display(glm.2)

glm(formula = laudatur ~ mies + koulu + clka, family = binomial)
            coef.est coef.se
(Intercept) -2.34     0.37
mies        -0.79     0.36
kouluA       0.69     0.45
kouluC       1.15     0.45
kouluD       0.20     0.46
```

```
clka         2.67     0.35
---
 n = 375, k = 6
residual deviance = 265.1, null deviance = 372.5 (difference = 107.4)
```

Kertoimien tulkinnat:

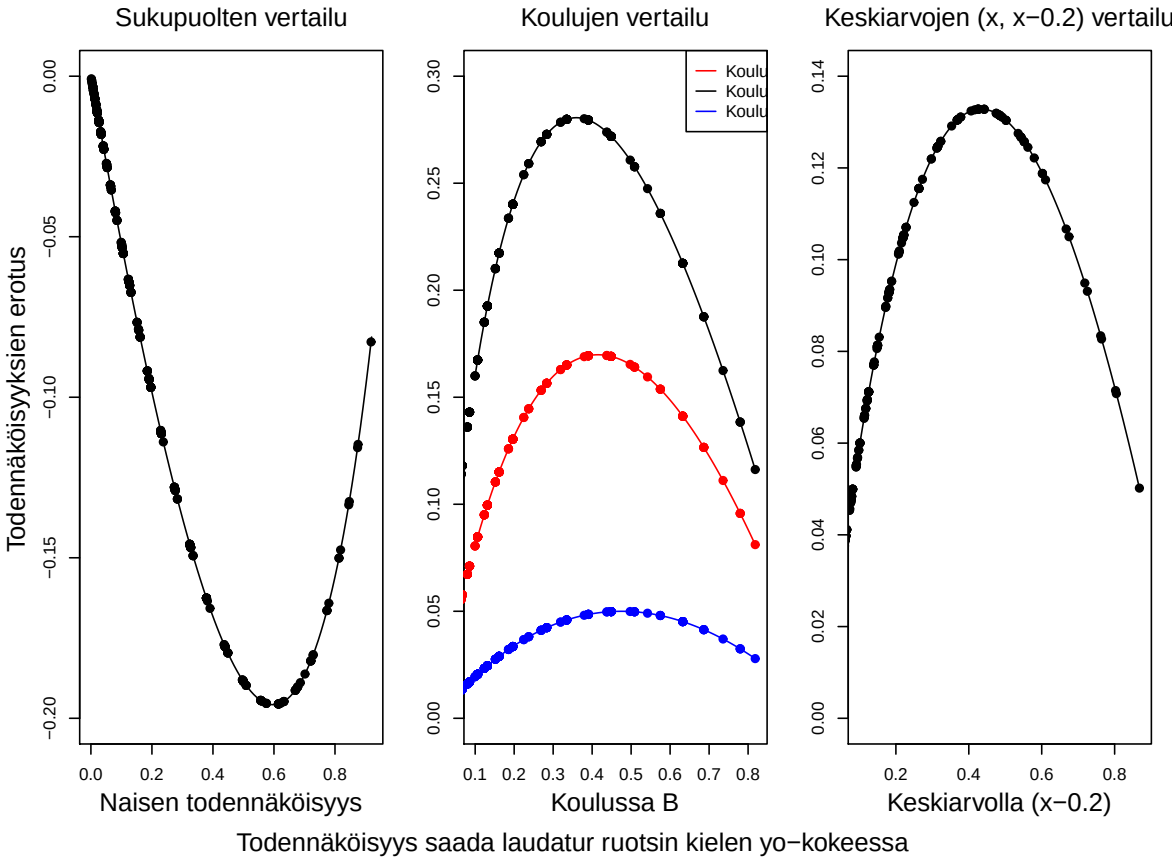
- Vakiosta saadaan todennäköisyys sille, että koulun B nainen, jonka lukuaineiden keskiarvo on sama kuin aineiston keskiarvo 8.6, saa laudaturin ruotsin kielessä, on $\text{logit}^{-1}(-2.34) = 0.088$. Vastaava vedonlyönsuhde on $e^{-2.34} = 0.096$.

- Prediktorin `mies` kerroin on negatiivinen, joten miesten todennäköisyys saada laudatur on pienempi kuin naisten, kun verrataan saman koulun ja keskiarvoltaan samanlaisia kokelaita. Kertoimen muunnos $e^{-0.79} = 0.45$ kertoo, että miesten vedonlyöntisuhde jaettuna naisten vedonlyöntisuhteella on 0.45, kun kokelaat ovat samasta koulusta ja heillä on sama lukuaineiden keskiarvo. Likimääräinen todennäköisyyksien vertailu saadaan jakamalla logit-mallin kerroin neljällä $-0.79/4 = -0.20$, mikä tarkoittaa, että miesten todennäköisyys on enintään 0.20:n verran pienempi. Huomaamme myös, että sukupuolten ero on tilastollisesti merkitsevä.

- Faktoriin `koulu` liittyvät kertoimet ovat kaikki positiivisia ja kertoimien suuruusjärjestys pienimmästä suurimpaan on D, A, C. Tulokset merkitsevät sitä, että koulun B kokelailla on pienin todennäköisyys saada laudatur, sitten tulevat koulut D, A, C tässä järjestyksessä. Tässä verrataan samaa sukupuolta ja keskiarvoltaan samanlaisia kokelaita.

  Prediktorin `kouluA` kertoimen eksponenttimuunnos on $e^{0.69} = 1.99$. Kun verrataan koulujen A ja B samaa sukupuolta olevia kokelaita, joilla vielä on sama lukuaineiden keskiarvo, saadaan tulos, että koulun A kokelaan vedonlyöntisuhde on n. kaksinkertainen koulun B oppilaaseen nähden. Koulujen C ja D OR:t koulun B suhteen ovat $e^{1.15} = 3.17$ ja $e^{0.20} = 1.22$ vastaavasti. Todennäköisyyksien erot kouluun B nähden ovat enintään 0.17 (koulu A), 0.29 (koulu C) ja 0.05 (koulu D). Näissä vertailuissa ainoastaan koulu C eroaa koulusta B tilastollisesti merkitsevästi. Muut vertailut koulujen välillä saadaan näiden lukujen erotuksina. Niiden merkitsevyyden laskemiseksi täytyy laskea myös mainittujen erotusten keskivirheet. Koska koulujen B ja C ero on suurin, nämä muut vertailut tuskin tuottavat merkitseviä eroja.

- Prediktorin `clka` kerroin on positiivinen, joten mitä parempi keskiarvo sitä suurempi todennäköisyys saada laudatur. Kertoimen eksponenttimuunnos $e^{2.67} = 14.5$ liittyy yhden numeron eroon keskiarvossa. Kahden kymmenyksen eroon liittyvä OR on $e^{0.2 \cdot 2.67} = 1.71$. Kun verrataan samasta koulusta samaa sukupuolta olevia kokelaita, kahden kymmenyksen ero keskiarvossa vastaa 71 %:n eroa vedonlyöntisuhteessa. Kahden kymmenyksen ero keskiarvossa vastaa todennäköisyyksien eroa, joka on korkeintaan $0.2 \cdot 2.67/4 = 0.13$.



Kuva 3.5: *Vasemmassa kuvassa on miesten ja naisten todennäköisyyksien erotukset. Keskellä koulujen erotukset koulun B kokelaiden suhteen. Oikealla ovat todennäköisyyksien erotukset, kun keskiarvot ovat x ja $x-0.2$. Pisteet vastaavat mallin antamia todennäköisyyksiä seuraavasti: vasemmalla kaikki kokelaat on oletettu naisiksi, keskellä kaikki kokelaat on oletettu tulevan koulusta B ja oikealla kaikkien lukuaineiden keskiarvosta on vähennetty 0.2.*

Miesten ja naisten vertailun voi tehdä myös seuraavaan tapaan. Oletetaan, että naisen todennäköisyys saada laudatur on π_F ja samasta koulusta ja keskiarvoltaan samanlaisen miehen vastaava todennäköisyys on π_M. Silloin mallin mukaan pätee yhtä pitävästi

$$\begin{aligned} \text{logit}(\pi_M) - \text{logit}(\pi_F) &= -0.79, \\ \text{logit}^{-1}(-0.79 + \text{logit}(\pi_F)) &= \pi_M. \end{aligned}$$

Sijoittamalla sopivia arvoja naisen todennäköisyydelle π_F saamme vastaavat miehen todennäköisyydet

Nainen	0.20	0.40	0.60	0.80
Mies	0.10	0.23	0.40	0.64

Samalla periaatteella saamme taulukon koulujen vertailemiseksi, kun kokelailla on sama sukupuoli ja sama lukuaineiden keskiarvo.

kouluB	0.20	0.40	0.60	0.80
kouluD	0.23	0.45	0.65	0.83
kouluA	0.33	0.57	0.75	0.89
kouluC	0.44	0.68	0.83	0.93

Ao. taulukossa verrataan todennäköisyyksiä, kun lukuaineiden keskiarvossa on kahden kymmenyksen ero, ja kokelaat ovat samaa sukupuolta ja samasta koulusta.

Keskiarvo	x	0.2	0.40	0.60	0.80
	$x + 0.2$	0.3	0.53	0.72	0.87

Kuvissa 3.5 on graafinen yhteenveto todennäköisyyksien vertailuista. Sukupuolten vertailussa huomataan, että sukupuolten ero on suurimmillaan, kun naisten todennäköisyys on n. 0.6, jolloin ero on n. 0.2 naisten hyväksi. Koulujen vertailussa koulun C ero kouluun B on suurimmillaan n. 0.27. Vastaavat suurimmat erot ovat koulun A osalta n. 0.16 ja koulun D osalta n. 0.05. Ero on suurimmillaan hiukan vaihtelevasti silloin kun koulun B kokelaiden todennäköisyys 0.4:n vaiheilla. Muiden kouluparien todennäköisyyksien erotukset ovat käyrien vertikaalisia etäisyyksiä. Keskiarvoerojen vertailussa ero on suurimmillaan n. 0.13.

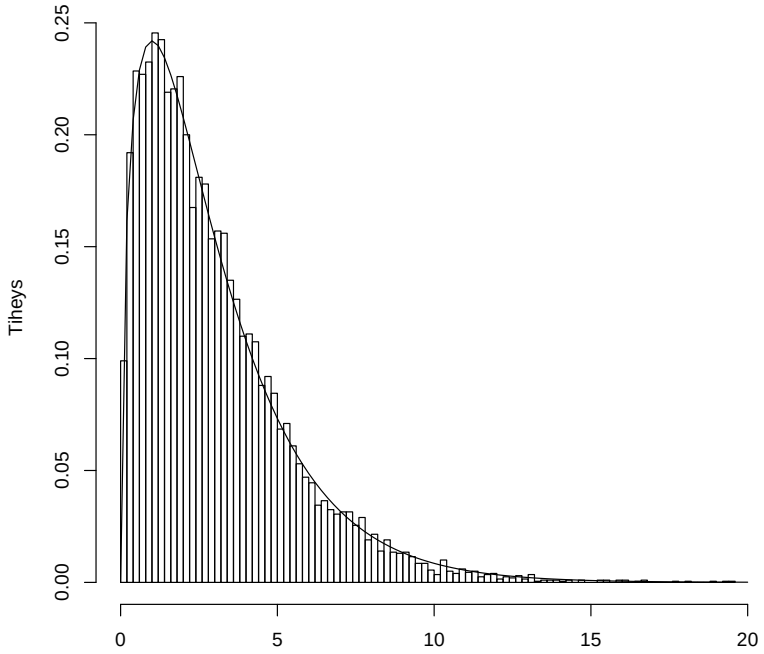
Luokittelevan syötemuuttujan merkitsevyys

Lineaarisen regression yhteydessä opimme, miten faktorin merkitsevyys saadaan kun tasoja on 3 tai enemmän. Miten sama tehdään logistisessa regressiossa? Oletetaan nyt, että $\boldsymbol{b}$ ja $\boldsymbol{S}$ on määritelty kuten lineaarisessa regressiossa. Silloin sopiva testisuure on neliömuoto

$$Q = \boldsymbol{b}'\boldsymbol{S}^{-1}\boldsymbol{b}. \tag{3.6}$$

Jos testattavia kertoimia on k kpl (b:n komponenttien lukumäärä), niin Q:n arvoa verrataan $\chi^2(k)$ jakaumaan. Katsotaan esimerkkiä.

```
 b <- coef(glm.2)
 S <- vcov(glm.2)
 Q <- b[3:5]%*%solve(S[3:5,3:5])%*%b[3:5]
 Q
         [,1]
[1,] 8.187022

 qchisq(.95, df=3)
[1] 7.814728

 1-pchisq(Q, df=3)
           [,1]
[1,] 0.04230059
```



Kuva 3.6: *Testisuureen* (3.6) *simuloitu nollahypoteesijakauma verrattuna* $\chi^2(3)$*-jakaumaan.*

Saamme tuloksen, että koulut poikkeavat merkitsevästi toisistaan. Huomaa, että testisuuretta ei tässä jaeta k:lla ja vertailujakauma on χ^2-jakauma. Lineaarisen regression tulos pätee annetuilla olettamuksilla täsmällisesti, mutta logistisessa regressiossa menettely on likimääräinen.

Edellä esitety testimenettely voidaan korvata simulointimenettelyllä saman tapaan kuin yhden kertoimen tapauksessa. Sen sijaan, että generoidaan z-testisuureita, generoidaan Q:n arvoja nollahypoteesin mukaisesta mallista ja sovitetaan sitten täysi malli.

```
## H0 mallin sovitus
glm.0 <- glm(laudatur ~ mies + clka, family=binomial)

coef.test <- function(M){
y <- rbinom(length(M$y), prob=fitted(M), size=1)
out <- glm(y ~ mies + koulu + clka, family=binomial)
d <- coef(out)
cov.sim <- cov.matrix(out)
d[3:5]%*%solve(cov.sim[3:5,3:5])%*%d[3:5]
}
```

```
q.sim <- replicate(10000, coef.test(glm.0))

quantile(q.sim, 0.95)
7.625914

mean(q.sim > c(Q))
 0.0367
```

Johtopäätös on sama kuin aikaisemmin (p-arvo on pienempi). Kuvassa 3.6 on testisuureen simuloitu tiheys.

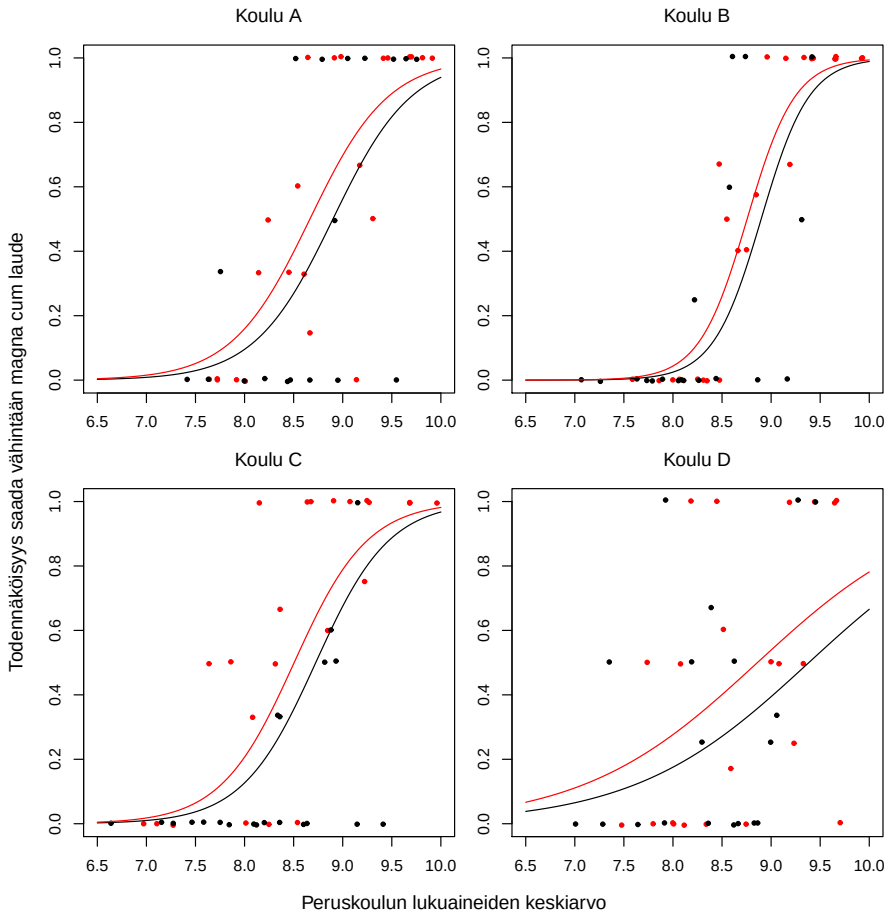
3.4 Interaktio logistisessa regressiossa

Käsittelemme edelleen aineistoa ruotsin kielen ylioppilaskirjoituksista. Tällä kertaa otamme vasteeksi muuttujan `ml`, joka saa arvokseen 1, kun kokelas on saa arvosanakseen joko magna cum laude approbaturin tai laudaturin. Eräiden kokeilujen jälkeen malliksi valikoitui sellainen, jossa syötemuuttujina ovat kuten ennenkin `mies`, `koulu` ja `clka`. Prediktoreina ovat `koulu`-faktorin lisäksi interaktiot `koulu*clka` l. kulmakertoimet vaihtelevat kouluittain.

```
ml <- as.numeric(arvosana > 4)

glm(formula = ml ~ mies + koulu * clka, family = binomial)
            coef.est coef.se
(Intercept) -0.79     0.33
mies        -0.59     0.27
kouluA       0.53     0.42
kouluC       0.94     0.41
kouluD       0.46     0.39
clka         4.08     0.83
kouluA:clka -1.59     1.03
kouluC:clka -1.42     1.02
kouluD:clka -2.97     0.92
---
  n = 375, k = 9
  residual deviance = 367.0, null deviance = 508.5 (difference = 141.5)
```

Kertoimien tulkinnat:

- Vakiosta saadaan muunnoksella todennäköisyys, että koulun B nainen, jonka lukuaineiden keskiarvo on sama kuin aineiston keskiarvo saa yo-kokeesta arvosanakseen vähintään magnan. Tämä todennäköisyys on $\text{logit}^{-1}(-0.79) = 0.31$.



Kuva 3.7: *Logistiset käyrät kouluittain ja sukupuolittain. Musta vastaa miehiä ja punainen naisia.*

- Prediktorin `mies` kertoimen tulkinta OR:nä on sama kuin aikaisemminkin $e^{-0.59} = 0.56$. Miesten vedonlyöntisuhde on 56 % naisten vedonlyöntisuhteesta, kun verrataan saman koulun ja keskiarvoltaan samanlaisia kokelaita.

  Todennäköisyystulkinta on ehkä havainnollisempi. Kerroin on negatiivinen, joten miehen todennäköisyys saada vähintään magna on pienempi kuin saman koulun ja keskiarvoltaan samanlaisen naisen todennäköisyys. Ero on naisten hyväksi korkeintaan $0.59/4 = 0.15$. Jos naisen todennäköisyys on π_F, niin koulun ja keskiarvon suhteen samanlaisen miehen todenäköisyys on $\text{logit}^{-1}(-0.59 + \text{logit}(\pi_F))$. Tässä pieni taulukko

Nainen	0.10	0.30	0.50	0.70	0.90
Mies	0.06	0.19	0.36	0.57	0.83

- Koska mallissa on koulun ja keskiarvon interaktiot, kouluun liittyvien prediktoreiden kertoimet ovat tulkittavissa vain siinä tapauksessa kun `clka` on 0. Sil-

loin vertaillaan vain niitä kokelaita, joilla on lukuaineiden keskiarvo sama kuin aineiston keskiarvo 8.6. OR:t ovat kouluittain 1.70 (A), 2.57 (C) ja 1.58 (D).

Voimme myös laskea magnan saamisen todennäköisyydet sukupuolen ja koulun mukaan:

	kouluA	kouluB	kouluC	kouluD
nainen	0.44	0.31	0.54	0.42
mies	0.30	0.20	0.39	0.29

Huomaamme, että ero naisten ja miesten välillä vaihtelee 0.11:sta (koulu B) 0.15:een (koulu C) naisten hyväksi. *Interaktiotermien takia nämä taulukko olettaa, että kokelaiden keskiarvot ovat samat kuin aineiston keskiarvo 8.6.*

- Prediktorin `clka` kertoimen eksponenttimuunnos on $e^{0.2 \cdot 4.08} = 2.26$. Tämä tulkitaan niin, että kahden kymmenyksen ero keskiarvossa kasvattaa OR:n 2.26-kertaiseksi koulun B kokelailla. Likimääräinen todennäköisyystulkinta on, että kahden kymmenyksen ero keskiarvossa liittyy todennäköisyyksien eroon, joka on enintään $0.2 \cdot 4.08/4 = 0.20$, kun verrataan koulun B samaa sukupuolta olevia kokelaita.

- Interaktioiden kertoimet ovat ikävä kyllä vaikeita tulkita. Verrataan koulun A kokelasta koulun B kokelaaseen, jolla on sama sukupuoli. Logit-skaalalla ero on

$$\text{logit}(\pi_A) - \text{logit}(\pi_B) = 0.53 - 1.59\texttt{clka}.$$

Kun tehdään eksponenttimuunnos puolittain, saamme OR:n, joka riippuu lukuaineiden keskiarvosta:

$$\text{OR}_{\text{A,B}}(\texttt{clka}) = e^{0.53 - 1.59\,\texttt{clka}}.$$

Seuraa, että

$$\text{OR}_{\text{A,B}}(\texttt{clka} + 1) = e^{-1.59}\text{OR}_{\text{A,B}}(\texttt{clka}).$$

Siis koulujen välinen OR riippuu lukuaineiden keskiarvosta ja kahden kymmenyksen kasvu keskiarvossa vähentää OR:ää kertoimella $e^{-1.59 \cdot 0.3} = 0.73$; prosentteina vähennys on 27%.

Voimme laskea keskiarvoon liittyvät todennäköisyystulkinnat kouluittain. Koulun B kulmakerroin on suurin ja loput järjestyksessä C, A, D. Kahden kymmenyksen ero lukuaineiden keskiarvossa liittyy todennäköisyyksien likimääräisiin eroihin samaa sukupuolta olevilla kokelailla seuraavasti:

$$\begin{aligned} 0.2 \cdot (4.08 - 1.59)/4 &= 0.12, \quad \text{koulussa A} \\ 0.2 \cdot (4.08 - 1.42)/4 &= 0.13, \quad \text{koulussa C} \\ 0.2 \cdot (4.08 - 2.97)/4 &= 0.06, \quad \text{koulussa D}. \end{aligned}$$

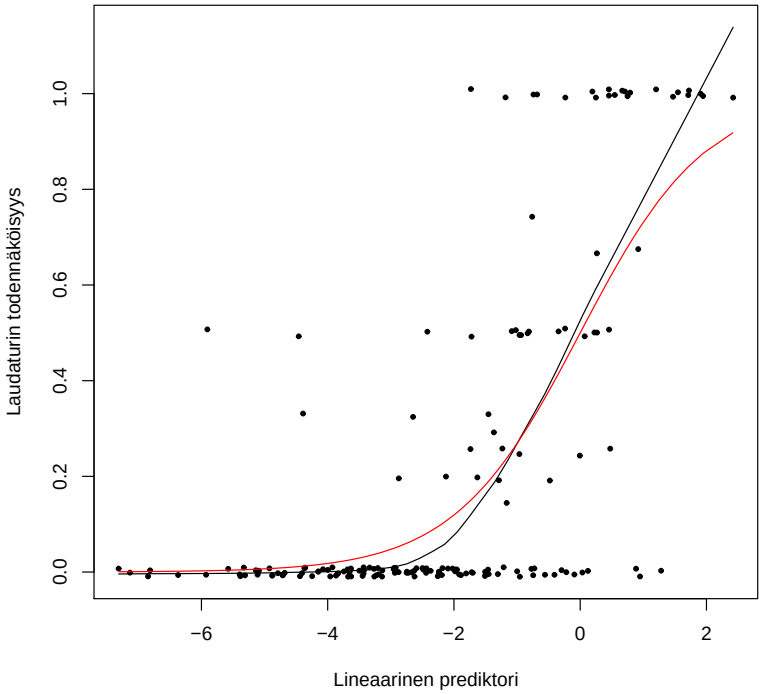
Muistetaan, että koulussa B vastaava luku on 0.20. Kuva 3.7 näyttää selvästi interaktion vaikutuksen. Koulun käyrä D poikkeaa selvästi muista. Koulujen A ja C käyrät ovat melkein samanmuotoiset. Koulun B käyrä kasvaa jyrkimmin.

Koulujen vertailuja voi tehdä myös laskemalla mikä keskiarvo riittää siihen, että saa magnan todennäköisyydellä 0.5:

	kouluA	kouluB	kouluC	kouluD
nainen	8.67	8.76	8.51	8.86
mies	8.90	8.90	8.73	9.38

3.5 Diagnostiikka

Sovitettu malli graafisesti



Kuva 3.8: *Pisteet ovat laudaturin saaneiden suhteellisia frekvenssejä lukuaineiden keskiarvon, sukupuolen ja koulun mukaan luokitellussa aineistossa. Punainen käyrä on additiivisesta logistisesta mallista, ja musta käyrä on havainnoista tasoitettu (loess) käyrä.*

Silloin kun mallissa on vain yksi kvantitatiivinen prediktori, kannattaa piirtää estimoidut todennäköisyydet tämän prediktorin suhteen, ks. kuva 3.2. Koska ruotsin kielen yo-kirjoitusaineistossa on kokelaita, joilla on samoja peruskoulun lukuaineiden keskiarvoja, voimme laskea laudatureiden suhteellisen osuuden eri keskiarvoilla (pisteet em. kuvassa). Koska ryhmien koot ovat pieniä (monissa on vain yksi tapaus)

pisteet voivat sijaita kaukanakin estimoidusta käyrästä. Mallin sopivuus aineistoon näyttää vähintäänkin kohtalaisen hyvältä.

Kun mallissa on useita prediktoreita, voi vastaavan kuvan piirtää valitsemalla vaaka-akselille mallin lineaarisen prediktorin arvot $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$. Kuva 3.8 liittyy luvussa 3.3 sovitettuun additiiviseen malliin. Koska tässä mallissa luokittelu on tehty paitsi lukuaineiden keskiarvon myös sukupuolen ja koulun mukaan, ryhmien koot ovat entistäkin pienempiä. Kuvaan on piirretty myös epäparametrisen regression tuottama sovitus. Tämä jälkimäinen sovitus on eräänlaisella liukuvan keskiarvon menetelmällä tehty laskelma, joka ei perustu mihinkään malliin. Huomaamme, että käyrät ovat aika lähellä toisiaan, mikä puhuu logistisen mallin puolesta. Pisteitä on jonkin verran täristetty, jotta ne erottuisivat paremmin toisistaan.

Luokitteluvirhe

Eräs yksinkertainen tapa arvioida logistisen mallin hyvyyttä on luokitella tapaukset kahteen luokkaan seuraavasti:

$$\begin{aligned} \text{Arvataan } y_i &= 1, \text{ kun } \text{logit}^{-1}(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}) > 0.5, \\ y_i &= 0, \text{ kun } \text{logit}^{-1}(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}) < 0.5. \end{aligned}$$

Luokitteluvirhe määritellään niin, että se on niiden tapausten suhteellinen osuus, jotka tulevat väärin luokitelluiksi. Virheluokittelufunktio `R`:ssä on

