Tampereen yliopisto

Anttoni Tornberg

LINEAARINEN REGRESSIOMALLI

ja sen soveltaminen estimoinnissa

Kandidaatintyö
Tekniikan ja luonnontieteiden tiedekunta
Tarkastaja: Jussi Kangas
Huhtikuu 2025

TIIVISTELMÄ

Anttoni Tornberg: Lineaarinen regressiomalli
Kandidaatintyö
Tampereen yliopisto
Tekniikka ja luonnontieteet, TkK
Huhtikuu 2025

Tässä työssä tutustutaan lineaariseen regressioanalyysiin tilastollisena analyysimenetelmänä. Tarkoituksena on tarkastella lineaarisen regressiomallin luomisen ja testaamisen taustalla olevaa matematiikkaa. Lopuksi teoriaa sovelletaan luomalla jääkiekkopelaajien tilastojen pohjalta lineaarinen regressiomalli pelaajien vuosipalkoista.

Teoriaosuudessa käsitellään regressioanalyysin suorittamista pienimmän neliösumman menetelmän avulla. Työssä oletetaan lukijan osaavan perusteet lineaarialgebrasta, eikä kaikkia siihen liittyviä käsitteitä avata tai todisteta. Teoriaosuudessa myös tarkastellaan mallin testaamista niin kokonaisuuden kuin yksittäisten selittäjien osalta.

Työn soveltavassa osiossa sovelletaan teoriaosuuden matematiikkaa oikeaan aineistoon ja verrataan sitä MATLAB:n fitlm-komennon tuloksiin. Tämän pohjalta luodaan esimerkinomaisesti lineaarinen regressiomalli NHL-pelaajien vuosipalkoista ja tarkastellaan, mitkä selittäjät vaikuttavat palkkoihin. Lopuksi mallin hyvyyttä tarkastellaan tilastollisten tunnuslukujen avulla, sekä arvioidaan mallin todenmukaisuutta ja käyttökelpoisuutta.

Avainsanat: regressioanalyysi,lineaarinen regressiomalli, pienimmän neliösumman menetelmä, data-analyysi

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

TEKOÄLYN KÄYTTÖ OPINNÄYTTEESSÄ

Opinnäytteessäni on käytetty tekoälysovelluksia:

☐ Ei
☒ Kyllä

Ilmoitukseni mukaan olen käyttänyt opinnäytteessäni tutkielmaprosessin aikana seuraavia tekoälysovelluksia:

- OpenAI ChatGPT

Tekoälysovellusten nimet ja versiot:

- ChatGPT-3.5
- ChatGPT-4o

Käyttötarkoitus: Tekoälyä on käytetty

- apuvälineenä ohjelmistojen, kuten MATLAB ja Latex, käytössä, esimerkiksi koodin kirjoittamiseen, virheiden etsintään ja komentojen luontiin.
- lähdeaineistojen tekstin kääntämisessä englannista suomeksi.
- tekstin tarkastuksessa ja muotoilussa.

Osiot, joissa tekoälyä on käytetty:

- Kirjoitusvaiheen alussa Latex-komentojen apuvälineenä. Liitteessä A komentojen apuvälineenä.
- Kuvan 2.1 piirtämisessä apuvälineenä.
- Oikolukuvaiheessa tekstin tarkastuksessa ja korjaamisessa.

Olen tietoinen siitä, että olen täysin vastuussa koko opinnäytteeni sisällöstä, mukaan lukien osat, joissa on hyödynnetty tekoälyä, ja hyväksyn vastuun mahdollisista eettisten ohjeiden rikkomuksista.

SISÄLLYSLUETTELO

1. Johdanto 1
2. Teoria 2
2.1 Lineaarinen regressiomalli 2
2.2 Pienimmän neliösumman menetelmä 3
2.3 Mallin testaaminen 7
3. Lineaarisen regressiomallin soveltaminen 12
3.1 Aineiston esittely 12
3.2 Mallin soveltaminen aineistoon 13
3.3 Mallin arviointi 17
4. Yhteenveto ja pohdintaa 21
Lähteet 23
Liite A: Työssä käytetty MATLAB-koodi 24

LYHENTEET JA MERKINNÄT

A^T	Matriisin A transpoosi
A^{-1}	Matriisin A käänteismatriisi
$E[x]$	Muuttujan x odotusarvo
$\beta_1, \ldots, \beta_k$	Regressiokertoimet
$\hat{y}$	Muuttujan y estimaatti
$\mathbf{x}$, $\mathbf{y}$	Vektoreita
∇	Gradientti
$x \sim \chi^2$	x noudattaa χ^2-jakaumaa
A	Matriisi

1. JOHDANTO

Regressioanalyysi on tilastotieteessä paljon käytetty menetelmä. Tutkijat useilla aloilla ovat kiinnostuneita siitä, millä tavoin asiat vaikuttavat toisiinsa. Johtopäätöksiä näistä riippuvuussuhteista voidaan tehdä empiiristen havaintojen ja tilastollisen analyysin avulla. Tähän analyysiin usein hyvin soveltuva työkalu onkin juuri regressioanalyysi, jossa tutkitaan selittävien muuttujien eli selittäjien $x_1, x_2, \ldots, x_n$ tilastollista vaikutusta selitettävään muuttujaan eli vasteeseen y. Kun tilastollinen vaikutus selvitetään, voidaan sen avulla ennustaa selitettävän muuttujan käyttäytymistä eri tilanteissa. Nykyään lähes kaikilla tilasto-ohjelmistoilla voidaan suorittaa regressioanalyysi, ja yhdistettynä datan keräämiseen se on erittäin käyttökelpoinen työkalu.

Regressiomallit voidaan jakaa kolmeen osaan: yhden selittäjän lineaariseen malliin, monen selittäjän lineaariseen malliin ja epälineaariseen malliin. Näistä kahta ensimmäistä käsitellään tässä työssä. Ensimmäinen laskennallinen menetelmä regressioanalyysin suorittamiseen oli tiettävästi pienimmän neliösumman menetelmä (PNS-menetelmä), jota kuuluisat matemaatikot Adrien-Marie Legendre ja Carl Friedrich Gauss tutkivat jo 1800-luvun alussa. Silloin mallia sovellettiin muun muassa taivaankappaleiden kiertoratojen arvioimiseen. [1, s. 2–3]

Lineaarista regressiomallia on mielekästä käyttää, kun havaintoparien välillä on korrelaatiota eli lineaarista riippuvuutta. Kuitenkin korrelaatio on todella harvoin täydellistä, jolloin kaikkia havaintopisteitä ei voida yhdistää suoralla viivalla. Tällöin kehitetään sellainen sovitesuora, joka minimoi kaikkien pisteiden etäisyyden suorasta. Tähän käytetyin menetelmä on pienimmän neliösumman menetelmä, jossa havaintopisteiden ja sovitesuoran väliset etäisyyksien neliöt minimoidaan. [2, s. 367] Tähän tutustutaan tarkemmin alaluvussa 2.2.

Tässä kandidaatintyössä tutustutaan regressioanalyysiin ja pienimmän neliösumman menetelmään. Työn teoriaosuudessa tarkastellaan lähtökohtia regressioanalyysiin sekä perehdytään pienimmän neliösumman menetelmään lineaarialgebran kautta. Samalla esitellään tietokoneohjelmistojen jokaisen tunteman “lineaarinen sovitus” -toiminnon taustalla olevaa matriisilaskentaa. Tämän lisäksi tarkastellaan regressiomallin matemaattista testaamista. Työn lopussa on soveltava osio, jossa luodaan esimerkinomaisesti lineaarinen regressiomalli jääkiekkopelaajien vuosipalkoista sekä tutkitaan eri tilastojen vaikutusta palkan suuruuteen.

2. TEORIA

Tässä luvussa esitellään teoriaa regressioanalyysista, pienimmän neliösumman menetelmästä ja mallin testaamisesta. Lähtökohtana regressioanalyysiin voidaan pitää tilannetta, jossa havaintopisteiden välillä havaitaan korrelaatiota eli lineaarista riippuvuutta, mutta halutaan tarkempaa tietoa muuttujien välisestä yhteydestä. Lineaarisessa regressioanalyysissa muodostetaan havaintopisteisiin lineaarinen sovite, jonka avulla voidaan estimoida selitettävän muuttujan käyttäytymistä.

2.1 Lineaarinen regressiomalli

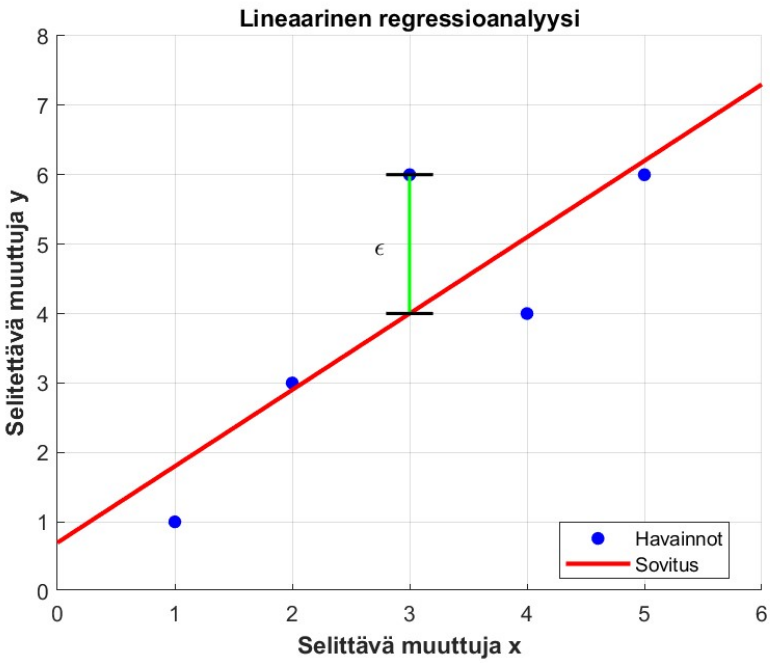
Kuvataan yhden selittäjän lineaarisessa regressiomallissa selitettävän muuttujan y käyttäytymistä yhtälöllä

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{2.1}$$

missä vakioita β_0 ja β_1 kutsutaan vakiotermiksi ja regressiokertoimeksi ja vakiota ϵ satunnaistermiksi [3, s. 7].

