HYÖKKÄYS ON PARAS PUOLUSTUS? NBA-JOUKKUEEN VOITTOMÄÄRÄN ENNUSTAMINEN TILASTOLLISEN ANALYYSIN AVULLA

Lappeenrannan–Lahden teknillinen yliopisto LUT

Kauppatieteiden kandidaatintutkielma

Liiketoiminta-analytiikka

2024

Kalle Visuri

Tarkastaja: Tutkijaopettaja Maija Hujala

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT-kauppakorkeakoulu

Kauppatieteet

Kalle Visuri

Hyökkäys on paras puolustus? NBA-joukkueen voittomäärän ennustaminen tilastollisen analyysin avulla

Kauppatieteiden kandidaatintutkielma

2024

31 sivua, 4 kuvaa, 3 taulukkoa ja 9 liitettä

Tarkastaja: Tutkijaopettaja Maija Hujala

Avainsanat: National Basketball Association, NBA, koripallo, data-analytiikka, urheiluanalytiikka

Tässä kandidaatintutkielmassa tutkittiin National Basketball Association koripalloliigan joukkueiden suoriutumista hyödyntäen historiallisia ottelutilastoja kausilta 2013/2014–2017/2018. Tutkimuksessa hyödynnettiin hyökkäys- ja puolustustilastoja, joilla pyrittiin selittämään joukkueen voittojen määriä NBA:n runkosarjassa. Lisäksi tutkimuksessa käytettiin regressiomallia, jolla pyrittiin ennustamaan yksittäisten NBA-joukkueiden voittojen määriä kaudella 2018/2019.

Tutkimuksen tulokset osoittivat, että NBA-joukkueen suoriutumista pystyään hyvin selittämään historiallisten ottelutilastojen avulla. Erityisesti joukkueen korien tekemisen tehokkuus sekä joukkuetta vastaan tehtyjen korien tehokkuus osoittautuivat merkittävimmiksi tekijöiksi joukkueen voittojen taustalla. Korien tekemisen tehokkuus on ollut merkittävä tekijä myös aikaisemmassa tutkimuksessa, jonka lisäksi myös levypallojen merkittävä vaikutus oli linjassa aikaisemman tutkimuksen kanssa. Tutkimuksen regressioon perustuva ennustemalli onnistui suhteellisen hyvin ennustamaan joukkueiden voittoja NBA-kaudella 2018/2019.

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT Business School

Business Administration

Kalle Visuri

Offense is the best defense? Predicting the number of wins for an NBA team through a statistical analysis

Bachelor's thesis

2024

31 pages, 4 figures, 3 tables and 9 appendices

Associate professor Maija Hujala

Keywords: National Basketball Association, NBA, basketball, data analytics, sports analytics

In this bachelor's thesis, the performance of National Basketball Association teams was examined using historical match statistics from seasons 2013/2014-2017/2018. The study used offensive and defensive statistics to explain the numbers of wins in regular season for teams in the NBA. Additionally, a regression model was used to predict the number of wins for individual NBA teams in the 2018/2019 season.

The results of the study indicated that the performance of NBA teams can be explained using historical statistics from games played. Particularly, the efficiency of scoring and the efficiency of preventing opponent scoring proved to be significant factors behind a team's number of wins. In previous studies, the efficiency of scoring has been significant as well, and the significant impact of rebounds was also consistent with previous studies. The predictive model based on regression was relatively successful in predicting the wins of teams for the NBA 2018/2019 season.

Sisällysluettelo

Liitteet

1 Johdanto

Digitalisaation myötä datan keräämisestä ja hyödyntämisestä organisaatioiden päätöksenteossa on kasvanut megatrendi, jota on vaikea olla huomaamatta nykypäivänä. Rashedin (2020, 119) mukaan dataohjautuva päätöksenteko ei ole tulevaisuudessa osalle yrityksistä ainoastaan mahdollisuus vaan siitä muodostuu pakollista, jos haluaa menestyä kilpailussa muita vastaan. Eri organisaatiot siis pyrkivät saavuttamaan kilpailuetua muihin analytiikan avulla, mikä pätee yhtä lailla urheiluorganisaatioihin. Nykypäivänä myös urheilussa käytetään analytiikkaa päätöksenteon tukena paljon, ja urheiluanalytiikka voidaan karkeasti määritellä tilastollisen data-analyysin hyödyntämisenä yksilön sekä joukkueen suorituskyvyn parantamiseksi (Baumer, Matthews & Nguyen 2023). Lisäksi urheiluorganisaatiot käyttävät analytiikkaa esimerkiksi palkkakaton optimointiin, pelaajien hankintaan, sekä otteluiden lippujen hinnoitteluun ja asiakassuhteiden hallintaan (Mondello & Kamke 2014).