```
error.rate <- function(y, mod)
  mean((fitted(mod) > 0.5 & y == 0) | (fitted(mod) < 0.5 & y == 1) )
```

missä `y` on dikotominen vastevektori ja `mod` on estimoitu malli. Luokitteluvirheen pitää olla aina alle 0.5, koska muuten malli, jossa on pelkkä vakio, mutta ei lainkaan prediktoreita tuottaa paremman tuloksen. Tämä ns. "nollamalli" antaa luokitteluvirheeksi $\min(p, 1-p)$, missä p on ykkösten suhteellinen osuus. Luvun 3.3 additiivisen mallin luokitteluvirhe on 14 %. Laudatureiden osuus aineistossa on 20 %, joka on siis myös nollamallin luokitteluvirhe. malli parantaa luokitteluvirhettä 6 %-yksikköä. Luokitteluvirhe ei tietenkään ole täydellinen mallin hyvyyden kriteeri. Sehän ei esim. erottele todennäköisyyksiä 0.6 ja 0.9 luokitellessaan tapauksen ykköseksi. Mutta se on helposti laskettavissa ja tulkittavissa.

Devianssi

Devianssi määritellään uskottavuusfunktion avulla. Se ilmaisee poikkeaman ns. saturoidusta mallista. Palaamme kurssin toisessa osassa devianssiin. Tässä kerrotaan lyhyesti muutama asia:

- Deviansilla verrataan malleja keskenään. Pienempi devianssi tarkoittaa parempaa yhteensopivuutta.

- Jos malliin lisätään prediktori, jonka arvot ovat satunnaista kohinaa, odotamme devianssin silti vähenevän keskimäärin 1:llä.
- Jos lisäämme malliin informatiivisen prediktorin, odotamme devianssin vähenevän yli ykkösen. kun lisäämme k informatiivista prediktoria, odotamme devianssin vähenevän enemmän kuin k:n verran.

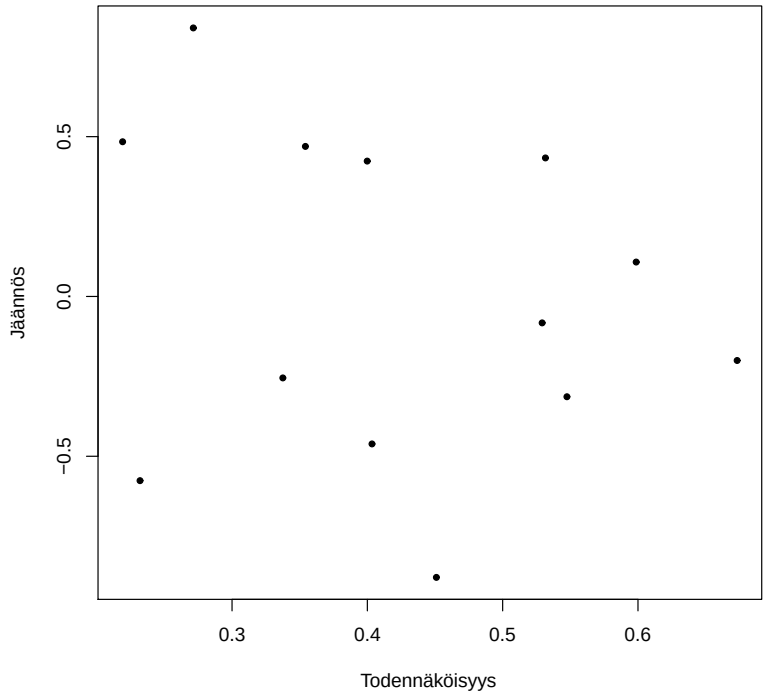
On paljon sovelluksia, joissa kaikki prediktorit ovat luokittelevia. Silloin sattuu useinkin, että aineisto voidaan esittää taulukkomuodossa. Tässä on esimerkki, jossa vasteena on ajokokeessa hylättyjen määrä `fail` sukupuolen, iän ja ajoharjoittelun suhteen. Muuttuja `total` kertoo ajokokeessa olleiden lukumäärän. Ajoharjoittelu on suoritettu joko autokoulussa tai kotiopetuksena. Ryhmiä pitäisi olla kaikkiaan $2 \cdot 4 \cdot 2 = 16$, mutta kuten huomaamme, kotiopetuksessa ei ole ollut kaikkia vanhempia ikäryhmiä. Tapaukset jakautuvat vain 13 ryhmään.

fail	total	sex	age	training
267	1187	male	18-19	school
47	156	male	20-29	school
6	23	male	30-39	school
6	10	male	40+	school
391	1083	female	18-19	school
106	272	female	20-29	school
18	56	female	30-39	school
35	53	female	40+	school
100	242	male	18-19	family
2	7	male	20-29	family
110	205	female	18-19	family
11	18	female	20-29	family
1	2	female	30-39	family

Additiivisen mallin sovitus:

```
glm(formula = fail/total ~ sex + age + training, family = binomial,
    weights = total)
             coef.est coef.se
(Intercept)     0.19     0.10
sexmale        -0.60     0.08
age20-29        0.21     0.11
age30-39       -0.07     0.25
age40+          1.32     0.27
trainingschool -0.79     0.10
---
  n = 13, k = 6
  residual deviance = 3.1, null deviance = 151.0 (difference = 147.9)
```

Huomaamme, että koodissa vaste on hylättyjen suhteellinen osuus, ja lisänä aikaisempaan on optio `weights = total`. Siis ryhmäkoot pitää antaa painoina. Tarkastellaan deviansseja tarkemmin. Ns. nolladevianssi tulee sellaisesta mallista, jossa on



Kuva 3.9: *Jäännökset todennäköisyyksien suhteen. Ajokoeaineistoon on sovitettu additiivinen malli.*

vain vakio. Jos prediktorit ovat täysin riippumattomia vasteesta, niin nolladevianssin arvoa voi verrata χ^2-jakaumaan $n-1$ vapausasteella. Jos nolladevianssin arvo ylittää kynnysarvon $\chi^2_{0.95;n-1}$, prediktorit ennustavat vasteen arvoja merkitsevästi. Tässä esimerkissä $n = 13$ ja $\chi^2_{0.95;12} = 21.0$, joten prediktorit todella ennustavat ajokokeessa hylkäämistä (ja tietysti samoin läpipääsyä). Tietyin säännöllisyysehdoin residuaalidevianssin arvoa voi verrata χ^2-jakaumaan vapausastein $n-k$; tässä n ei tarkoita ajokokeessa olleiden kokonaismäärää vaan ryhmien lukumäärää, so. $n = 13$. Jos estimoitu malli on oikea, niin residuaalidevianssin arvo ylittää kynnysarvon $\chi^2_{0.95;n-k}$ todennäköisyydellä 0.05. Tässä esimerkissä $\chi^2_{0.95;7} = 14.1$, joten residuaalidevianssi ei ole merkitsevä. Itse asiassa sen arvo 3.1 on varsin pieni.

Pearsonin jäännökset

Luokitellussa tilanteessa voimme laskea kunkin ryhmän odotetun frekvenssin ja verrata sitä havaittuun frekvenssiin. Merkitään ryhmän i frekvenssiä y_i:llä. Jos malli on oikea niin $y_i \sim \text{Bin}(n_i, \pi_i)$, missä n_i on ryhmän i koko ja $\pi_i = \text{logit}^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta})$. Tiedämme, että $\text{var}(y_i) = n_i\pi_i(1-\pi_i)$. Kun korvaamme $\boldsymbol{\beta}$:n estimaatillaan $\widehat{\boldsymbol{\beta}}$ ja

merkitsemme $\hat{p}_i = \text{logit}^{-1}(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})$, voimme laskea standardoidut jäännökset

$$\frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}, \quad i = 1, \ldots, n.$$

Standardoidut jäännökset noudattavat $N(0,1)$-jakaumaa. Kuten kuvasta 3.9 näemme, että kaikki jäännökset ovat hyvin pieniä.

Luku 4

Poisson-regressio

Poisson-jakaumaa käytetään lukumääräaineistojen mallittamiseen. Kuvassa 4.1 on tulipaloissa kuolleiden lukumäärät y_i Suomessa vuosittain 1986–2005 . Näemme, että palokuolemien määrä vaihtelee 90 kuoleman molemmin puolin vuodesta toiseen. Kuvassa on ehkä havaittavissa myös hienoinen kuolemien väheneminen. Oletamme, että

$$y_i \sim \text{Poisson}\,(\lambda_i), \quad i = 1, \ldots, n.$$

Koska parametrit $\lambda_i > 0$, rakennamme lineaarisen mallin logaritmi-skaalalle

$$\log \lambda_i = \boldsymbol{x}_i' \boldsymbol{\beta}.$$

Asetamme yksinkertaisen mallin, jossa prediktorina on vain aika $0, 1, \ldots, n-1$. Siis

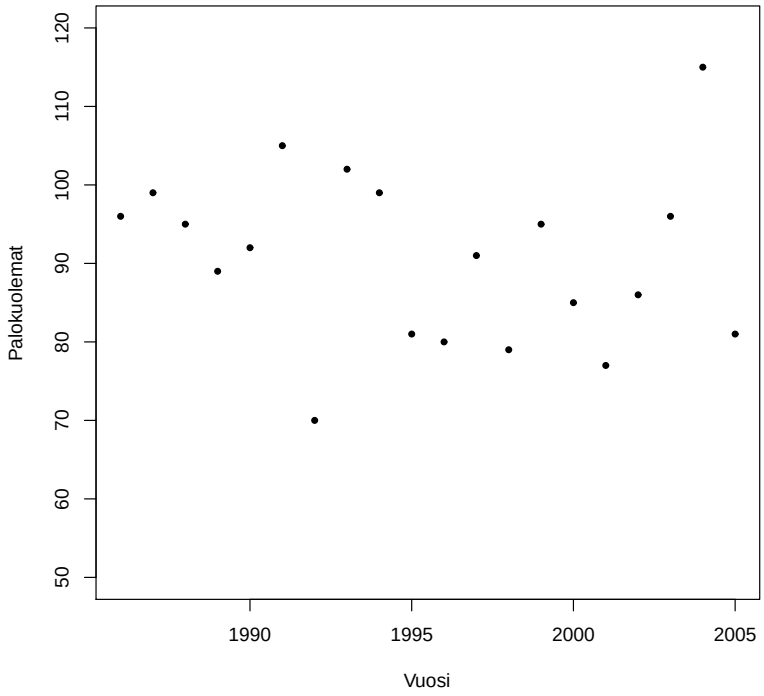
$$\log \lambda_i = \beta_0 + \beta_1 (i-1), \quad i = 1, \ldots, n.$$

Saamme parametreille yksinkertaisen tulkinnan

$$\begin{aligned} E(y_1) &= \lambda_1 = e^{\beta_0} \\ E(y_{i+1}) &= e^{\beta_0 + \beta_1 i} = e^{\beta_1} E(y_i). \end{aligned}$$

Vakion avulla saamme mallin mukaisen ennusteen kuolemien määrälle ensimmäisenä vuonna, ja regressiokerroin antaa eksponenttimuunnoksen jälkeen ennusteen suhteelliselle vuosimuutokselle. Jos β_1 on pieni, niin $100\beta_1$ antaa likimääräisen prosenttimuutoksen. Mallin sovitus R:llä saadaan seuraavasti:

```
glm(formula = kuolemat ~ aika, family = poisson("log"))
            coef.est coef.se
(Intercept)  4.534    0.045
aika        -0.003    0.004
---
  n = 20, k = 2
  residual deviance = 24.5, null deviance = 25.0 (difference = 0.5)
```



Kuva 4.1: *Tulipaloissa kuolleet Suomessa vuosina 1986–2005.*

Vakiosta saamme $e^{4.5534} \approx 93$, joka on mallin ennuste vuodelle 1986. Regressiokertoimen tulkinnan mukaan kuolemat ovat vähentyneet 0.3 % vuodessa, mutta sen keskivirhe on 0.004, joten se ei ole merkitsevä.

Altistus

Usein sovelluksissa olemme kiinnostuneita erityisesti tapahtumien intensiteetistä, esim. palokuolemista miljoonaa asukasta kohti. Tieliikenteessä voitaisiin laskea liikenneonnettomuuksissa kuolleita miljoonaa ajokilometriä kohti tms. Epidemiologiassa puhutaan altistuneiden tai altistuksen määrästä. Tällaisissa tapauksissa kirjoitamme $\lambda_i = m_i\theta_i$, missä θ_i on intensiteetti ja m_i altistuneiden tai altistuksen määrä useimmiten tiettynä aikana (esim. vuodessa) tai tietyssä paikassa (kansainvälisissä vertailuissa eri maat). Intensiteetin logaritmi saa muodon

$$\log \lambda_i = \log \theta_i + \log m_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \log m_i.$$

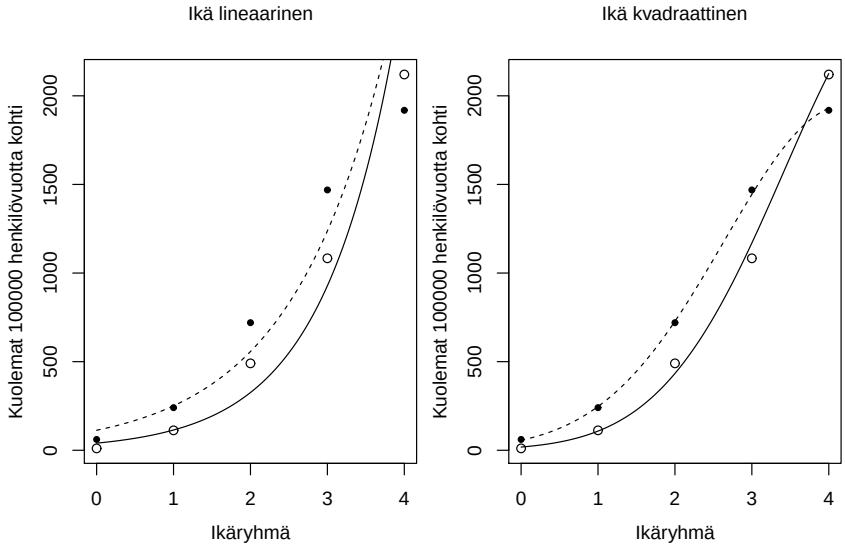
Formaalisti $\log m_i$ on prediktori, jonka kerroin asetetaan ykköseksi.

Sovitetaan palokuolemiin malli, jossa kuolemat suhteutetaan miljoonaa asukasta kohti:

```
glm(formula = kuolemat ~ aika, family = poisson("log"),
offset =  log(vakiluku/1e+06))
            coef.est coef.se
(Intercept)   2.9371    0.0450
aika         -0.0063    0.0041
---
  n = 20, k = 2
  residual deviance = 24.9, null deviance = 27.4 (difference = 2.5)
```

Optio *offset* saa aikaan sen, että regressiossa prediktorin `log(vakiluku/1e+06)` kertoimeksi asetetaan ykkönen. Vakion avulla saamme mallin mukaisen odotusarvon palokuolemien määrälle miljoonaa asukasta kohden v. 1986: $e^{2.941} = 18.9$. Prediktorin `aika` kerroin on nyt jonkin verran suurempi kuin aikaisemmin ja tarkoittaa, että väkilukuun suhteutettuna palokuolemat ovat vähentyneet keskimäärin 0.6 % vuodessa. Tämä on suurempi luku kuin aikaisemmin. Mutta edelleenkään kerroin ei ole merkitsevä.

Tupakointi ja kuolevuus



Kuva 4.2: *Kuolemat 100 000 henkilövuotta kohden. Avoimet ympyrät ja yhtenäiset käyrät vastaavat tupakoimattomia, ja mustat ympyrät ja katkoviivat tupakoijia.*

V. 1951 brittiläisille lääkäreille lähetettiin kysely, jossa tiedusteltiin, ovatko he tupakoijia vai eivät. Sen jälkeen kerättiin tiedot heidän kuolemistaan sepelvaltimotautiin seuraavien 10 vuoden aikana (Dobson, 2002, s. 154–156). Tässä esimerkissä tarkastellaan miespuolisia lääkäreitä ja heidän sepelvaltimokuolleisuuttaan. Aineisto on jaettu 5 ikäryhmään: 35–44, 45–54, 55-64, 65–74, 75–84, jotka on koodattu luvuiksi 0,1,2,3,4

(muuttuja `age`). Ikäryhmät on määritelty tutkimuksen alkaessa. Tupakointitieto on muuttujassa `smoker`. Kuolemat suhteutetaan henkilövuosiin (muuttuja `pyears`): kussakin ikäryhmässä eletyt vuodet erikseen tupakoijille ja tupakoimattomille. Sovitetaan aineistoon malli, jossa on iän ja tupakoinnin interaktio:

```
doctors <- read.table("http://users.jyu.fi/~junyblom/doctors.dat",
header=T)

doctors.glm <- glm(deaths~age*smoker, offset=log(pyears*1e-5),
family=poisson)
b <- coef(doctors.glm)
b
 (Intercept)           age     smokeryes age:smokeryes
   3.6926720     1.0468258     1.0346283    -0.2489752
```

Kuvan 4.2 vasemmalla puolella on sovitetut käyrät. Interaktio on selvästi tarpeellinen, sillä korkeimmassa ikäryhmässä tupakoivien kuolevuus on pienempi kuin tupakoimattomien. Kaiken kaikkiaan yhteensopivuus näyttää kuitenkin huonolta varsinkin korkeimmassa ikäryhmässä. Lisätään malliin iän kvadraattinen termi, jonka kerroin oletetaan samaksi sekä tupakoijilla että tupakoimattomilla. Tulkinnan kannalta neliötermi määritellään kaavan $\texttt{age2} = 0.5 * \texttt{age} * (\texttt{age} - 1)$ avulla. Yhteensopivuus näyttää nyt hyvältä (ks. kuvan 4.2 oikea puoli).

```
age2 <- 0.5*age*(age - 1)
glm(formula = deaths ~ age * smoker + age2, family = poisson,
   data = doctors, offset = log(pyears * 1e-05))
            coef.est coef.se
(Intercept)    2.90     0.29
age            1.78     0.14
smokeryes      1.13     0.28
age2          -0.40     0.05
age:smokeryes -0.31     0.10
---
 n = 10, k = 5
 residual deviance = 1.6, null deviance = 935.1 (difference =  933.4)
```

Tulosten tulkinta:

- Vakion avulla saamme ennusteen $e^{2.90} = 18.2$ kuolemaa 100 000 henkilövuotta kohden tupakoimattomille ikäryhmässä 35–44.

- Prediktorin `smokeryes` kertoimen tulkinta on se, että nuorimassa ikäryhmässä ($\texttt{age} = 0$) tupakoivien kuolevuus on $e^{1.13} = 3.11$ kertainen tupakoimattomiin verrattuna.

- Tupakoivien ja tupakoimattomien ero kaikissa ikäryhmissä saadaan ikäryhmittäin ja logaritmiasteikolla kaavasta

$$1.13 - 0.31\,\texttt{age},\ \texttt{age} = 0, 1, 2, 3, 4.$$

Siis tupakoivien kuolevuuden suhteellinen ero tupakoimattomien kuolevuuteen vähenee kertoimella $e^{-0.31} = 0.74$ ts. 26 % siirryttäessä korkeampaan ikäryhmään. Alla on tupakoivien kuolevuus verrattuna tupakoimattomiin ikäryhmittäin

```
b2 <- coef(doctors2.glm)
exp(Smo.d <- b2[3] + b2[5]*(0:4))
3.1 2.3 1.7 1.2 0.9
```

Huomaamme, että vanhimmassa ikäryhmässä kuolevuus on jo pienempi kuin tupakoimattomilla.

- Kahden peräkkäisen ikäryhmän ero tupakoimattomien osalta on logaritmiskaalalla

$$1.78 - 0.40\,\texttt{age}, \quad \texttt{age} = 1, 2, 3, 4,$$

ja sama tupakoivien osalta on

$$1.78 - 0.31 - 0.40\,\texttt{age} = 1.47 - 0.40\,\texttt{age}, \quad \texttt{age} = 1, 2, 3, 4.$$

Siis kuolevuuden vauhti hidastuu sekä tupakoivilla että tupakoimattomilla kertoimella $e^{-0.40} = 0.67$ (33 %) siirryttäessä korkeampaan ikäryhmään.

Jäännökset

Poisson- jakauman ominaisuuksiin kuuluu se, että odotusarvo ja varianssi ovat yhtä suuria. Siis jos $y_i \sim \text{Po}(m_i e^{\theta_i})$, niin $E(y_i) = m_i e^{\theta_i}$ ja $\text{var}(y_i) = m_i e^{\theta_i}$. Kun malli on sovitettu ja saatu $\log \hat{\theta}_i = \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}$, voimme laskea standardoidut jäännökset

$$r_i = \frac{y_i - m_i e^{\hat{\theta}_i}}{\sqrt{m_i e^{\hat{\theta}_i}}}.$$

Mikäli Poisson-malli on oikea, niin jäännösten pitäisi olla likimäärin (ei täsmälleen, koska samaa estimaattia $\widehat{\boldsymbol{\beta}}$ on käytetty niiden laskemiseen) riippumattomia, ja niiden odotusarvon pitäisi olla 0 ja keskihajonnan 1. Mallin sopivuutta, voidaan testata laskemalla neliösumman

$$Q = \sum_{i=1}^{n} r_i^2,$$

jota verrataan $\chi^2(n - p - 1)$-jakaumaan (vapausasteita vähennetään estimoitujen parametrien lukumäärän verran). Jäännökset kannattaa myös piirtää sovitteen ja prediktoreiden suhteen. Lääkäriaineistoon sovitetusta mallista saamme tulokset

