Tampereen yliopisto

Matleena Latvala

ENNUSTETARKKUUDEN MITTAAMINEN LINEAARISELLE REGRESSIOMALLILLE

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Huhtikuu 2023

Tiivistelmä

Matleena Latvala: Ennustetarkkuuden mittaaminen lineaariselle regressiomallille
Kandidaattitutkielma
Tampereen yliopisto
Matematiikka ja tilastollinen data-analyysi
Huhtikuu 2023

Tämän tutkielman tarkoituksena on esitellä, miten lineaarisen regressiomallin ennustetarkkuutta voidaan mitata. Tutkielmassa esitellään ensin regressioanalyysin perusteet, kuten yhden ja usean selittäjän lineaariset regressiomallit, jäännöstermit sekä luottamus- ja ennustevälit. Lisäksi esitellään, miten muodostaa mahdollisimman tarkka malli hyvän ennustetarkkuuden kannalta. Koska ennustetarkkuuden mittaaminen perustuu mallin sopivuuteen, esitellään myös kaksi tapaa millä tutkia mallin sopivuutta aineistoon.

Seuraavaksi tutkielmassa käsitellään lineaarisella regressiomallilla ennustamista sekä ennustetarkkuuden mittaamista. Jotta lineaarisen regressiomallin ennustetarkkuuden mittaaminen olisi mahdollisimman todenmukaista, tulee käytettävä aineisto jakaa kahteen osaan. Toisella osalla estimoidaan malli ja toisella testataan mallin ennustetarkkuutta. Ennustetarkkuuden mittareita on useita. Ne perustuvat ennustettujen sekä todellisten arvojen välisiin erotuksiin. Ennustetarkkuuden mittareita käytetään yleensä yhdessä, sillä yksinään mikään mittari ei kuvaa täysin ennustetarkkuutta. Yleisimmät ennustetarkkuuden mitat ovat keskimääräinen absoluuttinen virhe MAE sekä keskimääräinen neliövirhe MSE. Tutkielmassa esitellään lisäksi myös prosenttivirheet.

Viimeisenä tutkielmassa käytetään aiemmin esiteltyjä menetelmiä eturauhastutkimusaineistoon. Aineisto sisältää eturauhassyöpäpotilailta kerättyjä tietoja, kuten esimerkiksi potilaan pituus, paino, ikä sekä eturauhasen koko. Lisäksi aineistossa on paljon erilaisia veriarvoja. Kohdemuuttuja, jota tässä tutkimuksessa pyritään ennustamaan, on potilaan PSA-arvo, joka kuvaa eturauhassyövän mahdollisuutta. Aineisto jaetaan kahteen osaan. Toisella osalla muodostetaan viisi lineaarista regressiomallia, joissa käytetään selittävinä muuttujina kolmea tilastollisesti merkitsevää muuttujaa. Nämä muuttujat ovat potilaan BMI-arvo, eturauhasen pituus sekä alkalinen fosfataa-

si. Toisella osalla testataan mallien ennustetarkkuutta käyttämällä aiemmin esiteltyjä ennustetarkkuuden mittareita. Tuloksista huomataan, että millään mallilla ei ole erityisen hyvä ennustetarkkuus, eli PSA-arvoa ei pysty hyvin ennustamaan aineiston muuttujien avulla. Tuloksista saadaan myös selville, että muodostetuista viidestä mallista kaikista yksinkertaisimmilla malleilla on paras ennustetarkkuus.

Avainsanat: regressioanalyysi, ennustetarkkuus, ennustaminen

Sisällys

1 Johdanto

Tilastotieteessä voidaan ennustaa kahdella eri tavalla, aikasarja-analyysillä tai kausaalisilla malleilla. Aikasarja-analyysi perustuu tietyn ajanjakson aikana kerättyjen havaintojen mallien analysointiin, kun taas kausaaliset mallit perustuvat muuttujien välisten suhteiden analysointiin (Sanders 2015). Yleisimmin käytetty tilastollinen metodi, jolla analysoidaan muuttujien välisiä suhteita, on regressioanalyysi (Kim 2022). Regressioanalyysillä ennustamisessa on tavoitteena muodostaa selittävistä muuttujista funktio, jolla saadaan ilmaistua kohdemuuttuja.

Ennustamisessa tärkeää on ennustetarkkuuden mittaaminen, jolla voidaan arvioida mallilla ennustamisen tehokkuutta. Ennustetarkkuuden mittaamiseen käytetään useita eri mittareita. Tässä tutkimuksessa näytetään, miten ennustaa regressioanalyysin avulla ja kuinka lineaarisen regressiomallin ennustetarkkuutta voidaan mitata.

Ensin esitellään regressioanalyysin perusteet, kuten yhden ja useamman selittäjän lineaaristen regressiomallien muodot, jäännökset ja mallin estimoiminen. Näytetään, kuinka regressioanalyysillä muodostetaan mahdollisimman tarkka ennustemalli ja mitä oletuksia yhtälön muuttujista täytyy tällöin tehdä. Lisäksi esitellään, miten tutkia kuinka hyvin sovitettu malli sopii aineistoon ja miten luottamus- ja ennustevälit muodostetaan.

Seuraavaksi perehdytään paremmin lineaarisella regressiomallilla ennustamiseen sekä mallin ennustetarkkuuden mittaamiseen. Ensin esitellään, kuinka aineistoa tulisi käyttää, jotta uusien tietojen ennusteiden tarkkuus olisi mahdollisimman hyvä. Seuraavaksi näytetään, miten muodostetaan ennustetarkkuuden mittoja, kuten esimerkiksi keskimääräinen absoluuttinen virhe ja keskimääräinen neliövirhe.