Satunnaistermin ϵ odotusarvo on 0 ja varianssi on vakio σ^2, ja se kuvaa sitä muuttujan y vaihtelua, joka ei ole selitettävissä muuttujan x vaihtelulla [4, s. 17]. Vakiot β_0 ja β_1 ovat tuntemattomia, ja niille pyritään etsimään estimaatit eli arviot pienimmän neliösumman menetelmällä.

Tarkastellaan tilannetta, jossa $\epsilon = 0$. Tällöin yhtälö (2.1) on suoran yhtälön muodossa, jolloin puhutaan matemaattisesta mallista tilastollisen mallin sijaan [3, s. 2]. Todellisuudessa tilastot tai data eivät juuri koskaan noudata matemaattisia malleja, ja siksi satunnaistermi pidetään mukana regressiomallissa.



Kuva 2.1. *Esimerkkikuva yhden selittäjän lineaarisesta regressiomallista. Vihreällä on havainnollistettu jäännöstä ϵ havainnon ja sovitesuoran välillä.*

Määritelmä 2.1. Usean selittävän muuttujan lineaarista regressiomallia kuvataan yhtälöllä

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon,$$

missä parametrit $\beta_1, \ldots, \beta_k$ ovat muuttujiin $x_1, \ldots, x_k$ liittyviä regressiokertoimia ja parametri k on selittäjien lukumäärä [5, s. 9].

2.2 Pienimmän neliösumman menetelmä

Tässä luvussa tutustutaan lineaarisen regressiomallin regressiokertoimien estimoimiseen pienimmän neliösumman menetelmällä. Menetelmän tavoitteena on löytää sellaiset arvot kertoimille $\beta_0, \beta_1, \ldots, \beta_k$, että todellisten mittapisteiden ja sovitesuoran välisten poikkeamien neliöiden summa minimoituu [2, s. 367]. Kun poikkeamat neliöidään, itseisarvoltaan suuret poikkeamat korostuvat. Seurauksena pienimmän neliösumman menetelmä muodostaa regressiosuoran, joka ei ole kovin kaukana yhdestäkään havaintopisteestä [4, s. 19].

Määritelmä 2.2. Lineaarisen regressiomallin sovitesuoran yhtälö

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

antaa selitettävälle muuttujalle estimaatin $\hat{y}$ joillakin tuntemattomilla kertoimilla $\beta_0, \beta_1, \ldots, \beta_k$.

Määritelmä 2.3. Havainnon y_i ja sen estimaatin $\hat{y}$ välinen etäisyys ϵ määritellään yhtälöllä

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}. \tag{2.2}$$

Etäisyyttä kutsutaan myös jäännökseksi.

Määritelmä 2.4. Määritellään jäännösneliösumma SS_R (residual sum of squares) yhtälöllä

$$\begin{aligned} SS_R &= \sum_{i=1}^{n} (\epsilon_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2, \end{aligned} \tag{2.3}$$

missä parametri k on selittäjien lukumäärä ja parametri n havaintoparien lukumäärä.

Seuraavaksi etsitään kertoimille $\beta_0, \beta_1, \ldots, \beta_k$ sellaiset arvot, jotka minimoivat lausekkeen (2.3).

Lineaarialgebran hyödyntäminen PNS-menetelmässä

Jos aineiston havaintoparit toteuttavat suoran yhtälön, regressiokertoimet $\beta_0, \beta_1, \ldots$ ja β_k saadaan ratkaisemalla lineaarinen yhtälöryhmä, joka voidaan esittää matriisimuodossa

$$A\mathbf{x} = \mathbf{y}$$

eli

$$\mathbf{y} - A\mathbf{x} = \mathbf{0}.$$

Tässä yhtälöryhmässä

$$A = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \mathbf{x} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ ja } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \tag{2.4}$$

missä matriisin A sarakkeet, poislukien ensimmäinen, ovat selittäjien $x_1, \ldots, x_n$ arvoja. Vektorin $\mathbf{y}$ alkiot $y_1, \ldots, y_n$ ovat selittäjiä vastaavia vasteen arvoja. Indeksi n on aineiston havaintopisteiden lukumäärä ja indeksi k selittäjien lukumäärä. [5, s. 11]

Tilastollisessa analyysissa on kuitenkin hyvin harvoin tilannetta, jossa dataparit noudattavat matemaattista mallia, esimerkiksi suoran yhtälöä. Tällöin lineaarisen yhtälöryhmän ratkaisuvektoria $\mathbf{x}$ ei ole. Silloin etsitään sellainen vektori $\mathbf{x}$, joka on lähimpänä yhtälöryhmän ratkaisua. Tällainen ratkaisu löydetään, kun kaavan (2.3) mukainen neliösumma saa pienimmän arvonsa. [6, s. 226]

Määritelmä 2.5. Määritellään jäännösvektori yhtälöllä

$$\mathbf{e} = \mathbf{y} - A\mathbf{x}.$$

Jäännösvektoriksi vastaa yhtälön (2.2) jäännöksistä muodostettua vektoria.

Jos jäännösvektori $\mathbf{e} = \mathbf{0}$, lineaarisella yhtälöryhmällä on ratkaisu eikä pienimmän neliösumman ratkaisua tarvita.

Määritelmä 2.6. Pienimmän neliösumman ratkaisulla tarkoitetaan sellaista vektoria $\mathbf{x}$, joka minimoi jäännösvektorin normin neliön. Tämä voidaan kirjoittaa lausekkeena

$$\min_{\mathbf{x}}(\|\mathbf{e}\|^2) = \min_{\mathbf{x}}(\|\mathbf{y} - A\mathbf{x}\|^2),$$

missä etsitään sellainen vektori $\mathbf{x}$, että lauseke minimoituu. Merkitään tätä löydettyä pienimmän neliösumman ratkaisua vektorilla $\hat{\mathbf{x}}$ ja pienintä jäännösvektoria vektorilla $\mathbf{e_{min}}$.

Normin määritelmän mukaisesti

$$\|\mathbf{y} - A\mathbf{x}\| = \|A\mathbf{x} - \mathbf{y}\|,$$

joten voidaan käyttää oikeanpuoleista muotoa, kuten Boyd et al. käyttävät kirjassaan [6, s. 228].

Määritelmä 2.7. Merkitään funktion

$$f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|^2$$

minimikohdaksi $\mathbf{x} = \hat{\mathbf{x}}$.

Avataan toinen potenssi

$$\begin{aligned} f(\mathbf{x}) &= \|A\mathbf{x} - \mathbf{y}\|^2 \\ &= (A\mathbf{x} - \mathbf{y})^T(A\mathbf{x} - \mathbf{y}) \\ &= \mathbf{x}^T A^T A\mathbf{x} - 2\mathbf{y}^\mathbf{T}\mathbf{A}\mathbf{x} + \mathbf{y}^\mathbf{T}\mathbf{y}. \end{aligned}$$

Lasketaan gradientti jokaisesta termistä erikseen

$$\begin{aligned} \nabla f(\mathbf{x}) &= \nabla(\mathbf{x}^T A^T A\mathbf{x}) - \nabla(2\mathbf{y}^T A\mathbf{x}) + \nabla(\mathbf{y}^T\mathbf{y}) \\ &= 2A^T A\mathbf{x} - 2\mathbf{y}A^T \\ &= 2A^T(A\mathbf{x} - \mathbf{y}) \end{aligned}$$

[6, s. 229, s. 184].

Minimikohdassa $\mathbf{x} = \hat{\mathbf{x}}$ funktion gradientti on nolla. Merkitään tätä yhtälöllä

$$2A^T(A\hat{\mathbf{x}} - \mathbf{y}) = 0,$$

joka voidaan kirjoittaa muodossa

$$A^T A\hat{\mathbf{x}} = A^T\mathbf{y}.$$

Pienimmän neliösumman ratkaisu saadaan, kun kerrotaan molempia puolia matriisin $A^T A$ käänteismatriisilla $(A^T A)^{-1}$. Tämä on mahdollista vain, jos matriisi $A^T A$ on kääntyvä, eli matriisin $A^T A$ sarakkeet ovat lineaarisesti riippumattomia. Tällöin myös $\det A^T A \neq 0$.

Matriisin A sarakkeet on määritelty kaavarivillä (2.4). Sarakkeet ovat lineaarisesti riippuvia vain, jos jokin selittäjistä on muiden selittäjien lineaarikombinaatio. Tällainen selittäjä kannattaakin usein poistaa mallista, koska se ei tuo malliin mitään uutta informaatiota. Voidaan siis todeta, että lähes aina $\det(A^T A) \neq 0$, mistä seuraa, että matriisilla $A^T A$ on käänteismatriisi $(A^T A)^{-1}$, jolloin sillä voidaan kertoa yhtälöä puolittain.