```
fit <- fitted(doctors2.glm)
> r <- (deaths - fit)/sqrt(fit)
> n <- length(deaths)
> sum(r^2)
[1] 1.550251
```

Koska testisuureen arvo on huomattavasti pienempi kuin vapausasteet $n - p - 1 = 10 - 4 - 1 = 5$, niin testi ei ole merkitsevä.

Logistisen ja Poisson-regression erot

Sekä logistinen että Poisson-regressio ovat samanlaisia siinä suhteessa, että molemmissa vaste on lukumäärä, esim. aineistot ajokokeessa hylätyistä ja palokuolemista. Niitä kuitenkin sovelletaan erilaisissa tilanteissa:

- Jos vasteet y_i voidaan tulkita "onnistuneiden" lukumäärinä n_i kokeessa, niin silloin on tapana soveltaa logistista regressiota.
- Jos vasteella y_i ei luontevaa ylärajaa ts. sen ei voi ajatella syntyneen riippumattomien Bernoulli-kokeiden (0-1 -kokeiden) tuloksena, silloin on tapana soveltaa Poisson-regressiota.

Osa II

Teoriaa ja laajennuksia

Luku 5

Eksponenttiset jakaumaperheet

5.1 Momentti- ja kumulanttifunktio

Oletetaan, että satunaismuuttujan Y tiheysfunktio (tai pistetodennäköisyysfunktio diskreetissä tapauksessa) on f. Jatkossa käytetään tästä lyhyttä termiä tiheys. Oletetaan lisäksi, että momenttifunktio (l. momentit generoiva funktio)

$$M(t) = E(e^{tY})$$

on olemassa jollakin välillä $-h < t < h$, $h > 0$. Silloin kaikki momentit $E(Y^k)$, $k = 1, 2, \ldots$ ovat äärellisinä olemassa. Nämä momentit saadaan momenttifunktion derivaattojen[1] arvoista pisteessä $t = 0$, erityisesti

$$\begin{aligned} E(Y) &= \dot{M}(0), \\ E(Y^2) &= \ddot{M}(0). \end{aligned}$$

Sovellusten kannalta momenttifunktiota kätevämpi on kumulanttifunktio $K(t) = \log M(t)$. Koska $M(0) = 1$, niin

$$\begin{aligned} \dot{K}(0) &= \frac{\dot{M}(0)}{M(0)} = E(Y), \\ \ddot{K}(0) &= \frac{\ddot{M}(0)M(0) - \dot{M}(0)^2}{M(0)^2} = \ddot{M}(0) - \dot{M}(0)^2 = \mathrm{var}(Y). \end{aligned}$$

5.2 Eksponenttinen hajontaperhe

Oletetaan, että Y satunnaismuuttuja, jonka tiheys riippuu kahdesta parametrista θ ja τ, joista τ saattaa olla tunnettu. Tästä tiheydestä käytetään merkintää $f(y; \theta, \tau)$.

Määritelmä 5.1. Jakaumaperhe, jonka tiheys voidaan kirjoittaa muotoon

$$f(y; \theta, \tau) = c(y, \tau)e^{\tau^{-1}(y\theta - b(\theta))},$$

on **eksponenttinen hajontaperhe**, kun

[1] Koska yläpilkku on varattu transpoosin merkiksi, merkitään kahta ensimmäistä derivaattaa pisteillä: funktio f, ensimmäinen derivaatta $\dot{f}$ ja toinen derivaatta $\ddot{f}$.

- satunnaismuuttujan Y:n otosavaruus (so. kaikki mahdolliset arvot) on reaalilukujen osajoukko $\mathcal{Y}$, joka on sama parametreista θ, τ riippumatta;
- $\theta \in \Theta$, missä kanoninen parametriavaruus Θ on avoin reaalilukuväli (useimmiten ääreton);
- $\tau \in T$, T on positiivisten reaalilukujen osajoukko.

Huomautus 5.1. Hajontaparametri τ **ei useimmiten ole keskihajonta** $\sqrt{\text{var}(Y)}$.

Esimerkki 5.1 (Normaalijakauma)**.** Oletetaan, että $Y \sim N(\mu, \sigma^2)$. Silloin Y:n tiheys on

$$f(y; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right),$$

missä π on tavanomainen matemaattinen vakio. Tee neliöön korotus ja kirjoita tiheys muotoon

$$f(y; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{y^2}{2\sigma^2}\right) \exp\left[\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right)\right].$$

Normaalijakaumat muodostavat eksponenttisen hajontaperheen. Mitä tässä esimerkissä ovat $\mathcal{Y}$, θ, Θ, $b(\theta)$ ja τ?

Esimerkki 5.2 (Bernoulli- ja binomijakauma)**.** Oletetaan, että $Z \sim \text{Bin}(k, \pi)$. Silloin

$$f_Z(z; \pi, k) = \binom{k}{z} \pi^z (1-\pi)^{k-z}.$$

Tee sopivat laskut ja saata tiheys muotoon

$$f_Z(z; \pi, k) = \binom{k}{z} \exp\left[k\left(\frac{z}{k}\log\frac{\pi}{1-\pi} + \log(1-\pi)\right)\right].$$

Huomaamme, että satunnaismuuttujan $Y = Z/k$ jakaumaperhe on eksponenttinen hajontaperhe. Mitä tässä tapauksessa ovat $\mathcal{Y}$, θ, Θ, $b(\theta)$ ja τ?

Esimerkki 5.3. (Poisson-jakauma) Oletetaan, että $Z \sim \text{Po}(w\lambda)$, missä $w > 0$ on tunnettu luku. Silloin

$$f(z; \lambda) = \frac{(w\lambda)^z}{z!} e^{-w\lambda}.$$

Ota uudeksi satunnaismuuttujaksi $Y = Z/w$ ja kirjoita tämäkin sellaiseen muotoon, josta näkee, että Y:n jakaumaperhe on eksponenttinen hajontaperhe. Mitä tässä tapauksessa ovat $\mathcal{Y}$, θ, Θ, $b(\theta)$ ja τ?

Eksponenttisen hajontaperheen momentti- ja kumulanttifunktiot ovat helposti laskettavissa.

Lause 5.1. *Oletetaan, että satunnaismuuttujan Y jakauma kuuluu eksponenttiseen hajontaperheeseen. Silloin Y:n momenttifunktio ja kumulanttifunktio ovat vastaavasti*

$$\begin{aligned} M(t) &= \exp\left(\frac{b(\theta+\tau t) - b(\theta)}{\tau}\right), \\ K(t) &= \frac{b(\theta+\tau t) - b(\theta)}{\tau}. \end{aligned}$$

Todistus. Todistetaan jatkuva tapaus. Diskreetissä tapauksessa korvataan integraali summalla. Koska tiheys integroituu aina ykköseksi, niin

$$1 = \int_{\mathcal{Y}} c(y,\tau)e^{(y\theta - b(\theta))/\tau}\,\mathrm{d}y = e^{-b(\theta)/\tau}\int_{\mathcal{Y}} c(y,\tau)e^{y\theta/\tau}\,\mathrm{d}y.$$

Siis

$$e^{b(\theta)/\tau} = \int_{\mathcal{Y}} c(y,\tau)e^{y\theta/\tau}\,\mathrm{d}y,$$

mikä takaa integraalin olemassaolon kaikilla parametrien arvoilla $(\theta,\tau) \in \Theta \times T$. Yksinkertainen lasku osoittaa, että

$$\begin{aligned} M(t) &= E(e^{tY}) \\ &= e^{-b(\theta)/\tau}\int_{\mathcal{Y}} e^{ty}c(y,\tau)e^{y\theta/\tau}\,\mathrm{d}y \\ &= e^{-b(\theta)/\tau}\int_{\mathcal{Y}} c(y,\tau)e^{y(\theta+\tau t)/\tau)}\,\mathrm{d}y \\ &= e^{-b(\theta)/\tau + b(\theta+\tau t)/\tau}. \end{aligned}$$

Viimeinen yhtäsuuruus seuraa siitä, että Θ on avoin, joten on olemassa sellainen väli $-h < t < h$, että $(\theta+\tau t)/\tau \in \Theta$. □

Lause 5.2. *Eksponenttisessa hajontaperheessä pätee*

$$\begin{aligned} E(Y) &= \dot{b}(\theta), \\ \mathrm{var}(Y) &= \tau\ddot{b}(\theta). \end{aligned}$$

Todistus. Tulos seuraa siitä, että kumulanttifunktion kaksi ensimmäistä derivaattaa antavat odotusarvon ja varianssin. Tee laskut. □

Jatkossa merkitsemme myös $\dot{b}(\theta) = m(\theta)$ ja $\ddot{b}(\theta) = v(\theta)$.

Harjoitustehtävä 5.1. Laske esimerkkien 5.1–5.3 tapauksissa $\dot{b}(\theta)$ ja $\tau\ddot{b}(\theta)$. Totea myös, että ne ovat yhtä pitäviä vastaavien tutumpien odotusarvon ja varianssin kaavojen kanssa.

Koska $\tau > 0$ ja varianssi on positiivinen, niin myös $\ddot{b}(\theta) > 0$ jokaisella θ:n arvolla. Tästä seuraa, että $\dot{b}(\theta)$ on kasvava θ:n funktio, ja määrittelee bijektion θ:n arvojen ja odotusarvojen $\mu = \dot{b}(\theta) = m(\theta)$ välille. Siis $\theta = m^{-1}(\mu)$, ja varianssi $\mathrm{var}(Y) = \tau\ddot{b}(m^{-1}(\mu))$ riippuu odotusarvosta, paitsi jos $\ddot{b}(\theta)$ on vakiofunktio. Anna esimerkki tapauksesta, jossa $\ddot{b}(\theta)$ on vakiofunktio.

Esimerkki 5.4 (Gamma-jakauma)**.** Eräs tavanomainen määritelmä gamma-jakauman tiheydelle on

$$f(y;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}y^{\alpha-1}e^{-y/\beta}, \quad y > 0,\ \alpha > 0,\ \beta > 0.$$

Jakauman odotusarvo $\mu = \alpha\beta$. Otetaan tämä uudeksi parametriksi, ja jätetään muotoparametri α ennalleen. Tee tarvittavat laskut ja kirjoita tiheys muotoon

$$f(y;\mu,\alpha) = \frac{\alpha^\alpha y^{\alpha-1}}{\Gamma(\alpha)} \exp\left[\alpha\left(-\frac{y}{\mu} - \log\mu\right)\right].$$

Mitä ovat $\mathcal{Y}$, θ, Θ ja τ?

Harjoitustehtävä 5.2. Negatiivinen binomijakauma liittyy riippumattomien toistojen kokeeseen, joka lopetetaan, kun on saatu ν onnistumista. Kun merkitään onnistumisen todenäjöisyyttä π:llä, niin edeltävien epäonnistumisten lukumäärä Y noudattaa negatiivista binomijakaumaa $\mathrm{NB}(\nu,\pi)$, jonka tiheys on

$$f(y;\nu,\pi) = \binom{y+\nu-1}{y}\pi^\nu(1-\pi)^y.$$

Mitä ovat $\mathcal{Y}$, θ, Θ ja τ? Edellä ν olemme ajatelleet, että ν on positiivinen kokonaisluku, mutta jakauma on määritelty kaikilla positiivisilla ν:n arvoilla, kun kirjoitetaan

$$\binom{y+\nu-1}{y} = \frac{\Gamma(y+\nu)}{\Gamma(\nu)y!}.$$

Harjoitustehtävä 5.3. Laske ed. esimerkissä $E(Y)$ ja $\mathrm{var}(Y)$.

5.3 Yleistetty lineaarinen malli

Oletetaan, että $Y \sim \mathrm{EDF}(\theta,\tau)$ (Exponential Dispersion Family). Aikaisempien merkintöjen mukaisesti odotusarvo μ on kanonisen parametrin θ funktio $\mu = m(\theta)$. Oletamme edelleen, että on olemassa bijektiivinen linkkifunktio g, joka kuvaa odotusarvot kaikille reaaliarvoille[2] ja että $g(\mu)$ on prediktoreiden $x_1,\ldots,x_p$ lineaarinen funktio:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Esimerkki 5.5 (Normaalinen lineaarinen regressio)**.** Oletamme, että $Y \sim N(\mu,\sigma^2)$ ja että

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Tässä linkkifunktio on identtinen funktio $g(\mu) = \mu$.

Esimerkki 5.6 (Logistinen regressio)**.** Oletamme, että $Z \sim \mathrm{Bin}(n,\pi)$ ja että $Y = Z/n$. Silloin $E(Y) = \pi = \mu$, $0 < \mu < 1$. Valitaan linkkifunktioksi logit-funktio

$$g(\mu) = \mathrm{logit}(\mu) = \log\frac{\mu}{1-\mu} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

[2]Joissakin sovelluksissa käytetään myös linkkifunktiota, jonka arvojoukko ei ole koko reaaliväli. Tämä asettaa kuitenkin rajoituksia lineaarisen funktion kertoimille. Tällä kurssilla ei käsitellä näitä tapauksia.

Esimerkki 5.7 (Poisson-regressio)**.** Oletetaan $Z \sim \text{Po}(w\lambda)$, ja $Y = Z/w$. Silloin $E(Y) = \lambda = \mu > 0$. Valitaan linkkifunktioksi logaritmi-funktio

$$g(\mu) = \log \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Kaikissa kolmessa esimerkissä linkkifunktio on ns. kanoninen linkki:

$$g(\mu) = m^{-1}(\mu) = \theta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Esimerkki 5.8 (Probit-malli)**.** Tämä on esimerkki linkkifunktiosta, joka ei ole kanoninen. Teemme samat olettamukset kuin logistisessa regressiossa, mutta valitsemme linkkifunktioksi normaalijakauman kvantiilifunktion Φ^{-1}, missä Φ on $N(0,1)$-jakauman kertymäfunktio. Silloin

$$\Phi^{-1}(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Keksi oma linkkifunktiosi binomijakauman tilanteessa.

Harjoitustehtävä 5.4 (Gamma-regressio)**.** Mikä olisi sopiva linkkifunktio esimerkissä 5.4?

Harjoitustehtävä 5.5 (Negatiivinen binomiregressio)**.** Mikä olisi sopiva linkkifunktio esimerkissä 5.2?

5.4 Kanoninen linkki: uskottavuusfunktio

Oletetaan sellainen eksponenttinen jakaumaperhe $\text{EDF}(\theta, \phi\tau)$, missä kanoninen parametriavaruus $\Theta = \mathbb{R}$. Silloin on mahdollista määritellä kanoninen linkki

$$\theta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Lisäksi oletamme, että τ on tunnettu, ja ϕ on joko tuntematon skaalaparametri tai $\phi = 1$. Kuten olemme nähneet, normaali-, binomi- ja Poisson-jakaumat ovat tärkeitä erikoistatapauksia tästä tilanteesta.

Oletetaan nyt että $Y_i \sim \text{EDF}(\theta_i, \phi\tau_i)$ ovat riippumattomia ja että kanoninen linkki on käytössä. Silloin

$$\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta}, \quad i = 1, \ldots, n,$$

missä $\boldsymbol{x}_i' = (1, x_{i1}, \ldots, x_{ip})$, ja $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$. Uskottavuusfunktio on tiheyksien tulo havaintopisteissä $Y_i = y_i$, $i = 1, \ldots, n$, joten tässä tapauksessa

$$L(\boldsymbol{\beta}, \phi) = \prod_{i=1}^{n} c(y_i, \phi\tau_i) \exp\left(\frac{y_i\, \boldsymbol{x}_i'\boldsymbol{\beta} - b(\boldsymbol{x}_i'\boldsymbol{\beta})}{\phi\tau_i} \right). \tag{5.1}$$

Olemme aluksi kiinnostuneita vain $\boldsymbol{\beta}$:n estimoinnista. Oletamme myös, että kertoimet τ_i ovat tunnettuja. Uskottavuusyhtälöt saadaan laskemalla uskottavuusfunktion

logaritmin osoittaisderivaatat (gradientti) $\boldsymbol{\beta}$:n koordinaattien suhteen ja asettamalla ne nolliksi. Muistetaan merkintä $\dot{b} = m$. Uskottavuusyhtälöt ovat

$$\begin{aligned} \frac{\partial \log L(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{1}{\phi \tau_i}(y_i - m(\boldsymbol{x}_i'\boldsymbol{\beta}))\boldsymbol{x}_i = \mathbf{0} \\ &\Leftrightarrow \sum_{i=1}^{n} \frac{1}{\tau_i}(y_i - m(\boldsymbol{x}_i'\boldsymbol{\beta}))\boldsymbol{x}_i = \mathbf{0} \end{aligned} \tag{5.2}$$

Huomaamme tosiaan, että $\boldsymbol{\beta}$:n voi estimoida skaalasta ϕ riippumatta. Olkoon $\boldsymbol{m}(\boldsymbol{\beta})$ vektori $(n \times 1)$, joka sisältää alkiot $m(\boldsymbol{x}_i'\boldsymbol{\beta})$. Lisäksi oletetaan, että matriisin $\boldsymbol{X}$ i. rivi on $\boldsymbol{x}_i'$, $\boldsymbol{y}' = (y_1, \ldots, y_n)$, ja $\boldsymbol{T}$ on diagonaalimatriisi diag$[\tau_1, \ldots, \tau_n]$. Jatkossa oletamme aina, ellei toisin mainita, että $\boldsymbol{X}$ on täysiasteinen. Silloin voimme kirjoittaa yhtälön eqrefeq61 muotoon

$$\boldsymbol{X}'\boldsymbol{T}^{-1}\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{T}^{-1}\boldsymbol{m}(\boldsymbol{\beta}).$$

Esimerkki 5.9 (Lineaarinen regressio)**.** Koska tässä tapauksessa linkki on identtinen funktio, $m(\boldsymbol{x}_i'\boldsymbol{\beta}) = \boldsymbol{x}_i'\boldsymbol{\beta}$, ja siis $\boldsymbol{m}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$. Uskottavuusyhtälöksi tulee

$$\boldsymbol{X}'\boldsymbol{T}^{-1}\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{T}^{-1}\boldsymbol{X}\boldsymbol{\beta}, \tag{5.3}$$

jonka ratkaisu on $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{T}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{T}^{-1}\boldsymbol{y}$. Tärkeä erikoistapaus $\boldsymbol{T} = \boldsymbol{I}$ antaa ratkaisuksi $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Olemme olettaneet, että $\boldsymbol{X}$ on täysiasteinen. Matriisin $\boldsymbol{X}$ täysiasteisuus takaa, että käänteismatriisit $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ ja $(\boldsymbol{X}'\boldsymbol{T}^{-1}\boldsymbol{X})^{-1}$ ovat olemassa.

Harjoitustehtävä 5.6. a) Osoita, että yhtälön (5.3) ratkaisu, kun $\tau_1 = \cdots = \tau_n = 1$, minimoi neliö summan

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2.$$

Tämän takia ratkaisua sanotaan pienimmän neliösumman (p.n.s.) estimaatiksi.

b) Osoita, että yhtälön (5.3) ratkaisu yleisessä tapauksessa minimoi neliösumman

$$\sum_{i=1}^{n}\frac{1}{\tau_i}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2.$$

Tämän takia ratkaisua sanotaan painotetun pienimmän neliösumman estimaatiksi.

Harjoitustehtävä 5.7. Muodosta uskottavuusfunktio a) logistisessa regressiossa ja b) Poisson-regressiossa.

Harjoitustehtävä 5.8. Muodosta uskottavuusyhtälöt a) logistisessa regressiossa ja b) Poisson-regressiossa.

5.5 Kanoninen linkki: suurimman uskottavuuden estimaatit

Johdamme algoritmin, jonka avulla voimme laskea suurimman uskottavuuden estimaatit. Algoritmi on kätevä, koska se mahdollistaa estimoinnin yleisessä tapauksessa pelkästään painotetun pienimmän neliösumman algoritmin avulla (IWLS).

1. Valitse sopiva alkuarvo $\boldsymbol{\beta}_0$, ja laske apuarvot
$$u_{i0} = \frac{y_i - m(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}{v(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}, \quad i = 1, \ldots, n. \tag{5.4}$$

2. Muodosta diagonaalimatriisi
$$\boldsymbol{V}_0 = \mathrm{diag}[v(\boldsymbol{x}_1'\boldsymbol{\beta}_0)/\tau_1, \ldots, v(\boldsymbol{x}_n'\boldsymbol{\beta}_0)/\tau_n]$$
ja merkitse $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$.

3. Ratkaise $\boldsymbol{\delta}$ lineaarisesta yhtälöryhmästä
$$\boldsymbol{X}'\boldsymbol{V}_0\boldsymbol{X}\boldsymbol{\delta} = \boldsymbol{X}'\boldsymbol{V}_0\boldsymbol{u}_0.$$

4. Ratkaisu on $\boldsymbol{\delta}_1 = (\boldsymbol{X}'\boldsymbol{V}_0\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{V}_0\boldsymbol{u}_0)$, ja seuraava approksimaatio $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \boldsymbol{\delta}_1$.

5. Sijoita kohdassa 1 $\boldsymbol{\beta}_0$:n paikalle $\boldsymbol{\beta}_1$ ja toista laskut 1–4. Saat $\boldsymbol{\beta}_2$:n. Jatka kohtien 1–4 toistamista, kunnes k. iteraation korjaustermi $\boldsymbol{\delta}_k$ on riittävän lähellä nollavektoria. Vastaava $\boldsymbol{\beta}_k$ on sitten estimaattimme $\widehat{\boldsymbol{\beta}}$.

Algoritmia voidaan perustella seuraavilla tarkasteluilla. Lisätään ja vähennetään yhtälön (5.2) vasemmalle puolelle termi
$$\sum_{i=1}^{n} \frac{1}{\tau_i} m(\boldsymbol{x}_i'\boldsymbol{\beta}_0)\boldsymbol{x}_i,$$
ja järjestellään termejä sopivasti. Silloin
$$\sum_{i=1}^{n} \frac{1}{\tau_i}(y_i - m(\boldsymbol{x}_i'\boldsymbol{\beta}_0))\boldsymbol{x}_i = \sum_{i=1}^{n} \frac{1}{\tau_i}(m(\boldsymbol{x}_i'\boldsymbol{\beta}) - m(\boldsymbol{x}_i'\boldsymbol{\beta}_0))\boldsymbol{x}_i. \tag{5.5}$$
Yhtälön (5.5) oikealla puolella olevaa erotusta $m(\boldsymbol{x}_i'\boldsymbol{\beta}) - m(\boldsymbol{x}_i'\boldsymbol{\beta}_0)$ approksimoidaan $\boldsymbol{\beta}_0$:n ympäristössä gradientin avulla (muistetaan, että $\dot{m} = v$):
$$\frac{\partial m(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = v(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i.$$
Saamme approksimaation
$$m(\boldsymbol{x}_i'\boldsymbol{\beta}) - m(\boldsymbol{x}_i'\boldsymbol{\beta}_0) \;\approx\; v(\boldsymbol{x}_i'\boldsymbol{\beta}_0)\boldsymbol{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

Termejä uudelleen järjestelemällä yhtälön (5.5) oikea puoli on likimäärin

$$\sum_{i=1}^{n} \boldsymbol{x}_i \frac{v(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}{\tau_i} \boldsymbol{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \boldsymbol{X}'\boldsymbol{V}_0\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

Tämän tarkastelun nojalla muokkaamme yhtälön vasemman puolen muotoon

$$\sum_{i=1}^{n} \boldsymbol{x}_i \frac{v(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}{\tau_i} \frac{y_i - m(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}{v(\boldsymbol{x}_i'\boldsymbol{\beta}_0)} = \boldsymbol{X}'\boldsymbol{V}_0\boldsymbol{u}_0,$$

missä vektori $\boldsymbol{u}_0$ sisältää koordinaatit u_{i0} kaavasta (5.4). Olemme saaneet likimääräisen lineaarisen yhtälöryhmän, joka on matriisimuodossa

$$\boldsymbol{X}'\boldsymbol{V}_0\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \boldsymbol{X}'\boldsymbol{V}_0\boldsymbol{u}_0.$$

Tämän yhtälön ratkaisu antaa korjaustermin

$$\boldsymbol{\delta}_1 = (\boldsymbol{X}'\boldsymbol{V}_0\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}_0\boldsymbol{u}_0$$

ja uuden approksimaation $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \boldsymbol{\delta}_1$.

Harjoitustehtävä 5.9. Mitä ovat u_{0i} ja $\boldsymbol{V}_0$:n diagonaalialkiot a) logistisessa regressiossa (esimerkki 5.6) ja b) Poisson-regressiossa (esimerkki 5.7).

Harjoitustehtävä 5.10. Olemme jo nähneet esimerkissä 5.9, että estimointi ei edellytä iterointia normaalisessa lineaarisessa regressiossa. Laske siitä huolimatta u_{0i} ja $\boldsymbol{V}_0$:n diagonaalialkiot tässäkin tapauksessa ja totea, että ratkaisu löytyy jo ensimmäisellä iteraatiolla, jos sitä ryhtyy tekemään.

Yleinen suurimman uskottavuuden teoria antaa tuloksen, että suurimman uskottavuuden estimaatti $\widehat{\boldsymbol{\beta}}$ nodattaa suurissa otoksissa (l. asymptoottisesti) likimäärin multinormaalijakaumaa, jonka odotusarvovektori on oikea arvo $\boldsymbol{\beta}$, ja jonka kovarianssimatriisi saadaan seuraavasti. Ensin lasketaan ns. Fisherin informaatiomatriisi

$$\mathcal{I}(\boldsymbol{\beta}) = -E\left(\frac{\partial^2 \log L(\boldsymbol{\beta}, \phi)}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right).$$

Meidän tilanteessamme suluissa oleva ns. Hessen matriisi ei riipu vasteista, joten osittaisderivaattamatriisi riittää sellaisenaan. Yksinkertainen lasku tuottaa matriisin

$$\mathcal{I}(\boldsymbol{\beta}) = \phi^{-1}\sum_{i=1}^{n} \tau_i^{-1} v(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i\boldsymbol{x}_i' = \phi^{-1}\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{X},$$

missä

$$\boldsymbol{V}(\boldsymbol{\beta}) = \mathrm{diag}[v(\boldsymbol{x}_1'\boldsymbol{\beta})/\tau_1, \ldots, v(\boldsymbol{x}_n'\boldsymbol{\beta})/\tau_n]. \tag{5.6}$$

Likimäärin pätee (ks. esim. Andersen, 1994, luku 3).

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \phi(\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{X})^{-1}\right). \tag{5.7}$$

Harjoitustehtävä 5.11. Muodosta matriisi $\boldsymbol{V}(\boldsymbol{\beta})$ a) logistisessa regressiossa ja b) Poisson-regressiossa.

Seuraava päättelyketju takaa riittävän ehdon sille, että $\widehat{\boldsymbol{\beta}}$ antaa uskottavuusfunktion globaalin maksimin.

- Matriisi $\boldsymbol{X}$ on täysasteinen, ja $\boldsymbol{V}(\boldsymbol{\beta})$ on diagonaalimatriisi, jonka diagonaalialkiot ovat positivisia kaikilla $\boldsymbol{\beta}$:n arvoilla.

$\Rightarrow$ Matriisi $\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{X}$ on positiivisesti definiitti kaikilla $\boldsymbol{\beta}$:n arvoilla.

$\Rightarrow$ Matriisi $-\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{X}$ on negatiivisesti definiitti kaikilla $\boldsymbol{\beta}$:n arvoilla.

$\Rightarrow$ Uskottavuusfunktio $L(\boldsymbol{\beta}, \boldsymbol{\tau})$ on $\boldsymbol{\beta}$:n suhteen aidosti konkaavi, kun τ pidetään vakiona.

$\Rightarrow$ Jos uskottavuus yhtälöllä (5.2) on äärellinen ratkaisu $\widehat{\boldsymbol{\beta}}$, so. kaikki $\widehat{\boldsymbol{\beta}}$:n koordinaatiti ovat äärellisiä, niin $\widehat{\boldsymbol{\beta}}$ antaa globaalin maksimin uskottavuusfunktiolle l. se on suurimman uskottavuuden estimaatti.

Siis jos algoritmimme konvergoi ja tuottaa ratkaisun $\widehat{\boldsymbol{\beta}}$, niin tiedämme, että se on tavoiteltu suurimman uskottavuuden estimaatti. On kuitenkin mahdollista, että äärellistä ratkaisua ei ole olemassa. Käytännössä tämä tarkoittaa, että voi olla olemassa sellainen $\boldsymbol{b}$, että $L(a\boldsymbol{b}, \boldsymbol{\tau})$ on a:n suhteen kasvava, so. maksimi saavutetaan, kun $a \to \infty$. Silloin suurimman uskottavuuden estimaattia ei ole olemassa.[3]

5.6 Yleinen linkkifunktio

Oletetaan jakaumaperhe EDF$(\theta, \phi\tau)$, jossa odotusarvo $\mu = m(\theta)$ ja linkkifunktio $g(\mu) = \boldsymbol{x}'\boldsymbol{\beta}$. Silloin $\theta = m^{-1}(g^{-1}(\boldsymbol{x}'\boldsymbol{\beta})) = h(\boldsymbol{x}'\boldsymbol{\beta})$.

Harjoitustehtävä 5.12. Ratkaise funktio h probit-regressiossa, ks. esimerkki 5.8.

Uskottavuusfunktio on yleisessä tapauksessa

$$L(\boldsymbol{\beta}, \boldsymbol{\tau}) = \prod_{i=1}^{n} c(y_i, \tau_i) \exp\left(\frac{y_i h(\boldsymbol{x}_i'\boldsymbol{\beta}) - b(h(\boldsymbol{x}_i'\boldsymbol{\beta}))}{\tau_i}\right).$$

Uskottavuusyhtälöt saadaan derivoimalla (muista $\dot{b} = m$ ja $h(\cdot) = m^{-1}(g^{-1}(\cdot))$)

$$\frac{\partial \log L(\boldsymbol{\beta}, \boldsymbol{\tau})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{1}{\tau_i}(y_i - g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}))\dot{h}(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i = \boldsymbol{0}. \tag{5.8}$$

[3]Suurimman uskottavuuden estimaatin olemassaolo: ks. Andersen (1994), luku 3.

Menemättä yksityiskohtiin (ks. Fahrmeir and Tutz, 2001, s. 38–43) yhtälöt voidaan ratkaista painotetulla iterativisella pienimmän neliösumman menetelmällä kuten aikaisemminkin, kun määritellään uudelleen seuraavasti

$$\begin{aligned}
\mu_{i0} &= g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}_0) \\
u_{i0} &= \dot{g}(\mu_{i0})(y_i - \mu_{i0}) \\
v_{i0} &= \frac{1}{\dot{g}(\mu_{i0})^2 v(h(\boldsymbol{x}_i'\boldsymbol{\beta}_0))}, \quad i = 1, \ldots, n \\
\boldsymbol{V}_0 &= \text{diag}[v_{10}/\tau_1, \ldots, v_{n0}/\tau_n].
\end{aligned}$$

Luku 6

Lineaarinen regressio: teoria

6.1 Normaalijakauma

Kuten tunnettua standardin normaalijakauman tiheys on

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad -\infty < x < \infty. \tag{6.1}$$

Jos $Z \sim N(0,1)$, niin $Y = \mu + \sigma Z \sim N(\mu, \sigma^2)$, jonka tiheys on

$$\frac{1}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Oletetaan, että $Z_1, \ldots, Z_p$ ovat riippumattomia ja jokainen noudattaa standardia normaalijakaumaa. Silloin niiden yhteisjakauma on

$$f(z_1, \ldots, z_p) = \prod_{i=1}^{p} \varphi(z_i) = (2\pi)^{-\frac{p}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{p} z_i^2\right). \tag{6.2}$$

Kootaan satunnaismuuttujat Z_i vektoriksi $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ ja merkitään $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$. Muodostetaan uudet satunnaismuuttujat

$$Y_i = \mu_i + \sum_{j=1}^{p} a_{ij} Z_j, \quad i = 1, \ldots, p.$$

Tiedämme, että $E(Y_i) = \mu_i$, $\mathrm{var}(Y_i) = \sum_j a_{ij}^2$, $\mathrm{cov}(Y_i, Y_j) = \sum_k a_{ik}a_{jk}$. Matriisimerkinnöin pätee

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z},\ E(\boldsymbol{Y}) = \boldsymbol{\mu},\ \mathrm{cov}(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{A}' = \boldsymbol{\Sigma}. \tag{6.3}$$

Oletamme nyt, että $\boldsymbol{A}$ on epäsingulaarinen, jolloin myös $\boldsymbol{\Sigma}$ on epäsingulaarinen. Teemme muuttujan vaihdoksen $\boldsymbol{z} = \boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$ tiheyteen (6.2). Jacobin determinantti on matriisin $\boldsymbol{A}^{-1}$ determinantin itseisarvo, joka on sama kuin

$$1/\sqrt{|\boldsymbol{A}\boldsymbol{A}'|} = 1/\sqrt{|\boldsymbol{\Sigma}|}.$$

Saamme $\boldsymbol{Y}$:n yhteistiheydeksi

$$f(\boldsymbol{y}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right). \tag{6.4}$$

Sanomme, että $\boldsymbol{Y}$ noudattaa multinormaalijakaumaa $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Vastaavasti $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$. Jos $\boldsymbol{A}$ on singulaarinen, tiheysfunktiota ei ole olemassa, mutta sanomme siitä huolimatta, että $\boldsymbol{Y}$ noudattaa multinormaalijakaumaa $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Seuraava lista sisältää tärkeimmät multinormaalijakauman ominaisuudet:

1. Oletetaan, että $\boldsymbol{B}$ on $r \times p$. Silloin

   $$\boldsymbol{BY} \sim N_r(\boldsymbol{B\mu}, \boldsymbol{B\Sigma B}').$$

   Tässä sallimme myös tilanteen $r > p$, jolloin $\boldsymbol{B\Sigma B}'$ on singulaarinen. Jos $\boldsymbol{\Sigma}$ on epäsingulaarinen (itse asiassa positiivisesti definiitti, p.d.), milloin $\boldsymbol{B\Sigma B}'$ on p.d?

2. Oletetaan yhteensopiva ositus

   $$\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

   missä $\boldsymbol{Y}_1$ on $r \times 1$, $\boldsymbol{Y}_2$ on $(p-r) \times 1$, ja $\boldsymbol{\Sigma}_{12}' = \boldsymbol{\Sigma}_{21}$. Silloin marginaalijakaumat ovat multinormaalisia:

   $$\begin{aligned} \boldsymbol{Y}_1 &\sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \\ \boldsymbol{Y}_2 &\sim N_{p-r}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}). \end{aligned}$$

3. Ehdolliset jakaumat ovat myös multinormaalisia

   $$\begin{aligned} \boldsymbol{Y}_1 \,|\, \boldsymbol{Y}_2 &\sim N_r(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{Y}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}), \\ \boldsymbol{Y}_2 \,|\, \boldsymbol{Y}_1 &\sim N_{p-r}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{Y}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}). \end{aligned}$$

Harjoitustehtävä 6.1. Kohdan (2) tulokset seuraavat kohdan (1) tuloksesta. Miten?

Harjoitustehtävä 6.2. Osoita, että yhteistiheys (6.4) hajoaa marginaalijakaumien tuloksi, kun $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$. Silloin $\boldsymbol{Y}_1$ ja $\boldsymbol{Y}_2$ ovat ...?

6.2 Normaalijakauman johdannaisia

Määritelmä 6.1 (χ^2-jakauma)**.** Oletetaan, että $Z_1, \ldots, Z_k$ ovat riippumattomia ja kukin noudattaa $N(0,1)$-jakaumaa. Silloin

$$Q = Z_1^2 + \cdots + Z_k^2 \sim \chi^2(k).$$

Kaava (6.3) antaa tuloksen $\boldsymbol{Z} = \boldsymbol{A}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$. Käytä sitä ja todista seuraava lause.

Lause 6.1. *Jos* $\boldsymbol{Y}$, $p \times 1$, *noudattaa multinormaalijakaumaa* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *ja* $\boldsymbol{\Sigma}$ *on epäsingulaarinen, niin*

$$(\boldsymbol{Y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2(p).$$

Lause 6.2. *Oletetaan, että* $\boldsymbol{Z}$, $p \times 1$, *noudattaa* $N_p(\boldsymbol{0}, \boldsymbol{I})$*-jakaumaa, ja että* $\boldsymbol{A}$ *on* $p \times p$ *symmetrinen matriisi, jonka ominaisarvot ovat* $\lambda_1, \ldots, \lambda_p$. *Silloin* $\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z}$ *on jakautunut kuten*

$$\lambda_1 U_1^2 + \cdots + \lambda_p U_p^2$$

missä $U_1, \ldots, U_p$ *ovat riippumattomia ja kukin noudattaa* $N(0,1)$*-jakaumaa. Neliömuodon* $\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z}$ *jakauma riippuu* $\boldsymbol{A}$*:sta vain sen ominaisarvojen kautta.*

Todistus. Koska $\boldsymbol{A}$ on symmetrinen, niin on olemassa sellainen ortogonaalinen $p \times p$ matriisi $\boldsymbol{Q}$, $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{I}$, että $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$. Tee sijoitus neliömuotoon $\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z}$. Mikä on $\boldsymbol{Q}'\boldsymbol{Z} = \boldsymbol{U}$:n jakauma? Tee tämä sijoitus ja päättele väite. □

Tärkeä erikoistapaus on se, kun $\boldsymbol{A}$ on idempotentti: $\boldsymbol{A}^2 = \boldsymbol{A}$. Silloin $\boldsymbol{A}$:n ominaisarvot ovat nollia ja ykkösiä. Todista seuraava lause.

Lause 6.3. *Oletetaan, että* $\boldsymbol{Z}$, $p \times 1$, *noudattaa* $N_p(\boldsymbol{0}, \boldsymbol{I})$*-jakaumaa, ja että* $\boldsymbol{A}$ *on* $p \times p$ *symmetrinen ja idempotentti matriisi, jolla on* r *kpl ominaisarvoja 1 ja* $p - r$ *kpl ominaisarvoja 0. Silloin* $\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z} \sim \chi^2(r)$.

Lauseen 6.3 symmetrisen matriisin $\boldsymbol{A}$ ominaisarvohajotelma $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ antaa tuloksen, että $\boldsymbol{A}$:n aste on sama kuin $\boldsymbol{\Lambda}$:n aste (miksi?). Koska $\boldsymbol{\Lambda}$ on diagonaalimatriisi, niin sen aste sama kuin nollasta poikkeavien ominaisarvojen lukumäärä. Päättele tämän perusteella, että lauseessa 6.3 olevat vapausasteet r voidaan ilmaista, myös niin että $\operatorname{rank}(\boldsymbol{A}) = r$ ja $\operatorname{trace}(\boldsymbol{A}) = r$. Jälkimmäinen on matriisin $\boldsymbol{A}$ jälki eli lävistäjäalkioiden summa. Viimeksi mainittu on helpoin tapa laskea (ei numeerisesti vaan analyyttisesti!) vapausasteet.

Lause 6.4. *Oletetaan, että* $\boldsymbol{Z}$, $p \times 1$, *noudattaa* $N_p(\boldsymbol{0}, \boldsymbol{I})$*-jakaumaa, ja että matriisit* $\boldsymbol{P}$ *ja* $\boldsymbol{Q}$ *ovat* $p \times p$ *symmetrisiä ja idempotentteja matriiseja. Jos* $\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{0}$, *niin neliömuodot* $\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}$ *ja* $\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z}$ *ovat riippumattomia.*

Todistus. Koska $\boldsymbol{P}$ ja $\boldsymbol{Q}$ ovat symmetrisiä ja idempotentteja, niiden ominaisarvot ovat nollia ja ykkösiä. Siksi voimme kirjoittaa $\boldsymbol{P} = \boldsymbol{U}_r\boldsymbol{U}_r'$ ja $\boldsymbol{Q} = \boldsymbol{V}_s\boldsymbol{V}_s'$, missä $\boldsymbol{U}_r$ on $p \times r$, $\boldsymbol{U}_r'\boldsymbol{U}_r = \boldsymbol{I}$, ja $\boldsymbol{V}_s$ $p \times s$, $\boldsymbol{V}_s'\boldsymbol{V}_s = \boldsymbol{I}$. Edelleen tiedämme, että $\operatorname{rank}(\boldsymbol{P}) = r < p$ ja $\operatorname{rank}(\boldsymbol{Q}) = s < p$, Olettamuksen mukaan

$$\boldsymbol{0} = \boldsymbol{P}\boldsymbol{Q} = \boldsymbol{U}_r\boldsymbol{U}_r'\boldsymbol{V}_s\boldsymbol{V}_s'.$$

Kertomalla oikeanpuoleisin tulo vasemmalta $\boldsymbol{U}_r'$:lla ja oikealta $\boldsymbol{V}_s$:llä, saamme yhtälön $\boldsymbol{U}_r'\boldsymbol{V}_s = \boldsymbol{0}$. Siis $\boldsymbol{U}_r$:n sarakkeet ovat kohtisuorassa $\boldsymbol{V}_s$:n sarakkeita vastaan, mistä seuraa, että $\boldsymbol{U}_r'\boldsymbol{Z}$ ja $\boldsymbol{V}_s'\boldsymbol{Z}$ ovat korreloimattomia ja multinormaalisuuden nojalla myös riippumattomia. Väite seuraa nyt siitä, että

$$\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{U}_r\boldsymbol{U}_r'\boldsymbol{Z},$$
$$\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{V}_s\boldsymbol{V}_s'\boldsymbol{Z}.$$

□

Määritelmä 6.2 (t-jakauma)**.** Oletetaan, että $Z \sim N(0,1)$, $Q \sim \chi^2(k)$ ja että Z ja Q ovat riippumattomia. Silloin

$$T = \frac{Z}{\sqrt{Q/k}} \sim t(k).$$

Määritelmä 6.3 (Monimuuttujainen t-jakauma)**.** Oletetaan, että $\boldsymbol{Y} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, $Q \sim \chi^2(k)$ ja että $\boldsymbol{Y}$ ja Q ovat riippumattomia. Silloin

$$\boldsymbol{T} = \boldsymbol{\mu} + \frac{1}{\sqrt{Q/k}}\boldsymbol{Y} \sim t_p(k; \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Monimuuttujaisen t-jakauman ominaisuuksia:

- Kun vapausasteet $k > 1$, niin odotusarvo on olemassa ja $E(\boldsymbol{T}) = \boldsymbol{\mu}$. Kun $k > 2$, niin kovarianssimatriisi on olemassa ja $\mathrm{cov}(\boldsymbol{T}) = k(k-2)^{-1}\boldsymbol{\Sigma}$.
- Jos $\boldsymbol{T} \sim t_p(k; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ ja $\boldsymbol{B}$ on $q \times p$, niin $\boldsymbol{BT} \sim t_q(k, \boldsymbol{B\mu}, \boldsymbol{B\Sigma B}')$. Matriisin $\boldsymbol{B}$ ei tarvitse olla neliömatriisi.
- Kun $\boldsymbol{T} \sim t_p(k, \boldsymbol{\mu}, \boldsymbol{I})$, $k > 2$, niin sen koordinaatit T_i ja T_j, $i \neq j$ ovat korreloimattomia mutta eivät riippumattomia.

Määritelmä 6.4 (F-jakauma)**.** Oletetaan, että $Q_1 \sim \chi^2(m)$, $Q_2 \sim \chi^2(n)$ ja että Q_1 ja Q_2 ovat riippumattomia. Silloin

$$T = \frac{Q_1/m}{Q_2/n} \sim F(m,n).$$

Samaan tapaan kuin todistetaan lause 6.1 , voidaan todistaa seuraava lause.

Lause 6.5. *Jos* $\boldsymbol{T} \sim t_p(k; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ *ja* $\boldsymbol{\Sigma}$ *on epäsingulaarinen, niin*

$$\frac{(\boldsymbol{T}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{T}-\boldsymbol{\mu})}{p} \sim F(p,k).$$

6.3 Estimointi

Siirrymme nyt merkinnöissä yksinkertaisempaan tilastotieteessä tavanomaiseen käytäntöön ja merkitsemme satunnaismuuttujia ja -vektoreita samalla symbolilla kuin niiden arvoja. Kirjoitamme tavallisen lineaarisen regressiomallin lyhyesti matriisisalgebran merkinnöin seuraavasti

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Ellei toisin mainita, oletamme, että matriisin $\boldsymbol{X}$ ensimmäinen sarake koostuu ykkösistä, ts. oletamme, että mallissa on vakio β_0. Vastaavasti uskottavuusfunktio on

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right], \\ &= (2\pi)^{-\frac{n}{2}}\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right]. \end{aligned}$$

Edellisessä luvussa saimme tuloksen, että $\boldsymbol{\beta}$:n suurimman uskottavuuden estimaatti saadaan varianssista σ^2 riippumatta kaavasta

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Kun sijoitetaan tämä uskottavuusfunktioon $\boldsymbol{\beta}$:n paikalle ja otetaan uskottavuusfunktiosta logaritmi, saamme

$$\log L(\widehat{\boldsymbol{\beta}}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{2\sigma^2}. \tag{6.5}$$

Derivoidaan tämä σ^2:n suhteen ja asetetaan derivaatta nollaksi:

$$\frac{\partial \log L(\widehat{\boldsymbol{\beta}}, \sigma^2)}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0.$$

Ratkaisu on

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n}$$

Harjoitustehtävä 6.3. Osoita, että $\hat{\sigma}^2$ antaa funktiolle (6.5) maksimin.

Saamamme σ^2:n estimaatti on kuitenkin harhainen, kuten jatkossa nähdään. Onkin tapana tehdä harhan korjaus ja määritellä parametrin σ^2 estimaatti kaavalla

$$\begin{aligned} s^2 &= \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n-p-1} \\ &= \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

Sen neliöjuurta s sanotaan jäännöshajonnaksi tai joskus mallin keskivirheeksi.

6.4 Sovite, jäännökset ja selitysaste

Merkitään kuten aikaisemminkin matriisin $\boldsymbol{X}$ i. riviä $\boldsymbol{x}_i'$:llä. Silloin

$$\begin{aligned} \boldsymbol{x}_i'\boldsymbol{\beta} &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \\ \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}. \end{aligned}$$

Mallin *sovitteet* (fitted values) ja *jäännökset* (residuals) ovat vastaavasti

$$\begin{aligned} \hat{y}_i &= \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \\ e_i &= y_i - \hat{y}_i, \quad i = 1, \ldots, n. \end{aligned}$$

Kokoamalla nämä vektoreiksi voidaan kirjoittaa matriisimerkinnöin

$$\begin{aligned}\widehat{\boldsymbol{y}} = \begin{bmatrix}\hat{y}_1\\ \vdots \\ \hat{y}_n\end{bmatrix} &= \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\\ \boldsymbol{e} = \begin{bmatrix}e_1\\ \vdots \\ e_n\end{bmatrix} &= \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\\ &= \big[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big]\boldsymbol{y}. \qquad (6.6)\end{aligned}$$

Harjoitustehtävä 6.4. Osoita matriisilaskuilla, että $\widehat{\boldsymbol{y}}'\boldsymbol{e} = 0$ ja $\boldsymbol{e}'\boldsymbol{X} = \boldsymbol{0}'$. Tilastollinen tulkinta kohtisuoruudelle on korreloimattomuus. Siis jäännökset eivät korreloi sen paremmin sovitteiden kuin selittäjienkään kanssa. Täsmällisemmin sanottuna *Pearsonin korrelaatikerroin on nolla*. Epälineaarinen riippuvuus on mahdollista ks. esim. kuva 2.3.

Jäännösvarianssi voidaan kirjoittaa jäännösten avulla

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-p-1} = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-p-1}. \qquad (6.7)$$

Vasteen kokonaisvaihtelu keskiarvon suhteen on neliösumma

$$SYY = \sum_{i=1}^n (y_i - \bar{y})^2,$$

missä $\bar{y}$ on keskiarvo $n^{-1}\sum y_i$. Seuraavaksi johdetaan hyödyllinen neliösummahajotelma, joka jakaa tämän kokonaisvaihtelun mallin selittämään osaan ja jäännösvaihteluun. Suoraan laskemalla saamme

$$\begin{aligned}SYY &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2\\ &= \sum_{i=1}^n e_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^n e_i(\hat{y}_i - \bar{y})\\ &= \sum_{i=1}^n e_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.\end{aligned}$$

Harjoitustehtävä 6.5. Miksi edellisessä laskussa ristitulosumma häviää?

Mallin selittämä vaihtelu ja jäännösvaihtelu ovat vastaavasti

$$\begin{aligned}SS_{reg} &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,\\ RSS &= \sum_{i=1}^n e_i^2.\end{aligned}$$

Selitysaste (coefficient of determination) määritellään osamääränä

$$R^2 = \frac{SS_{reg}}{SYY} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Harjoitustehtävä 6.6. Kirjoita kaavana sovitteiden ja havintojen (otos-)korrelaatiokerroin. Tee osoittajaan sijoitus $y_i = \hat{y}_i + e_i$ ja osoita, että tuloksena on selitysasteen neliöjuuri.

6.5 Estimaattien otosjakaumat

Edellä olemme johtaneet p.n.s. estimaatit $\widehat{\boldsymbol{\beta}}$. Nyt tarkastelemme niiden ominaisuuksia.

Harjoitustehtävä 6.7. Sijoita kaavaan $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ vastevektorin $\boldsymbol{y}$ paikalle mallin antama esitysmuoto $\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ja tee tarvittavat laskut niin, että saat tulokseksi $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

Harjoitustehtävä 6.8. Käytä hyväksesi edelliisen tehtävän laskuja ja tietoa $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$. Johda tulos

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Harjoitustehtävä 6.9. Kun oletetaan virhetermien multinormaalisuus

$$\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \tag{6.8}$$

saadaan tulos

$$\widehat{\boldsymbol{\beta}} \sim N_{p+1}\big(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\big).$$

Kirjoita perustelu näkyviin.

Siirrytään seuraavaksi johtamaan jäännösvarianssin s^2 jakaumaa.

Harjoitustehtävä 6.10. Otetaan käyttöön merkintä $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Muista kaava (6.6). Perustele seuraavat päättelyt:

$$\begin{aligned}
&\boldsymbol{Q} \Rightarrow \boldsymbol{e} = \boldsymbol{Q}\boldsymbol{y},\\
&\boldsymbol{Q} \Rightarrow \boldsymbol{Q}\boldsymbol{X} = \boldsymbol{0},\\
&\therefore \boldsymbol{e} = \boldsymbol{Q}\boldsymbol{\varepsilon},
\end{aligned}$$

ja nämäkin päättelyt

$$\begin{aligned}
&\boldsymbol{Q} \Rightarrow \boldsymbol{Q}^2 = \boldsymbol{Q},\\
&\boldsymbol{Q} \Rightarrow \boldsymbol{Q}' = \boldsymbol{Q},\\
&\mathrm{trace}(\boldsymbol{Q}) = n - p - 1.
\end{aligned}$$

Käytä tietoa(6.8) ja kirjoita vielä perustelu lopulliselle tulokselle

$$\frac{\boldsymbol{\varepsilon}'\boldsymbol{Q}\boldsymbol{\varepsilon}}{\sigma^2} \sim \chi^2(n - p - 1).$$

Lopuksi osoitetaan, että estimaatit $\widehat{\boldsymbol{\beta}}$ ja s^2 ovat (stokastisesti) riippumattomia.

Harjoitustehtävä 6.11. Kirjoitetaan aluksi

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{0}\end{bmatrix} + \begin{bmatrix}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\\ \boldsymbol{Q}\end{bmatrix}\boldsymbol{\varepsilon},$$

Perustele, että

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \boldsymbol{e}\end{bmatrix}$$

on multinormaalinen. Laske sen kovarianssimatriisi

$$\operatorname{cov}\left(\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \boldsymbol{e}\end{bmatrix}\right) = \begin{bmatrix}\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \mathbf{0}\\ \mathbf{0}' & \sigma^2\boldsymbol{Q}\end{bmatrix}.$$

Kirjoita perustelut sille, että vektorit $\widehat{\boldsymbol{\beta}}$ ja $\boldsymbol{e}$ ovat (stokastisesti) riippumattomia. Mistä seuraa, että $\widehat{\boldsymbol{\beta}}$ ja s^2 ovat (stokastisesti) riippumattomia?

Kokoamme tulokset lauseeksi.

Lause 6.6. *Oletetaan lineaarinen malli $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I})$. Silloin suurimman uskottavuuden estimaateille pätee*

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right), \tag{6.9}$$

$$\frac{(n-p-1)s^2}{\sigma^2} \sim \chi^2(n-p-1). \tag{6.10}$$

Lisäksi $\widehat{\boldsymbol{\beta}}$ ja s^2 ovat (stokastisesti) riippumattomia.

Seurauslause 6.1. *Edellisen lauseen olettamuksista seuraa, että*

$$\frac{1}{s}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim t_{p+1}(n-p-1, \mathbf{0}, (\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Harjoitustehtävä 6.12. Kirjoita perustelut yo. seurauslauseelle käyttämällä monimuuttujaisen t-jakauman määritelmää 6.2. Tarkoituksena on löytää tarvittava multinormaalinen vektori ja siitä riippumaton satunnaismuuttuja, joka noudattaa tarvittava χ^2-jakaumaa.

6.6 Regressiokertoimien luottamusvälit

Yksinkertaisuuden vuoksi merkitään $\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ ja $\boldsymbol{C} = (c_{jk})$.

Harjoitustehtävä 6.13. Perustele seurauslauseen 6.1 ja määritelmän 6.2 avulla tulos

$$\frac{\hat{\beta}_j - \beta_j}{s} \sim t_1(n-p-1, 0, c_{jj}),$$

ja tämän perusteella vielä tulos

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \sim t(n-p-1).$$

Harjoitustehtävä 6.14. Oletetaan Studentin t-jakauman kvantiili $t_{\frac{\alpha}{2};n-p-1}$, jonka yläpuolelle jää jakaumasta osuus $\alpha/2$. Johda luottamusväli (engl. confidence interval)

$$\hat{\beta}_j \pm s\sqrt{c_{jj}}\, t_{\frac{\alpha}{2};n-p-1},$$

jonka *peittotodennäköisyys* (engl. coverage probability) on $1-\alpha$. Minkä tapahtuman todennäköisyydestä on kysymys? Mikä on satunnaista?

Estimaatin $\hat{\beta}_j$ estimoitua keskihajontaa $s\sqrt{c_{jj}}$ sanotaan $\hat{\beta}_j$:n *keskivirheeksi* (engl. standard error), ja sen merkintä jatkossa on s.e.$(\hat{\beta}_j)$. Luottamusväli siis voidaan kirjoittaa muotoon

$$\hat{\beta}_j \pm \text{s.e.}(\hat{\beta}_j) t_{\frac{\alpha}{2};n-p-1}. \tag{6.11}$$

Esimerkki 6.1. Aineistossa `fuel1974` (Weisberg, 2005) on tietoja Yhdysvaltojen osavaltioista vuodelta 1974 (lukuunottamatta Alaskaa ja Havaijia):

- `fuel` = keskimääräinen moottoriajoneuvojen polttoaineen kulutus, gallona/asukas
- `tax` = polttoainevero, cent/gallona
- `inc` = keskimääräiset tulot, 1000$/asukas
- `dlic` = ajokortillisten osuus, %
- `road` = liittovaltion tukeman tiestön pituus, 1000 mailia