Viimeisenä käytetään menetelmiä, joita on aiemmin tutkimuksessa esitelty, oikeaan aineistoon. Aineistona tutkimuksessa on eturauhastutkimusaineisto, jossa on eturauhassyöpäpotilailta kerättyjä henkilökohtaisia tietoja. Jaetaan aineisto kahteen osaan, joista toisella muodostetaan useampi malli eturauhassyövän mahdollisuuden ennustamiseen, ja toisella tutkitaan, kuinka hyvä ennustetarkkuus malleilla on. Lopuksi tulkitaan tuloksia.

2 Regressioanalyysi

Regressioanalyysi on ehkä yleisimmin käytetty tilastollinen metodi, jolla tutkitaan muuttujien välisiä suhteita. Sen on kehittänyt 1800-luvulla Sir Francis Dalton, joka tutki vanhempien ja heidän lastensa pituuden välistä suhdetta. Hän huomasi, että lasten pituudet lähestyivät väestön keskiarvoa. (Kim 2022) Regressioanalyysiä käytetään tyypillisesti mallin parametrien estimoimiseen, hypoteesitestaukseen sekä ennustamiseen. Tavoitteena on mallintaa kohdemuuttuja selittävistä muuttujista muodostetun lineaarisen mallin avulla.

2.1 Yhden selittäjän lineaarinen regressiomalli

Regressioanalyysin perustana toimii yhden selittäjän lineaarinen regressiomalli, jossa on yksi selittävä muuttuja X sekä kohdemuuttuja Y, jonka vaihtelua tullaan selittämään (Kim 2022). Yleinen muoto yhden selittäjän regressiomallille on

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

missä $\epsilon_i \sim N(0, \sigma^2)$. Mallissa on kohdemuuttujan ja selittävän muuttujan lisäksi regressiokertoimet β_0, β_1 sekä jäännöstermi ϵ_i, joka oletetaan riippumattomaksi. Regressiokerroin β_0 on vakiotermi, joka kertoo mikä on kohdemuuttujan Y_i odotusarvo, kun selittävä muuttuja X_i on 0. Selittäjän X_i regressiokerroin β_1 kertoo, miten paljon kohdemuuttuja Y_i muuttuu yhdellä yksiköllä selittävän muuttujan X_i kasvaessa yhdellä yksiköllä.

2.2 Usean selittäjän lineaarinen regressiomalli

Kun mallissa on useampi selittävä muuttuja, kutsutaan mallia usean selittäjän lineaariseksi regressiomalliksi. Mallissa on tällöin useampi selittävä muuttuja X. Yleinen muoto monen selittäjän lineaariselle regressiomallille on

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i,$$

missä $i = i, ..., n$, $n > p$ ja $\epsilon_i \sim N(0, \sigma^2)$. Mallissa on kohdemuuttuja Y_i, selittävät muuttujat $X_{1i}, .., X_{pi}$ ja jäännöstermit ϵ_i, jotka oletetaan riippumattomiksi. Kertoimet $\beta_0, ..., \beta_p$ mittaavat kunkin selittävän muuttujan vaikutusta kun on otettu huomioon

mallin kaikkien muiden selittävien muuttujien vaikutukset. Näin ollen kertoimet mittaavat selittävien muuttujien marginaalivaikutuksia (Hyndman & Athanasopoulos 2018). Malli voidaan esittää myös matriisimuodossa

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

missä $\epsilon \sim MVN(\mathbf{0}, \sigma^2\mathbf{I})$. Matriisimuotoisessa mallissa $\mathbf{X}$ on suunnittelu- tai mallimatriisi, β on parametreistä koostuva vektori ja ϵ on vektori, joka koostuu virhetermeistä. Virhetermit oletetaan riippumattomiksi.

2.3 Jäännöstermit

Havaitun arvon ja siihen liittyvän ennustetun arvon välistä eroa kutsutaan jäännöstermiksi (Draper ja Smith 1998). Jäännöstermit määritellään siis

$$e_i = Y_i - \hat{Y}_i,$$

missä $i = 1, 2, ..., n$, Y_i on havaittu arvo ja $\hat{Y}_i$ on vastaava sovitettu arvo, joka saadaan käyttämällä sovitettua regressiomallia. Jäännökset kuvaavat siis sitä määrää, jota regressioyhtälö ei ole pystynyt selittämään. Regressioanalyysia suorittaessa olemme tehneet jäännöksistä tiettyjä oletuksia: ne ovat riippumattomia, niiden odotusarvo on 0, ja ne noudattavat normaalijakaumaa vakiovarianssilla σ^2.

Matriisimuodossa jäännöstermien vektori määritellään

$$\epsilon = \mathbf{Y} - \mathbf{X}\beta,$$

missä $\epsilon \sim MVN\left(0,\sigma^2 I\right)$.

2.4 Pienimmän neliösumman menetelmä

Regressiomallin parametrien $\beta_1, \beta_2, ..., \beta_p$ arvojen estimointiin käytetään usein pienimmän neliösumman menetelmää. Menetelmä perustuu siihen, että neliöidään jäännöstermit ja lasketaan ne yhteen. (Kim 2022) Valitaan siis sellaiset arvot $\beta_1, \beta_2, ..., \beta_p$, että minimoidaan

$$\sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} \left(y_t - \beta_0 - \beta_1 x_{1,t} - \beta_{2,t} - ... - \beta_p x_{p,t}\right)^2.$$

Parhaiden kertoimien estimaattien löytämistä kutsutaan usein mallin sovittamiseksi dataan. (Hyndman & Athanasopoulos 2018)

2.5 Mallin muodostus

Regressiomallin muodostuksessa tavoitteena on muodostaa sellainen malli, joka selittää parhaiten muiden muuttujien avulla mallinnettavan muuttujan vaihtelun. (Kim 2022) On olemassa useita tapoja jolla voi valita mallin. Kaikilla tavoilla ei tule samoja tuloksia. Malli halutaan pitää mahdollisimman suppeana, sillä jos mukana on tarpeettomia muuttujia, tulkinta on vaikeampaa ja ylimääräiset muuttujat heikentävät ennustetarkkuutta. On tärkeää valita oikeat selittävät muuttujat ja käyttää oikeaa matemaattista mallia tarkkojen ennusteiden saamiseksi.