Lause 2.8. *Olkoon A kokoa $n \times (k+1)$ oleva matriisi ja $\mathbf{y} \in \mathbb{R}^n$. Yhtälöryhmän $A\mathbf{x} = \mathbf{y}$ pienimmän neliösumman ratkaisu*

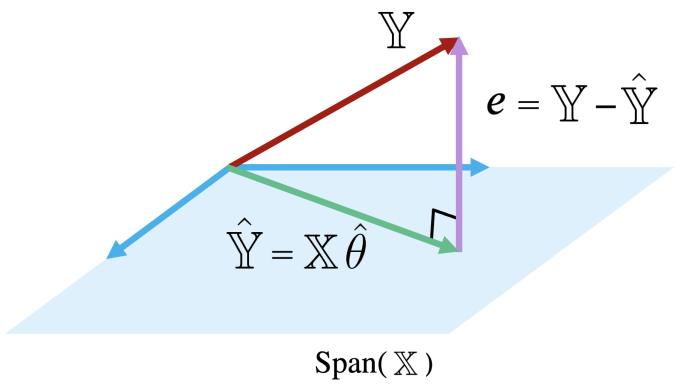
$$\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{y}, \tag{2.5}$$

missä

$$\hat{\mathbf{x}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}.$$

Kertoimia $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ kutsutaan PNS-estimaateiksi.

Tiedetään, että lyhin vektori tason ja pisteen välillä on tasoa vastaan kohtisuora vektori. Täten pienimmän neliösumman ratkaisu voidaan myös tulkita sellaisena vektorina $\hat{\mathbf{x}}$, että vektorit $A\hat{\mathbf{x}}$ ja $\mathbf{e}_{\mathbf{min}}$ ovat ortogonaaliset. Tällöin vektori $\mathbf{e}_{\mathbf{min}}$ on matriisin A sarakeavaruutta $\mathcal{R}(A)$ vastaan kohtisuorassa. [7, s. 185–186] Kuva 2.2 havainnollistaa tätä hyvin.



Kuva 2.2. *Havainnollistava kuva pienimmän neliösumman menetelmästä. Matriisi $\mathbb{X}$ vastaa työssä käytettyä matriisia A ja vektori $\hat{\theta}$ PNS-ratkaisuvektoria $\hat{\mathbf{x}}$. Vektori $\mathbf{e}$ on minimoitu jäännösvektori, jota merkitään tässä työssä vektorilla $\mathbf{e}_{\mathbf{min}}$ Sininen taso kuvaa matriisin A sarakeavaruutta. [8]*

Voidaan myös merkitä, että

$$A^{\dagger} = (A^T A)^{-1} A^T.$$

Tällöin $\hat{\mathbf{x}} = A^{\dagger}\mathbf{b}$, missä matriisi $A^{\dagger}$ on matriisin A pseudoinverssi [6, s. 229].

Määritelmä 2.9. Selitettävän muuttujan y estimaatti $\hat{y}$ saadaan yhtälön

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k \tag{2.6}$$

mukaisesti, joka voidaan myös esittää matriisiyhtälönä

$$\hat{\mathbf{y}} = A\hat{\mathbf{x}}. \tag{2.7}$$

[5, s. 13]. Kun tähän yhtälöön sijoitetaan selittäjien x_i arvoja, saadaan tulokseksi estimaatti selitettävälle muuttujalle.

2.3 Mallin testaaminen

Tässä luvussa tarkastellaan lineaarisen regressiomallin tilastollista testaamista. Ennen luodun mallin käyttöönottoa on syytä testata, onko saatu malli käyttökelpoinen. On mahdollista, että selittävät muuttujat eivät selitä vasteen vaihtelua tai selittävät sitä huonosti. Testeillä voidaan myös selvittää, mitkä muuttujat ovat merkitseviä ja mitkä eivät, joten testit antavat tärkeää informaatiota regressiomallista.

Havainnon y_i ja sen estimaatin $\hat{y}$ välinen etäisyys on usean selittäjän tapauksessa määritelty yhtälöllä (2.2). PNS-menetelmässä kaikkien näiden etäisyyksien neliösumma minimoitiin. Tämä minimoitu jäännösvektori

$$\mathbf{e_{min}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - A\hat{\mathbf{x}}$$

yhtälön (2.7) mukaisesti.

Jäännösneliösumma on määritelty yhtälöllä (2.3). Muutetaan jäännösneliösumma matriisimuotoon, jolloin sen laskennallinen määrittäminen on helpompaa. Yhtälö saadaan muotoon

$$\begin{aligned} SS_R &= \mathbf{e_{min}}^T \mathbf{e_{min}} \\ &= (\mathbf{y} - A\hat{\mathbf{x}})^T (\mathbf{y} - A\hat{\mathbf{x}}) \\ &= \mathbf{y^T y} - \hat{\mathbf{x}} A^T \mathbf{y}. \end{aligned} \tag{2.8}$$

Viimeisellä rivillä kaikkia välivaiheita ei avata, mutta niihin voi tutustua tarkemmin Myersin et al. kirjasta [5, s. 18].

Jäännösneliösumma kertoo siis kokonaispoikkeaman havaintopisteiden ja sovitesuoran välillä. Sen merkitystä ei suoraan matriisimuodosta helposti näe, mutta muoto on erittäin käyttökelpoinen laskentaa ajatellen. Jäännösneliösumman avulla voidaan estimoida jäännöksen varianssia σ^2, joka kertoo, kuinka paljon jäännös vaihtelee aineistossa.

Jäännösneliösumman ja jäännöksen varianssin suhteesta tiedetään, että

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-k-1}$$

[2, s. 373]. Tällöin odotusarvo

$$E\left[\frac{SS_R}{\sigma^2}\right] = n - k - 1,$$

josta saadaan ratkaistua varianssi

$$\sigma^2 = E\left[\frac{SS_R}{n-k-1}\right].$$

Huomataan, että $\frac{SS_R}{n-k-1}$ on siis varianssin harhaton estimaattori.

Määritelmä 2.10. Jäännöksen varianssia voidaan estimoida otoksen avulla laskettavalla suureella, joka määritellään kaavalla

$$s^2 = \frac{SS_R}{n-k-1} \tag{2.9}$$

[5, s. 18]. Tätä kutsutaan jäännöksen otosvarianssiksi.

Sijoitettamalla tähän yhtälöt (2.3) ja (2.8), saadaan otosvarianssiksi

$$\begin{aligned} s^2 &= \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2 \\ &= \frac{1}{n-k-1}(\mathbf{y}^\mathbf{T}\mathbf{y} - \hat{\mathbf{x}}A^T\mathbf{y}). \end{aligned}$$

Regressioanalyysissa varianssin halutaan olevan pieni, koska silloin malli selittää vasteen vaihtelua paremmin kuin suurella varianssilla.

Määritelmä 2.11. Kokonaisneliösumma (total sum of squares)

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

missä $\bar{y}$ on aineiston vasteen y otoskeskiarvo. Kokonaisneliösumma kuvaa havaintojen vaihtelua keskiarvonsa ympärillä [4, s. 20] [2, s. 392].

Vektorilaskentaa hyödyntäen kokonaisneliösumman yhtälöksi saadaan

$$SS_T = (\mathbf{y} - \bar{y})^T((\mathbf{y} - \bar{y})). \tag{2.10}$$

Määritelmä 2.12. Mallin selitysosuus (R-squared) määritellään yhtälöllä

$$R^2 = \frac{SS_T - SS_R}{SS_T} = 1 - \frac{SS_R}{SS_T} \tag{2.11}$$

[2, s. 393]. Selitysosuus kertoo, kuinka monta prosenttia muuttujan y vaihtelusta voidaan selittää selittäjien vaihtelulla.

Mitä lähempänä selitysosuus on arvoa yksi, sitä parempi regressiomalli on. Jäännösneliösummalle vastaava optimaalinen arvo on kaavan (2.11) mukaisesti nolla, jolloin kaikki havaintopisteet ovat regressiosuoralla. Selitysosuuden avulla päästään testaamaan sitä, kuinka hyvin malli onnistuu kuvaamaan selitettävän muuttujan vaihtelua. Sen kaavassa ei kuitenkaan huomioida selittäjien lukumäärää. Tällöin selittäjän lisääminen malliin voi kasvattaa selitysosuutta riippumatta selittäjän merkitsevyydestä. Tästä syystä määritellään mallille korjattu selitysosuus, joka pienenee, kun merkitsemättömiä selittäjiä lisätään malliin.

Määritelmä 2.13. Mallin korjattu selitysosuus (adjusted R-squared) määritellään yhtälöllä

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-k}\right)(1 - R^2), \tag{2.12}$$

missä parametri n on havaintoparien lukumäärä ja k selittäjien lukumäärä. Jos aineiston koko on suuri verrattuna selittäjien lukumäärään eli $\left(\frac{n-1}{n-k}\right) \approx 1$, niin $R^2_{adj} \approx R^2$. [5, s. 22-23]

Jos siis aineiston havaintopisteiden määrä on suuri, ei ole juurikaan merkitystä, käytetäänkö selitysosuutta R^2 vai korjattua selitysosuutta R^2_{adj}. Tästä syystä tässä työssä käytetään selitysosuudesta yksinkertaista merkintää R^2.

Määritelmä 2.14. Usean selittäjän regressiomallin testauksessa hypoteesit määritellään seuraavasti:

$$H_0 : \beta_1 = \ldots = \beta_k = 0$$
$$H_1 : \beta_i \neq 0 \text{ jollain indeksin } i \text{ arvolla}$$

Testattava nollahypoteesi siis olettaa, että mikään selittäjistä ei pysty selittämään muuttujan y vaihtelua. Vaihtoehtoinen hypoteesi puolestaan sanoo, että vähintään yksi selittäjistä on merkitsevä. [5, s. 20]

Nollahypoteesia testataan testisuureen

$$F_0 = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \sim F(k, n - k - 1) \tag{2.13}$$

avulla [5, s. 20]. Jos malli onnistuu selittämään vasteen vaihtelua erittäin hyvin, niin $R^2 \approx 1$. Lausekkeesta nähdään, että tällöin nimittäjä lähestyy nollaa, jolloin F-testisuureen arvo kasvaa. Vastaavasti jos malli on huono, niin $R^2 \approx 0$. Tällöin osoittaja lähestyy nollaa, jolloin myös F-testisuureen arvo lähestyy nollaa.