Analytiikan hyödyntäminen urheilussa on kasvanut viime vuosikymmenten aikana kasvanut merkittävästi, mikä näkyy esimerkiksi aiheeseen liittyvän tieteellisen tutkimuksen määrän kasvuna. Monissa lajeissa kehittyneiden mittausmenetelmien myötä saatavilla oleva datan määrä on valtava, mikä on johtanut monimutkaistenkin datarakenteiden hyödyntämiseen päätöksenteossa. Tällaisen datan avulla pystytään luomaan tarkempia tilastollisia malleja, joita voidaan hyödyntää yksilön tai joukkueen suorituskyvyn parantamiseen. (Groll, Manisera, Schauberger & Zuccolotto 2018; Sarlis & Tjortjis 2020) Kehittyneen data-analyysin hyödyntäminen urheilussa näkyy konkreettisesti esimerkiksi koripallossa, jossa joukkueet heittävät nykypäivänä enemmän kolmen pisteen heittoja sekä vähemmän pitkän etäisyyden kahden pisteen heittoja kuin aikaisemmin (Baumer et al. 2023).

Vaikka urheiluanalytiikan hyödyntäminen ja tieteellisen tutkimuksen määrä aiheeseen liittyen on kasvanut, Abezan, O'Reillyn, Nadeaun ja Abdourazakoun (2021) mukaan asiaa tulisi tutkia enemmän. Koska dataa kerätään urheilussa valtavia määriä, se tarjoaa uniikin alustan tieteelliselle tutkimukselle (Morgulev, Azar & Lidor 2018). Urheiluanalytiikkaan liittyvä tutkimus onkin tärkeää urheiluorganisaatioille, jotta ne saavuttaisivat kilpailuetua ja menestyisivät paremmin muita organisaatioita ja niiden joukkueita vastaan. Bradbury (2019) esittää, että esimerkiksi koripallossa joukkueen menestyminen kentällä on tutkittu olevan

positiivisesti yhteydessä organisaation liikevaihtoon, mikä lisää erityisesti seurojen omistajien sekä muiden rahoittajien kiinnostusta joukkueen menestymiseen urheiluanalytiikan aikakaudella.

1.1 Tutkielman tavoitteet ja tutkimuskysymykset

Nykypäivänä National Basketball Association (NBA) -koripallosarjassa tilastoidaan erilaisia tilastorivejä jopa 2850 kappaletta, mikä on seurausta koripalloanalytiikan kasvaneesta hyödyntämisestä sarjassa ja sen seuraorganisaatioissa (Colás 2020). NBA:ssa otteludatan kerääminen on edistyksellisellä tasolla, ja jokaisella peliareenalla on käytössä esimerkiksi kamerajärjestelmä, joka mittaa lukuisia tapahtumia pelin aikana (Sarlis & Tjortjis 2020). Morgulev et al (2018) mukaan NBA:ssa dataa kerätäänkin niin valtava määrä, että kaikkea ei edes ole mahdollista hyödyntää järkevällä tavalla päätöksenteossa.

NBA-joukkueiden suoriutumisesta ja siihen vaikuttavista pelillisistä tekijöistä on tehty useita tieteellisiä tutkimuksia. Teramoto ja Cross (2010) esittävät tutkimuksessaan, että puolustus ja hyökkäys ovat yhtä tärkeitä elementtejä joukkueen menestymiseen runkosarjassa, mutta puolustuksen merkitys kasvaa kauden edetessä pudotuspeleihin. Ottenin ja Millerin (2015) tutkimuksessa merkittävin tekijä joukkueen menestymisessä oli pelitilanneheittojen onnistumisprosentti, jonka tutkijat määrittelivät koostuvan kahden ja kolmen pisteen heitoista. Li, Wang ja Li (2021) ovat kuitenkin omassa tutkimuksessaan jakaneet pelitilanneheitot erikseen kahden ja kolmen pisteen heitoiksi, joista kolmen pisteen heittojen onnistumisprosentti on ollut merkittävin muuttuja joukkueen menestymisessä.

Tämän tutkielman tavoitteena on selvittää, miten hyvin koripallosarja NBA:n historiassa kerätyt joukkueiden ottelutilastot selittävät joukkueiden suoriutumista. Lisäksi halutaan selvittää, miten historiallisten ottelutilastojen avulla pystytään ennustamaan joukkueen suoriutumista. Tätä varten on oleellista tarkastella, miten hyvin joukkueiden historialliset hyökkäystä ja puolustusta ilmentävät ottelutilastot pystyvät selittämään joukkueiden suoriutumista, jota mitataan tässä tutkimuksessa voittojen kokonaismäärällä runkosarjassa. Tutkimusongelmaan pyritään vastaamaan seuraavan päätutkimuskysymyksen avulla:

> *Kuinka historiallisen otteludatan avulla voidaan selittää NBA-joukkueen suoriutumista?*

Päätutkimuskysymyksen lisäksi pyritään saamaan vastaus seuraaviin alatutkimuskysymyksiin:

> *Mitkä ovat merkittävimpiä tilastollisia tekijöitä, jotka vaikuttavat NBA-joukkueen suoriutumiseen*?
> *Kuinka hyvin historialliseen otteludataan perustuva tilastollinen malli pystyy ennustamaan joukkueiden suoriutumista tulevilla kausilla*?