Yksi tapa muodostaa malli on poistovalinta. Poistovalinnan tavoitteena on regressiomallin pienentäminen siten, että jäljellä olevilla muuttujilla on merkittävä suhde ennustettavaan muuttujaan. Poistovalinnassa aloitetaan täydestä mallista, jossa on mukana kaikki muuttujat. Sitten mallista otetaan pois vähiten merkitsevä muuttuja, eli muuttuja, jonka p-arvo on suurin. Kun tämä muuttuja on poistettu, asennetaan malli uudelleen aineistoon ilman tätä muuttujaa ja poistetaan taas vähiten merkitsevä muuttuja. Tätä jatketaan niin kauan, kunnes mallissa on enää merkitseviä muuttujia.

2.6 Oletukset muuttujista

Kun lineaarista regressiomallia käytetään ennustamiseen, tehdään implisiittisesti joitakin oletuksia yhtälön muuttujista. Mallista oletetaan, että se on kohtuullinen approksimaatio todellisuudesta, eli muuttujien välinen suhde täyttää lineaarisen yhtälön. Jäännöstermeistä oletetaan, että niiden odotusarvo on 0, jotta ennusteet eivät ole systemaattisesti harhaisia. Lisäksi ne eivät ole automaattisesti korreloituneita, koska tällöin ennusteet ovat tehottomia, sillä datasta löytyisi tällöin vielä enemmän tietoa, jota voisi hyödyntää. (Hyndman ja Athanasopoulos 2018) Jäännökset eivät myöskään liity ennustaviin muuttujiin, sillä muuten mallin systemaattiseen osaan tulisi sisällyttää enemmän tietoa. On myös hyödyllistä, että jäännökset ovat normaalisti jakautuneita vakiovarianssilla σ^2, sillä ennustevälit ovat tällöin helpommat muodostaa. (Hyndman & Athanasopoulos 2018)

2.7 Mallin sopivuuden tarkastelu

Regressioanalyysissä ennustetarkkuuden mittaaminen perustuu siihen, kuinka hyvin malli sopii ennustettavaan aineistoon. Tässä kappaleessa käydään läpi analyysejä,

joilla tutkia kuinka hyvin malli sopii aineistoon.

2.7.1 Satunnaisvirheiden analyysi

Sovitetun mallin satunnaisvirheiden jakautumisen analysointi on regressioanalyysissä hyvä tapa katsoa, kuinka hyvin malli on sovitettu aineistoon (John et al. 2005). Koska satunnaisvirheet ovat erotus ennustetun arvon ja todellisen arvon välillä, satunnaisvirheiden jakauman tarkastelu kertoo, kuinka paljon ja millä tavalla ennusteet poikkeavat todellisista arvoista. Tämä voi auttaa arvioimaan, kuinka hyvin malli ennustaa tulevaa ja millaisia virheitä malli tekee. Satunnaisvirheiden jakauman tarkastelusta voi myös nähdä onko mallissa systemaattisia virheitä.

Satunnaisvirheiden jakaumaa voi tarkastella monilla eri tavoilla. Yksi yleisimmistä tavoista on muodostaa jakaumasta histogrammi, joka kertoo miten satunnaisvirheet jakautuvat eri arvoille. Histogrammista myös näkee, ovatko satunnaisvirheet normaalijakautuneita. Jos satunnaisvirheet ovat normaalijakautuneita, on ennustevälien muodostus helpompaa (Hyndman & Athanasopoulos 2018).

Toinen tapa tutkia satunnaisvirheiden jakaumaa on luoda sirontakaavioita satunnaisvirheiden ja ennustemuuttujien välille. Satunnaisvirheiden odotetaan olevan satunnaisia ja näyttämättä mitään systemaattista kuviota. Jos sirontakaaviot kuitenkin näyttävät jonkun kuvion, suhde on epälineaarinen ja mallia tulisi muuttaa. (Hyndman & Athanasopoulos 2018) Satunnaisvirheiden ja sovitettujen arvojen sirontakuviossa ei myöskään tulisi näkyä kuviota, sillä tällöin satunnaisvirheiden varianssi ei välttämättä ole vakio. Tällöin ennustemuuttujan muunnos, kuten esimerkiksi logaritmi olisi tarpeen. Satunnaisvirheitä voi myös tutkia sirontakuviolla (scatterplot), joka auttaa tunnistamaan poikkeavia havaintoja.

2.7.2 Selitysaste

Toinen tapa katsoa, sopiiko regressiomalli aineistoon, on selitysaste R^2. Sitä voidaan kutsua myös determinaatiokertoimeksi. R^2 -arvo mittaa regression aiheuttaman vaihtelun osuutta Y:ssä. (Kim 2022) R^2 määritellään:

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{\mathrm{Y}}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{\mathrm{Y}}\right)^2}$$

R^2 voi saada arvoja väliltä $[0, 1]$. Jos R^2 on lähempänä yhtä, se osoittaa voimakasta regressiota. Jos se on lähempänä nollaa, se osoittaa heikompaa regressiota. Kun se on pieni, voi olla, että tiedoissa on paljon satunnaista luontaista vaihtelua (Kim 2022).