Nollahypoteesin testaamisessa valitaan jokin merkitsevyystaso α, joka kuvaa todennäköisyyttä hylätä nollahypoteesi, vaikka se olisikin tosi [2, s. 307]. Merkitsevyystasoksi asetetaan siis etukäteen jokin pieni luku, esimerkiksi $\alpha = 0,05 = 5\%$ tai $\alpha = 0,01 = 1\%$.

Määritelmä 2.15. Todennäköisyyttä

$$p_F = P(F_{k,n-k-1} > F_0). \tag{2.14}$$

kutsutaan F-testisuureen p-arvoksi ja se lasketaan F-jakaumasta vapausastein k ja $n-k-1$ [2, s. 315].

Jos p-arvo on suurempi kuin merkitsevyystaso, nollahypoteesi jää voimaan ja todetaan, että malli ei selitä vasteen vaihtelua. Jos taas p-arvo on pienempi, niin nollahypoteesi hylätään ja vaihtoehtoinen hypoteesi jää voimaan. Tällöin vähintään yksi selittäjistä on merkitsevä.

Kun mallin yleinen käyttökelpoisuus on selvitetty, usein ollaan kiinnostuneita siitä, mitkä selittäjät ovat mallin kannalta tärkeimpiä. Esimerkiksi jos selittäjän data on vaikeasti saatavilla tai datan kerääminen on kallista, voidaan olla kiinnostuneita siitä, kuinka tärkeä selittäjä on mallin kannalta ja voisiko sen poistamista mahdollisesti harkita.

Määritelmä 2.16. Testattaessa yksittäisen selittäjän x_j merkitsevyyttä, vastaavan regressiokertoimen β_j hypoteesit määritellään seuraavasti:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

[5, s. 24].

Jos nollahypoteesia ei hylätä, kerrointa vastaava selittävä muuttujan poistamista mallista kannattaa harkita. Nollahypoteesia testataan t-testisuurella

$$t_0 = \frac{\hat{\beta}_j}{SE_{\hat{\beta}_j}},$$

missä $SE_{\hat{\beta}_j}$ on estimaatin $\hat{\beta}_j$ keskivirhe, eli se lasketaan yhtälöllä

$$SE_{\hat{\beta}_j} = \sqrt{s^2 C_{jj}}.$$

Tässä vakio s^2 on määritelmän 2.10 mukainen jäännöksen otosvarianssi ja $C_{jj} = ((A^T A)^{-1})_{jj}$ eli matriisin $(A^T A)^{-1}$ diagonaalialkio riviltä ja sarakkeelta j. [5, s. 24]

T-testisuureeksi saadaan

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{s^2 C_{jj}}} = \frac{\hat{\beta}_j}{\sqrt{s^2((A^T A)^{-1})_{jj}}}. \tag{2.15}$$

Määritelmä 2.17. Todennäköisyyttä

$$p_t = P(t_{n-k-1} > |t_0|) \tag{2.16}$$

kutsutaan t-testisuureen p-arvoksi, joka lasketaan t-jakaumasta vapausasteella $n-k-1$ [2, s. 321].

Yhtälö (2.16) saadaan muotoon

$$p_t = 2 \cdot \min(P(t_{n-k-1} > t_0), P(t_{n-k-1} < t_0)), \tag{2.17}$$

missä valitaan näistä kahdesta todennäköisyydestä pienempi ja kerrotaan se kahdella. Jos t-testisuure t_0 saa negatiivisen arvon, valitaan todennäköisyys $P(t_{n-k-1} < t_0)$, eli että jakauman arvo on pienempi kuin t_0. Positiivisella t-testisuureen arvolla taas valitaan todennäköisyys $P(t_{n-k-1} > t_0)$.

P-arvoa verrataan valittuun merkitsevyystasoon α, joka on yksittäisen selittäjän testaamisessa usein 5% tai 10% [1, s. 172]. Jos p-arvo on suurempi, nollahypoteesi jää voimaan ja todetaan, että selittäjää ei pidetä merkitsevänä. Tällöin sen poistamista mallista kannattaa harkita. Jos taas p-arvo on pienempi, oletetaan vaihtoehtoinen hypoteesi todeksi ja todetaan selittäjä tilastollisesti merkitseväksi.

3. LINEAARISEN REGRESSIOMALLIN SOVELTAMINEN

Teorialuvussa käsiteltiin matemaattista teoriaa lineaarisen regressioanalyysin taustalla. Nyt niiden tietojen pohjalta voidaan ratkaista lineaarisen regressiomallin regressiokertoimille estimaatit PNS-menetelmällä, tutkia mallin käyttökelpoisuutta mallin tunnuslukujen avulla ja tutkia yksittäisten selittäjien merkitsevyyttä. Tässä luvussa sovelletaankin teorialuvun aiheita aineistoon. Työssä käytetään laskentaan MATLAB:ia ja myös verrataan matemaattisen laskennan ja MATLAB:n fitlm-komennon antamia tuloksia toisiinsa.

3.1 Aineiston esittely

Sovellettavaksi aineistoksi on valittu jääkiekkoaiheinen data. Selitettävänä muuttujana on NHL-pelaajien keskivuosiansio, AAV (Annual Average Value), joka lasketaan jokaista sopimusta kohden kaavalla

$$\text{AAV} = \frac{\text{sopimuksen kokonaisarvo}}{\text{sopimuksen pituus}}.$$

Tässä työssä tutkitaan, miten muut pelaajan tilastot vaikuttavat keskivuosipalkkaan. Otoksena on NHL:ssä pistepörssin 198 parasta pelaajaa kaudelta 2024-2025. Tilastot on kerätty 12.2.2025 mennessä pelatuista otteluista, joten otanta on noin 55 pelattua ottelua. Työssä käytetyt selittäjät on kirjoitettu taulukkoon 3.1.

Taulukko 3.1. *Aineiston selittäjät ja niiden kuvailu.*

Selittäjä	**Kuvaus**
P/GP	tehopisteiden määrä pelattua ottelua kohden
TOI/GP	keskimääräinen peliaika ottelussa
Signing Age	ikä sopimuksentekohetkellä
Length	nykyisen sopimuksen pituus vuosina
G60	maalien lukumäärä 60 pelattua minuuttia kohden
PM	plus-miinus -tilasto
GWG	tehtyjen voittomaalien lukumäärä
DFF%	pelaajan kentälläolon vaikutus vaarallisten maalipaikkojen suhteeseen

Lähteenä P/GP-, TOI/GP- ja GWG-tilastoille on nhl.com [9] ja muille tilastoille puckpedia.com

[10]. Alkuperäiselle datalle tehdään seuraavat muutokset:

1. Yhdistetään molempien lähteiden tilastot samaan taulukkoon.
2. Poistetaan tuhaterottimena olevat pilkut.
3. Muutetaan TOI/GP-tilaston alkuperäinen muoto (minuutit:sekunnit) desimaaliluvuksi.
4. Vaihdetaan pelaajien järjestystä niissä tilanteissa, joissa alkuperäinen järjestys oli lähteiden välillä eri.
5. Poistetaan AAV-tilastosta dollarimerkit.
6. Vaihdetaan plusmiinustilaston nimen "+/-" tilalle nimeksi "PM".

Työssä muodostetaan aineistosta usean selittäjän lineaarinen regressiomalli lausekkeen (2.1) mukaisesti, lasketaan regressiokertoimille estimaatit ja tarkastellaan mallin selitysosuutta sekä yksittäisten selittäjien merkitsevyyttä.

Lopuksi, mikäli malli todetaan selityskykyiseksi, estimoidaan mallilla NHL:n supertähtilaitahyökkääjien Mikko Rantasen ja Mitchell Marnerin sopimusten keskivuosipalkat. Rantasen ja Marnerin nykyiset sopimukset päättyvät heinäkuussa 2025, ja tällä päivämäärällä, 3.3.2025, kummallakaan heistä ei ole jatkosopimusta tehtynä. Yleisessä keskustelussa heidän tulevien sopimustensa keskivuosipalkkojen arvioidaan olevan noin 13–14 miljoonaa dollaria.

3.2 Mallin soveltaminen aineistoon

Kun aineistoon on tehty edellä kuvatut muutokset, voidaan aineiston pohjalta luoda lineaarinen regressiomalli. Tässä alaluvussa muodostetaan lineaarinen malli työn teoriaosuudessa esitellyn matematiikan avulla MATLAB:lla ja verrataan tuloksia MATLAB:n fitlm-komennon arvoihin.

Selittäjien ja vastemuuttujan välinen korrelaatiokerroin on esitetty taulukossa 3.2.

Taulukko 3.2. *Selittäjien ja vastemuuttujan AAV välinen korrelaatiokerroin.*

Selittäjä	**Korrelaatiokerroin**
P/GP	0,53330
TOI/GP	0,49764
Signing Age	0,12676
Length	0,75164
G60	-0,06781
PM	0,01203
GWG	0,17208
DFF%	0,12701

Taulukosta 3.2 huomataan, että kiinnostavimmat selittäjät korrelaatiokertoimen perusteella ovat P/GP, TOI/GP ja Length, koska korrelaatiokerroin niiden ja selittäjän välillä on kauimpana nollasta. Vähiten mielenkiintoisia ovat taas PM ja G60.