Tutkimus toteutetaan määrällisenä eli kvantitatiivisena tutkimuksena, jossa aineistona käytetään NBA:n keräämiä ottelutilastoja viiden vuoden tarkasteluperiodilla kausilta 2013/2014–2017/2018. Analyysimenetelmänä käytetään lineaarista regressioanalyysia. Regressioanalyysissä hyödynnetään erilaisia koripallossa mitattavia ottelutilastoja selittäjinä, ja selitettävänä muuttujana on joukkueen saavuttama voittojen määrä runkosarjassa yhden kauden aikana. NBA:n runkosarjassa jokainen joukkue pelaa 82 ottelua, eli yhden joukkueen runkosarjan aikana saavuttama voittomäärä on 0–82 ottelua. Runkosarjassa voitettujen otteluiden määrä kertoo paljon joukkueen suoriutumisesta muihin joukkueisiin verrattuna ja on tärkeä muuttuja erityisesti pudotuspelien näkökulmasta. Joukkueet, jotka ovat runkosarjassa voittaneet eniten otteluita, aloittavat pudotuspelit heikointa pudotuspeleihin päässyttä joukkuetta vastaan. (NBAstuffer 2023) Lopulta tilastollisen mallin suorituskykyä testataan kauden 2018/2019 dataan syöttämällä malliin tarvittavat tilastot jokaisesta joukkueesta kyseiseltä kaudelta ja testataan, kuinka paljon joukkueen voittojen kokonaismäärä runkosarjassa eroaa toteutuneesta.

1.2 Tutkimuksen rajaukset

Tutkimuskohteeksi on valittu NBA ja sen joukkueet, koska kyseessä on tunnetuin koripallosarja maailmassa, ja analytiikka on todella merkittävässä roolissa sarjan ja sen seuraorganisaatioiden toiminnassa (Mandić, Jakovljević, Erčulj & Štrumbelj 2019; Morgulev et al. 2018). Tutkimustuloksia on kuitenkin epämielekästä verrata muihin koripallosarjoihin, koska sarjojen välillä on vaihtelua joissain säännöissä, kuten peliajassa. Tutkimuksessa hyödynnetään eri kausilta jokaisen NBA-joukkueen tilastoja, koska se antaa paremman pohjan mallille ja parantaa sen luotettavuutta.

Tutkimuksessa käytettävien vuosien valintaperusteena on käytetty viiden peräkkäisen kauden hyödyntämistä mallin muodostamiseen, mutta koronapandemian vuoksi kausien 2019/2020 sekä 2020/2021 tilastoja ei voida käyttää, koska näillä kausilla pelattuja otteluita on COVID-19-pandemian vuoksi vähemmän, ja täten ne eivät ole vertailukelpoisia muihin kausiin nähden (NBA 2020a; NBA 2020b). Lisäksi näillä kausilla pelejä on pelattu tyhjille katsomoille, ja esimerkiksi kotietujen puuttuessa kaudet ovat olleet poikkeuksellisia normaaliin verrattuna (Gong 2022). Edellä mainituista syistä, tarkasteluperiodiksi on valittu NBA-kaudet 2013/2014 – 2017/2018.

1.3 Tutkielman rakenne

Tutkielma koostuu yhteensä viidestä pääluvusta. Johdannon jälkeen toisena lukuna on teoriaosuus, jossa käydään läpi data- ja urheiluanalytiikkaa käsitteinä, sekä koripallon ja erityisesti NBA:n tilastojen hyödyntämiseen liittyvää aikaisempaa tutkimusta. Kolmannessa luvussa esitellään käytettävä tutkimusaineisto sekä -menetelmät. Lisäksi luvussa käydään läpi, miten dataa on käsitelty tutkimusta varten. Neljännessä luvussa esitellään tutkimustulokset, ja testataan luodun ennustemallin toimivuutta eri otoksessa. Viimeisessä eli viidennessä luvussa käydään tutkielma tiivistetysti läpi yhteenvedon muodossa sekä esitellään johtopäätökset. Lisäksi pohditaan potentiaalisia jatkotutkimusaiheita, ja arvioidaan tutkimuksen luotettavuutta sekä yleistettävyyttä.

2 Kirjallisuuskatsaus

Tässä luvussa käsitellään aluksi data-analytiikkaa, jonka jälkeen tutkitaan tarkemmin urheiluanalytiikkaa ja sen erityispiirteitä. Urheiluanalytiikasta syvennytään edelleen analytiikan hyödyntämiseen koripallossa yleisesti, sekä erityisesti NBA:ssa. Tämän lisäksi esitellään aikaisempaa tutkimusta NBA:sta tehdyistä tutkimuksista, joissa hyödynnetään historiallisia ottelutilastoja.

2.1 Data-analytiikka

Data-analytiikkaa on haastavaa määritellä yksikäsitteisesti, ja esimerkiksi termit analytiikka sekä data-analytiikka on määritelty hyvin samankaltaisesti eri tutkimuksissa. Tukey (1962) määrittelee data-analyysin koostuvan eri keinoista kerätä dataa, tilastollisista menetelmistä, joilla dataa analysoidaan sekä eri tekniikoista, joilla tuloksien pohjalta voidaan tehdä johtopäätöksiä. Angelov, Gu, ja Kangin (2017) määrittelevät data-analytiikan prosessiksi, jossa tilastollisia menetelmiä hyödynnetään datan kuvailemiseen, havainnollistamiseen sekä arvioimiseen. Runklerin (2017) mukaan data-analytiikka voidaan yksinkertaisesti määritellä tietokoneavusteisesti suurten datamäärien analysoimiseksi päätöksentekoa varten. Davenport ja Kim (2013, 3) taas määrittelevät laajemmin analytiikan datan kattavana käyttönä, tilastollisten ja kvantitatiivisten eli määrällisten menetelmien hyödyntämisenä, selittävien ja ennustavien mallien luomisena sekä faktoihin perustuvana päätöksentekona.