R^2-luvun käytössä on kuitenkin myös ongelmia. Jos malliin lisää selittäviä muuttujia, luku kasvaa vaikka muuttujat eivät olisi tilastollisesti merkitseviä (Draper & Smith 1998). Tällöin on parempi käyttää muokattua selitysastetta R_m^2, koska se on paremmin tulkittavissa (Kim 2022). R_m^2 määritellään:

$$R_m^2 = R^2 - \frac{1}{n-2}\left(1 - R^2\right)$$

2.8 Luottamus- ja ennustevälit

Luottamusväli kertoo mille alueelle tuntemattomien parametrien arvot osuvat tietyllä luotettavuustasolla. $100 - \alpha$ luottamusväli arvolle Y_0, kun mallissa on vain yksi selittävä muuttuja, on

$$\hat{Y_0} \pm t_{1-\frac{\alpha}{2},n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{\left(X_0 - \bar{\mathrm{X}}\right)^2}{\sum\left(X_i - \bar{\mathrm{X}}\right)^2}\right)},$$

missä MSE on keskimääräinen neliövirhe ja $\hat{Y_0} = \hat{\beta_0} + \hat{\beta_1} X_0$.

Ennusteväli kuvaa ennustetun arvon epävarmuutta. Se määrittelee välin, jolla ennustettava arvo tulee olemaan tietyllä todennäköisyydellä. (Kim 2022) Ennusteväli lasketaan muuten samalla tavalla kuin luottamusvälit, mutta kaavaan lisätään neliöjuuren alle yksi.

$100 - \alpha$ ennusteväli arvolle Y_0, kun mallissa on vain yksi selittävä muuttuja, on

$$\hat{Y_0} \pm t_{1-\frac{\alpha}{2},n-2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{\left(X_0 - \bar{\mathrm{X}}\right)^2}{\sum\left(X_i - \bar{\mathrm{X}}\right)^2}\right)},$$

missä MSE on keskimääräinen neliövirhe ja $\hat{Y_0} = \hat{\beta_0} + \hat{\beta_1} X_0$.

95% ennustevälin kaava matriisimuodossa, kun mallissa on useampi selittävä muuttuja:

$$\hat{y} \pm 1.96\hat{\sigma_e}\sqrt{1 + (\mathbf{x}*)'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}*},$$

missä $\hat{y} = \mathbf{x}*'\hat{\beta}$, $\mathbf{x}*$ on vektori, joka sisältää ennustavien muuttujien arvot ja $\mathbf{X}$ on suunnittelumatriisi. Jos ennusteväli on suuri, se tarkoittaa että ennusteen tarkkuus on epävarmempi. Jos taas ennusteväli on kapea, se tarkoittaa että ennuste on luotettavampi.

3 Ennustetarkkuuden mittaaminen

Regressiomalleilla ennustamisessa oletetaan, että ennustettava muuttuja liittyy muihin muuttujiin. Ennusteprosessi sisältää muuttujien suhteiden tunnistamisen, niiden ilmaisemisen matemaattisessa muodossa, ja näiden tietojen käyttäminen ennusteen luomiseen. Aikasarjamalleihin verrattuna regressiomallit ovat yleensä hankalampia käyttää, koska niitä on vaikeammin saatavilla ennusteohjelmistoissa. (Sanders 2015)

Regressiomallit voivat kuitenkin olla parempia esimerkiksi tarkastellessa monen muuttujan välisiä suhteita. Regressiomallit vaativat paljon analysointia mm. muuttujien välisten suhteiden paljastamisessa ja tämä voi parantaa ennusteen tarkkuutta joissain tilanteissa. (Sanders 2015) Usean selittäjän regressiomallilla ennustaminen onkin tehokas työkalu ennustamiseen.

Tärkeä osa ennustamista on ennustetarkkuuden mittaaminen. Sen avulla voidaan arvioida ennustemallien tehokkuutta ja tarkkuutta. Ennustetarkkuuden mittaamiseen on olemassa useita eri mittareita. Yksikään mittari ei kuitenkaan kuvaa täysin ennustamisen tarkkuutta. Tämän takia useita mittareita käytetään usein yhdessä, jotta ennustetarkkuudesta saisi tarkemman kuvan.

3.1 Uuden tiedon ennustaminen

On tärkeää arvioida ennusteiden tarkkuus aitojen ennusteiden avulla. Näin ollen ainoastaan jäännösten suuruus ei ole luotettava osoitus siitä, kuinka suuria todelliset ennustevirheet todennäköisesti ovat. Ennusteiden tarkkuus voidaan määrittää vain ottamalla huomioon, kuinka hyvin malli toimii uusilla tiedoilla. (Hyndman & Athanasopoulos 2018)

Uusilla tiedoilla tarkoitetaan sellaisia havaintoja, jota ei ole käytetty mallin muodostamiseen. Tämä tarkoittaa, että mallia muodostettaessa tulisi aineistosta ottaa osa sivuun mallin testaamista varten, eikä käyttää kaikkea mallin muodostukseen. Koska testidataa ei käytetä ennusteiden määrittämisessä, sen pitäisi antaa luotettava osoitus siitä, kuinka hyvin malli todennäköisesti ennustaa uusien tietojen perusteella. (Hyndman & Athanasopoulos 2018)

Se, kuinka paljon dataa tulisi käyttää mallin muodostukseen sekä testaukseen riippuu ennustemenetelmän tyypistä sekä käytettävissä olevien havaintojen määrästä. (John et al. 2005) Jos havaintoja on paljon, voidaan mallista käyttää testaamiseen

jopa kolmannes. Jos havaintoja on vähän, täytyy käyttää enemmän havaintoja itse mallin muodostamiseen jotta siitä tulisi mahdollisimman luotettava.