```
fuel <- read.table("http://users.jyu.fi/~junyblom/fuel1974.dat",
header=TRUE)

c.tax <- tax - mean(tax)
c.inc <- inc - mean(inc)
c.dlic <- dlic - mean(dlic)
c.road <- road - mean(road)

lm.fuel <- lm(fuel ~ c.tax + c.inc + c.dlic + c.road)

lm(formula = fuel ~ c.tax + c.inc + c.dlic + c.road)
            coef.est coef.se
(Intercept) 576.77     9.57
c.tax       -34.79    12.97
c.inc       -66.59    17.22
c.dlic       13.36     1.92
c.road       -2.43     3.39
---
n = 48, k = 5
residual sd = 66.31, R-Squared = 0.68
```

```
b <- coef(lm.fuel)
se <- se.coef(lm.fuel)

tab <- rbind(b + qt(.025, df=48-5)*se, b + qt(.975, df=48-5)*se)
rownames(tab) <- c("2.5%", "97.5%")
round(tab,2)

      (Intercept)    tax     inc  dlic  road
2.5%       557.47 -60.95 -101.32  9.49 -9.26
97.5%      596.07  -8.63  -31.86 17.24  4.41
```

6.7 Luottamusväli bootstrap-menetelmällä

Jos on syytä epäillä mallin virhetermien normaalisuutta, voi varmuuden vuoksi laskea luottamusvälit bootstrap-menetelmällä, joka ei edellytä virheen jakauman tuntemista. Yksinkertaisin tapa soveltaa bootstrap-menetelmää lineaarisessa mallissa on seuraavanlainen (Efron and Tibshirani, 1993, Sec. 9.5):

- Poimi otos palauttaen tapauksista $(y_i, \boldsymbol{x}_i')$, so. havaintomatriisin riveistä.
- Laske otoksesta kertoimien estimaatit $\boldsymbol{\beta}^*$.
- Toista edellisiä kohtia B kertaa. Saat estimaatit $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_B^*$.
- Etsi koordinaattikohtaisesti $\alpha/2$ ja $1-\alpha/2$ -kvantiilit, joista saat luottamusvälit, joiden peittotodennäköisyys on likimäärin $1-\alpha$.

Tässä R-koodit, joka tekee edellä kuvatun tehtävän. Huom. ao. koodissa oleva `X` ei sisällää ykkössaraketta!