Data-analytiikan eri määritelmissä on yleisesti kuitenkin ytimessä datan hyödyntäminen päätöksentekoa varten soveltamalla kerättyyn dataan erilaisia tilastollisia menetelmiä. Vaikka data-analytiikka ei ole käsitteenä uusi, sen käyttö on yleistynyt vasta 2000-luvun alkupuolella, kun saatavilla olevan datan määrä on kasvanut sekä tietokoneet ovat kehittyneet tehokkaammiksi (Runkler 2017).

2.1.1 Big data

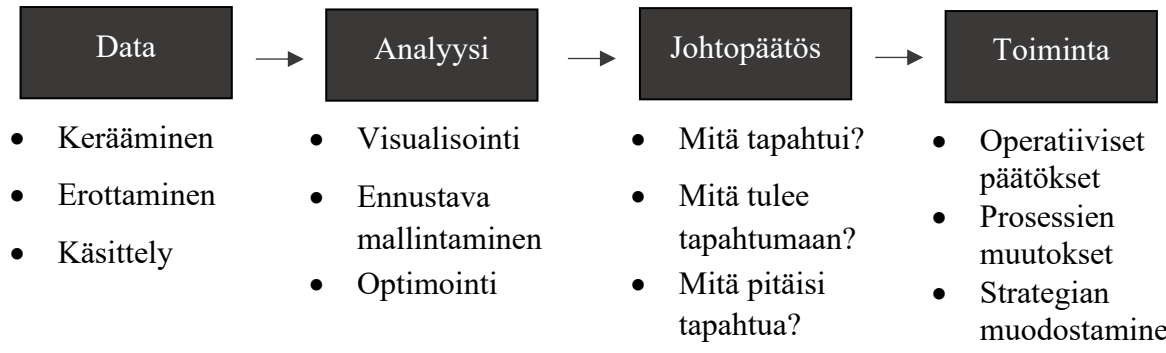
Nykypäivänä saatavilla olevan datan määrä on kasvanut todella isoksi, ja tämän kehityskulun myötä on kehitetty käsite big data. Big datalla tarkoitetaan valtavaa, monimutkaista ja reaaliaikaista dataa, joka vaatii erittäin kehittyneitä datanhallinta- ja analysointitekniikoita. (Dubey, Gunasekaran, Childe, Blome & Papadopoulos 2019) Big data on Hurwitzin, Nugentin, Halperin ja Kaufmanin (2013) mukaan yksi ajankohtaisimmista teknologiatrendeistä, jonka avulla liiketoimintaorganisaatiot pystyvät koosta tai toimialasta riippumatta kehittämään liiketoimintaansa paremmaksi. Big data on kuitenkin aiheuttanut mahdollisuuksien lisäksi myös haasteita niin organisaatioille kuin tutkijoille (Chen & Zhang 2014).

Useat tutkijat, kuten Ali, Qadir, Rasool, Sathiaseelan, Zwitter ja Crowcroft (2016), Tsai, Lai, Chao ja Vasilakos (2015), Qin (2014) sekä Li, Thomas ja Osei-Bryson (2016) ovat tutkimuksissaan keskustelleet big datalle kolmesta keskeisestä ominaisuudesta, jotka erottavat sen normaalista datasta. Edellä mainittujen tutkimusten perusteella ensimmäinen ominaispiirre big datalle on volyymi, eli datan suuri määrä. Tsain et al. (2015) mukaan ongelmana datan suuressa määrässä on nykyisten tietokoneiden laskentateho, eli riittävän suuria datasettejä ei pystytä käsittelemään yhdellä normaalilla tietokoneella, jonka lisäksi monet datankäsittelyyn tehdyt analyysimenetelmät eivät toimi suoraan tällaisten erittäin suurien datasettien analysoimiseen.

Toisena ominaisuutena tutkimuksissa nostettiin esiin big datan nopeus. Qin (2014) korostaa nopeuden merkitystä sekä haastetta aikaherkissä tilanteissa, joissa dataa täytyisi pystyä hyödyntämään reaaliaikaisesti. Tutkimuksissa big datan kolmanneksi ominaisuudeksi on määritetty datan moninaisuus. Ali et. al (2016) mukaan moninaisuudella kuvataan eri lähteistä tulevaa dataa, eli datatyypit voivat vaihdella ja olla hyvinkin erilaisia. Nämä haasteet vaativat Li et al. (2016) mukaan skaalautuvia analyysiratkaisuja, jotta big dataa voitaisiin hyödyntää päätöksenteossa. Tietokoneiden ja parempien analyysimenetelmien kehityksen myötä big data voi muodostua monelle organisaatiolle erittäin arvokkaaksi tulevaisuudessa. Erityisesti yritykset, jotka keräävät paljon asiakasdataa, käsittelevät todella suuria datasettejä, joista voi löytyä erittäin arvokasta tietoa.