3.2 Keskimääräinen absoluuttinen virhe

Yksi useimmin käytetty ennustetarkkuuden mitta on keskimääräinen absoluuttinen virhe MAE. (Hyndman & Athanasopoulos 2018) Absoluuttista virhettä käytetään usein silloin, kun halutaan tarkastella ennusteen virhettä ilman, että suuret virheet saavat suurempaa painoarvoa. Keskimääräinen absoluuttinen virhe määritellään

$$MAE = \frac{\sum_{i=1}^{n} \left(\left|Y_i - \hat{Y}_i\right|\right)}{n}$$

eli keskimääräinen absoluuttinen virhe on ennustettujen ja todellisten arvojen välisten erotusten absoluuttisten arvojen keskiarvo. Keskimääräinen absoluuttinen virhe mittaa virheen suuruutta samoissa yksiköissä, kuin ennustettava muuttuja, joten virhettä on helppo tulkita. Pienempi tulos esittää parempaa ennustetarkkuutta.

3.3 Keskimääräinen neliövirhe

Parempaa ennustetarkkuutta mittaa myös keskimääräinen neliövirhe MSE (Draper & Smith 1998). Se mittaa keskimääräistä virhettä ennustetun sekä todellisen arvon välillä. Keskimääräinen neliövirhe määritellään

$$MSE = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n},$$

eli keskimääräinen neliövirhe on ennustettujen ja todellisten arvojen välisten erotusten neliöt summattuna yhteen ja jaettuna havaintojen lukumäärällä. Kuten keskimääräinen absoluuttinen virhe, myös keskimääräinen neliövirhe mittaa virheen suuruutta samoissa yksiköissä, kuin ennustettava muuttuja, joten sitä on helppo tulkita. Mitä pienempi keskimääräinen neliövirhe on, sitä parempi on mallin ennustetarkkuus.

Ennustetarkkuutta mitattaessa käytetään myös usein keskimääräisen neliövirheen muunnosta RMSE:tä, joka on keskimääräisen neliövirheen neliöjuuri. (Hyndman & Athanasopoulos 2018) Tässäkin mitassa ennustetarkkuus paranee sitä mukaa kun luku pienenee. RMSE määritellään

$$RMSE = \sqrt{MSE}.$$

3.4 Prosenttivirheet

Prosenttivirheitä käytetään usein vertailemaan ennusteiden suorituskykyä, sillä ne ovat yksiköttömiä. (Hyndman & Athanasopoulos 2018) Yleisimmin käytetyt prosenttivirheet ovat keskimääräinen absoluuttinen prosenttivirhe MAPE sekä keskimääräinen neliöprosenttivirhe MSPE. Myös prosenttivirheissä pienempi arvo kertoo tarkemmasta ennusteesta. MAPE mittaa keskimääräistä prosentuaalista virhettä ennusteessa suhteessa todelliseen arvoon. MAPE määritellään

$$MAPE = \frac{100}{n} \sum_{i=0}^{n} \left| \frac{\left(Y_i - \hat{Y}_i\right)}{Y_i} \right|.$$

MSPE mittaa keskimääräistä neliöllistä prosentuaalista virhettä ennusteessa suhteessa todelliseen arvoon. MSPE määritellään

$$MSPE = \frac{100}{n} \sum_{i=0}^{n} \left(\frac{\left(Y_i - \hat{Y}_i\right)}{Y_i} \right)^2.$$

4 Menetelmien soveltaminen aineistoon

4.1 Aineiston esittely

Aineisto on peräisin Finnprostate-tutkimuksesta vuodelta 1998, jossa tutkittiin edenneen eturauhassyövän hoitoa. Tutkimuksen tavoitteena oli selvittää, onko kahden eri hoito-ohjelman välillä eroja hoidon tehokkuuteen tai elämänlaatuun vaikuttavissa tekijöissä. Mukaan otettiin noin 600 potilasta, jotka jaettiin kahteen ryhmään. Toiselle ryhmälle annettiin yleisesti käytettyjä eturauhassyöpälääkkeitä jatkuvasti ja toiselle ryhmälle ajoittain.

Tässä tutkimuksessa käytetään aineistosta sitä versiota, jossa oli karsittu Finnprostaten tutkimuksessa ne potilaat, jotka eivät soveltuneet tutkimukseen. Havaintoja jäi tällöin jäljelle 568. Alkuperäisessä aineistossa on 21 muuttujaa. Muuttujat ovat potilaiden henkilökohtaisia tietoja, kuten potilasnumero, pituus, paino ja ikä, sekä terveystietoja, kuten erilaisia veriarvoja sekä eturauhasen koko. Tutkittava muuttuja on PSA-arvo, joka kuvaa eturauhassyövän mahdollisuutta. Taulukossa 4.1. on esitelty suurin osa muuttujista sekä niiden tunnuslukuja.

4.2 Menetelmät

Tutkimuksessa käytetään R-ohjelmistoa. Aineisto jaetaan ensin kahteen osaan. Toisella osalla estimoidaan viisi erilaista regressiomallia ja toisella osalla testataan mallien ennustetarkkuutta. Mallien muodostamisessa käytetään R-ohjelmiston lm-funktiota, joka muodostaa parametrien estimaatit ja keskihajonnat käyttäen pienimmän neliösumman menetelmää. Lm-funktio myös kertoo suoraan esimerkiksi mallin jäännösten tunnuslukuja, mallin selitysasteen sekä parametrien tilastollisen merkittävyyden.

PSA-arvo muunnetaan malleihin sen logaritmiksi, jotta tuloksia olisi helpompi lukea, ja jolloin PSA-arvojen jakauma olisi enemmän normaalijakautuneempi. Lisäksi potilaan painon ja pituuden avulla lasketaan potilaan BMI-arvo, joka on paino jaettuna pituuden neliöllä. Ensiksi tutkitaan muuttujien tilastollista merkittävyyttä PSA-arvoon ja valitaan sen perusteella sopivat muuttujat malliin.