Matriisi A muodostetaan kaavarivin (2.4) mukaan. Varmistetaan, että matriisi $A^T A$ on kääntyvä laskemalla sen determinantti. Determinantti

$$\det(A^T A) = 1,453 \times 10^{25},$$

mikä vahvistaa, että matriisi $A^T A$ on kääntyvä.

Pienimmän neliösumman ratkaisuvektori saadaan yhtälön (2.5) mukaan. Vektorin alkioina ovat PNS-estimaatit, joista saadaan muodostettua lineaarisen sovitteen yhtälö. PNS-estimaatit on esitetty taulukossa 3.3.

Taulukko 3.3. *Selittäjät ja niiden β-kertoimien PNS-estimaatit matemaattisesti laskettuna ja MATLAB:n fitlm-komennolla saatuna.*

Selittäjä	**Laskettu PNS- estimaatti**	**Fitlm-komennon PNS-estimaatti**
β_0	-5 701 231	-5 701 231
P/GP	4 640 795	4 640 795
TOI/GP	99 745	99 745
Signing Age	131 432	131 432
Length	672 698	672 698
G60	-743 486	-743 486
PM	-22 435	-22 435
GWG	-1 143	-1 143
DFF%	-2 211	-2 211

Mallin selitysosuus R^2 ja korjattu selitysosuus R^2_{adj} saadaan yhtälöillä (2.11) ja (2.12). F-testisuure lasketaan yhtälöllä (2.13). Sen p-arvo lasketaan MATLAB:lla F-jakaumasta yhtälön (2.14) mukaisesti. Tulokset näistä on esitetty taulukossa 3.4.

Taulukko 3.4. *Selitysosuuden, korjatun selitysosuuden sekä F-testisuureen ja sen p-arvon arvot matemaattisesti laskettuna ja fitlm-komennolla saatuna.*

Tunnusluku	**Laskettu arvo**	**Fitlm-komennon antama arvo**
R^2	0,71364	0,714
R^2_{adj}	0,70309	0,702
F_0	58,87775	58,9
p_F	0	$2{,}62\times10^{-47}$

Korjattu selitysosuus on noin yhden prosenttiyksikön pienempi kuin varsinainen selitysosuus, joten ei ole juurikaan merkitystä, kumpaa arvoa käytetään. Todetaan malli selityskykyiseksi, kun selittäjät pystyvät selittämään noin 70% vasteen vaihtelusta ja F-testisuureen p-arvo on erittäin pieni.

Mallin avulla voidaan estimoida pelaajien palkkoja, jos tiedetään pelaajien tilastolukemia selittävien muuttujien osalta. Estimoidaan mallin avulla Rantasen ja Marnerin keskipalkat. Estimoinnissa muodostetaan kahdeksan selittäjän lineaarisen regressiomallin yhtälö kaavan (2.6) mukaisesti si-

joittamalla vakioiden $\hat{\beta}$ tilalle taulukon 3.3 PNS-estimaatit. Palkkojen estimaatit ovat taulukossa 3.5. Ne on laskettu käyttäen kuluvan kauden 2024-2025 tilastoja. Signing Age on pelaajan tämänhetkinen ikä. Sopimuksen pituus Length arvioidaan molemmilla olevan maksimipituus kahdeksan vuotta, mutta lasketaan Marnerille myös estimaatti neljän vuoden mittaisella jatkosopimuksella.

Taulukko 3.5. *Estimaatit Mikko Rantasen ja Mitchell Marnerin keskivuosipalkoista.*

Pelaaja & sopimuksen pituus	**AAV**
Rantanen & 8 vuotta	9 820 000
Marner & 8 vuotta	10 730 000
Marner & 4 vuotta	8 040 000

Mallin mukaan Marnerin palkka pitkässä sopimuksessa on noin miljoona dollaria suurempi kuin Rantasella. Lisäksi lyhyellä sopimuksella palkka olisi matalampi kuin pitkässä, mikä ei kuulosta järkevältä. Tarkastellaan tätä tarkemmin alaluvussa 3.3.

Tarkastellaan mallin yksittäisiä selittäjiä tarkemmin yhtälön (2.15) mukaisella t-testisuurella ja sen avulla lasketulla p-arvolla. Nämä on esitetty taulukossa 3.6.

Taulukko 3.6. *T-testisuure ja sen p-arvo kaikille selittäjille matemaattisesti laskettuna ja fitlm-komennolla saatuna.*

Selittäjä	**Laskettu t_0**	**Fitlm-komennon antama t_0**	**Laskettu p_t**	**Fitlm-komennon antama p_t**
P/GP	7,22176	7,22176	$1{,}22 \times 10^{-11}$	$1{,}22 \times 10^{-11}$
TOI/GP	1,78134	1,78134	0,07646	0,07646
Signing Age	4,81737	4,81737	$2{,}98 \times 10^{-6}$	$2{,}98 \times 10^{-6}$
Length	12,7013	12,7013	0	$4{,}003 \times 10^{-27}$
G60	-1,79196	-1,79196	0,07473	0,07473
PM	-2,26818	-2,26818	0,02445	0,02445
GWG	-0,01668	-0,01668	0,98671	0,98671
DFF%	-0,15415	-0,15415	0,87765	0,87765

T-testisuureen p-arvojen perusteella voidaan hyväksyä tai hylätä selittäjää koskeva nollahypoteesi, joka on esitetty määritelmässä 2.16. Valitaan merkitsevyystasoksi 5%, eli hyväksytään nollahypoteesi kaikille muuttujille, joiden t-testisuureen p-arvo $p_t > 0,05$. Mitä pienempi t-testisuureen p-arvo on, sitä luotettavampana selittäjää voidaan pitää. Mallin kannalta luotettavimmat selittäjät ovat siis P/GP, Signing Age ja Length. Epäluotettavimmat taas ovat GWG ja DFF%.

Tarkasteltaessa taulukon 3.3 PNS-estimaatteja, huomataan, että selittäjien G60, PM, GWG, ja DFF% PNS-estimaatit ovat negatiivisia, mikä tarkoittaa, että tilastolukeman kasvaessa keskipalkka

pienenee. Tämä ei kuulosta todellisuutta ajatellen järkevältä. Koska vähiten merkitsevät selittäjät vaikuttavat myös olevan virheellisiä todellisuuteen peilaten, näiden selittäjien poistaminen on järkevää. Poistaminen tehdään eliminaatiomenetelmällä (backward elimination). Menetelmässä selittäjiä poistetaan yksi kerrallaan huonoimmasta aloittaen. Tähän voi tutustua tarkemmin Xinin et al. kirjasta [1, s. 171].

Aloitetaan poistamalla mallista selittäjä GWG. Poistamisen jälkeen mallin $R^2 = 0,714$ ja $p_F = 3 \times 10^{-48}$, sekä selittäjän DFF% $p_t = 0,87$. Seuraavaksi mallista poistetaan selittäjä DFF%. Poistamisen jälkeen huomataan, että edelleenkään selittäjän poistaminen ei laske selitysosuutta, vaan $R^2 = 0,714$. Vähiten merkitseviksi muuttujiksi jäävät TOI/GP ($p_t = 0,678$) ja G60 ($p_t = 0,632$). Sovelletaan eliminaatiomenetelmää ja poistetaankin seuraavaksi selittäjä G60, vaikka sen p-arvo onkin hieman parempi kuin selittäjän TOI/GP. Tämä tehdään siitä syystä, että selittäjän G60 PNS-estimaatin negatiivisuus vähentää mallin todenmukaisuutta. Taulukoissa 3.7 ja 3.8 on esitetty uusi malli eliminaatiomenetelmän suorittamisen jälkeen.

Taulukko 3.7. *Selittäjien PNS-estimaatit, t-testisuureen arvot ja niiden p-arvot fitlm-komennolla saatuna selittäjien eliminaatiomenetelmän suorittamisen jälkeen.*

Selittäjä	**PNS-estimaatti**	$\mathbf{t_0}$	$\mathbf{p_t}$
β_0	-6 948 271	-6,23	$2{,}83\times10^{-9}$
P/GP	4 104 620	7,55	$1{,}76\times10^{-12}$
TOI/GP	154 370	3,26	0,0013
Signing Age	130 324	4,81	$2{,}98\times10^{-6}$
Length	669 952	12,70	3×10^{-27}
PM	-25 060	-2,64	0,009

Taulukko 3.8. *Mallin selitysosuus, korjattu selitysosuus, F-testisuure ja sen p-arvo fitlm-komennolla saatuna eliminaatiomenetelmän suorittamisen jälkeen.*

Tunnusluku	**Arvo**
R^2	0,708
R^2_{adj}	0,701
F_0	93,3
p_F	$1{,}82 \times10^{-49}$

Taulukoista huomataan, että uuden mallin selittäjien t-testisuureiden p-arvot ovat hyvin pieniä, joten selittäjät todetaan merkitseviksi. Lisäksi mallin selitysosuus on vain alle yhden prosenttiyksikön pienempi kuin alkuperäisessä mallissa, eli mallit selittävät vasteen vaihtelua lähes yhtä hyvin. Käytetään tätä mallia Rantasen ja Marnerin palkkojen uudelleenestimointiin.

Taulukko 3.9. *Estimaatit Mikko Rantasen ja Mitchell Marnerin keskivuosipalkoista*

Pelaaja & sopimuksen pituus	AAV
Rantanen & 8 vuotta	10 080 000
Marner & 8 vuotta	10 410 000
Marner & 4 vuotta	7 730 000

Kun verrataan taulukoiden 3.5 ja 3.9 tuloksia, huomataan, että Rantasen palkka hieman nousi ja Marnerin hieman laski, mutta isoa muutosta keskipalkoissa ei mallin päivittäminen tehnyt.