2.1.2 Data-analytiikka prosessina

Liberatoren ja Luon (2010) mukaan data-analytiikka pystytään prosessina jakamaan neljään osaan, jotka on esitelty kuvassa 1. Tutkijoiden mukaan prosessi alkaa raa'an datan keräämisestä, erottamisesta sekä käsittelystä. Kuten edeltävässä kappaleessa käytiin läpi, suurten datamassojen vuoksi datan erottaminen ja käsittely on tärkeä vaihe, jotta olennainen data saadaan erotettua sekä järjesteltyä analyysivaihetta varten (Tsai et al. 2015). Tämä myös auttaa tietokoneiden laskentakyvyn riittämiseen, sekä tilastollisten testien parempaan toimivuuteen.



Kuva 1. Data-analyyttinen prosessi, mukaillen Liberatore & Luo (2010)

Prosessin toisessa vaiheessa käytetään erilaisia analyyttisia lähestymistapoja sekä tekniikoita datan tutkimiseen sekä arviointiin. Yksi yleisistä analyysimenetelmistä on datan visualisointi, eli kuvaileva analyysi. (Liberatore & Luo 2010) Visualisoinnin tarkoituksena on Tsai et al. (2015) mukaan hyödyntää erilaisia kuvaajia, kaavioita sekä taulukoita, jotta tietoa voidaan esittää intuitiivisemmin sekä tehokkaammin usein taulukkomuotoiseen raakadataan verrattuna. Liberatoren ja Luon (2010) prosessissa analyysikeinona voidaan käyttää visualisoinnin lisäksi ennustavaa analyysia, jossa tarkoituksena on luoda ennustemalleja datan avulla. Usein ennustavassa analyysissä käytetään hyödyksi tilastollisia menetelmiä, kuten regressioanalyysia, koneoppimiseen perustuvia malleja, klusterianalyysia tai tekoälyihin liittyvät tekniikoita, kuten neuroverkkoja (Mohbey, Pandey & Rajput 2020). Liberatore ja Luo (2010) nostivat myös yhdeksi analyysikeinoksi optimoinnin, jossa pyritään löytämään datan perusteella optimaalinen ratkaisu annetun datan sekä vallitsevien rajoitusten puitteissa.

Prosessin kolmannessa vaiheessa on tarkoitus muodostaa johtopäätöksiä suoritetun analyysin perusteella. Edellä mainituilla keinoilla on kaikilla erilaisia tuloksia, joista johtopäätöksiä pystytään muodostamaan. (Liberatore & Luo 2010) Yleisesti kuvailevassa analyysissä esitetään tietoa siitä, mitä on tapahtunut menneisyydessä. Ennustavassa mallinnuksessa keskitytään siihen, mitä tapahtuu tulevaisuudessa, jos jokin trendi jatkuu tai kun tietyt olosuhteet täyttyvät. (Davenport & Harris 2017, 30–31) Optimointia hyödynnetään erityisesti teollisuudessa tuotanto- ja logistiikkaprosesseissa, kun esimerkiksi voiton määrä halutaan maksimoida (Tang & Meng 2021). Nämä analyysikeinot täydentävät hyvin toisiaan, ja antavat päätöksentekijälle hyvän kokonaiskuvan, jonka perusteella tehdä datavetoisia päätöksiä.

Yksinään johtopäätöksillä ei ole suurta arvoa, jos niitä ei pystytä kääntämään konkreettisiksi toimiksi. Liberatoren ja Luon (2010) prosessin viimeisessä vaiheessa johtopäätöksiä on tarkoitus hyödyntää esimerkiksi organisaation operatiivisissa päätöksissä, prosessien muutoksissa tai strategian muodostamisessa. Davenport ja Kim (2013, 16) nostavat kuitenkin esille, että datavetoinen päätöksenteko vie usein aikaa sekä resursseja, jolloin datavetoista päätöksentekoa kannattaa hyödyntää toistuvissa päätöksentekotilanteissa. Täten yksinkertaisempia ja kertaluontoisia päätöksiä varten ei ole järkevää alkaa keräämään dataa ja suorittaa dataanalyyttista prosessia.

2.2 Urheiluanalytiikka

Urheiluanalytiikka ja sen hyödyntäminen on kasvanut viimeisen 30 vuoden aikana niin urheiluorganisaatioissa kuin tieteellisessä kirjallisuudessa (Baumer et al. 2023). Watanabe, Shapiro ja Drayer (2021) esittävät, että urheiluanalytiikka on lähtenyt liikkeelle 1950-luvulla, kun yhdysvaltalaiset baseball-joukkueet alkoivat systemaattisesti keräämään otteludataa pelaajiensa suorituksista. Samaan aikaan 1950-luvun Englannissa, Charles Reep alkoi laskemaan jalkapallopeleissä maaliin johtaneiden syöttöjen määriä, minkä seurauksena Reep ja Benjamin (1968) tekivät yhden ensimmäisistä urheiluanalytiikkaan liittyvistä tieteellisistä tutkimuksista. Tutkimuksessa kehitettiin niin kutsuttu ”pitkän pallon” pelityyli, joka vuosikymmeniä leimasi englantilaista jalkapalloa (Morgulev et al. 2018).