Kun aineistoon on sovitettu regressiomallit, tutkitaan ja vertaillaan niiden ennustetarkkuutta. Tutkitaan mallien jäännöksiä ja ennustevälejä, sekä lasketaan malleis-

ta selitysaste, keskimääräinen neliövirhe sekä keskimääräinen absoluuttinen virhe. Päätellään niiden avulla, kuinka hyvä ennustetarkkuus malleilla on.

Taulukko 4.1. Aineiston muuttujat ja niiden tunnusluvut

Muuttuja	Minimi	Maksimi	Keskiarvo
PSA (ng/mL)	0.9	5123.0	150.0
Testosteroni (nmol/L)	0.7	41.7	15.1
Alkalinen fosfataasi, AFOS (U/L)	73.0	6610.0	291.7
Kreatininiini (umol/L)	46.0	577.0	99.8
Hematokriitti (%)	23.0	53.0	141.7
Hemoglobiini (g/L)	80.0	182.0	141.7
Valkosolut (E9/L)	2.2	17.4	7.0
Punasolut (E12/L)	2.7	6.3	4.6
Korpuskulaarinen tilavuus (fL)	71.0	108.0	90.8
Korpuskulaarinen hemoglobiini (pg)	22.0	37.0	30.7
Eturauhasen leveys (cm)	2.0	9.0	4.6
Eturauhasen pituus (cm)	2.0	12.0	4.6
Potilaan syntymävuosi	1907	1952	1928
Potilaan paino (kg)	50.0	170.0	78.0
Potilaan pituus (cm)	148.0	198.0	172.7

4.3 Tutkimustulokset ja niiden tulkinta

Kun tutkitaan aineistoa, huomataan että vain kolme muuttujaa ovat tilastollisesti merkitseviä, kun halutaan ennustaa PSA-arvoa. Nämä muuttujat ovat alkalinen fosfataasi AFOS, eturauhasen pituus sekä potilaan BMI-arvo. Kun aineistosta poistetaan muut muuttujat ja puuttuvat arvot, jää aineistoon 538 havaintoa. Tämä aineisto jaetaan kahteen osaan, jolloin molemmissa aineiston osissa on 269 havaintoa.

Muodostetaan näiden kolmen tilastollisesti merkitsevien muuttujien avulla viisi lineaarista regressiomallia aineiston osasta, joka on tarkoitettu mallien estimointiin. Malleilla pyritään ennustamaan PSA-arvon logaritmin suuruus. Ensimmäisessä mallissa selittävänä muuttujana on alkalinen fosfataasi AFOS, toisessa mallissa eturauhasen pituus ja kolmannessa mallissa potilaan BMI-arvo.

Havainnollistetaan yhden selittäjän lineaarisia regressiomalleja kuvaajilla. Ku-

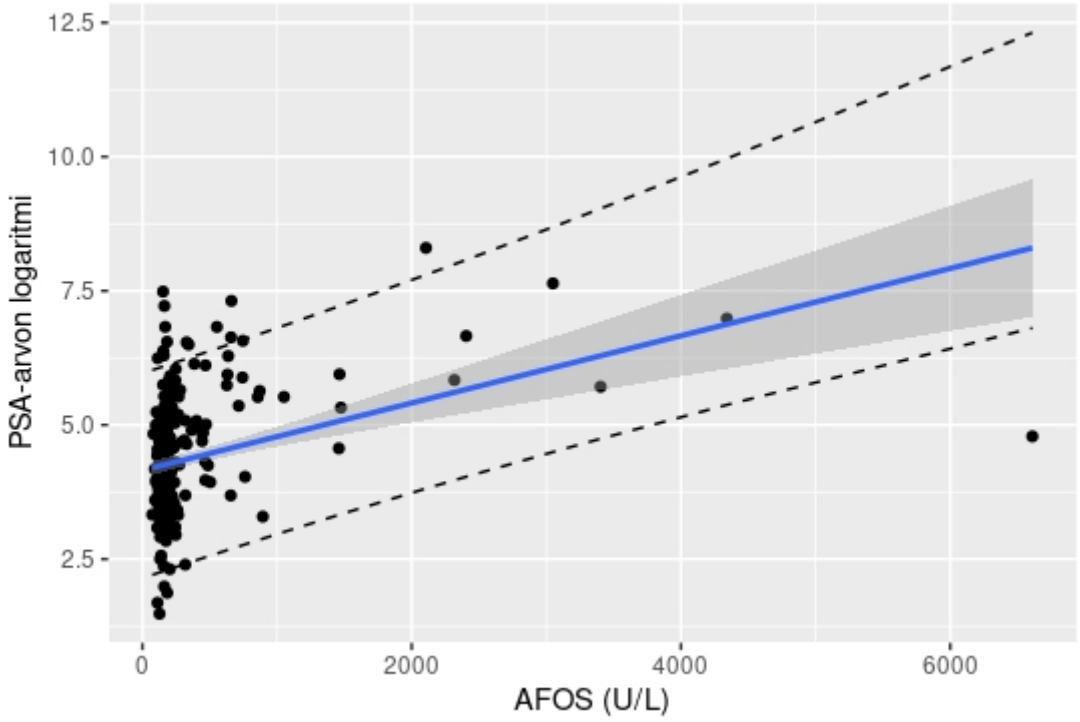
vaajissa on X-akselilla selittävä muuttuja ja Y-akselilla kohdemuuttuja. Sininen viiva on muodostettu ennustemalli ja katkonaiset viivat ovat ennustevälit kohdemuuttujalle, eli tässä tapauksessa PSA-arvon logaritmille. Enustevälit kuvaavat välejä, joille ennustettavien arvojen on ennustettu osuvan 95% todennäköisyydellä. Harmaa alue on 95% luottamusväli. Pisteet ovat havaintoja testausaineistosta.

Malli 1.

$$Y_i = 4.0536 + 0.0008 \cdot X_i + \epsilon_i,$$

missä X_i on AFOS-arvo.