3.3 Mallin arviointi

Mallin luomisen jälkeen ollaan usein kiinnostuneita siitä, kuinka hyvin mallilla kyetään kuvaamaan todellisuutta. Tässä luvussa tarkastellaan edellisessä luvussa luotua mallia käytettävyyden ja mahdollisten ongelmien näkökulmasta niin yksittäisten selittäjien kuin mallin kokonaisuuden näkökulmasta.

Mallin selittäjien arviointi

Tutkittaessa taulukon 3.2 korrelaatiokertoimia mallin luomisen jälkeen, huomataan, ettei niistä voi suoraan tehdä johtopäätöksiä selittäjien merkitsevyydestä. Esimerkiksi plusmiinustilaston PM korrelaatiokerroin on kaikista lähimpänä nollaa, mutta kuitenkin mallissa se todetaan merkitseväksi. Lisäksi korrelaatiokertoimen ja PNS-estimaatin etumerkkikin voi olla eri usean selittäjän mallissa.

Tarkastellaan mallin selittäjien PNS-estimaatteja taulukosta 3.7. PNS-estimaatin arvo kertoo, kuinka paljon keskipalkka muuttuu selittäjän arvon muuttuessa yhdellä yksiköllä.

1. Selittäjä P/GP: vaikutus keskipalkkaan on luonnollinen – mitä enemmän pelaaja tekee pisteitä ottelua kohti, sitä suurempi on keskipalkka. Selittäjän merkittävyys korostuu erityisesti, kun otoksena on 198 pistepörssin parasta pelaajaa.
2. Selittäjä TOI/GP: kasvu vaikuttaa myös palkkaan nostavasti. Peliajan nouseminen tarkoittaa usein myös vastuun kasvamista, mikä luonnollisestikin näkyy usein pelaajan palkassa.
3. Selittäjä Signing Age: arvon kasvaessa myös keskipalkka kasvaa. Tämän selittää NHL:n sopimusjärjestelmä – ensimmäiset kolme vuotta pelataan tulokassopimuksella, jonka AAV on noin miljoona dollaria. Tämän jälkeen seura voi tarjota pelaajalle sopimuksen ilman muiden seurojen kilpailutusta. Kun pelaaja on 27-vuotias tai pelannut NHL:ssä vähintään seitsemän vuotta, hän voi kilpailuttaa sopimuksensa kaikilla seuroilla, jolloin AAV luonnollisesti nousee [11]. Lineaarinen malli ei kuitenkaan ole selittäjälle Signing Age optimaalinen, koska malli ennustaa esimerkiksi 50-vuotiaana sopimuksen tekeville pelaajille korkeampaa palkkaa kuin 27-vuotiaille, mikä ei ole realistista. Esimerkiksi logaritminen malli voisi olla parempi ainakin yksinään selittäjälle Signing Age.
4. Selittäjä Length: PNS-estimaatti on positiivinen, mikä tarkoittaa vuosipalkan kasvamista so-

pimuksen pidentyessä. Todellisuudessa kuitenkin usein samalla pelisuorituksella lyhyempi sopimus antaa korkeamman keskivuosipalkan kuin pitkä, koska seuralle lyhyessä sopimuksessa on pienempi riski. Jos esimerkiksi pelaaja pyytäisi tarjousta vuoden ja kahdeksan vuoden sopimuksista, on erittäin todennäköistä, että vuoden sopimukseen seura pystyy tarjoamaan korkeampaa vuosipalkkaa. Tämä mallin vinouma johtuu osin siitä, että parhaille pelaajille solmitaan suurimmaksi osaksi pitkiä sopimuksia, jolloin tilastollisesti näyttää siltä, että pitkä sopimus lisäisi vuosipalkkaa. Myös otoksessa yli 60%:lla pelaajista on vähintään viiden vuoden mittainen sopimus, ja vain alle 4%:lla vuoden mittainen sopimus, eli pitkät sopimukset ovat yliedustettuina. Vääristymä näkyy myös taulukossa 3.9 verrattaessa Marnerin neljän ja kahdeksan vuoden välisten sopimusten keskipalkkoja.

5. Selittäjä PM: estimaatti on negatiivinen, eli mallin mukaan hyvä plusmiinustilasto vaikuttaa palkkaan negatiivisesti, mikä ei kuulosta järkevältä. Koska paremmin palkatut pelaajat tekevät kuitenkin paljon tehopisteitä, kuten P/GP-tilastosta nähdään, vaikuttaisi mallin mukaan siltä, että he ovat myös kentällä paljon vastustajan tehdessä maaleja. Tämä itse asiassa hyökkäyspelivoittoisessa NHL:ssä pitääkin usein paikkansa. Toisaalta myös joukkueen menestys ja kokonaismaaliero vaikuttavat PM-tilastoon merkittävästi, jolloin huonossa joukkueessa huippupelaajilla voi olla erittäin huono PM-lukema ja vastaavasti hyvässä joukkueessa huonommalla pelaajalla hyvä PM-lukema. Myös tilaston painotus on mallissa varsin pieni – esimerkiksi merkittävä 20 yksikön ero tilastossa vaikuttaa palkkaan vain noin 500 000 dollaria. Mallin tavoitteena on kuitenkin estimoida pelaajille palkkoja, joten ei ole järkevää, että hyvästä PM-lukemasta "rankaistaan"pelaajaa. Mallia on syytä tarkastella vielä ilman selittäjää PM.

Mallin kokonaisuuden arviointi

Poistetaan mallista selittäjä PM ja luodaan neljän selittäjän lineaarinen malli. Tarkastellaan mallin selitysosuutta kokonaisuudessaan. MATLAB:n fitlm-komennon tulokset ovat esitetty taulukossa 3.10.

Taulukko 3.10. *Selitysosuus, korjattu selitysosuus, t-testisuure ja sen p-arvo ilman selittäjää PM.*

	Arvo
R^2	0,698
R^2_{adj}	0,692
F_0	111
p_F	$4{,}73 \times 10^{-49}$

Selitysosuuden arvo on noin 70%, eli vain noin prosenttiyksikön matalampi kuin edeltävässä viiden selittäjän mallissa, taulukossa 3.8. Mallin selityskyky ei siis juurikaan putoa PM-selittäjän poiston myötä, joten tarkastellaan kokonaisuutta tällä mallilla. Myös F-testisuureen p-arvo on erittäin pieni, joten nollahypoteesi hylätään.

Tarkastellaan fitlm-komennon antamia arvoja PNS-estimaateista, jotka on esitetty taulukossa 3.11.

Taulukko 3.11. *PNS-estimaatit, t-testisuureen arvot ja niiden p-arvot ilman selittäjää PM.*

Selittäjä	**PNS-estimaatti**	$\mathbf{t_0}$	$\mathbf{p_t}$
β_0	-6 764 819	-5,99	$1{,}01\times10^{-8}$
P/GP	3 705 840	6,99	$4{,}44\times10^{-11}$
TOI/GP	148 127	3,09	0,002
Signing Age	135 549	4,95	$1{,}64\times10^{-6}$
Length	682 144	12,79	$1{,}60\times10^{-27}$

PNS-estimaateissa ei ole merkittäviä muutoksia selittäjän poiston jälkeen. Estimoidaan Rantasen ja Marnerin palkat tällä mallilla. Palkat on esitetty taulukossa 3.12.

Taulukko 3.12. *Estimaatit Rantasen ja Marnerin keskivuosipalkoista ilman selittäjää PM.*

Pelaaja & sopimuksen pituus	**AAV**
Rantanen & 8 vuotta	1 020 000
Marner & 8 vuotta	10 400 000
Marner & 4 vuotta	7 600 000

Nyt Marnerin ja Rantasen pitkät sopimukset ovat hieman lähempänä toisiaan kuin taulukossa 3.9, jossa PM-selittäjä oli mukana.

Palkan estimointiin tätä neljän selittäjän mallia voi käyttää tietyin rajoituksin. Taulukon 3.12 pitkien sopimusten estimoidut keskipalkat ovat oikeansuuntaiset, mutta on syytä tarkastella mallia vielä tarkemmin. Testataan mallin toimivuutta vielä estimoimalla lyhyen sopimuksen palkka Rasmus Kuparille, joka on 13.3.2025 kauden pistetilastossa sijalla 546 [9]. Kupari on seuransa 4. ketjun pelaaja, joka tunnetaan hyvästä peruspelaamisesta, mutta tehopisteitä ei paljoa tule. Oletetaan sopimus kahden vuoden mittaiseksi ja sopimuksentekoiäksi 24 vuotta. Selittäjä $P/GP = 0,14$ ja $TOI/GP = 10,33$ 17.3.2025 mennessä pelattujen ottelujen otannalla [9]. Näillä arvoilla $AAV = -98\ 000$ \$, joka on selkeästi virheellinen. Toisaalta Kuparin tehdessä 8 vuoden sopimuksen, malli antaa keskivuosipalkaksi noin 4 miljoonaa dollaria, joka on taas liikaa Kuparin tasoiselle pelaajalle. Mallia ei siis voi käyttää Kuparin kaltaisten pohjaketjujen pelaajien palkkoihin luotettavasti.

Malli ei myöskään huomioi muita pelitilastoja pistetilaston lisäksi, eli esimerkiksi puolustuspelaajille se ei anna juurikaan arvoa. Toisaalta mallin otosvalinnassa, pistepörssin 198 parhaassa pelaajassa, puolustuspelaajat ovat aliedustettuina, joten puolustustilastojen puuttuminen on luonnollinen seuraus tästä.