Morgulev et al. (2018) mukaan tietokoneiden kehittymisen myötä myös urheiluanalytiikassa on alettu keräämään laajemmin dataa sekä hyödyntämään monimutkaisempia analyysimenetelmiä päätöksien tekemiseen. Vaikka urheiluanalytiikassa on perinteisesti ollut kyse

yksilön tai joukkueen suorituskyvyn parantamisesta datan ja tilastollisten menetelmien avulla, käyttävät urheiluorganisaatiot analytiikkaa myös kentän ulkopuolella kasvattaakseen esimerkiksi katsojamääriä joukkueen peleissä (Baumer et al. 2023; Watanabe et al. 2021). Davenport (2014) kuitenkin huomauttaa, että yleisesti ottaen urheiluorganisaatiot ovat analytiikan hyödyntämisessä paljon muita toimialoja jäljessä, koska usein niillä ei ole varaa yhtä suuriin analytiikkaosastoihin suurempiin organisaatioihin verrattuna.

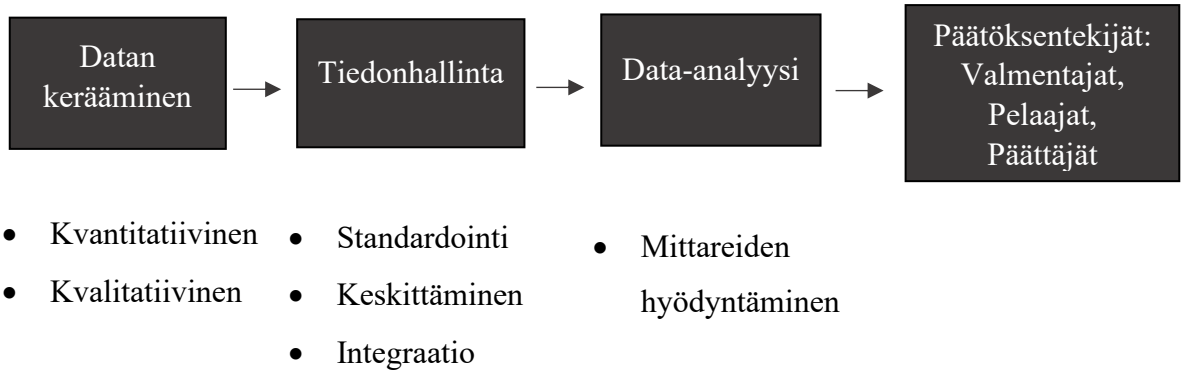
2.2.1 Urheiluanalytiikan hyödyntäminen eri lajeissa

Urheiluanalytiikan läpimurtona pidetään Michael Lewisin (2003) julkaiseman teoksen ”Moneyball: the art of winning an unfair game” jälkeistä ajanjaksoa (Elitzur 2020). Kyseessä on tapaustutkimus, joka kertoo Major League Baseball (MLB)-sarjan Oakland Athletics-joukkueesta. Joukkue onnistui vuoden 2002 kaudella data-analytiikkaa hyödyntämällä suoriutumaan erittäin hyvin muita joukkueita vastaan siitä huolimatta, että joukkueella oli sarjan toiseksi alhaisin pelaajabudjetti käytettävissään. Joukkueen valmentajat ja päätöksentekijät hyödynsivät erityisesti pelaajien tilastoja sekä ottelutilastoja ja onnistuivat optimoimaan pienen budjettinsa menestyvän joukkueen rakentamiseen. Vaikka joukkue ei voittanut mestaruutta, se onnistui kuitenkin saavuttamaan kilpailuetua ja pärjäämään muita paremman pelaajabudjetin omaavia joukkueita vastaan. (Lewis 2003) Teoksen julkaisun jälkeen monet muut MLB-joukkueet alkoivat myös hyödyntämään data-analytiikkaa, mikä johti Oakland Athletics-joukkueen kilpailuedun katoamiseen (Duquette, Cebula & Mixon 2019). Elitzur (2020) mukaan nykypäivänä jokainen MLB-joukkue hyödyntää analytiikkaa, mikä kertoo urheiluanalytiikan yleisestä käytöstä erityisesti baseballissa.

MLB:n lisäksi lukuisissa muissa lajeissa hyödynnetään data-analytiikkaa. Erityisesti Yhdysvalloissa useissa ammattiurheilulajeissa, kuten koripallossa NBA:ssa, jääkiekossa National Hockey League (NHL)-sarjassa ja amerikkalaisessa jalkapallossa National Football League (NFL)-sarjassa analytiikka on merkittävä osa seuraorganisaatioiden päivittäistä toimintaa (Abeza et al. 2021; Davenport 2014). Lisäksi data-analytiikkaa hyödynnetään myös eurooppalaisissa jalkapalloseuroissa, vaikka jalkapallo ei ole yhtä tilastorikas laji esimerkiksi baseballiin verrattuna (Schumaker, Solieman & Chen 2010).