Kuva 4.1. Malli 1.

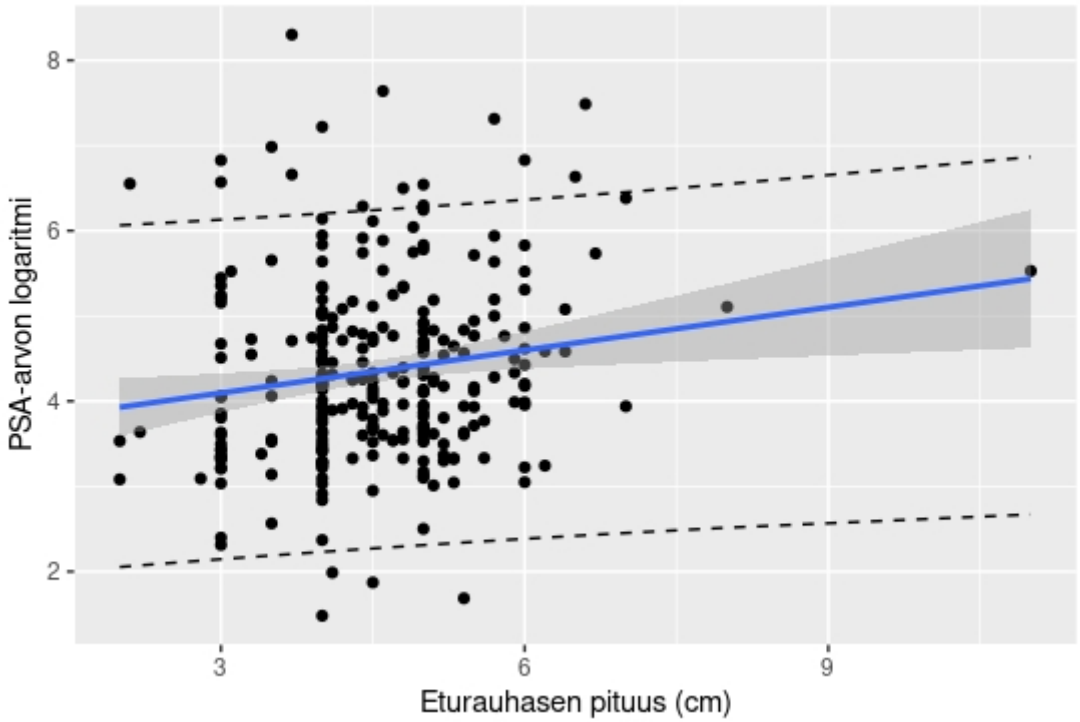


Malli 2.

$$Y_i = 3.9014 + 0.0788 \cdot X_i + \epsilon_i,$$

missä X_i on eturauhasen pituus.

Kuva 4.2. Malli 2.

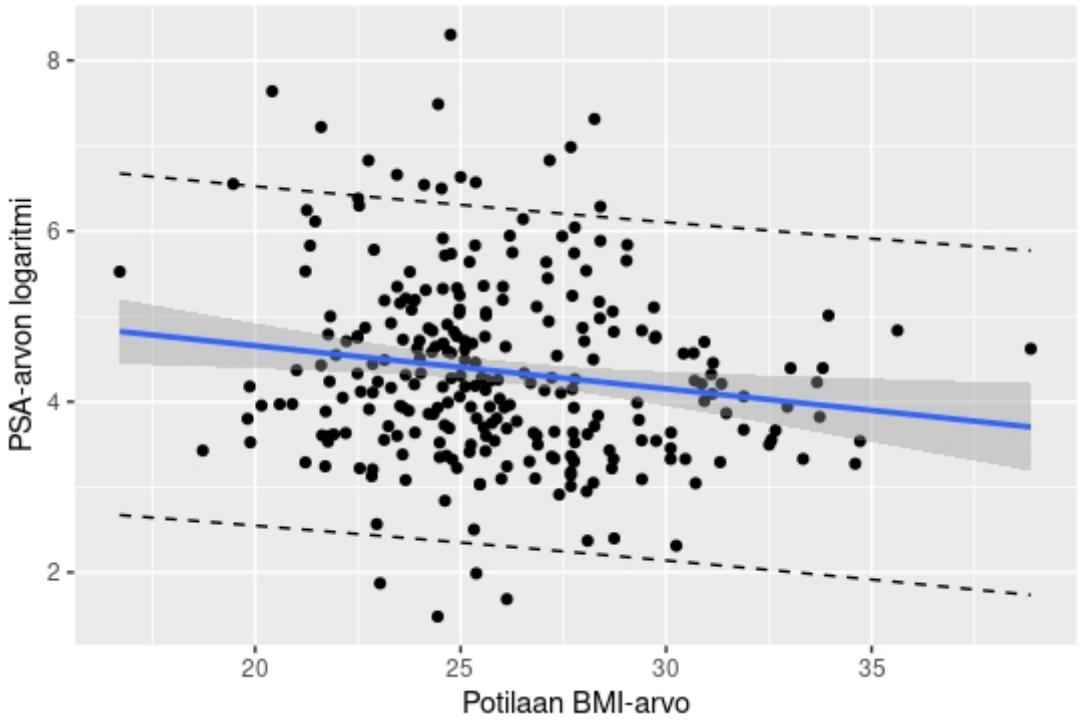


Malli 3.

$$Y_i = 5.3661 - 0.0415 \cdot X_i + \epsilon_i,$$

missä X_i on potilaan BMI-arvo.

Kuva 4.3. Malli 3.



Kaikista kuvaajista huomataan, että vain pieni osa havainnoista osuu estimoiduille malleille. Mallit eivät selkeästi sovi aineistoon kovin hyvin. Valitaan neljänteen malliin selittäviksi muuttujiksi näistä kolmesta muuttujasta kaksi eniten tilastollisesti merkitsevää muuttujaa, jotka olivat mallien perusteella AFOS ja BMI-arvo. Viidennessä mallissa on kaikki kolme muuttujaa.