Malli toimii parhaiten paljon pisteitä tekevien pelaajien pitkien sopimusten kohdalla. Mallia voisi esimerkiksi käyttää noin 27-vuotiaiden tähtipelaajien pitkän kilpailutettujen jatkosopimusten

estimointiin tai noin 22-25-vuotiaiden pelaajien ei-kilpailutettuihin, esimerkiksi 4-6 vuoden mittaisiin, sopimuksiin. Toimivuutta parantaisi vielä se, että sopimuksen päättymisikä rajoitettaisiin esimerkiksi 38 vuoteen, jolloin yli 30-vuotiaille pelaajille ei voisi enää tarjota kahdeksan vuoden sopimusta, niin kuin todellisuudessakaan ei ole tapana. Tämä "ikärajan" asettaminen laskisi myös yli 30-vuotiaiden pelaajien palkkoja sopimuksen pituuden laskiessa, jolloin ikävuosien tuoma lisä palkkaan kumoutuu. Tällöin mallia voidaan soveltaa paremmin myös yli 30-vuotiaiden pelaajien palkkoihin.

Estimoidaan esimerkin vuoksi palkka Patrick Kanelle. P/GP-tilastolukema ja TOI/GP otetaan mallissa käytetystä aineistosta, eli kauden 2024-2025 tilastoista 12.2. mennessä pelatuista otteluista. Signing Age = 36 ja rajoitetaan ikä 38 ikävuoteen. Tällöin Length = 2. Vuosipalkaksi malli antaa Kanelle 4 600 000$, mikä on hyvä arvio. Kesällä 2024 hän solmi yhden vuoden ja neljän miljoonan dollarin arvoisen sopimuksen, joten mallin tulos on hyvin linjassa tämän kanssa.

4. YHTEENVETO JA POHDINTAA

Tässä työssä tutustuttiin lineaariseen regressioanalyysiin, sen suorittamiseen pienimmän neliösumman menetelmällä ja mallin tilastolliseen testaamiseen. Teoriaosuudessa käsiteltiin ohjelmistojen taustalla olevaa matematiikkaa, ja verrattiin laskennallisesti määritettyjä tuloksia MATLAB:n fitlm-komentoon. Lopuksi mallilla tutkittiin NHL-pelaajien keskivuosipalkkoja ja keskivuosipalkkoihin vaikuttavia selittäjiä.

Luvun 3.2 taulukoissa 3.3, 3.4 ja 3.6 on vertailtu laskennallisia tuloksia ja fitlm-komennon tuloksia. Taulukosta 3.4 huomataan, että selitysosuuden R^2 arvoissa on hyvin pieniä eroja, jotka todennäköisesti johtuvat pienestä pyöristyksestä fitlm-komennon laskujen aikana. Myös laskettu p_F-arvo eroaa hieman fitlm-komennon arvosta, mikä johtuu siitä, että MATLAB:n asetuksissa vastauksen tarkkuus on määritetty noin 15 desimaaliin, jolloin sillä tarkkuudella laskennallinen $p_F = 0$. Kokonaisuutta tarkasteltaessa voidaan todeta, että fitlm-komennon tulokset ovat seurausta täsmälleen samasta matematiikasta, jota teorialuvussa käsiteltiin.

Luvussa 3 tutkittiin NHL-pelaajien keskivuosipalkkoja lineaarisella regressiomallilla. Otokseksi valittiin 198 parasta pelaajaa NHL:n pistetilastosta kauden 2024-2025 ensimmäisten noin 55 ottelun otannalla. Aluksi vasteen vaihtelua tarkasteltiin kahdeksan selittäjän mallilla, mutta lopulta t-testien ja realiteettitarkastelun jälkeen malliin jäi neljä selittäjää, jotka kykenivät selittämään noin 70% keskivuosipalkkojen vaihtelusta. Neljän selittäjän poistamisen jälkeen korjattu selitysosuus laski vain noin 1%-yksikköä, joten mallin selityskyky säilyi likimain samana.

Vaikka lopullinen malli onnistui ennustamaan osan palkoista melko hyvin, löydettiin sille myös puutteita. Esimerkiksi mallissa selittäjänä on pelitilastoista ainoastaan P/GP- ja TOI/GP-tilastot. P/GP kertoo ainoastaan hyökkäyspeliosaamisesta ja tuloksen tekemisestä. NHL:ssä on kuitenkin muitakin merkittäviä osaamisalueita, kuten puolustuspelaaminen, alivoimaosaaminen, maalivahtipelaaminen ja esimerkiksi joukkueen johtajuus. Tällaisia asioita ei mallissa huomioida ollenkaan, vaikka oikeissa sopimusneuvotteluissa kaikki nämä vaikuttavat palkkaan. Malli saattoi jopa ennustaa negatiivista vuosipalkkaa nuorelle, erittäin vähän pisteitä tekevälle pelaajalle lyhyen sopimuksen tapauksessa. Tähän yhtenä syynä on otos – aineistossa mukana oli NHL:n 198 pistetilaston parasta pelaajaa, joka on vain alle 25% kaikista NHL:ssä pelanneista pelaajista kaudella 2024-2025. Mallia voisi siis kehittää ottamalla malliin mukaan kaikkien NHL-pelaajien tilastot, jolloin myös muita tilastoja voitaisiin saada malliin luotettavammin. Tällöin malli myös todennäköisesti huomioisi matalan palkkatason pelaajia paremmin. Lisäksi on hyvä pohtia, voisiko jokin epälineaarinen malli olla lineaarista mallia parempi keskipalkkojen ennustamisessa.

Rantasen ja Marnerin arvioituja 13–14 miljoonan dollarin vuosipalkkoja ei tällä mallilla saavutet-

tu, vaikka palkkalukemat olivatkin oikeansuuntaiset. Tähän yhtenä syynä varmastikin on NHL:n palkkakaton nousu, mikä nostaa pelaajien palkkoja tulevaisuudessa, ja odotukset on asetettu sen mukaisesti. Aineiston pelaajadatassa tämä palkkakaton nousu ei kuitenkaan vielä näy, vaan sopimukset on tehty vanhan palkkakaton mukaisesti.

Erityisen mielenkiintoinen asia palkkatarkastelussa on se, että NHL:n sopimukset ovat usein vuosien mittaisia. Se, että työn aineisto oli kaudelta 2024-2025, mutta osa sopimuksista on saatettu tehdä esimerkiksi seitsemän vuotta sitten, tarkoittaa, etteivät tämän kauden pelisuoritukset konkreettisesti vaikuta vuosipalkkaan ollenkaan. Tässä työssä voidaankin ajatella, että mallilla tarkastellaan käänteisesti pelisuorituksia aikaisemmin tehtyihin sopimuksiin verrattuna. Palkan asettaa aina pelaajan suoritustasolle odotuksia, ja verrattaessa pelaajan todellista palkkaa ja mallin estimaattia, voidaan tarkastella, onko pelaaja täyttänyt palkallaan asetetut odotukset. Jos haluaisi tarkastella itse palkan muodostumiseen vaikuttavia tilastoja, tulisi kerätyn aineiston selittäjien olla esimerkiksi kunkin pelaajan sopimuksentekoa edeltäneeltä kaudelta. Tällöin nähtäisiin konkreettisesti, miten pelisuoritteet vaikuttavat toteutuneeseen keskipalkkaan. Tästä aineistosta voitaisiin tehdä lineaarinen regressiomalli ja mallin avulla estimoida tulevalle kaudelle tehtäviä sopimuksia kuluvan kauden pelitilastoilla. Tällöin mallin selittäjät vaikuttaisivat konkreettisesti palkkaan, ja mallilla voitaisiin saada parempia ennusteita palkoille. Tällaisen spesifin aineiston löytäminen maksuttomana on kuitenkin haastavaa. Vaikka tehopisteitä tekevien pelaajien palkoille saatiin oikeansuuntaisia ennusteita, ei työssä käytetty aineisto sovellu edellä mainituista syistä tutkittavaan ongelmaan optimaalisesti. Sen avulla saatiin kuitenkin kattava esimerkki lineaarisen regressioanalyysin soveltamisesta, mikä oli tässä työssä keskeistä.

LÄHTEET

[1] Y. Xin ja S. Xiaogang. *Linear regression analysis: theory and computing*. World Scientific, 2009.

[2] S. M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press, 2020.

[3] C. R. Rao, C. Heumann, H. Toutenburg ja Shalabh. *Linear Models and Generalizations: Least Squares and Alternatives*. Springer Nature, 2007.

[4] P. Pere. *Tilastomenetelmien perusteet*. Tampereen yliopisto, 2020.

[5] R. H. Myers, D. C. Montgomery, G. Vining ja T. J. Robinson. *Generalized Linear Models: With Applications in Engineering and the Sciences*. John Wiley & Sons, Incorporated, 2010.

[6] S. Boyd ja L. Vandenberghe. *Introduction to Applied Linear Algebra*. Cambridge University Press, 2018.

[7] J. Rämö, T. Kaarakka, J. Häsä ja L. Oinonen. *Vektorit ja matriisit*. Tampereen yliopisto, 2021.

[8] Data100. *Ordinary Least Squares*. viitattu 3.3.2025. URL: `https://ds100.org/course-notes/ols/ols.html`.

[9] NHL. *nhl.com*. Viitattu: 12.2.2025. URL: `https://www.nhl.com/stats/skaters`.

[10] Puckpedia. *puckpedia.com*. Viitattu 12.2.2025. URL: `https://puckpedia.com/players/search?q=2024-25-leaders`.

[11] Sportsnet. *sportsnet.com*. viitattu 3.3.2025. URL: `https://www.sportsnet.ca/hockey/nhl/salary-contract-glossary/`.