```
bootstrap <- function(y,X){
n <- length(y)
s <- sample(1:n, replace=TRUE)
out <- lm(y[s] ~ X[s,])
coef(out)
}

X <- cbind(c.tax, c.inc, c.dlic, c.road)

b.boot <- replicate(10000, bootstrap(fuel,X))

tab.boot <- apply(b.boot, 1, quantile, c(.025, .975))
colnames(tab.boot) <- c("(Intercept)",colnames(X))
round(tab.boot, 2)

      (Intercept)  c.tax   c.inc c.dlic c.road
```

```
2.5%       558.52 -57.18 -100.99   8.45  -9.16
97.5%      596.30  -8.15  -41.81  18.64   4.14
```

Prediktorin `inc` kertoimen luottamusväli poikkeaa eniten teorian mukaisesta. Muiden kohdalla erot ovat selvästi vähäisempiä.

6.8 Usean regressiokertoimen samanaikainen testaus

Yksittäisen kertoimen merkitsevyyden näkee, kun laskee luottamusvälin tai p-arvon, jonka useimmat ohjelmistot antavat automaattisesti. Funktio `display()` antaa p-arvon optiolla `detail = TRUE`. Kolmi- tai useampitasoisen faktorin merkitsevyyttä olemme jo aikaisemmin testanneet ns. F-testillä. Käymme seuraavaksi läpi sen perustelut.

Oletetaan, että lineaarisessa regressiomallissa, jossa on kaikkiaan p prediktoria vakion lisäksi. Haluamme testata k prediktorin merkitsevyyttä samanaikaisesti. Ne voivat olla esim. $k+1$ -tasoisen faktorin kertoimia tai kahden faktorin interaktioita tms. Merkintöjen kannalta on yksinkertaisinta olettaa, että kysymyksessä on k ensimäistä kerrointa $\beta_1, \ldots, \beta_k$ (käytännössä näin ei tietenkään tarvitse olla). Nollahypoteesi on

$$H_0 : \beta_1 = \cdots = \beta_k = 0.$$

Kertoimien estimaatit ovat $\hat{\beta}_1, \ldots, \hat{\beta}_k$. Kootaan estimaatit vektoriin $\widehat{\boldsymbol{\beta}}_k$ ja kovarianssit matriisiksi $\sigma^2 \boldsymbol{C}_{kk}$.

Harjoitustehtävä 6.15. a) Perustele seurauslauseen 6.1 ja monimuuttujaisen t-jakauman ominaisuuksien perusteella, että

$$\frac{1}{s}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) \sim t_k(n-p-1; \mathbf{0}, \boldsymbol{C}_{kk}).$$

b) Perustele lauseen 6.5 nojalla, että

$$\frac{(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)'(s^2 \boldsymbol{C}_{kk})^{-1}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)}{k} \sim F(k, n-p-1).$$

Tulosta voisi periaatteessa käyttää yhteisen luottamusalueen määrittämisessä koko vektorille $\boldsymbol{\beta}_k$, mutta jos $k \geq 3$, niin sen visualisointi ei helposti onnistu.

Harjoitustehtävä 6.16. Kaikesta huolimatta määrittele kaavana ko. luottamusalue.

Harjoitustehtävä 6.17. Kuvittele, että asetat nollahypoteesin mukaisesti $\boldsymbol{\beta}_k = \mathbf{0}$ ed. tehtävän kaavassa ja lasket

$$F_{\text{obs}} = \widehat{\boldsymbol{\beta}}_k'(s^2 \boldsymbol{C}_{kk})^{-1} \widehat{\boldsymbol{\beta}}_k / k.$$

Oleta, että sinulla on $F(k, n-p-1)$-jakauman kriittinen arvo $F_{\alpha,k,n-p-1}$, jonka yläpuolella on jakaumasta osuus α. (esim. 5 %). Mitä päätelmiä voit tehdä tilanteissa $F_{\text{obs}} > F_{\alpha,k,n-p-1}$ ja $F_{\text{obs}} < F_{\alpha,k,n-p-1}$? Käytä sekä "luottamusvälikieltä"että "testikieltä".

Testin voi tehdä myös seuraavasti:

- Sovita malli, jossa on kaikki prediktorit ja ota talteen jäännösneliösumma $RSS = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ ja siihen liityvät vapausasteet $n - p - 1$.
- Sovita malli, josta puuttuvat kertoimia $\beta_1, \ldots, \beta_k$ vastaavat prediktorit ja ota talteen jäännösneliösumma $RSS_0 = (\boldsymbol{y} - \boldsymbol{X}_0\widehat{\boldsymbol{\beta}}_0)'(\boldsymbol{y} - \boldsymbol{X}_0\widehat{\boldsymbol{\beta}}_0)$ ja vastaavat vapausasteet $n - (p - k) - 1 = n - p + k - 1$. Tämä neliösumma ei koskaan ole pienempi kuin RSS.
- Laske erotus $RSS_0 - RSS$ ja vapausasteiden erotus $(n-p+k-1)-(n-p-1) = k$.
- Laske osamäärä
$$F = \frac{(RSS_0 - RSS)/k}{RSS/(n-p-1)}$$
ja vertaa F jakaumaan vapausastein $(k, n - p - 1)$.

Molemmat menetelmät, sekä estimaatteihin perustuva että jäännösneliösummiin perustuva, johtavat numeerisesti täsmälleen samaan lopputulokseen. Tämä jätetään lukijalle harjoitustehtäväksi.

6.9 Ryhmien monivertailu

Kun vertaillaan useampaa kuin kahta ryhmää, halutaan joskus ryhmien eroille sellaiset luottamusvälit, jotka ovat voimassa samanaikaisesti. Oletetaan, että ryhmiä on $k + 1$ kpl ja että ryhmien erot referenssitason suhteen ovat $\hat{\beta}_j$, $j = 1, \ldots, k$. Luottamusvälit poikkeamille referenssitasosta ovat

$$\hat{\beta}_j \pm \mathrm{se}(\hat{\beta}_j) t_{\frac{\alpha}{2};n-p-1}, \quad j = 1, \ldots, k,$$

missä p on prediktoreiden kokonaismäärä vakiota lukuun ottamatta. Muiden ryhmien valisille poikkeamille saadaan luottamusvälit seuraavasti:

$$\hat{\beta}_j - \hat{\beta}_r \pm \mathrm{se}(\hat{\beta}_j - \hat{\beta}_r) t_{\frac{\alpha}{2};n-p-1},$$
$$\mathrm{se}(\hat{\beta}_j - \hat{\beta}_r) = s\sqrt{c_{jj} + c_{rr} - 2c_{jr}} \quad j = 1, \ldots, k-1;\ r = j+1, \ldots, k.$$

Harjoitustehtävä 6.18. Anna perustelu erotuksen $\hat{\beta}_j - \hat{\beta}_r$ keskivirheelle.

Tavoitteena on löytää sellainen α, että kaikki luottamusvälit ovat yhtaikaa voimassa todennäköisyydellä $1 - \alpha^*$ (esim. $1 - \alpha^* = 0.95$). Luottamusvälejä on kaiken kaikkiaan $m = (k+1)k/2$ kpl. Oletetaan nyt, että $\boldsymbol{K}$ on sellainen $m \times (p+1)$ matriisi, että $\boldsymbol{K}\widehat{\boldsymbol{\beta}}$ sisältää kaikki parivertailut eikä mitään muuta.

Monimuuttujaisen t-jakauman ominaisuuksien perusteella pätee

$$\frac{1}{s}\boldsymbol{K}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim t_m(n - p - 1, \boldsymbol{K}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}')$$

Seuraavaksi muodostetaan diagonaalimatriisi $\boldsymbol{D}$, jonka diagonaalialkiot ovat matriisin $\boldsymbol{K}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}'$ diagonaalialkioiden neliöjuuria. Silloin matriisin $\boldsymbol{A} = \boldsymbol{D}^{-1}\boldsymbol{K}$ avulla saa standardoituja poikkeamia. Esim. jos meillä on $k+1 = 3$ ryhmää

$$\frac{1}{s}\boldsymbol{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \begin{bmatrix} \frac{\hat{\beta}_1 - \beta_1}{s\sqrt{c_{11}}} \\ \frac{\hat{\beta}_2 - \beta_2}{s\sqrt{c_{22}}} \\ \frac{(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)}{s\sqrt{c_{11}+c_{22}-2c_{12}}}. \end{bmatrix}$$

Yritämme sitten ratkaista α:n niin, että

$$P\left(-t_{\frac{\alpha}{2};n-p-1} < \frac{1}{s}\boldsymbol{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) < t_{\frac{\alpha}{2};n-p-1}\right) = 1 - \alpha^*,$$

missä epäyhtälöt ovat koordinaattikohtaisia. Jokaisen yksittäisen luottamusvälin peittotodennäköisyys on siis $1-\alpha$ mutta kaikki ovat samanaikaisesti voimassa todennäköisyydellä $1-\alpha^*$. Tietysti pätee $1-\alpha > 1-\alpha^*$. Analyyttistä ratkaisua on käytännössä mahdotonta löytää, mutta jakaumaa $t_m(n-p-1; \boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}')$ on helppo simuloida (se ei sisällä mitään tuntematonta!).

Esimerkki 6.2. Palataan luvun 2.5 esimerkkiin:

```
 lm.add <- lm(ruotsi.pist ~ mies + koulu)
display(lm.add)

lm(formula = ruotsi.pist ~ mies + koulu)
            coef.est coef.se
(Intercept) 233.33     4.01
mies        -21.03     3.62
kouluB      -10.48     5.09
kouluC        2.10     5.18
kouluD       -0.87     5.16
---
n = 375, k = 5
residual sd = 34.03, R-Squared = 0.10

tab <- simultaneous.ci(level=0.95, lm.obj=lm.add, factor=koulu)

 Individual coverage probability  0.9886882

 round(tab, 1)
              Difference 0.025 0.975
kouluA-kouluB       10.5  -2.5  23.4
kouluA-kouluC       -2.1 -15.3  11.1
kouluA-kouluD        0.9 -12.3  14.0
kouluB-kouluC      -12.6 -24.9  -0.3
kouluB-kouluD       -9.6 -21.9   2.7
kouluC-kouluD        3.0  -9.5  15.5
```

Yksittäisen välin luottamuskerroin on 0.99, mutta simultaaninen luottamuskerroin on 0.95.

Vaikka sukupuolen ja koulun interaktio ei ole aivan merkitsevä, katsotaan, miten voimme tehdä koulujen monivertailun naisille ja interaktioille:

```
lm.int <- lm(ruotsi.pist ~ mies * koulu)
display(lm.int)
b <- coef(lm.int)

## Naisille

tab <- simultaneous.ci(level=0.95, lm.obj=lm.int, names=names(b)[3:5],
ref="kouluA")

 Individual coverage probability  0.9899149

             Difference 0.025 0.975
kouluA-kouluB        3.6 -12.6  19.8
kouluA-kouluC       -9.8 -26.7   7.1
kouluA-kouluD        1.5 -15.1  18.2
kouluB-kouluC      -13.4 -29.5   2.7
kouluB-kouluD       -2.0 -17.8  13.8
kouluC-kouluD       11.4  -5.2  27.9

## Interaktioille

tab <- simultaneous.ci(level=0.95, lm.obj=lm.int, names=names(b)[6:8],
ref="mies:kouluA")

Individual coverage probability  0.9890787

round(tab, 1)
                       Difference 0.025 0.975
mies:kouluA-mies:kouluB       19.9  -7.4  47.1
mies:kouluA-mies:kouluC       20.3  -7.0  47.7
mies:kouluA-mies:kouluD        0.3 -27.1  27.7
mies:kouluB-mies:kouluC        0.5 -24.6  25.5
mies:kouluB-mies:kouluD      -19.6 -44.7   5.6
mies:kouluC-mies:kouluD      -20.0 -45.3   5.2
```

Huomaamme, että molemmissa ryhmissä kaikki luottamusvälit sisältävät nollan. Kuva 2.7 näyttää, että sukupuolten ero on keskimäärin suurempi kuin koulujen välinen. Kun oletetaan additiivinen malli koulujen B ja C ero korostuu ja tulee merkitseväksi. Koulujen erot miesten ja naisten ryhmissä erikseen eivät tule merkitseviksi monivertailussa, mikä selittyy sillä, että koulujen erot ovat pienehköjä ja mahdollinen interaktiokin vähäistä.

6.10 Ennustaminen

Regressiomallia voidaan käyttää ennustamiseen. Oletetaan, että meillä on käytössä estimaatit $\widehat{\boldsymbol{\beta}}$ ja s^2. Haluamme ennustaa vasteen arvon, kun selittäjien arvot on annettu, esim. $X_1 = x_{a1}, \cdots, X_p = x_{ap}$. Merkitään $\boldsymbol{x}_a' = (1, x_{a1}, \cdots, x_{ap})$. Tulevaisuuden arvon y_a (piste-)ennusteeksi on luonnollista valita $\boldsymbol{x}_a'\widehat{\boldsymbol{\beta}} = \hat{y}_a$. Voimme arvioida ennusteen tarkkuutta, jos voimme olettaa, että $y_a \sim N(\boldsymbol{x}_a'\boldsymbol{\beta}, \sigma^2)$, so. tulevaisuuden arvo tulee saman mallin mukaisesti kuin aikaisempi aineisto. Ennuste on harhaton (miksi?)

$$E(y_a - \hat{y}_a) = 0$$

Kun vielä oletetaan, että y_a on riippumaton aikaisemmista havainnoista $y_1, \ldots, y_n$, niin (miksi?)

$$\text{var}(y_a - \hat{y}_a) = \sigma^2(1 + \boldsymbol{x}_a'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_a),$$

Kun korvataan σ estimaatillaan s, samanlainen päättely kuin regressiokertoimien luottamusvälejä johdettaessa antaa tulokseksi *ennustevälin*

$$\hat{y}_a \pm s\sqrt{1 + \boldsymbol{x}_a'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_a}\, t_{\frac{\alpha}{2};\, n-p-1},$$

jonka peittotodennäköisyys on $1 - \alpha$.

Ennusteväliä ei pidä sekoittaa ennusteen odotusarvon $\boldsymbol{x}_a'\boldsymbol{\beta}$:n *luottamusväliin*

$$\boldsymbol{x}_a'\widehat{\boldsymbol{\beta}} \pm s\sqrt{\boldsymbol{x}_a'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_a}\, t_{\frac{\alpha}{2};\, n-p-1},$$

joka yleensä huomattavasti lyhyempi. Väärin käytettynä se antaa harhaan johtavan käsityksen ennusteen tarkkuudesta.

Esimerkki 6.3. Seuraava aineisto on teoksesta, Freund et al. (2006). Mittayksiköt on muutettu metrisiksi:

- vol = tilavuus, m^3,
- dbh = rinnakorkeusläpimitta, cm,
- d5 = läpimitta 5m:n korkeudessa, cm,
- ht = pituus, m.

```
puut <- read.table("http://users.jyu.fi/~junyblom/puut.dat",
header=T)

puut.lm <- lm(formula = vol ~ dbh + ht + d5, data = puut)

summary(puut.lm)

Call:
lm(formula = vol ~ dbh + ht + d5, data = puut)
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.148799 -0.047473 -0.003616  0.043133  0.141557

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.07453    0.40046  -7.677 9.42e-07 ***
dbh          0.01812    0.01144   1.585 0.132611
ht           0.06445    0.01515   4.254 0.000606 ***
d5           0.06323    0.01340   4.717 0.000232 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08765 on 16 degrees of freedom
Multiple R-squared: 0.9591,     Adjusted R-squared: 0.9514
F-statistic: 124.9 on 3 and 16 DF,  p-value: 2.587e-11
```

Ennustetaan tilavuutta 2 puulle ja lasketaan 90%:n ennustevälit.

dbh	ht	d5
30	30	25
50	32	45