2.2.2 Urheiluanalyyttinen prosessi

Morgulev et al. (2018) ovat tutkimuksessaan esittäneet urheiluanalyyttisen prosessin, jossa on nähtävissä samankaltaisuuksia Liberatoren ja Luon (2010) data-analyyttisen prosessin kanssa. Kuvan 2 mukaisesti Morgulev et al. (2018) esittää, että prosessissa lähdetään liikkeelle kvantitatiivisen eli määrällisen ja kvalitatiivisen eli laadullisen datan keräämisestä. Urheiluanalytiikassa hyödynnetään usein kvantitatiivisena datana historiallisia ottelutilastoja, jonka lisäksi kvalitatiivisena datana hyödynnetään paljon esimerkiksi videoita otteluista ja urheilijoiden suorituksista (Schumaker et al. 2010).



Kuva 2. Urheiluanalytiikan viitekehys, mukaillen Morgulev et al. (2018)

Prosessin toisessa vaiheessa Morgulev et al. (2018) korostavat tiedonhallinnan merkitystä. Urheiluanalyyttisessä prosessissa on tärkeää keskittää eri lähteistä tulevaa tietoa ja integroida kerätty data yhtenäiseksi kokonaisuudeksi, jolloin kokonaiskuva on helpompi hahmottaa, sekä päätöksenteko on helpompaa (Patel, Shah & Shah 2020). Koska urheilussa kerätään valtavia määriä dataa eri lähteistä, Tan (2019) esittää, että big datalla on nykypäivänä erittäin keskeinen rooli urheiluorganisaatioiden päätöksenteossa ja kilpailumenestyksen saavuttamisessa.

Prosessin kolmannessa vaiheessa hyödynnetään dataa erilaisten mittarien ja mallien luomiseen. Monissa lajeissa on jo vakiintuneita mittareita, jotka kuvaavat pelaajan tai joukkueen suorituskykyä eri osa-alueilla, kuten hyökkäyksessä tai puolustuksessa. (Morgulev et al. 2018) Vaikka vakiintuneita mittareita on paljon olemassa, Schumaker et al. (2010) mukaan monia urheilussa käytettäviä mittareita voidaan kuitenkin hyödyntää väärin, tai pahemmassa

tapauksessa mittarit eivät tosiasiallisesti edes mittaa suorituskykyä oikealla tavalla, kun lopullisena tavoitteena on suoriutua paremmin muita vastaan. Mittareita ja dataa hyödynnetään urheilussa erityisesti selite- ja ennustemallien rakentamiseen, jotka tarjoavat arvokasta tietoa esimerkiksi joukkueiden valmentajille ja pelaajille (Patel et al. 2020). Päätöksenteko siis perustuukin historiallisen suoriutumisen dataan, johon sovelletaan tilastollisia menetelmiä ja täten löydetään potentiaalisesti arvokasta tietoa omista pelaajista sekä joukkueesta.

2.3 Data-analytiikka NBA:ssa

NBA:ssa data-analytiikan hyödyntäminen on lähtenyt liikkeelle kaudelta 1995/1996, kun sarjassa otettiin 16 joukkueen käyttöön Advanced Scout-ohjelmisto. Ohjelmisto kerää otteluista erilaisia pelitilastoja, kuten heittoyrityksiä, ja luo peleistä dataa, jossa nämä eri pelitilastot ovat muuttujina. Pelien jälkeen tilastot ovat näkyvissä joukkueiden valmentajille. (Morgulev et al. 2018) Tämän lisäksi Bhandari, Colet, Parker, Pines, Pratap ja Ramanujam (1997) mukaan ohjelmisto luo datan perusteella päätöksenteon tueksi malleja, ja pyrkii esimerkiksi löytämään optimaalisia viisikoita joukkueelle perustuen kenttäpelaajien heittoprosentteihin eri kentällisissä, eli kentällä olevien pelaajien eri yhdistelmissä.

Chartier (2017) mukaan datan keruuta NBA:ssa kiihdytti vuonna 2013 tehty linjaus siitä, että jokaisella areenalla otetaan käyttöön SportVU-kamerajärjestelmä. Järjestelmä on kuusikamerainen, ja nämä kamerat seuraavat pelaajien ja pallon liikettä kentällä tehokkaasti ottamalla 25 kuvaa sekunnissa. Tämä tarjoaa monimutkaisemman analyysin pelaajien ja pallon liikkeestä, jonka perusteella valmentajat voivat mallintaa parempia hyökkäystaktiikoita, mikä johtaa suurempien pistemäärien tekemiseen. (Wu & Bornn 2017) Ottelutilastoja ja videodataa yhdistämällä pystytäänkin analysoimaan esimerkiksi mistä kohtaa hyökkäysaluetta joukkue tai pelaaja onnistuu korinteossa parhaiten. Vuonna 2016 NBA päätti vaihtaa tilastojen toimittajaa, ja teki sopimuksen yhtiöiden Second Spectrum ja Sportradar kanssa NBA-datan keräämisestä ja jakamisesta jopa yli 80:een eri maahan esimerkiksi vedonlyöntiyhtiöille. Second Spectrumin koneoppimista hyödyntävä kamerajärjestelmä korvasi SportVU:n optisen seurannan kaudella 2017/2018. (NBA 2016) Vuonna 2023 NBA ja Second Spectrum sopivat kumppanuutensa jatkosta ja uusien teknologien kehittämisestä, jolloin Sport Spectrumista tuli myös virallinen NBA:n koripalloanalytiikan tarjoaja (NBA 2023b).