LIITE A: TYÖSSÄ KÄYTETTY MATLAB-KOODI

```
% Piirretään kuva 2.1.
x = [1 2 3 4 5];
y = [1 3 6 4 6];

p = polyfit(x, y, 1);

x_ext = linspace(min(x) - 1, max(x) + 1, 100);
y_ext = polyval(p, x_ext);

figure;
scatter(x, y, 'bo', 'filled'); % Siniset pisteet datapisteitä
hold on;

plot(x_ext, y_ext, 'r-', 'LineWidth', 2); % Punainen viiva regressiosuora

x_val = 3;
y_val = [4, 6];

plot([x_val, x_val], y_val, 'g', 'LineWidth', 1.5);
hold on;


musta_pituus = 0.2; % Lisätään mustat pienet viivat
plot([x_val-musta_pituus, x_val+musta_pituus], [y_val(1), y_val(1)], ...
    'k', 'LineWidth', 1.5);
plot([x_val-musta_pituus, x_val+musta_pituus], [y_val(2), y_val(2)], ...
    'k', 'LineWidth', 1.5);

% Teksti
teksti = '\epsilon_i';
x_teksti = x_val-0.25;
y_teksti = mean(y_val);
text(x_teksti, y_teksti, teksti, 'HorizontalAlignment', 'center', ...
    'VerticalAlignment', 'middle',FontSize=12);

axis([0 6 0 8]);
grid on;
hold off;

% Otsikot
xlabel('Selittava␣muuttuja␣x', FontWeight='bold');
ylabel('Selitettava␣muuttuja␣y', FontWeight='bold');
```

```
title('Lineaarinen␣regressioanalyysi');
grid on;
legend('Havainnot', 'Sovitus', 'Location', 'Best');
hold off;
saveas(gcf,'kuva.jpg')
%-----------------------------------------------------------------

% Aineiston analysointi
format long
% Ladataan aineisto MATLAB:iin
data = readtable('NHLdata.xlsx', VariableNamingRule='preserve');

% Tallennetaan havaintoparien ja selittäjien lukumäärä
n=198;
k=8;
% Tallennetaan aineistosta muuttujaan y selitettävän muuttujan arvoiksi
% keskimääräiset vuosipalkat
y=data{:,"AAV"};

% Tallennetaan x-muuttujiin mallissa käytettävät selittäjät
x1=data{:,"P/GP"};
x2=data{:,"TOI/GP"};
x3=data{:,"Signing Age"};
x4=data{:,"Length"};
x5=data{:,"G60"};
x6=data{:,"PM"};
x7=data{:,"GWG"};
x8=data{:,"DFF"};

% Tallennetaan selittäjät matriisin sarakkeiksi
X=[x1,x2,x3,x4,x5,x6,x7,x8];

clear corr
% Lasketaan korrelaatiokertoimet selittäjien ja vasteen välille
c=corr(X,y);

% Muodostetaan matriisi A lausekkeen 2.4 mukaisesti
A=[ones(n,1),X];

% Varmistetaan, että matriisi A^T * A on kaantyva, eli det(A^T * A)
% on nollasta poikkeava ja muodostetaan kaanteismatriisi
det(A'*A);
invATA=inv(A'*A);

% Muodostetaan PNS-ratkaisu kaavan 2.5 mukaisesti, vektorin alkiot ovat
% PNS-estimaatit. (tulo (A^T * A)^( - 1) * A^T lasketaan komennolla)
x_hat=(A'*A)\A'*y;

% Lasketaan jäännösneliösumma kaavan 2.8 mukaan
SS_R=y'*y-x_hat'*A'*y;

% Lasketaan kokonaisneliösumma kaavan 2.9 mukaan
```

```
SS_T=(y-mean(y))'*(y-mean(y));

% Lasketaan selitysosuus kaavan 2.11 mukaan
R_squared=1-(SS_R/SS_T);

% Lasketaan korjattu selitysosuus kaavan 2.12 mukaan
adjR_squared=1-((n-1)/(n-k))*(1-R_squared);

% Lasketaan F-testisuure kaavan 2.13 mukaan
F_0=(R_squared/k)/((1-R_squared)/(n-k-1));

% Lasketaan p-arvo kaavan 2.13 mukaan
p_f=1-fcdf(F_0,k,n-k-1);

% Lasketaan jäännöksen otosvarianssi kaavan 2.9 mukaan
s_squared=SS_R/(n-k-1);

% lasketaan jokaiselle selittäjälle t-testisuure kaavan 2.15 mukaan ja
% p-arvo kaavan 2.17 mukaan

for i = 2:(k+1)
    % Lasketaan t-testisuureelle arvot
    t_0 = x_hat(i) / sqrt(s_squared * invATA(i, i));

    % Lasketaan p-arvot
    p_t=2*min(1-tcdf(t_0,n-k-1),tcdf(t_0,n-k-1));
end

% Lasketaan verrokkitulokset MATLAB:n fitlm-komennolla
lm1=fitlm(X, y);

% Lasketaan Rantasen ja Marnerin palkoille estimaatit
PNSest1=lm1.Coefficients.Estimate;
p1=find(contains(data.Player, 'Mikko␣Rantanen'));
x11 = data{p1, "P/GP"};
x21 = data{p1,"TOI/GP"};
x31 = 28;
x41 = 8;
x51 = data{p1,"G60"};
x61 = data{p1,"PM"};
x71 = data{p1,"GWG"};
x81 = data{p1,"DFF"};
P1AAV1 = PNSest1(1)+PNSest1(2)*x11+PNSest1(3)*x21+PNSest1(4)*x31+PNSest1(5)*
    x41+PNSest1(6)*x51+PNSest1(7)*x61+PNSest1(8)*x71+PNSest1(9)*x81;

p2=find(contains(data.Player, 'Mitchell␣Marner'));
x12 = data{p2, "P/GP"};
x22 = data{p2,"TOI/GP"};
x32 = 27;
x42 = 8;
x52 = data{p2,"G60"};
x62 = data{p2,"PM"};
```

```
x72 = data{p2,"GWG"};
x82 = data{p2,"DFF"};
P2AAV1a = PNSest1(1)+PNSest1(2)*x12+PNSest1(3)*x22+PNSest1(4)*x32+PNSest1(5)*
    x42+PNSest1(6)*x52+PNSest1(7)*x62+PNSest1(8)*x72+PNSest1(9)*x82;

% Marnerille lyhyempi sopimus
x42=4;
P2AAV1b = PNSest1(1)+PNSest1(2)*x12+PNSest1(3)*x22+PNSest1(4)*x32+PNSest1(5)*
    x42+PNSest1(6)*x52+PNSest1(7)*x62+PNSest1(8)*x72+PNSest1(9)*x82;

% Poistetaan selittäjiä eliminaatiolla
X2 = [x1,x2,x3,x4,x5,x6,x8];
fitlm(X2,y);

X3 = [x1,x2,x3,x4,x5,x6];
fitlm(X3, y);

X4 = [x1,x2,x3,x4,x6];
lm2 = fitlm(X4, y);


% Estimoidaan uudella mallilla Rantasen ja Marnerin sopimuksia
PNSest2 = lm2.Coefficients.Estimate;
% Rantanen
P1AAV2 = PNSest2(1)+PNSest2(2)*x11+PNSest2(3)*x21+PNSest2(4)*x31+PNSest2(5)*
    x41+PNSest2(6)*x61;
x42 = 8;
% Marner
P2AAV2a = PNSest2(1)+PNSest2(2)*x12+PNSest2(3)*x22+PNSest2(4)*x32+PNSest2(5)*
    x42+PNSest2(6)*x62;
% Marner lyhyempi
x42 = 4;
P2AAV2b = PNSest2(1)+PNSest2(2)*x12+PNSest2(3)*x22+PNSest2(4)*x32+PNSest2(5)*
    x42+PNSest2(6)*x62;

% Lasketaan, kuinka moni sopimuksista on vähintään 5 vuotta pitkiä
length_vec=data.Length;
count1=sum(length_vec>4);
% Lasketaan osuus kaikista pelaajista
count1/n;

% lasketaan, kuinka moni sopimuksista on vain 1 vuoden mittainen
count2=sum(length_vec<2);
count2/n;

% Poistetaan PM mallista
X5 = [x1,x2,x3,x4];
lm3 = fitlm(X5,y);

% Estimoidaan Rantasen ja Marnerin palkat ilman PM-tilastoa
PNSest3 = lm3.Coefficients.Estimate;
P1AAV3 = PNSest3(1)+PNSest3(2)*x11+PNSest3(3)*x21+PNSest3(4)*x31+PNSest3(5)*
```

```
    x41;
x42 = 8;
P2AAV3a = PNSest3(1)+PNSest3(2)*x12+PNSest3(3)*x22+PNSest3(4)*x32+PNSest3(5)*
    x42;
x42 = 4;
P2AAV3b = PNSest3(1)+PNSest3(2)*x12+PNSest3(3)*x22+PNSest3(4)*x32+PNSest3(5)*
    x42;

% Estimoidaan palkka Rasmus Kuparille
x13=0.14;
x23=10.33;
x33=24;
x43=2;
P3AAV = PNSest3(1)+PNSest3(2)*x13+PNSest3(3)*x23+PNSest3(4)*x33+PNSest3(5)*x43
    ;
x43=8;
P3AAV = PNSest3(1)+PNSest3(2)*x13+PNSest3(3)*x23+PNSest3(4)*x33+PNSest3(5)*x43
    ;

% Estimoidaan Patrick Kanelle palkka siten, että
% Signing Age + Length <= 38
p4ind=find(contains(data.Player, 'Patrick␣Kane'));
x14=data{p4ind, 'P/GP'};
x24= data{p4ind,"TOI/GP"};
x34=36;
x44=2;
P4AAV3 = PNSest3(1)+PNSest3(2)*x14+ PNSest3(3)*x24+PNSest3(4)*x34+PNSest3(5)*
    x44;
```