```
uusi <- data.frame(matrix(c(30,50,30,32,25,45), 2,3))
colnames(uusi) <- colnames(puut)[1:3]
puut.pred <- predict.lm(puut.lm,newdata=uusi,interval="prediction",
level=0.9)
round(puut.pred,2)
   fit  lwr  upr
1 0.98 0.79 1.18
2 2.74 2.56 2.92
```

■

Luku 7

Yleinen lineaarinen malli

Ryhdymme tarkastelemaan aikaisempaa yleisempää lineaarista mallia. Kun aikaisemmin oletettiin, että virheet ovat riippumattomia ja vakiovarianssisia, nyt oletamme, että $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{V}$, $\boldsymbol{V} \neq \boldsymbol{I}$. Oletamme aluksi, että $\boldsymbol{V}$ on tunnettu epäsingulaarinen matriisi.

Uskottavuusfunktio on nyt

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2}\sigma^{-n}|\boldsymbol{V}|^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right] \tag{7.1}$$

Regressiokertoimien suurimman uskottavuuden estimaatiksi saadaan

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}, \tag{7.2}$$

ja suurimman uskottavuuden estimaatti jäännösvarianssille on

$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}). \tag{7.3}$$

Kaavaa (7.2) muistuttaa kaavaa (5.3). Erona on se, että jälkimmäisen matriisi $\boldsymbol{T}$ on diagonaalimatriisi. Tässä $\boldsymbol{V}$ on mikä tahansa positiivisesti definiitti matriisi.

Harjoitustehtävä 7.1. Laske $\text{cov}(\widehat{\boldsymbol{\beta}})$.

Harhaton estimaatti s^2 saadaan, kuten aikaisemminkin, korvaamalla n vapausasteiden mukaisella arvolla $n - p - 1$. Lause 6.6 pätee sillä muutoksella, että jäännösvarianssi saadaan kaavasta (7.3) ja että

$$\widehat{\boldsymbol{\beta}} \sim N[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}].$$

Regressiokertoimien keskivirheet saadaan kovarianssimatriisin lävistäjäalkioiden neliöjuurina, kun σ korvataan estimaatillaan. Luottamusvälit ja testit ovat voimassa näillä muutoksilla. Tulokset edellyttävät, että $\boldsymbol{V}$ on täysin tunnettu. Käytännössä $\boldsymbol{V}$ riippuu usein joistakin lisäparametreista $\boldsymbol{\theta}$. Silloin annetulla $\boldsymbol{\theta}$:n arvolla lasketut kaavojen (7.2) ja (7.3) mukaiset estimaatit ovat $\boldsymbol{\theta}$:n funktioita. Sijoittamalla nämä

arvot uskottavuusfunktioon (7.1) saamme (laske!) ns. profiiliuskottavuuden, joka logaritmoituna on

$$\log \hat{L}(\boldsymbol{\theta}) = \text{vakio} - \frac{n}{2}\log \hat{\sigma}^2_{\boldsymbol{\theta}} - \frac{1}{2}\log |\boldsymbol{V}_{\boldsymbol{\theta}}|, \tag{7.4}$$

missä varianssin ja kovarianssimatriisin riippuvuus $\boldsymbol{\theta}$:sta on merkitty näkyviin. Kun suurimman uskottavuuden estimaatti $\widehat{\boldsymbol{\theta}}$ on saatu, saadaan myös kertoimien estimaatti $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\theta}}}$ ja $\hat{\sigma}^2 = \hat{\sigma}^2_{\widehat{\boldsymbol{\theta}}}$. Jälkimmäisessä ajatellaan (7.3):n mukaisesti, että nimittäjässä on n. Se tietysti voidaan korvata arvolla $n - p - 1$.

Usein käytetään $\boldsymbol{\theta}$:n ja σ^2:n estimoinnissa uskottavuuden (7.4) sijasta ns. rajoitettua uskottavuutta (engl. restricted likelihood, Patterson and Thompson, 1971). Se saadaan, kun ensin eliminoidaan vasteen $\boldsymbol{y}$ riippuvuus $\boldsymbol{\beta}$:sta. Eliminointi voidaan tehdä muunnoksella $\boldsymbol{N}'\boldsymbol{y}$, missä $n \times (n - p - 1)$ matriisi $\boldsymbol{N}$ on sellainen, että

$$\boldsymbol{N}'\boldsymbol{X} = \boldsymbol{0}, \quad \boldsymbol{N}'\boldsymbol{N} = \boldsymbol{I}.$$

Silloin $\boldsymbol{N}'\boldsymbol{y} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{N}'\boldsymbol{V}_{\boldsymbol{\theta}}\boldsymbol{N})$. Käytännössä matriisia $\boldsymbol{N}$ ei tarvitse eksplisiittisesti konstruoida. Muutamien matriisialgebran laskujen jälkeen saamme rajoitetun logaritmisen uskottavuuden muotoon

$$\begin{aligned} \log L_{\text{reml}}(\sigma^2, \boldsymbol{\theta}) &= \text{vakio} - \frac{n-p-1}{2}\log \sigma^2 - \frac{1}{2}(\log |\boldsymbol{V}_{\boldsymbol{\theta}}| + \log |\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}|) \\ &\quad -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}})'\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}). \end{aligned} \tag{7.5}$$

Tämän uskottavuuden maksimoivia estimaatteja sanotaan REML-estimaateiksi (restricted maxumum likelihood estimates). Estimaattien $\widehat{\boldsymbol{\theta}}$ ja $\hat{\sigma}^2_{\widehat{\boldsymbol{\theta}}}$ harha on yleensä pienempi kuin tavallisten suurimman uskottavuuden estimaattien harha. Esim. jos $\boldsymbol{V}$ on täysin tunnettu, niin (7.5) antaa aikaisemmin mainitun harhattoman estimaatin s^2 varianssille σ^2.

Huomautus 7.1. Kun $\boldsymbol{\theta}$ estimoidaan, $\boldsymbol{\beta}$ estimoidaan kaavan (7.2) avulla, missä $\boldsymbol{V}$ korvataan matriisilla $\boldsymbol{V}_{\widehat{\boldsymbol{\theta}}}$, missä $\widehat{\boldsymbol{\theta}}$ on joko tavallinen suurimman uskottavuuden estimaatti tai REML-estimaatti. Sama pätee myös estimaattiin $\hat{\sigma}^2$ kaavassa (7.3). Nimittäjään laitetaan kuitenkin $n - p - 1$, kun estimointi on REML.

Joissakin tapauksissa $\widehat{\boldsymbol{\beta}}$ on sama riippumatta siitä, käytetäänkö tavallista suurimman uskottavuuden menetelmää vai REML-estimointia (yleensä silloin myös tavallinen p.n.s. antaa saman tuloksen, ks. luku 7.3), mutta keskivirheet kuitenkin yleensä poikkeavat toisistaan. Valmisohjelmia käytettäessä kannattaa tarkistaa, mikäli mahdollista, miten keskivirheet on laskettu.

Harjoitustehtävä 7.2. Kaavan (7.5) perusteella a) laske $\hat{\sigma}^2_{\widehat{\boldsymbol{\theta}}}$, ja b) muodosta profiiliuskottavuus REML-estimoinnissa.

Käytännön sovelluksissa ongelmaksi muodostuu usein käänteismatrisin $\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}$ laskeminen. Seuraavaksi esitämme kolme erikoistapausta, joissa käänteismatriisin saa helposti analyyttisesti laskettua.

7.1 Painotettu pienimmän neliösumman menetelmä

Oletetaan, että $\boldsymbol{V} = \mathrm{diag}[v_1^2, \ldots, v_n^2]$. Virhetermit ovat siis riippumattomia, mutta niiden varianssit vaihtelevat. Tällainen tilanne syntyy esim. silloin kun havaitsemme keskiarvot $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$, mutta emme yksittäisiä havaintoja $y_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_{ij}$, missä virheet ovat riippumattomia vakiovarianssisia. Nyt $\mathrm{var}(\bar{y}_i) = \sigma^2/n_i$. Tietokoneohjelmissa pitää yleensä antaa painokerroin, joka on verrannollinen varianssin käänteislukuun. Tässä tapauksessa painokerroin on otoskoko n_i.

Joskus havaitaan empiirisesti, että

$$\begin{aligned} \mathrm{var}(y_i) &= \mathrm{var}(\varepsilon_i) \\ &\propto g\big(E(y_i)\big) = g(\boldsymbol{x}_i'\boldsymbol{\beta}), \end{aligned}$$

missä g on jokin kasvava funktio: $e^{x_i'\beta}$, tai $(\boldsymbol{x}_i'\boldsymbol{\beta})^2$ tai jokin muu. Parametrit voi estimoida iteratiivisesti lähtemällä jostakin alkuestimaatista $\widehat{\boldsymbol{\beta}}_0$, jolloin painoiksi asetetaan $1/g(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_0) = 1/g(\widehat{y}_i)$. Painotettu p.n.s. antaa uuden estimaatin $\widehat{\boldsymbol{\beta}}_1$. Voi tyytyä tähän tai jatkaa iterointia siihen asti kunnes estimaatti ei enää muutu. Koska painokertoimet on estimoitu aineistosta, regressiokertoimien luottamusvälit ovat vain likiarvoja samoin kuin testien p-arvot.

7.2 Aikasarjaregressio

Eräs tavallinen tilanne, missä mallin virheet korreloivat, syntyy silloin, kun havainnot muodostavat aikasarjan. Aikasarjaregressiossa meidän täytyy kuitenkin miettiä uudelleen vasteiden ja prediktoreiden keskinäiset suhteet. Tähän asti olemme tarkastelleet tilannetta, missä havainnot $(y_i, \boldsymbol{x}_i)$ ovat riippumattomia eri i:n arvoilla. Yhteisjakauman voi silloin kirjoittaa muotoon

$$\prod_{i=1}^{n} f(y_i, \boldsymbol{x}_i) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{x}_i) f(\boldsymbol{x}_i),$$

missä f on geneerinen tiheysfunktion merkintä. Kun oletetaan, että tiheys $f(\boldsymbol{x}_i)$ ei riipu kiinnostavista regressioparametreista, voimme perustaa tilastollisen päättelyn pelkästään ehdollisiin tiheyksiin $f(y_i \mid \boldsymbol{x}_i)$.

Aikasarjojen kohdalla joudumme monimutkaisempaan yhteistiheyteen

$$f(y_1, \boldsymbol{x}_1, \ldots, y_n, \boldsymbol{x}_n),$$

jonka voi kuitenkin aina kirjoittaa ehdollisten tiheyksien tuloksi

$$f(y_1, \boldsymbol{x}_1, \ldots, y_n, \boldsymbol{x}_n) = f(y_1, \boldsymbol{x}_1) \prod_{t=2}^{n} f(y_t, \boldsymbol{x}_t \mid y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1), \tag{7.6}$$

missä olemme ottaneet juoksevan indeksin merkiksi t:n muistuttamaan, että kysymyksessä on aikasarja ja että ajallinen jarjestys $t = 1, \ldots, n$ on oleellinen asia. Voimme

edelleen kirjoittaa ehdolliset tiheydet (7.6) muotoon

$$\begin{aligned} f(y_t, \boldsymbol{x}_t \mid y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1) &= f(y_t, \mid \boldsymbol{x}_t, y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1) \\ &\quad \times f(\boldsymbol{x}_t \mid y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1). \end{aligned}$$

Teemme tässä nyt yksinkertaistavan olettamuksen

$$f(\boldsymbol{x}_t \mid y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1) = f(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_1), \quad t = 1, \ldots, n. \tag{7.7}$$

Olettamuksemme tarkoittaa, että $\boldsymbol{x}_t$ on ehdollisesti riippumaton menneistä vasteen arvoista $y_{t-1}, \ldots, y_1$, kun prediktorien arvot $\boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_1$ on annettu. Tämä on vahva olettamus eikä se suinkaan ole aina voimassa. Mutta kun tämä olettamus on voimassa, ja kun prediktoreiden tiheys ei riipu kiinnostavista regressioparametreista, voimme rajoittua tilastollisissa tarkasteluissa ehdolliseen tiheyteen

$$f(y_t \mid \boldsymbol{x}_t, y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1).$$

Oletamme nyt, että (7.7) pätee, ja rakennamme aikasarjamallin. Yksinkertainen malli virheiden aikariippuvuudelle on autoregressiivinen malli

$$\varepsilon_t = \phi\varepsilon_{t-1} + \eta_t, \quad t = 2, \ldots, n, \tag{7.8}$$

missä $|\phi| < 1$. Termit η_t ovat keskenään riippumattomia ja noudattavat normaalijakaumaa $N(0, \sigma^2)$. Lisäksi niiden oletetaan olevan riippumattomia ε_1:sta. Prosessin aloitus voidaan tehdä *stationaarisen* jakauman mukaan

$$\varepsilon_1 \sim N(0, \sigma^2/(1-\phi^2)). \tag{7.9}$$

Voidaan osoittaa, että nyt kaikki virheet ε_t noudattavat samaa normaalijakaumaa $N(0, \sigma^2/(1-\phi^2))$ ja että $\mathrm{cor}(\varepsilon_{t_1}, \varepsilon_{t_2}) = \phi^{|t_1 - t_2|}$.

Harjoitustehtävä 7.3. Johda edellisen perusteella $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{V}_\phi$.

Harjoitustehtävä 7.4. Johda matriisi $\boldsymbol{L}_\phi$, joka toteuttaa yhtälön

$$\boldsymbol{L}_\phi \boldsymbol{\varepsilon} = \begin{bmatrix} \sqrt{1-\phi^2}\varepsilon_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}.$$

Osoita, että $\mathrm{cov}(\boldsymbol{L}_\phi \boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$.

Harjoitustehtävä 7.5. Osoita, että a) $\boldsymbol{V}_\phi = \boldsymbol{L}_\phi^{-1}(\boldsymbol{L}_\phi')^{-1}$, ja b) $\boldsymbol{V}_\phi^{-1} = \boldsymbol{L}_\phi' \boldsymbol{L}_\phi$.

Harjoitustehtävä 7.6. Osoita, että $|\boldsymbol{V}_\phi^{-1}| = 1 - \phi^2$ ja siis $|\boldsymbol{V}_\phi| = 1/(1-\phi^2)$.

Kirjoitamme aluksi kuten ennenkin

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \varepsilon_t, \quad t = 1, \ldots, n.$$

Olettamus (7.7) saadaan voimaan, kun oletetetaan, että prediktorit $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ ovat riippumattomia virheistä $\varepsilon_1, \eta_2, \ldots, \eta_n$. Käyttämällä rekursiota (7.8) ja alkuarvoa (7.9) saamme muodon (tee laskut)

$$\begin{aligned} y_1 &= \boldsymbol{x}_1'\boldsymbol{\beta} + \varepsilon_1, && (7.10)\\ y_t &= \phi y_{t-1} + (\boldsymbol{x}_t - \phi\boldsymbol{x}_{t-1})'\boldsymbol{\beta} + \eta_t, \quad t = 2, \ldots, n. && (7.11) \end{aligned}$$

Voimme nyt päätellä, että

$$\begin{aligned} y_1 \mid \boldsymbol{x}_1 &\sim N(\boldsymbol{x}_1'\boldsymbol{\beta}, \sigma^2/(1-\phi^2)),\\ y_t \mid \boldsymbol{x}_t, y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1 &\sim N(\phi y_{t-1} + (\boldsymbol{x}_t - \phi\boldsymbol{x}_{t-1})'\boldsymbol{\beta}, \sigma^2). \end{aligned}$$

Menneisyyden avulla voimme ennustaa y_t:n arvoa kaavalla

$$E(y_t \mid \boldsymbol{x}_t, y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1) = \phi y_{t-1} + (\boldsymbol{x}_t - \phi\boldsymbol{x}_{t-1})'\boldsymbol{\beta}.$$

Regressiokertoimien tulkinnan saa seuraavasti. Olkoon $\boldsymbol{a}_j$ sellainen, että sen koordinaatit ovat nollia muualla paitsi, että paikassa j on 1. Silloin $\boldsymbol{a}_j'\boldsymbol{\beta} = \beta_j$, ja

$$E(y_t \mid \boldsymbol{x}_t + \boldsymbol{a}_j, y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1) - E(y_t \mid \boldsymbol{x}_t, y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1) = \boldsymbol{a}_j'\boldsymbol{\beta} = \beta_j.$$

Siis yhden yksikön muutos x_{tj}:ssä ennustaa β_j:n muutosta y_t:ssä. Emme kuitenkaan voi laskea ennustetta kahden askeleen päähän eli y_{t+1}:n ennustetta kaavasta

$$E(y_{t+1} \mid \boldsymbol{x}_t, y_{t-1}, \boldsymbol{x}_{t-1}, \ldots, y_1, \boldsymbol{x}_1)$$

mallimme perusteella, jos emme tunne prediktoriprosessin dynamiikkaa (miksi?).

Parametrit estimoimme suurimman uskottavuuden menetelmällä. Ehdollisten jakaumien (7.10) ja (7.11) perusteella uskottavuusfunktion logaritmi saa muodon

$$\begin{aligned} \log L(\boldsymbol{\beta}, \sigma^2, \phi) &= \text{vakio} - \frac{n}{2}\log\sigma^2 + \frac{1}{2}\log(1-\phi^2) - \frac{1}{2\sigma^2}\Bigg[(1-\phi^2)(y_1 - \boldsymbol{x}_1'\boldsymbol{\beta})^2 + \\ &\qquad \sum_{t=2}^{n}\Big((y_t - \phi y_{t-1}) - (\boldsymbol{x}_t - \phi\boldsymbol{x}_{t-1})'\boldsymbol{\beta}\Big)^2\Bigg]. \end{aligned}$$

Kun kiinnittää ϕ:n, $\boldsymbol{\beta}$:n voi estimoida painotetulla pienimmän neliösumman menetelmällä (Vaste? Prediktorit? Painot?). Estimaatin $\widehat{\boldsymbol{\beta}}_\phi$ avulla saadaan $\hat{\sigma}^2_\phi$ ja lopulta logaritminen profiiliuskottavuus

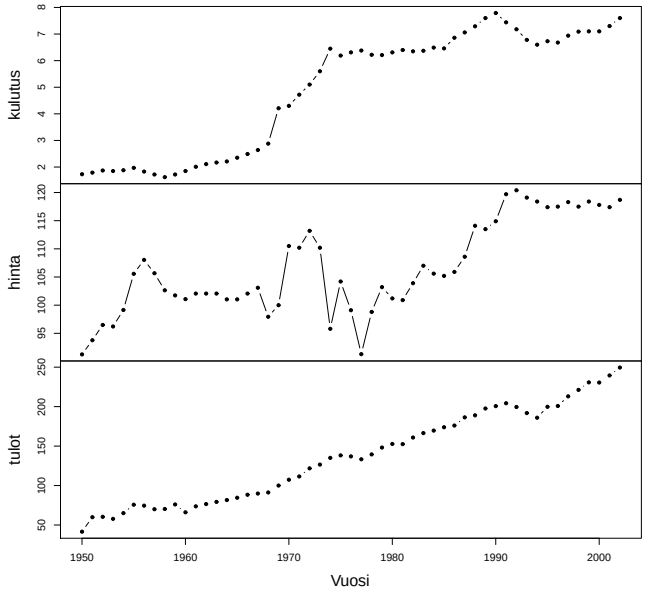
$$\log \hat{L}(\phi) = \text{vakio} - \frac{n}{2}\log(\hat{\sigma}^2_\phi) + \frac{1}{2}\log(1-\phi^2). \qquad (7.12)$$

Estimointimenetelmänä käytetään yleensä suurimman uskottavuuden menetelmää. Myös REML-estimointi on mahdollista, vaikka se ei ehkä ole tyypillistä aikasarjaregresiossa. Logaritminen profiiliuskottavuus on silloin, ks. kaava (7.5) ja harjoitustehtävä 7.2,

$$\log \hat{L}_{\text{reml}}(\phi) = \text{vakio} - \frac{n-p-1}{2}\log(\hat{\sigma}^2_\phi) + \frac{1}{2}\log(1-\phi^2) - \frac{1}{2}\log(|\boldsymbol{X}'\boldsymbol{L}'_\phi\boldsymbol{L}_\phi\boldsymbol{X}|). \quad (7.13)$$

Kun ϕ estimoidaan, luottamusvälit ja testien p-arvot ovat vain likiarvoja.

Esimerkki 7.1. Aineistona on alkoholin kulutus Suomessa litroina 15 vuotta täyttänyttä henkilöä kohti, alkoholin reaalinen hintaindeksi ja kotitalouksien käytettävissä olevat reaaliset tulot (indeksi). Kiinnostuksen kohteena on, miten hinta ja tulot ennustavat kulutuksen muutoksia. Aineisto kattaa vuodet 1950–2002.

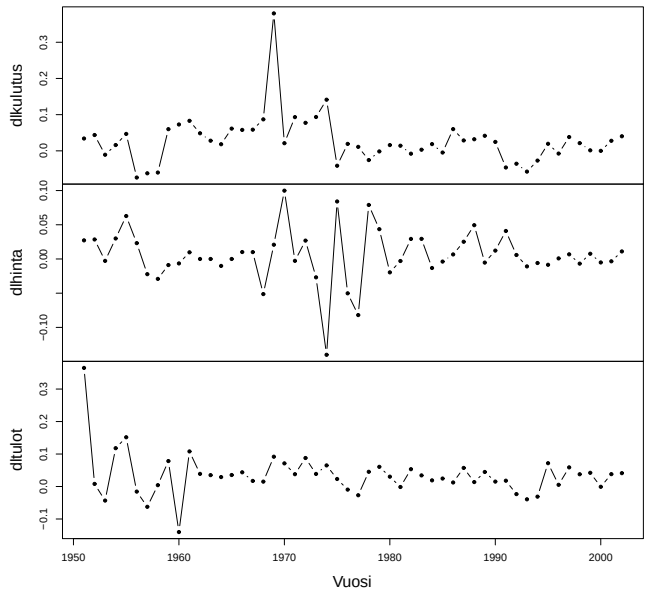


Kuva 7.1: *Alkoholin kulutus litroina 15 vuotta täyttänyttä henkilöä kohti, alkoholin reaalinen hinta ja kotitalouksien käytettävissä olevat reaaliset tulot vuosina 1950–2002.*

Kuvasta 7.1 näemme, että kaikissa sarjoissa näkyy selvä kasvava trendi selvimmin kulutuksessa ja tuloissa. Tarkemmin katsoen huomaamme myös, että kulutuksen kasvu vuodesta 1968 vuoteen 1969 on suhteellisen suuri. Paremmin tilanteen näkee kuvasta 7.2, jossa on muuttujien logaritmiset erotukset l. likimääräiset suhteelliset vuosimuutokset

$$\log\left(\frac{y_t}{y_{t-1}}\right) = \log\left(1 + \frac{y_t - y_{t-1}}{y_{t-1}}\right) \approx \frac{y_t - y_{t-1}}{y_{t-1}}.$$

Kuva paljastaa, että v. 1969 kulutus kasvoi yli 30 % vuodesta 1968. V. 1969 alusta alkaen alkoholilakia muutettiin, niin, että keskiolutta voitiin myydä elintarvikeliikkeissä. Lisäksi Alkon myymäläverkkoa laajennettiin myös maaseudulle.



Kuva 7.2: *Suhteelliset vuosimuutokset (logaritmiset erotukset) alkoholin kulutuksessa reaalihinnassa ja reaalituloissa vuosina 1950–2002.*

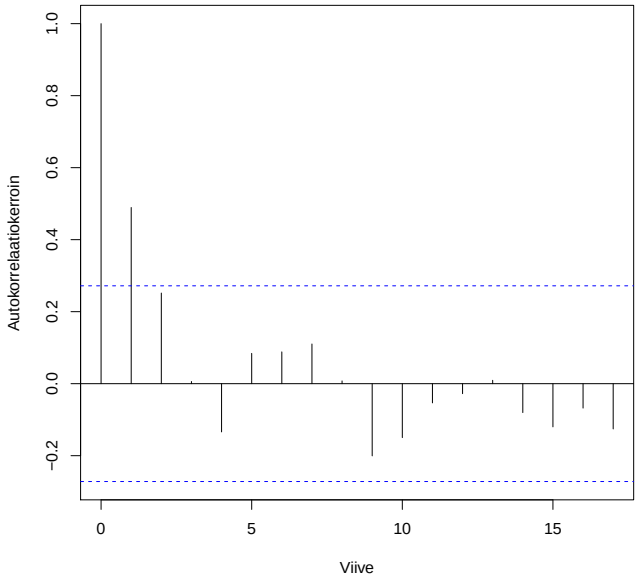
Graafisten tarkasteluiden perusteella aloitamme mallilla, missä vasteena on kulutuksen logaritmiset erotukset ja prediktoreina logaritmiset erotukset sekä hinnasta että tuloista. Nämä muunnokset poistavat trendin kaikista sarjoista ja on sen takia paremmin sopusoinnussa em. autoregressivisen virhemallin kanssa. Lisäksi otamme mukaan dummy muuttujan, joka poimii kulutushyppäyksen 1968–69. Ajamme aluksi tavanomaisen regression, so. oletamme, että $\phi = 0$.

```
lm(formula = dlkulutus ~ dlhinta + dltulot + dd69)
            coef.est coef.se
(Intercept)  0.015    0.006
dlhinta     -0.474    0.152
dltulot      0.258    0.088
dd69         0.351    0.041
---
n = 52, k = 4
residual sd = 0.040, R-Squared = 0.66
```

Kuvat 7.3 ja 7.4 osoittavat merkittävää autokorrelaatiota jäännöksissä erityisesti

viipeellä 1. Jäännösten autokorrelaatiokertoimen r_k viipeellä k määritelee kaava

$$r_k = \frac{\sum_{t=1}^{n-k} e_{t+k} e_t}{\sum_{t=1}^{n} e_t^2}.$$



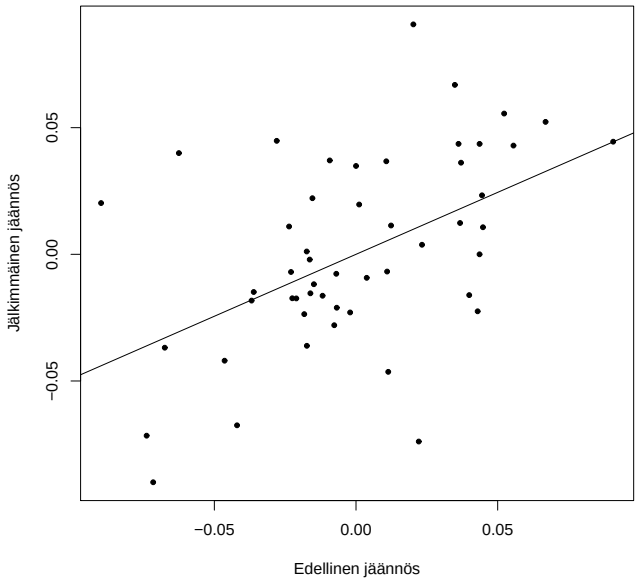
Kuva 7.3: *Jäännösten autokorrelaatot tavallisesta regressiosta.*

Seuraavaksi parametrit (myös ϕ) estimoidaan maksimoimalla uskottavuusfunktiot (7.12) ja (7.13). Kaikilla kolmella tavalla estimoidut kertoimet ovat ao. taulukossa (ols-rivillä on asetettu $\phi = 0$).

```
     (Intercept) dlhinta dltulot  dd69   phi
ols        0.015  -0.474   0.258 0.351 0.000
mle        0.018  -0.473   0.171 0.321 0.564
reml       0.018  -0.473   0.170 0.321 0.578
```

Huomaamme, että tavallinen MLE ja REML tuottavat lähes samat regressiokertoimien estimaatit. Parametrin ϕ estimaatit poikkevat jonkin verran. Ainoastaan tuloihin liittyvä tavallisen regression kerroin poikkeaa selvästi vastaavista MLE- ja REML-estimaateista.

Tavallisesta regressiosta saadut keskivirheet poikkevat enemmän suurimman uskottavuuden estimaateista (MLE ja REML antavat käytännössä samat). Ehkä hieman yllättäen AR-mallilla estimoitujen kertoimien keskivirheet ovat pienempiä kuin tavallisella regressiolla estimoitujen kertoimien keskivirheet. Vakion keskivirhe on ainoa poikkeus.



Kuva 7.4: *Jäännösten viivepisteparvi (viive = 1). Suoran kulmakerroin on autokorrelaatiokerroin* $r_1 = 0.489$.

	(Intercept)	dlhinta	dltulot	dd69
se.ols	0.006	0.152	0.088	0.041
se.mle	0.011	0.105	0.065	0.029
se.reml	0.011	0.104	0.065	0.029

Kertoimien tulkinnat:

1. *Vakio* kertoo kulutuksen muutoksen, kun hinta ja tulot pysyvät ennallaan (dlhinta = dltulot = 0) ja dd69 = 0, so. paitsi v. 1969. Siis hinnasta ja tuloista riippumaton sekä v. 1969 lain muutoksesta riippumaton kulutuksen kasvu on $100 \times (e^{0.018} - 1) = 1.9$ % vuodessa. AR-mallin mukaan kerroin ei aivan ole merkitsevä.

2. *Hintaan* liittyvä kerroin (hintajousto) saa tulkinnan, että muiden prediktoreiden pysyessä samoina, 1 %:n hinnan kasvu ennustaa samana vuonna 0.47 %:n laskua kulutuksessa: $100 \times (1.01^{-0.473} - 1) = -0.469$.

3. *Tuloihin* liittyvä kerroin (tulojousto) kertoo, että muiden prediktoreiden pysyessä vakioina, 1 %:n kasvu ennustaa samana vuonna 0.17 %:n kasvua kulutuksessa.

4. *Dummy*-muuttujaan liittyvä kerroin on suuri 0.321. Jos oletamme, että hinnan ja tulojen lisäksi ei ole muita sekoittavia muuttujia, niin voinemme tulkita, että alkoholilain muutos on aiheuttanut $100 \times (e^{0.321}) = 37.9$ %:n tasomuutoksen kulutuksessa v. 1969.

Simuloimalla estimoitua mallia voimme tutkia ϕ:n estimaatin harhaa ja laskea sille bootstrap-luottamusvälin sekä ottaa huomioon myös ϕ:s estimaatin epävarmuus regressiokertoimien estimaateissa. Osoittautuu, että MLE-estimaatissa on harhaa 3–4 sadasosaa alaspäin (so. estimaatti on keskimäärin liian pieni). Prosenttipistemenetelmä antaa 95 %:n välin $(0.25, 0.73)$. Harhan takia suositeltavampi BC_a-väli (Efron and Tibshirani, 1993, Sec. 14.3) on $(0.31, 0.76)$. Vastaavat REML-estimoinnista saadut välit ovat sekä prosenttipiste- että BC_a-menetelmällä pyöristettynä $(0.28, 0.77)$. Nämä ovat ehkä luotettavimmat. Joka tapauksessa ϕ on merkitsevästi positiivinen. Käytännön kannalta ϕ ei ole niin kiinnostava parametri kuin regressiokertoimet.

Simuloimme estimoitua mallia ja estimoimme simuloidusta aineistosta uudelleen ϕ:n, kertoimet $\boldsymbol{\beta}$ ja jäännösvarianssin σ^2. Voimme käyttää bootstrap-t luottamusvälejä (ks. luku 3.2) tai tavallista prosenttipistemenetelmää (ks. luku 6.7). Kertoimien luottamusvälit ovat

	std. teoria		prosenttipiste		bootstrap-t	
	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
(Intercept)	0.00	0.04	0.00	0.04	-0.01	0.04
dlhinta	-0.68	-0.26	-0.68	-0.27	-0.68	-0.26
dltulot	0.04	0.30	0.04	0.30	0.04	0.30
dd69	0.26	0.38	0.26	0.38	0.26	0.38

Merkittäviä eroja menetelmien välillä ei tällä kertaa kuitenkaan ole.

7.3 Hierarkkinen malli

Monissa tilanteissa aineisto muodostuu varsinaisten tilastoyksiköiden klustereista tai ryhmistä. Esim. koulutustutkimuksissa aineisto voidaan poimia ryväsotannalla, jossa aluksi poimitaan otos kouluista ja kussakin koulussa sitten otos oppilaista. Tällaisissa tilanteissa syntyy vasteiden välille riippuvuutta: saman koulun oppilailla on keskenään vähemmän vaihtelua verrattuna jonkin toisen koulun oppilaisiin. Tuonnempana tarkastelemme esimerkkiä, joissa tutkitaan taimien pituuskasvua. Tietty määrä taimia istutetaan samaan ruukkuun, joita on useita. Myös samassa ruukussa kasvatetut taimet korreloivat keskenään, kun taas eri ruukiussa kasvatetut ovat riippumattomia toisistaan.

Yleisesti voimme kuvata tällaista tilannetta mallilla, jossa ryhmään i kuuluvat vasteet toteuttavat yhtälön

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \eta_i + \varepsilon_{ij}, \quad i = 1, \ldots, k;\ j = 1, \ldots, n_i.$$

Kaiken kaikkiaan vasteita on $n = n_1 + \cdots + n_k$. Lisäksi oletamme, että virheet η_i ja ε_{ij} ovat kaikki riippumattomia keskenään, $\eta_i \sim N(0, \sigma_\eta^2)$ ja $\varepsilon_{ij} \sim N(0, \sigma^2)$. Selvästi vektorit $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i\,n_i})'$ ovat keskenään riippumattomia, mutta ryhmän sisällä on riippuvuutta. Itse asiassa

$$\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma_\eta^2 \mathbf{1}_i\mathbf{1}_i' + \sigma^2 \boldsymbol{I}_i),$$

missä $\mathbf{1}_i$ on $n_i \times 1$ ja $\boldsymbol{I}_i$ on $n_i \times n_i$. Matriisiin $\boldsymbol{X}_i$ on pinottu rivit $\boldsymbol{x}'_{ij}, j = 1, \ldots, n_i$. Uskottavuusfunktioksi saadaan (miksi!)

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2_\eta, \sigma^2) &= \prod_{i=1}^{k} (2\pi)^{-\frac{n_i}{2}} |\sigma^2_\eta \mathbf{1}_i \mathbf{1}'_i + \sigma^2 \boldsymbol{I}_i|^{-\frac{1}{2}} \\ &\quad \times \exp\left[-\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta})' (\sigma^2_\eta \mathbf{1}_i \mathbf{1}'_i + \sigma^2 \boldsymbol{I}_i)^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) \right]. \end{aligned}$$

Jatkossa merkitsemme $\sigma^2_\eta / \sigma^2 = \rho$.

Harjoitustehtävä 7.7. Osoita, että

$$(\rho \mathbf{1}_i \mathbf{1}'_i + \boldsymbol{I}_i)^{-1} = \boldsymbol{I}_i - \frac{\rho}{1 + n_i \rho} \mathbf{1}_i \mathbf{1}'_i.$$

Harjoitustehtävä 7.8. Osoita edellisen tehtävän tilanteessa, että

$$|\rho \mathbf{1}_i \mathbf{1}'_i + \boldsymbol{I}_i| = 1 + n_i \rho.$$

Näiden harjoitustehtävien perusteella saamme uskottavuusfunktion logaritmin muotoon

$$\begin{aligned} \log L(\boldsymbol{\beta}, \rho, \sigma^2) &= \text{vakio} - \frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^{k} \Big[\log(1 + n_i \rho) \\ &\quad - \frac{1}{2\sigma^2} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta})' \Big(\boldsymbol{I}_i - \frac{\rho}{1 + n_i \rho} \mathbf{1}_i \mathbf{1}'_i \Big) (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) \Big]. \quad (7.14) \end{aligned}$$

Kun ρ kiinnitetään saamme estimaatit $\widehat{\boldsymbol{\beta}}_\rho$ ja $\hat{\sigma}^2_\rho$.

Informaatiota satunnaiskomponenttivektorista $\boldsymbol{\eta}$ saamme ehdollisen odotusarvon $E(\boldsymbol{\eta} \mid \boldsymbol{y})$ kautta (ks. luku 6.1). Tässä tilanteessa yksittäinen η_i riippuu vain $\boldsymbol{y}_i$:stä, joten saamme (laske!)

$$E(\eta_i \mid \boldsymbol{y}_i) = \frac{n_i \rho}{1 + n_i \rho} (\bar{y}_i - \bar{\boldsymbol{x}}'_i \boldsymbol{\beta}),$$

missä

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad \bar{\boldsymbol{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}.$$

Kun $\boldsymbol{\beta}$ ja ρ korvataan estimaateillaan saamme

$$\hat{\eta}_i = \frac{n_i \hat{\rho}}{1 + n_i \hat{\rho}} (\bar{y}_i - \bar{\boldsymbol{x}}'_i \widehat{\boldsymbol{\beta}})$$

Harjoitustehtävä 7.9. Johda (7.14):n perusteella kaavat a) $\widehat{\boldsymbol{\beta}}_\rho$:lle ja b) $\hat{\sigma}^2_\rho$:lle.

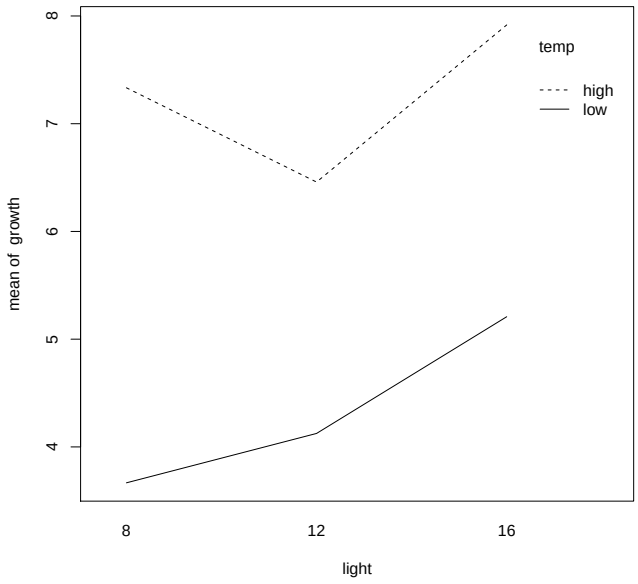
Logaritminen profiiliuskottavuus on

$$\log \hat{L}(\rho) = \text{vakio} - \frac{n}{2} \log \hat{\sigma}^2_\rho - \frac{1}{2} \sum_{i=1}^{k} \log(1 + n_i \rho).$$

Vastaava rajoitettu logaritminen profiiliuskottavuus on

$$\begin{aligned}\log \hat{L}(\rho) &= \text{vakio} - \frac{n-p-1}{2}\log\hat{\sigma}^2_\rho \\ &\quad -\frac{1}{2}\sum_{i=1}^{k}\log(1+n_i\rho) - \frac{1}{2}\log\Big|\sum_{i=1}^{k}\boldsymbol{X}_i'\Big(\boldsymbol{I}_i - \frac{\rho}{1+n_i\rho}\mathbf{1}_i\mathbf{1}_i'\Big)\boldsymbol{X}_i\Big|.\end{aligned}$$

Esimerkki 7.2. Krzanowski (1998, p. 152) kuvaa kokeen, jossa tutkitaan päivänvalon ja yölämpotilan vaikutusta taimien pituuskasvuun. Käsittelyjä on kaikkiaan 6 kpl: 8, 12 ja 16 tunnin altistus päivänvalolle kahdessa kasvihuoneessa, joista toisessa on alhainen ja toisessa korkea yölämpötila. Taimet, joita oli 72 kpl, istutettiin satunnaisesti ruukkuihin, 4 kuhunkin. Kolmen ruukun taimet käsiteltiin samalla tavalla. Käsittelyjä oli 6 kpl. Taimien kasvu (cm) mitattiin viikon kuluttua. Edellä kuvatun teorian mukaisesti samaan ruukkuun sijoitetut taimet muodostavat yhden klusterin.



Kuva 7.5: *Valon määrän ja yölämpötilan interaktio*

Teorian mukaisen matriisin $\boldsymbol{X}_i$ ilmoittavat, millaisen käsittelyn ruukun i taimet saavat. Koska kaikki saman ruukun taimet käsitellään samalla tavalla, matriisin $\boldsymbol{X}_i$ rivit ovat samoja. Lisäksi kaikissa on yhtä monta riviä (4 kpl). Näistä seikoista johtuu, että sekä tavallinen että yleistetty p.n.s. tuottavat samat estimaatit käsittelyvaikutuksille. Estimaattien keskivirheet kuitenkin poikkevat toisistaan.

Varianssien estimaatit