Malli 4.

$$Y_i = 5.1696 + 0.0008 \cdot X_{1i} - 0.0426 \cdot X_{2i} + \epsilon_i,$$

missä X_{1i} on AFOS-arvo ja X_{2i} potilaan BMI-arvo.

Malli 5.

$$Y_i = 4.6766 + 0.0009 \cdot X_{1i} + 0.1208 \cdot X_{2i} - 0.0458 \cdot X_{3i} + \epsilon_i,$$

missä X_{1i} on AFOS-arvo, X_{2i} eturauhasen pituus ja X_{3i} potilaan BMI-arvo.

Vertaillaan nyt mallien ennustetarkkuutta aineiston testausosuuden avulla. Lasketaan malleista ennustetarkkuuden mittoja, joita ovat keskimääräinen absoluuttinen virhe MAE ja keskimääräinen neliövirhe MSE. Lasketaan myös prosenttivirhe MAPE sekä keskimääräisen neliövirheen neliöjuuri RMSE. Tulokset näkyvät taulukossa 4.2.

Taulukko 4.2. Ennustetarkkuuden mitat

	MAE	MSE	RMSE	MAPE
Malli 1	0.848	1.274	1.129	20.9 %
Malli 2	0.840	1.186	1.089	20.5 %
Malli 3	0.843	1.188	1.090	20.6 %
Malli 4	0.857	1.287	1.134	21.1 %
Malli 5	0.881	1.332	1.154	21.8 %

Tuloksista huomataan, ettei malleilla ole suurta eroa keskenään ennustetarkkuudessa. Kaikkien mallien ennustetarkkuus on aika heikkoa, eikä aineiston muilla muuttujilla selkeästi pysty tarkasti ennustamaan potilaan PSA-arvoa. Kuitenkin näistä viidestä mallista mallilla 2, jossa selittävänä muuttujana oli eturauhasen pituus, on paras ennustetarkkuus, sillä jokainen ennustetarkkuuden mitan arvo on pienin verrattuna muiden mallien mittojen arvoihin. Mallilla 5, jossa selittävänä mallina oli kaikki kolme tilastollisesti merkittävää muuttujaa, on heikoin ennustetarkkuus, sillä sen ennustetarkkuuden mittojen arvot ovat kaikista suurimpia. Huomataan myös, että kaikista yksinkertaisimmilla malleilla oli selkeästi paremmat tulokset kuin monimutkaisemmilla.

5 Yhteenveto

Tutkielman alussa esiteltiin regressioanalyysin perusteita eli yhden ja usean selittäjän lineaariset regressiomallit, jäännöstermien määrittely sekä pienimmän neliösumman menetelmä. Lisäksi kerrottiin, miten muodostaa hyvän ennustetarkkuuden kannalta mahdollisimman hyvä malli. Yhtenä esimerkkinä mallin muodostukseen esiteltiin poistovalinta. Mallin muodostuksen jälkeen esiteltiin kaksi tapaa millä tutkia mallin sopivuutta. Viimeisenä regressioanalyysistä määriteltiin luottamus- ja ennustevälit.

Seuraavassa luvussa perehdyttiin tarkemmin regressiomalleilla ennustamiseen sekä niiden ennustetarkkuuden mittaamiseen. Luotettavimman ennustetarkkuuden tuloksen saa sillä, että jakaa aineiston kahteen osaan, joista toisella muodostetaan malli ja toisella testataan ennustetarkkuutta. Ennustetarkkuuden mittaamiseen on useita eri mittareita, joita yleensä käytetään yhdessä, sillä yksikään mittari ei täysin kuvaa ennustetarkkuutta. Mittarit perustuvat pääosin ennustettujen ja todellisten arvojen välisiin erotuksiin. Yleisimmät ennustetarkkuuden mitat ovat keskimääräinen absoluuttinen virhe ja keskimääräinen neliövirhe.

Viimeisessä luvussa käytettiin aiemmin esiteltyjä menetelmiä ja laskettiin ennustetarkkuuden mittoja eturauhassyöpäaineistosta muodostetuille malleille, joilla pyrittiin ennustamaan PSA-arvon suuruutta. Aineistosta muodostettiin viisi eri mallia, joista kolme olivat yhden selittäjän lineaarisia regressiomalleja, ja kaksi olivat usean selittäjän lineaarisia regressiomalleja. Yhdelläkään mallilla ei ollut erityisen hyvää ennustetarkkuutta, mutta kun malleja vertaili, huomattiin että yksinkertaisemmilla malleilla oli huomattavasti paremmat ennustetarkkuudet kuin malleilla, joissa oli useampi selittäjä.

Lähteet

[1] Draper, N. ja Smith, H. (1998) *Applied regression analysis*. 3. painos. URL: https://ebookcentral.proquest.com/lib/tampere/detail.action?docID=7103889&query=applied+regression+analysis

[2] Hyndman, R. J. ja Athanasopoulos, G. (2018) *Forecasting: principles and practice*. URL: https://otexts.com/fpp2/

[3] John, J. A., Whitaker, D ja Johnson, D.G. (2005) *Statistical thinking in Business* URL: https://ebookcentral.proquest.com/lib/tampere/detail.action?docID=1633513&query=forecasting+accuracy+measures

[4] Kim, H.-J. (2022) "DATA.STAT.460 Regression analysis". *Luentomoniste, Tampereen yliopisto*

[5] Sanders, N. (2015) *Forecasting Fundamentals*. URL: https://ebookcentral.proquest.com/lib/tampere/detail.action?docID=4742536&query=forecasting+fundamentals