```
      Residual Random comp. Ratio
ML       0.934        0.125 0.134
REML     0.934        0.305 0.326
```

Huomaamme, että jäänösvarianssin estimaatit ovat samat (johtuu pyöristyksestä). Sen sijaan satunnaiskomponentin l. ruukkujen vaikutus estimoituu selvästi suuremmaksi REML-estimoinnissa.

Kertoimien estimaatit ja keskivirheet eri menetelmillä ovat

```
                coef.est se.ols se.ml se.reml
(Intercept)        7.333  0.310 0.346   0.424
templow           -3.667  0.439 0.489   0.599
light12           -0.875  0.439 0.489   0.599
light16            0.583  0.439 0.489   0.599
templow:light12    1.333  0.621 0.692   0.847
templow:light16    0.958  0.621 0.692   0.847
```

Odotetusti keskivirheet ovat suurempia, kun otetaan ruukkujen tuoma vaihtelu huomioon. Suurimmat keskivirheet tuottaa REML. Käytännössä useimmiten luotetaan REML-estimaatteihin. Interaktiokuvasta 7.5 näemme, että korkeamman yölämpötilan kohdalla valonmäärän muutos 8 tunnista 12 tuntiin näyttäisi vähentävän kasvua, mikä tuntuu epäuskottavalta. Interaktiotesti ei olekaan merkitsevä, esim. REML-estimoinnissa saamme p-arvon 0.27. Kun sovitetaan malli ilman interaktioita, saamme varianssien estimaateiksi

```
     Residual Random comp. Ratio
ML      0.934        0.204 0.219
REML    0.934        0.329 0.352
```

Kertoimien estimatit keskivirheineen ovat

```
            coef.est se.ols se.ml se.reml
(Intercept)    6.951  0.259 0.312   0.354
templow       -2.903  0.259 0.312   0.354
light12       -0.208  0.317 0.382   0.433
light16        1.062  0.317 0.382   0.433
```

Korkeammassa lämpötilassa 8 tunnin valaistuksessa kasvu on keskimäärin 7.0 cm, ja 95%:n luottamusväli on (6.2, 7.7). Alempi yölämpötila vähentää kasvua n. 2.9 cm. Vastaava luottamusväli on (2.1, 3.7) cm. Kun valaistuksen määrä kasvaa 8 tunnista 12 tuntiin, sen vaikutus estimoituu negatiiviseksi. Estimaatti ei kuitenkaan ole merkitsevä. Sen sijaan kasvu 8 tunnista ja 16 tuntiin vaikuttaa kasvuun n. 1.1 cm, ja vastaava luottamusväli on (0.1, 2.0) cm.

Mallin simulointi antaa ρ:n ML-estimaatin harhaksi -0.077 ja harhan luottamusväliksi $(-0.080, -0.074)$. Satunnaiskomponentin varianssi estimoituu nollaksi n. 22 %:ssa simuloinneista. Osamäärän ρ prosenttipisteluottamusväli on $(0, 0.53)$, mutta BC_a-väli on huomattavan erilainen $0.02, 1.07$.

Vastaavat tulokset REML-estimoinnista ovat: harhan estimaatti on 0.026, ja sen luottamusväli on $(0.021, 0.032)$. Nollien osuus estimaateissa on 4 %. Luottamusvälit ovat $(0, 1.07)$ (prosenttipiste) ja $(0, 1.14)$ (BC_a). Nyt molemmat luottamusvälit sisältävät nollan. Tästä huolimatta lienee syytä uskoa, että ruukkujen vaikutus on syytä ottaa huomioon laskettaessa regressiokertoimien luottamusvälejä.

On ehkä valaisevaa tarkastella esimerkin aineistoa myös seuraavalla tavalla. Olkoon ruukun i vasteet $y_{i1}, y_{i2}, y_{i3}, y_{i4}$. Olettamusten mukaa ne toteuttavat mallin

$$y_{ij} = \boldsymbol{u}_i'\boldsymbol{\beta} + \eta_i + \varepsilon_{ij},$$

missä vektori $\boldsymbol{u}_i$, $(p+1)\times 1$, kertoo millaisen käsittelyn ruukun i taimet saavat. Ruukkukeskiarvot $\bar{y}_{i\cdot}$ ovat riippumattomia ja noudattavat $N(\boldsymbol{u}_i'\boldsymbol{\beta}, \sigma_\eta^2 + \sigma^2/4)$ (perustele!). Voimme siis estimoida $\boldsymbol{\beta}$:s tavallisella p.n.s.-menetelmällä. Lisäksi saamme jäännösten avulla harhattoman estimaatin varianssille $\sigma_\eta^2 + \sigma^2/4$ ja estimaattien keskivirheet. Osoittautuu, että nämä keskivirheet ovat samat kuin REML-estimoinnilla saadut. Tällainen yhtäpitävyys lienee ollut pontimena REML-estimoinnin kehittelyssä: Tietyissä tasapainoisissa tilanteissa se tuottaa automaattisesti samat tulokset kuin näihin tasapainoisiin tilanteisiin erikseen räätälöidyt "luonnolliset" ratkaisut. Toisaalta epätasapainoisiin tilanteisiin on esitetty monenlaisia ratkaisuja, joiden sijaan REML-estimointi tarjoaa yleisen lähestymistavan. Em. luottamusvälit on laskettu tästä analyysista. Samaan tulokseen johtaa REML-keskivirheet ja kriittiset arvot t-jakaumasta 14 vapausasteella ($t_{0.025;14} = 2.14$).

Luku 8

Binaarinen regressio

8.1 Linkkifunktio ja uskottavuusfunktio

Oletetaan, että vasteet $y_i \sim \text{Bin}(n_i, \pi_i)$, $i = 1, \ldots, k$ missä n_i toistojen lukumäärä ja $0 < \pi_i < 1$ on onnistumisen todennäköisyys. Kuten olemme nähneet kanoniset parametrit ovat

$$\theta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right), \quad i = 1, \ldots, k.$$

Tiedämme myös, että

$$\begin{aligned} E(y_i) &= n_i\pi_i = n_i\frac{e^{\theta_i}}{1+e^{\theta_i}} = n_i m(\theta_i) \\ \text{var}(y_i) &= n_i\pi_i(1-\pi_i) = n_i\frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = n_i v(\theta_i), \quad i = 1, \ldots, k. \end{aligned}$$

Tavallisimmat linkkifunktiot ovat

$$\begin{aligned} \text{logit}(\pi_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}, \quad \text{logit}, \\ \Phi^{-1}(\pi_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}, \quad \text{probit}, \\ \log(-\log(1-\pi_i)) &= \boldsymbol{x}_i'\boldsymbol{\beta}, \quad \text{komplementaarinen log-log}. \end{aligned}$$

Linkkifunktioksi käy periaatteessa mikä tahansa kvantiilifunktio, mutta käytännössä kätevimpiä ovat sellaiset kvantiilifunktiot, jotka kuvaavat välin $(0, 1)$ väliksi $(-\infty, \infty)$. Edellä mainitut linkit ovat tällaisia.

Oletetaan nyt, että linkkifunktio on g ja sen käänteisfunktio g^{-1}. Silloin uskottavuusfunktio on

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{k} \binom{n_i}{y_i} \pi_i^{y_i}(1-\pi_i))^{n_i-y_i} \\ &= \prod_{i=1}^{k} \binom{n_i}{y_i} [g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta})]^{y_i}[1-g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta})]^{n_i-y_i}. \end{aligned}$$

Suurimman uskottavuuden estimaatin likimääräinen (asymptoottinen) jakauma on

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \mathcal{I}(\boldsymbol{\beta})^{-1}\right),$$

missä

$$\mathcal{I}(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}}\right].$$

Harjoitustehtävä 8.1. Johda $\mathcal{I}(\boldsymbol{\beta})$, kun linkkifunktio on kvantiilifunktio F^{-1}, ts. $g^{-1} = F$, missä F on kertymäfunktio.

Harjoitustehtävä 8.2. Sovella ed. tehtävän tulosta probit-malliin. Merkinnät yksinkertaistuvat, kun huomaat, että normaalijakauma tiheys φ toteuttaa differentiaaliyhtälön $\mathrm{d}\varphi(x)/\mathrm{d}x = -x\varphi(x)$, ks. kaava (6.1).

Kun korvataan $\boldsymbol{\beta}$ estimaatillaan $\widehat{\boldsymbol{\beta}}$, saadaan keskivirheet matriisin $\mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1}$ lävistäjäalkioiden neliöjuurina $\sqrt{\hat{c}_{jj}}$. Silloin pätee likimäärin

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{c}_{jj}}} \sim N(0,1),$$

mistä saadaan testit ja luottamusvälit tavalliseen tapaan.

8.2 Devianssi

Edellä kerrottu parametrien estimointi perustuu binomijakaumaa noudattaviin vasteen arvoihin $y_i \sim \mathrm{Bin}(n_i, \pi_i)$, $i = 1, \ldots, k$. Toisaalta regressiokertoimia on $p+1$ kpl. Täytyy siis olla $p+1 \leq k$. Tämä antaa mahdollisuuden verrata logistista regressiomallia, jossa $p+1 < k$, sellaiseen *kyllästettyyn l. saturoituun* malliin, missä todennäköisyydet π_i estimoidaan vapaasti ilman rajoituksia. Saturoidun mallin estimaatit ovat yksinkertaisesti suhteellisia osuuksia $\bar{y}_i = y_i/n_i$ (totea tämä yksinkertaisella laskulla).

Kyllästetyn mallin uskottavuusfunktion maksimiarvo on

$$\hat{L}_{sat} = \prod_{i=1}^{k} \binom{n_i}{y_i} \bar{y}_i^{y_i} (1 - \bar{y}_i)^{n_i - y_i} \tag{8.1}$$

Oletetaan linkkifunktio g, ja merkitään $g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}) = \pi_i(\boldsymbol{\beta})$. Silloin uskottavuusfunktion maksimiarvo on

$$\hat{L} = L(\widehat{\boldsymbol{\beta}}) = \prod_{i=1}^{k} \binom{n_i}{y_i} \pi_i(\widehat{\boldsymbol{\beta}})^{y_i} (1 - \pi_i(\widehat{\boldsymbol{\beta}}))^{n_i - y_i}.$$

Devianssi on näiden uskottavuuksien osamäärän logaritmi kerrottuna kahdella. Kun

merkitään $\hat{\pi}_i = \pi_i(\widehat{\boldsymbol{\beta}})$, devianssi on

$$\begin{aligned} D &= 2\log\left(\frac{\hat{L}_{sat}}{\hat{L}}\right) \\ &= 2\sum_{i=1}^{k}\left[y_i\log\left(\frac{\bar{y}_i}{\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{1-\bar{y}_i}{1-\hat{\pi}_i}\right)\right] \\ &= 2\sum_{i=1}^{k}\left[y_i\log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right] \end{aligned} \tag{8.2}$$

missä $\hat{y}_i = n_i\hat{\pi}_i$ on binaarisen regressiomallin mukainen odotettu frekvenssi. Lisäksi pätee sopimus $0 \cdot \log 0 = 0$.

Devianssi tarjoaa mahdollisuuden yhteensopivuustestiin. Asetetaan hypoteesit

$$\begin{aligned} &H_0 : g(\pi_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \quad i = 1, \ldots, k. \\ &H_A : \text{todennäköisyydet } \pi_i, i = 1, \ldots, k \text{ ovat mielivaltaisia.} \end{aligned}$$

Jos H_0 on tosi, niin D noudattaa likimäärin $\chi^2(k-p-1)$-jakaumaa, kun ryhmäkoot n_i ovat isoja. Approksimaatio saattaa olla huono, jos odotetut frekvenssit $\hat{y}_i$ ovat pieniä esim. alle ykkösen. Näissä tapauksissa voi simuloida sovitettua mallia ja laskea devianssin jakaumaa. ***Mutta jos aineisto on binaarisessa muodossa,*** *(*$n_1 = \cdots = n_k = 1$***), testiä ei voi käyttää ollenkaan em.*** H_0***:n testaamiseen.***

Kun oletetaan, että tietty p:n prediktorin binaarinen regressiomalli on oikea malli, voimme testata kertoimien osajoukkoon liittyvää nollahypoteesia samaan tapaan kuin lineaarisessa regressiomallin puitteissa.

$$\begin{aligned} &H_0 : \beta_{q+1} = \cdots = \beta_p = 0 \\ &H_A : \text{ jokin } \beta_j \neq 0, \quad j = q+1, \ldots, p. \end{aligned}$$

Tämä testi tehdään valitun linkkifunktion puitteissa, so. *olettamalla* että $g(\pi_i) = x_i'\boldsymbol{\beta}$ on oikea malli. Testi konstruoidaan maksimoimalla sekä p:n prediktorin malli että suppeampi q:n prediktorin malli

$$\begin{aligned} \hat{L}_p &= \max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \\ \hat{L}_q &= \max_{\beta_{q+1}=\cdots=\beta_p=0} L(\boldsymbol{\beta}). \end{aligned}$$

Kun H_0 on tosi

$$2\log\left(\frac{\hat{L}_p}{\hat{L}_q}\right) \sim \chi^2(p-q)$$

likimäärin. Tämä testisuure voidaan kirjoittaa yksinkertaisesti kahden devianssin erotuksena

$$\begin{aligned} 2\log\left(\frac{\hat{L}_p}{\hat{L}_q}\right) &= 2\log\left(\frac{\hat{L}_{sat}}{\hat{L}_q}\,\frac{\hat{L}_p}{\hat{L}_{sat}}\right) \\ &= D_q - D_p, \end{aligned}$$

missä $\hat{L}_{sat}$ on saturoidun mallin uskottavuus, ja D_p ja D_q ovat laajemman ja suppeamman mallin devianssit vastaavasti. Tämä testi on käyttökelpoinen myös tilanteessa $n_1 = \cdots = n_k = 1$. Aikaisemmin luvussa 3.3 esitetty testi on vaihtoehtoinen testi. Testien p-arvot poikkeavat jonkin verran toisistaan, mutta useimmiten johtavat samaan johtopäätökseen.

Lineaarisen regressiomallin F-testiä vastaava testi saadaan, kun $q = 0$. Tällä testillä testataan, riippuuko vaste prediktoreista (**R**:n *null deviance*).

8.3 Jäännökset ja diagnostiikka

Vaikka logistisessa regressiossa ei ole teoreettista virhetermiä, voimme silti laskea jäännöksiä (l. poikkeamia) sovitetusta mallista. Pearsonin jäännökset määritellään kaavan

$$r_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

avulla. Nimi tulee siitä, että ns. Pearsonin testisuure X_P^2 on niiden neliösumma

$$X_P^2 = \sum r_i^2 = \sum \left[\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \frac{\big(n_i - y_i - n_i(1-\hat{\pi}_i)\big)^2}{n_i(1-\hat{\pi}_i)} \right]$$

Kun malli oikea, niin $X_P^2 \sim \chi^2(k-p-1)$ likimäärin, kun ryhmäkoot ovat suuria, $n_i \to \infty$. Tämä Pearsonin χ^2-testi on vaihtoehto devianssiin (8.2) perustuvalle testille.

Devianssin (8.2) perusteella voidaan myös määritellä jäännökset seuraavasti

$$d_i \;=\; \text{sign}(y_i - n_i\hat{\pi}_i) \left\{ 2 \left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right] \right\}^{1/2}.$$

Selvästi $\sum d_i^2 = D$. Kun malli on oikea Pearsonin X_P^2 ja devianssi D ovat likipitäen samat. Kun ryhmien koot ovat riittävän suuria jäännökset noudattavat likimäärin $N(0,1)$-jakaumaa.

8.4 Ylihajonta logistisessa regressiossa

Mikäli vapausasteilla jaettu jäännöneliösumma

$$\frac{1}{k-p-1}\sum_{i=1}^{k} r_i^2$$

ylittää huomattavasti ykkösen, on syytä epäillä ylihajonnan mahdollisuutta. Kuten aikaisemmin todettiin tilastollinen testi voidaan perustaa neliösummaan, jota verrataan $\chi^2(k-p-1)$-jakaumaan. Mikä testin arvo ylittää kriittisen arvon $\chi^2_{0.95,k-p-1}$, niin aineistossa merkitsevästi enemmän vaihtelua kuin malli edellyttää. Tilannetta voi yrittää korjata etsimällä lisää prediktoreita, jotka sieppaavat tämän ylihajonnan. Aina

tämä ei ole mahdollista. Tällaisissa tilanteissa mahdollinen ratkaisu, on valita binomijakauman sijasta, jokin muu jakauma. Tavallisin vaihtoehto on beta-binomiaalinen jakauma.

Tarkastellaan seuraavaksi beta-binomiaalista regressiomallia. Teemme oletukset

$$y \mid \pi \sim \text{Bin}(r, \pi), \quad \pi \sim \text{Beta}(\alpha, \delta).$$

Silloin

$$\begin{aligned} P(y = m) &= \int_0^1 \binom{r}{m} \pi^m (1-\pi)^{r-m} \frac{\Gamma(\alpha+\delta)}{\Gamma(\alpha)\Gamma(\delta)} \pi^{\alpha-1}(1-\pi)^{\delta-1} \, \mathrm{d}\pi \\ &= \binom{r}{m} \frac{\Gamma(\alpha+\delta)}{\Gamma(\alpha)\Gamma(\delta)} \frac{\Gamma(\alpha+m)\Gamma(\delta+r-m)}{\Gamma(r+\alpha+\delta)}. \end{aligned}$$

Beta-jakauman odotusarvo ja varianssi ovat vastaavasti

$$\begin{aligned} E(\pi) &= \frac{\alpha}{\alpha+\delta}, \\ \text{var}(\pi) &= \frac{\alpha\delta}{(\alpha+\delta)^2(\alpha+\delta+1)} = \frac{E(\pi)(1-E(\pi))}{\alpha+\delta+1}. \end{aligned}$$

Suoraan laskemalla saamme (tee laskut!), että

$$\begin{aligned} E(y) &= E[E(y \mid \pi)] = r\frac{\alpha}{\alpha+\delta} \\ \text{var}(y) &= E[\text{var}(y \mid \pi)] + \text{var}[E(y \mid \pi)] \\ &= r\frac{\alpha\delta}{(\alpha+\delta)^2}\left(1 + \frac{r-1}{\alpha+\delta+1}\right) \end{aligned}$$

Tämän perusteella voi uudelleen parametroida niin, että

$$\text{logit}\left(\frac{\alpha}{\alpha+\delta}\right) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p = \eta$$

ja $\tau = \alpha + \delta$. Silloin $\alpha = \tau e^\eta/(1+e^\eta)$ ja $\delta = \tau/(1+e^\eta)$. Usein valitaan parametriksi $\phi = 1/(\tau+1)$. (Mikä on ϕ:n vaihteluväli?). Silloin

$$\text{var}(y) = r\frac{\alpha\delta}{(\alpha+\delta)^2}[1 + (r-1)\phi].$$

Uskottavuusfunktio beta-binomiaalisessa regressiossa on

$$L(\boldsymbol{\beta}, \tau) = \text{vakio} \times \prod_{i=1}^{k} \frac{\Gamma\left(\frac{\tau e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} + y_i\right)}{\Gamma\left(\frac{\tau e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right)} \frac{\Gamma\left(\frac{\tau}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} + n_i - y_i\right)}{\Gamma\left(\frac{\tau}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right)} \frac{\Gamma(\tau)}{\Gamma(n_i+\tau)}.$$

Uskottavuusfunktio näyttää monimutkaiselta gamma-funktioineen, mutta käyttämällä toistuvasti kaavaa $\Gamma(t+1) = t\Gamma(t)$ kaikki gammafunktiot häviävät (näytä!). Kun beta-binomiaalisen mallin sovittaa, monet ohjelmat antavat jäännösdevianssin. Se ei

kuitenkaan ole likimäärin χ^2-jakautunut, sillä kyllästetty malli (8.1) ei ole erikoistapaus beta-binomiaalisesta regressiomallista. Sen jakaumaa voi kuitenkin selvittää simuloimalla sovitettua beta-binomiaalista mallia.

Beta-binomiaalinen malli ei ole ainoa ylihajontamalli. Collett (1991, Luku 6) on hyvä katsaus aiheeseen.

Esimerkki 8.1. Esimerkin aineisto on alunperin artikkelista Crowder (1978), (ks. myös Collett, 1991, s. 4 ja 196–199). *Orobanche*-juuriloinen kasvaa kukkivien kasvien juurissa. Eräässä kokeessa sen siementen itävyyttä tutkittiin. Kokeessa oli mukana kaksi lajiketta *Orobanche aegyptiaca* 75 (10 siemenpussia)ja *Orobanche aegyptiaca* 73 (11 siemenpussia). Kunkin pussin siemenet idätettiin erikseen alustalla, jolla oli samanvahvuinen liuos, joka oli uutettu joko pavun tai kurkun juurista. Ao. aineistossa `germ` on itäneiden lukumäärä, `total` on siementen määrä pussissa; `plant` saa arvot `b` ja `c` sen mukaan, miten siemenet on idätetty: Pavun juuresta uutetulla liuoksella idätetyt merkitään `b`:llä ja kurkun juuresta uutetut `c`:llä. Faktori `species` erottelee loislajikkeet.

```
 glm(formula = cbind(germ, total - germ) ~ plant * species,
family  = binomial)

                  coef.est coef.se
 (Intercept)      -0.412    0.184
 plantc            0.540    0.250
 species75        -0.146    0.223
 plantc:species75  0.778    0.306
 ---
   n = 21, k = 4
   residual deviance = 33.3, null deviance = 98.7 (difference = 65.4)
```

Ylihajonta on selvästi merkitsevä. Tässä esimerkissä voimme ajatella, että samanaikaisesti idätettävien siementen välille syntyy positiivista korrelaatiota. Se puolestaan kasvattaa itäneiden varianssia verrattuna siihen tilanteeseen, että jokaisen siemenen itäminen on riippumatonta muiden itämisestä.[1] Toisaalta jo pussittamisvaiheessa siemenet ovat saattaneet valikoitua samasta lähteestä niin, että samaan pussiin joutuneiden siementen itäminen on positiivisesti korreloitunutta.

Sovitetaan beta-binomiaalinen malli

```
> oro.bb <- beta.binomial(oro.glm, gradient=TRUE)
                 coef.est coef.se
(Intercept)       -0.444   0.218
plantc             0.522   0.297
species75         -0.097   0.274
plantc:species75   0.798   0.378

Overdispersion parameter phi  0.013
 Log-likelihood  -541.94
```

[1] Oleta, että binaariset satunnaismuuttujat X_i ovat sellaisia, että $\mathrm{cor}(X_i, X_j) = \rho > 0$, $i \neq j$.

Lasketaan solujen todennäköisyydet logistisella regressiolla ja beta-binomiaalisella regressiolla

```
> pr.glm <- tapply(fitted(oro.glm), list(plant, species), mean)
> pr.bb <- tapply(fitted(oro.bb), list(plant, species), mean)
>
> round(pr.glm, 2)
    73   75
b 0.40 0.36
c 0.53 0.68

> round(pr.bb, 2)
    73   75
b 0.39 0.37
c 0.52 0.69
```

Estimaatit eivät juurikaan poikkea. Teemme vertailut beta-binomiaalisen mallin perusteella. Pavun juuresta uutetussa liuoksessa lajikkeilla on suunnilleen sama itämistodennäköisyys. Ero ei ole merkitsevä, sillä logit-skaalalla ero on -0.097 ja sen keskivirhe 0.274. Näyttäisi kuitenkin siltä, että kurkun juuresta uutetulla liuoksella saadaan parempi itävyys molemille lajikkeille. Kuitenkaan lajikkeen 73 kohdalla ero ei ole merkitsevä: Ero logit skaalalla on 0.522, ja sen keskivirhe on 0.297. Interaktio on selvästi merkitsevä: Itäminen poikkeaa merkitsevästi papu- ja kurkkualustoilla lajikkeen 75 osalta.

Tarkastelemme lopuksi beta-binomiaalisen mallin devianssia. Tavanomainen kaava antaa 30.9. Kun estimoitua mallia simuloidaan saamme, että n. 16 % simuloiduista devianssin arvoista ylittaa aineistosta lasketun. Siis havaittu arvo ei ole ollenkaan poikkeuksellisen suuri. Sen perusteella ei beta-binomiaalista mallia voi hylätä.

Luku 9

Lukumääräinen vaste

Luvussa 4 tutustuimme Poisson-regressioon, missä vasteena on lukumäärä. Tässä luvussa esitetään joitakin siihen liittyviä teoreettisia tuloksia ja sen laajennusta ylihajontatilanteeseen. Poisson-regressiomallin, jossa on log-linkki, kirjoitamme kuten aikaisemminkin muodossa

$$\begin{aligned} y_i &\sim \text{Po}(m_i\theta_i), \quad i = 1, \ldots, n, \\ \log\theta_i &= \boldsymbol{x}_i'\boldsymbol{\beta}, \end{aligned}$$

missä m_i viittaa altistuksen määrään tapauksessa i. Poisson-regressiota olemme myös käsitelleet jo luvun 5 esimerkissa 5.7 sekä harjoitustehtävissä 5.7, 5.8, 5.9 ja 5.11, jotka on syytä palauttaa mieleen tässä.

9.1 Devianssi

Kuten binaarisen vasteen tapauksessa, myös Poisson-regressiossa voidaan määritellä kyllästetty malli, missä intensiteetit θ_i saavat arvonsa riippumatta selittäjien arvoista. On helppo osoittaa, että θ_i:n suurimman uskottavuuden estimaatti on yksinkertaisesti $\bar{y}_i = y_i/m_i$, $i = 1, \ldots, n$. Kyllästetyn mallin ja Poisson-regressiomallin uskottavuudet ovat vastaavasti

$$\begin{aligned} \hat{L}_{sat} &= \prod_{i=1}^{n} \frac{(m_i\bar{y}_i)^{\,y_i}}{y_i!}\, e^{-m_i\bar{y}_i} \\ \hat{L} &= \prod_{i=1}^{n} \frac{(m_i e^{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}})^{y_i}}{y_i!}\, \exp\left(-m_i e^{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}}\right). \end{aligned}$$

Merkitään nyt $\hat{\theta}_i = e^{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}}$. Devianssi määritellään samoin kuin logistisessa regressiossa

$$\begin{aligned} D &= 2\log\left(\frac{\hat{L}_{sat}}{\hat{L}}\right) \\ &= 2\sum\left[y_i\log\frac{\bar{y}_i}{\hat{\theta}_i} - m_i(\bar{y}_i - \hat{\theta}_i)\right] \\ &= 2\sum y_i\log\frac{y_i}{m_i\hat{\theta}_i}. \end{aligned}$$

Viimeinen yhtäsuuruus pätee vain jos mallissa on vakio. Silloin uskottavuusyhtälöistä ensimmäinen antaa (ks. harjoitustehtävä 5.8) $\sum(y_i - m_i e^{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}}) = \sum(m_i\bar{y}_i - m_i\hat{\theta}_i) = 0$. Devianssi noudattaa likimäärin $\chi^2(n-p-1)$-jakaumaa. Approksimaatio saattaa olla huono, jos odotetut frekvenssit $\hat{\lambda}_i = m_i\hat{\theta}_i = m_i e^{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}}$, ovat pieniä. Devianssia käytetään yhteensopivuustestinä hypoteeseille

$$\begin{aligned} H_0: &\quad \log\lambda_i = \log m_i + \boldsymbol{x}_i'\boldsymbol{\beta} \quad i = 1,\ldots,n \\ H_A: &\quad \text{odotusarvot } \lambda_i \text{ ovat mielivaltaisia.} \end{aligned}$$

Kertoimien osajoukkoon liittyviä nollahypoteeseja

$$\begin{aligned} &H_0: \beta_{q+1} = \cdots = \beta_p = 0 \\ &H_A: \text{ jokin } \beta_j \neq 0, \quad j = q+1,\ldots,p. \end{aligned}$$

testataan devianssien avulla kuten logistisessa regressiossa. Jos D_q on rajoitetun mallin devianssi ja D_p kertoimia rajoittamattoman mallin devianssi, niin $D_q - D_p \sim \chi^2(p-q)$ likimäärin kun H_0 on tosi. Tämä testi tehdään olettamalla p selittäjän regressiomalli oikeaksi.

9.2 Jäännökset

Pearsonin jäännökset ovat määritelmän mukaan

$$r_i \;=\; \frac{y_i - m_i\hat{\theta}_i}{\sqrt{m_i\hat{\theta}_i}},$$

missä $\hat{\theta}_i = e^{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}}$. Nimi tulee jälleen siitä, että Pearsonin χ^2-statistiikka on

$$X_P^2 = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n \frac{(y_i - m_i\hat{\theta}_i)^2}{m_i\hat{\theta}_i}.$$

Jos malli on oikea, $X_P^2 \sim \chi^2(n-p-1)$ ja on numeerisesti lähellä devianssia D. Kuten logistisessa regressiossa voidaan määritellä myös devianssijäännös

$$d_i = \text{sign}(y_i - m_i\hat{\theta}_i)\left\{2\left[y_i\log\frac{y_i}{m_i\hat{\theta}_i} - (y_i - m_i\hat{\theta}_i)\right]\right\}^{\frac{1}{2}}.$$

Diagnostiset tarkastelut tehdään samaan tapaan kuin lineaarisessa regressiossa.

9.3 Ylihajonta Poisson-regressiossa

Samoin kuin aikaisemmin binomijakaman kohdalla, Poisson-jakauman varianssi on odotusarvon funktio. Tämä johtaa aika ajoin tilanteeseen, missä vasteen varianssi on

suurempi kuin mitä Poisson-malli antaa odottaa. Asian havaitsee jäännöstarkasteluista. Jäännösten odotusarvo on 0, ja varianssi on likimäärin 1. Vapausasteilla korjatun neliösumman

$$\frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2$$

pitäisi olla likipitäen 1, jos malli on oikea. Kuten olemme nähneet voidaan tilastollinen testi perustaa jäännösten neliösummaan, jota verrataan $\chi^2(n-p-1)$-jakaumaan. Mikä testin arvo ylittää kriittisen arvon $\chi^2_{0.95,n-p-1}$, niin aineistossa merkitsevästi enemmän vaihtelua kuin malli edellyttää. Yksi parannuskeino on jälleen lisäselittäjien löytäminen. Toinen keino on samantapainen kuin se, mikä johtaa logistisesta regressiosta beta-binomiaaliseen regressioon. Poisson-regressiosta voidaan siirtyä negatiiviseen binomijakaumaan.

Oletamme, että

$$\begin{aligned} y \mid \lambda &\sim \text{Po}(m\lambda), \\ \lambda &\sim \text{Gamma}(\nu, \psi), \end{aligned}$$

missä m on altistuksen määrä. Silloin

$$\begin{aligned} P(y=k) &= \int_0^\infty \frac{m^k\lambda^k}{k!} e^{-m\lambda} \frac{1}{\Gamma(\nu)\psi^\nu} \lambda^{\nu-1} e^{-\lambda/\psi}\,\mathrm{d}\lambda \\ &= \frac{\Gamma(k+\nu)m^k}{\Gamma(\nu)\psi^\nu k!}\left(m+\frac{1}{\psi}\right)^{-k-\nu} \\ &= \frac{\Gamma(k+\nu)}{\Gamma(\nu)k!}\left(\frac{1}{1+m\psi}\right)^\nu \left(\frac{m\psi}{1+m\psi}\right)^k. \end{aligned}$$

Saatu todennäköisyys on negatiivinen binomitodennäköisyys, ks. harjoitustehtävä 5.2, so. $y \sim \text{NB}(\nu, \pi)$, missä $\pi = 1/(1+m\psi)$.

Iteratiivisen odotusarvon ja varianssin kaavasta saamme

$$\begin{aligned} E(y) &= E[E(y \mid \lambda)] = mE(\lambda) = m\nu\psi \\ \text{var}(y) &= E[\text{var}(y \mid \lambda)] + \text{var}[E(y \mid \lambda)] \\ &= mE(\lambda) + \text{var}(m\lambda) \\ &= m\nu\psi + m^2\nu\psi^2 = m\nu\psi(1+m\psi) \end{aligned}$$

Voimme asettaa log-linkin odotusarvon parametriosalle $\log(\nu\psi) = \boldsymbol{x}'\boldsymbol{\beta}$, ja olettaa, että ψ on vakio. Tämä malli on lähellä alkuperäista Poisson-regressiota: $E(y) = me^{\boldsymbol{x}'\boldsymbol{\beta}}$ ja $\text{var}(y) = (1+m\psi)me^{\boldsymbol{x}'\boldsymbol{\beta}}$. Kun altistus on vakio, jolloin yleensä asetetaan $m = 1$, niin varianssi on verrannollinen odotusarvoon. Poisson-regressiossa verrannollisuuskerroin on 1. Toinen mahdollisuus on asettaa ν vakioksi, jolloin $\text{var}(y) = me^{\boldsymbol{x}'\boldsymbol{\beta}}(1+m\nu^{-1}e^{\boldsymbol{x}'\boldsymbol{\beta}})$. Tässä mallissa varianssi kasvaa odotusarvon neliössä.

Harjoitustehtävä 9.1. Jos negatiivisessa binomimallissa $\text{NB}(\nu, \pi)$ oletetaan $\text{logit}(\pi) = \boldsymbol{x}'\boldsymbol{\beta}$, millainen regressiomalli syntyy verrattuna kahteen jo esitettyyn malliin?

Esimerkki 9.1. Käsittelemme esimerkkinä aineistoa, joka on peräisin artikkelista Bissell (1972). Vasteena on virheiden lukumäärä kangasrullissa, joiden pituudet vaihtelevat. tehtävänä on arvoida, paljonko virheitä on keskimäärin pituusyksikköä kohti. Pituus on annettu metreissä. Valitsemme yksiköksi virheiden lukumäärän 100 metriä kohti. Poisson-sovitus antaa tulokset:

```
glm(formula = faults ~ 1, family = poisson, offset = log(length/100),
   x = TRUE, y = TRUE)
            coef.est coef.se
(Intercept) 0.412    0.059
---
 n = 32, k = 1
 residual deviance = 64.5, null deviance = 64.5 (difference = 0.0)
```

Virheitä on siis keskimäärin $e^{0.412} = 1.51$ kpl 100 metriä kohti. Luottamusväliksi (95 %) tulee $(1.34, 1.69)$. Huomaamme, että residuaalidevianssi on suuri 64.5 verrattuna vapausasteisiin = 31. Sovitamme seuraavaksi negatiivisen binomimallin. Koska mallissa on vain vakio, molemmat mallityypit tuottavat saman kertoimen:

```
            coef.est coef.se
(Intercept)    0.412   0.086

Overdispersion parameter (psi)  0.174
 Log-likelihood  -160.731
```

Kertoimen estimaatti on sama (3 desimaalilla), mutta keskivirhe on lähes puolitoistakertainen. Luottamusväli on $(1.28, 1.79)$. Se on jonkin verran pidempi kuin Poisson-regression antama.

9.4 Kontingenssitaulut ja todennäköisyysmallit

Kontingenssitaulujen ns. log-lineaarinen analyysi voidaan käsitellä Poisson-regression erikoistapauksena. Se perustuu Poisson-jakauman ja multinomijakauman väliseen yhteyteen. Oletetaan, että satunnaismuuttujat $y_i \sim \text{Po}(\lambda_i)$, $i = 1, \ldots, n$, ovat riippumattomia. Valitaan mielivaltaiset kokonaisluvut $k_i \geq 0$, $\sum k_i = N$. Lasketaan ehdollinen todennäköisyys

$$\begin{aligned} P\left(y_i = k_i,\ i = 1, \ldots, n \,\middle|\, \sum y_i = N\right) &= \frac{P\left(y_1 = k_1, \ldots, y_n = k_n, \sum y_i = N\right)}{P\left(\sum y_i = N\right)} \\ &= \frac{\prod \frac{\lambda_i^{k_i}}{k_i!} e^{-\lambda_i}}{\frac{(\sum \lambda_i)^N}{N!} e^{-\sum \lambda_i}} \\ &= \frac{N!}{k_1! \cdots k_n!} \left(\frac{\lambda_1}{\sum \lambda_i}\right)^{k_1} \cdots \left(\frac{\lambda_n}{\sum \lambda_i}\right)^{k_n}, \end{aligned}$$

missä toinen yhtäsuuruus seuraa siitä, että $\sum y_i \sim \text{Po}(\sum \lambda_i)$. Merkitään $\pi_i = \lambda_i / \sum \lambda_j$, $i = 1, \ldots, n$. Tästä tuloksesta seuraa, että $(y_1, \ldots, y_n)$ noudattaa ehdolla $\sum y_i = N$

multinomijakauma $\mathrm{MN}(N; \pi_1, \ldots, \pi_n)$. Tulos voidaan helposti yleistää tilanteeseen, missä $y_{ij} \sim \mathrm{Po}(\lambda_{ij})$, $i = 1, \ldots, n$ ja $j = 1, \ldots, r$. Kun ehdollistetaan summiin $\sum_j y_{ij} = N_i$ saadaan riippumattomia multinomiaalisia vektoreita $(y_{i1}, \ldots, y_{ir}) \sim \mathrm{MN}(N_i; \pi_{i1}, \ldots, \pi_{ir})$, $\pi_{i1} = \lambda_{i1} / \sum_j \lambda_{ij}, \ldots, \pi_{ir} = \lambda_{ir} / \sum_j \lambda_{ij}$. Jos frekvenssit muodostavat kolmi- tai useampiulotteisia tauluja, voimme tarkastella monimutkaisempia ehdollisia jakaumia kuten seuraavasta esimerkistä ilmenee.

Esimerkki 9.2. Tarkastellaan taulua, jossa frekvenssit on luokiteltu kolmen dikotomisen muuttujan suhteen (Bishop, 1969).

Klinikka	Hoito määrä ennen syntymää	Eloonjääneet	Kuolleet	Summa
A	Vähän	176	3	179
	Paljon	293	4	297
B	Vähän	197	17	214
	Paljon	23	2	25
		689	26	715

Merkitsemme y_{ijr}:llä taulun frekvenssejä,. Indeksi i erottelee henkiin jääneet ja kuolleet $i = 1, 2$ vastaavasti. Indeksi j erottelee hoidon määrän: $j = 1$ tarkoittaa vähän hoitoa ja $j = 2$ paljon hoitoa saaneita. Klinikka A ja B liittyvät kolmanteen indeksiin: $r = 1$ tarkoittaa klinikkaa A ja $r = 2$ klinikkaa B. Voimme ajatella, että taulu on syntynyt jollakin seuraavista otantamenetelmistä.

(i) Otokseen on poimittu kaikki tietyn kuukauden aikana syntyneet. Silloin mikään taulun luvuista tai reunasummista ei ole kiinnitetty. Oletamme, että taulun frekvenssit ovat riippumattomia Poisson-muuttujia, $y_{ijr} \sim \mathrm{Po}(\lambda_{ijr})$.

(ii) Päätettiin, että aineisto koostuu 715 ensiksi syntyneestä. Silloin summa 715 on kiinnitetty, ja taulun frekvenssit ovat multinomiaalisia todennäköisyyksin $\pi_{ijr} = \lambda_{ijr}/\lambda_{...}$, $\lambda_{...} = \sum_{ijr} \lambda_{ijr}$.

(iii) Päätettiin poimia 476 ensiksi syntynyttä klinikalta A ja 239 klinikalta B. Silloin reunasummat 476 ja 239 ovat kiinnitetyt, ja klinikan A frekvenssit y_{ij1} ovat multinomiaalisia todennäköisyyksin $\pi_{ij|1} = \lambda_{ij1}/\lambda_{..1}$, $\lambda_{..1} = \sum_{ij} \lambda_{ij1}$. Vastaavasti klinikan B frekvenssit ovat multinomiaalisia todennäköisyyksin $\pi_{ij|2} = \lambda_{ij2}/\lambda_{..2}$, $\lambda_{..2} = \sum_{ij} \lambda_{ij2}$. Klinikoiden frekvenssivektorit ovat toisistaan riippumattomia.

(iv) Päätettiin poimia

179 vähän hoitoa saanutta klinikasta A,
297 paljon hoitoa saanutta klinikasta A,
214 vähän hoitoa saanutta klinikasta B,
25 paljon hoitoa saanutta klinikasta B,

so. klinikka × hoito marginaalit ovat kiinnitetyt. Silloin eloonjääneet ovat riippumattomia binomiaalisia satunnaismuuttujia. Esim. klinikalla A vähän hoitoa

saaneista eloonjääneet noudattavat Bin($\pi_{1|11}, 179$)-jakaumaa, $\pi_{1|11} = \lambda_{111}/\lambda_{\cdot 11}$, $\lambda_{\cdot 11} = \lambda_{111} + \lambda_{211}$. Samalla klinikalla paljon hoitoa saaneista eloonjääneet noudattavat Bin($\pi_{1|21}, 297$)-jakaumaa, $\pi_{1|21} = \lambda_{121}/\lambda_{\cdot 21}$, $\lambda_{\cdot 21} = \lambda_{121} + \lambda_{221}$ jne. Kuolleet ovat myös binomiaalisia todenäköisyyksin $\pi_{2|11} = 1 - \pi_{1|11}$, $\pi_{2|12} = 1 - \pi_{1|12}$ jne.

9.5 Kaksiulotteinen taulu

Tarkastellaan $I \times J$ taulua, jonka frekvenssit ovat y_{ij}, $i = 1, \ldots, I$ ja $j = 1, \ldots, J$. Poisson-mallissa oletetaan, että frekvenssit ovat riippumattomia $y_{ij} \sim \text{Po}(\lambda_{ij})$. Selittäjiä ovat rivi- ja sarakemuuttujat. Kirjoitamme

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}.$$

Terminologia on sama kuin kaksisuuntaisessa varianssianalyysissä: α_i on 1. faktorin tason i päävaikuutus, β_j on toisen faktorin tason j pääävaikutus ja $(\alpha\beta)_{ij}$ on niiden interaktio. Oletamme kuten ennenkin, että $\alpha_I = \beta_J = 0$ ja $(\alpha\beta)_{Ij} = (\alpha\beta)_{iJ} = 0$, $i = 1, \ldots, I$ ja $j = 1, \ldots, J$.

Kyllästetyssä Poisson-mallissa suurimman uskottavuuden estimaatit parametreille λ_{ij} ovat yksinkertaisesti havaitut frekvenssit $\hat{\lambda}_{ij} = y_{ij}$. Toisaalta on tapana ajatella että kontingenssistaulun kokonaisfrekvenssi $N = \sum_{ij} y_{ij}$ on kiinteä (jos se ei alunperin ole, ehdollistamme havaittuun arvoon). Silloin syntyy multinomijakauma, jonka solutodennäköisyydet ovat, kuten aikaisemmin nähtiin, $\pi_{ij} = \lambda_{ij} / \sum_{ij} \lambda_{ij}$. Helposti näemme, että multinomimallissa solutodennäköisyyksien suurimman uskottavuuden estimaatit ovat $\hat{\pi}_{ij} = y_{ij}/N$. Poisson-mallista saamme saman arvon, sillä $\hat{\lambda}_{ij} / \sum_{ij} \hat{\lambda}_{ij} = y_{ij}/N = \hat{\pi}_{ij}$.

Kiinnostava nollahypoteesi on rivi- ja sarakemuuttujien riippumattomuus. Mallimme parametrien avulla tämä nollahypoteesi tarkoittaa, että kaikki interaktiot häviävät, ts. jokainen $(\alpha\beta)_{ij} = 0$. Poisson-mallin uskottavuusyhtälöt voidaan nyt kirjoittaa muotoon

$$\begin{aligned}
\sum_{ij} y_{ij} &= y_{\cdot\cdot} = e^{\mu} \left(\sum e^{\alpha_i} \right) \left(\sum e^{\beta_j} \right), \\
\sum_{i} y_{ij} &= y_{\cdot j} = e^{\mu} \left(\sum e^{\alpha_i} \right) e^{\beta_j}, \quad j = 1, \ldots, J-1, \\
\sum_{j} y_{ij} &= y_{i\cdot} = e^{\mu} e^{\alpha_i} \left(\sum_{j} e^{\beta_j} \right), \quad i = 1, \ldots, I-1.
\end{aligned}$$

Laskujen jälkeen saamme tulokseksi, että odotusarvon $E(y_{ij}) = e^{\mu + \alpha_i + \beta_j}$ estimaatti (ns. odotettu frekvenssi) on

$$e^{\widetilde{\mu} + \widetilde{\alpha}_i + \widetilde{\beta}_j} = \frac{y_{i\cdot} y_{\cdot j}}{y_{\cdot\cdot}}.$$

Multinomimallissa riippumattomuushypoteesi tarkoittaa, että solutodennäköisyydet π_{ij} voidaan kirjoittaa reunatodennäköisyyksien $\pi_{i\cdot} = \sum_j \pi_{ij}$ ja $\pi_{\cdot j} = \sum_i \pi_{ij}$ tuloksi. Näiden todennäköisyyksien suurimman uskottavuuden estimaatit multinomimallissa ovat $\widetilde{\pi}_{i\cdot} = y_{i\cdot}/N$ ja $\widetilde{\pi}_{\cdot j} = y_{\cdot j}/N$. Multinomimallin odotetut frekvenssit ovat siis

$N\widetilde{\pi}_{i.}\widetilde{\pi}_{.j} = y_{i.}y_{.j}/y..$, joiden huomaamme olevan samat kuin Poisson-mallista saadut. Voimme tehdä johtopäätöksen, että rivi- ja sarakemuuttujien riippumattomuutta voidaan testata Poisson-regression keinoin. Samantapainen päättely osoittaa, että vaikka kokonaisfrekvenssin lisäksi rivi- tai sarakemuuttujaa vastaavat reunafrekvenssitkin olisivat kiinteät, Poisson-regressio on pätevä riippumattomuushypoteesin testauksessa.

9.6 Kolmiulotteinen taulu

Tarkastellaan nyt kolmiulotteista taulua, jonka frekvenssit ovat y_{ijr}, $i = 1, \ldots, I$, $j = 1, \ldots, J$ ja $r = 1, \ldots, R$. Oletetaan, että $y_{ijr} \sim \text{Po}(\lambda_{ijr})$. Kyllästetyn mallin voimme kirjoittaa

$$\log \lambda_{ijr} = \mu + \alpha_i + \beta_j + \gamma_r + (\alpha\beta)_{ij} + (\alpha\gamma)_{ir} + (\beta\gamma)_{jr} + (\alpha\beta\gamma)_{ijr}.$$

Oletamme, että $\alpha_I = \beta_J = \gamma_R = 0$. Lisäksi kaikki sellaiset interaktiotermit häviävät, joissa $i = I$ tai $j = J$ tai $r = R$.

Jos kaikki kolmen faktorin interaktiot $(\alpha\beta\gamma)_{ijr}$ ovat nollia, so.

$$\log \lambda_{ijr} = \mu + \alpha_i + \beta_j + \gamma_r + (\alpha\beta)_{ij} + (\alpha\gamma)_{ir} + (\beta\gamma)_{jr}$$

näemme, että kun yhden faktorin taso kiinnitetään, esim. 3. faktori tasolle r, niin kahden ensimmäisen faktorin riippuvuuteen vaikuttaa vain interaktio $(\alpha\beta)_{ij}$. Siis kahden ensimmäisen faktorin riippuvuus on samanlainen jokaisella kolmannen faktorin tasolla. Vastaava päättely voidaan tehdä kiinnittämällä vuorotellen myös 1. faktori ja 2. faktori.

Jos kaikki kolmen faktorin interaktiot ovat nollia ja niiden lisäksi myös kaikki interaktiot $(\alpha\beta)_{ij} = 0$, niin kaksi ensimmäistä faktoria ovat riippumattomia jokaisella kolmannen faktorin tasolla. Voimme myös sanoa, että kahden ensimmäisen faktorin riippuvuus johtuu niiden riippuvuudesta kolmannesta faktorista. Kysymyksessä on ns. ehdollinen riippumattomuus.

Jos kontingenssitaulussa osa marginaaleista on kiinteitä, niin taulun voi analysoida Poisson-regression avulla, kunhan mallissa on mukana kiinteitä marginaaleja vastaavat parametrit. Seuraava esimerkki kolmiulotteisesta taulusta kertoo mitä tällä tarkoitetaan.

Huomautus 9.1. Jos 1. faktoria vastaavat marginaalit $\sum_{jr} y_{ijr} = N_{i..}$ ovat kiinteitä, mallissa pitää olla päävaikutukset α_i, $i = 1, \ldots, I$. Jos taas 1. ja 2. faktoria vastaavat marginaalit $\sum_r y_{ijr} = N_{ij.}$ ovat kiinteitä, malliin on otettava mukaan päävaikukset α_i ja β_j sekä interaktiot $(\alpha\beta)_{ij}$, $i = 1, \ldots, I$ ja $j = 1, \ldots, J$.

Kirjallisuutta

Andersen, E. B. (1994). *The Statistical Analysis of Categorical Data*. Springer, New York, Second edition.

Bishop, Y. M. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics*, 25:383–399.

Bissell, A. F. (1972). A negative binomial model with varying element sizes. *Biometrika*, 59:435–441.

Collett, D. (1991). *Modelling Binary Data*. Chapman & Hall, London.

Crowder, M. J. (1978). Beta-binomial anova for proportions. *Applied Statistics*, 27:34–37.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, Second edition.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, Second edition.

Freund, R. J., Wilson, W. J., and Sa, P. (2006). *Regression Analysis: Statistical modeling of a Response Variable*. Academic Press, Burlington, Second edition.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Krzanowski, W. J. (1998). *An Introduction to Statistical Modelling*. Arnold, London.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Weisberg, S. (2005). *Applied Linear Regression*. Wiley-Interscience, Hoboken, Third edition.