Erityisesti joukkueiden valmentajat hyödyntävät paljon perinteisiä ottelutilastoja päätöksenteossa pelien aikana, ja pelien jälkeen valmistautuessa seuraavaan peliin. Lisäksi pidemmällä aikavälillä valmentajat seuraavat joukkueen ja sen pelaajien suoriutumista tilastojen valossa. (Mandić et al. 2019) Valmentajien lisäksi Song ja Shi (2020) mukaan myös pelaajat, fanit ja media ovat kiinnostuneita erilaisten lopputulosten kuten yksittäisten pelien ennustamisesta. Thabtah, Zhang ja Abdelhamid (2019) mukaan urheiluvedonlyönnin suuret rahavirrat osoittavat, että erilaisten ennusteiden suosio on kasvanut 2010-luvulla. Monet yksityishenkilöt hyödyntävät vedonlyöntipäätöksissään historiallisia tilastoja ja analyytikoiden tekemiä ennusteita esimerkiksi otteluiden lopputuloksista (Houghton, Nowlin & Walker 2019).

NBA-peleistä kerättyä dataa on julkisesti saatavilla NBA:n omalla statistiikkasivustolla, josta esimerkiksi joukkuedataa on saatavilla kaudesta 1996/1997 lähtien. Sivustolla on saatavilla pelaajista ja joukkueista edistynyttä dataa, kuten yhdistelmätilastoja, jotka kuvaavat paremmin pelaajan tai joukkueen suoriutumista puolustuksessa tai hyökkäyksessä. (NBA 2023a) NBA:n omien statistiikkasivustojen lisäksi dataa on julkisesti saatavilla useista eri lähteistä, esimerkiksi ESPN (2023), Basketball-Reference (2023) ja Yahoo Sports (2023) tarjoavat niin yksittäisten pelaajien kuin joukkueiden tilastoja erittäin kattavasti (Kubatko, Oliver, Pelton & Rosenbaum 2007).

2.4 Aikaisempi tutkimus NBA – joukkueiden suoriutumisesta

NBA:sta on tehty lukuisia data-analytiikkaa ja tilastollisia menetelmiä soveltavia tutkimuksia liittyen joukkueiden ja yksilöiden suorituskykyyn. Voittavien ja häviävien joukkueiden erona keskeisinä tekijöinä on Cabarkapa, Deane, Fry, Jones, Cabarkapa, Philipp, Yu ja Abián-Vicén (2022) tutkimuksessa löydetty pelitilanneheittojen onnistumisprosentti sekä puolustuslevypallojen määrä. Otten ja Miller (2015) tutkimus tukee Caparkaba et al. (2022) löydöksiä, pitäen pelitilanneheittojen onnistumisprosenttia merkittävämpänä erottavana tekijänä voittavien ja häviävien joukkueiden välillä. Lisätutkimuksena Li et al. (2021) ovat erotelleet pelitilanneheitot kahden ja kolmen pisteen heitoiksi, joista kolmen pisteen heittojen onnistumisprosentti on ollut merkittävin erottava tekijä voittavien ja häviävien joukkueiden välillä. Thabtah et al. (2019) sekä Horvat, Job, Logozar ja Livada (2023) ovat tutkimuksissaan ennustaneet yksittäisten NBA-pelien voittajaa, mutta ennustemalleja runkosarjan voittojen määristä ei juurikaan ole tehty.

Monet perinteisiä pelitilastoja hyödyntävät tutkimukset ovat saaneet kritiikkiä liittyen pelin tempoon. Lorenzo, Gómez, Ortega, Ibáñez ja Sampaio (2010) ovat esittäneet, että pelitilastoja tarkastellessa on merkitystä sillä, kuinka paljon joukkueella on pallonhallintoja pelin aikana. Jos joukkueella A on 45 onnistunutta pelitilanneheittoa sekä 100 pallonhallintaa, ja joukkueella B on myös 45 onnistunutta pelitilanneheittoa, mutta vain 85 pallonhallintaa, eivät pelitilanneheittojen onnistunut määrä kuvaa joukkueiden suoriutumista vertailukelpoisesti, koska toisella joukkueella on ollut enemmän mahdollisuuksia tehdä pisteitä. Oliverin (2004, 43) mukaan nopeatempoista koripalloa pelaavien NBA-joukkueiden peleissä pallonhallintoja voi olla yli 100, kun taas hidastempoisemmissa peleissä pallonhallintoja voi olla alle 80.

Lisäksi myös perinteisen pelitilanneheittoprosentin hyödyntämistä on kritisoitu, koska se sisältää kahden sekä kolmen pisteen heitot. Caporale ja Collier (2015) ovat tutkimuksessaan todenneet, että kahden ja kolmen pisteen heitot pystytään optimoimaan niiden odotusarvojen perusteella. Tutkijoiden mukaan joukkueet, jotka parhaiten optimoivat kahden ja kolmen pisteen heittojen suhteelliset osuudet, suoriutuvat keskimäärin muita joukkueita paremmin. Pelitilanneheittoprosenttia on parannettu Oliverin (2004) teoksessa, jossa on esitelty tehokas pelitilanneheittoprosentti. Tässä mittarissa on painotettu suuremmin kolmen pisteen heittoja, koska ne ovat tuovat yhden lisäpisteen kahden pisteen heittoihin verrattuna (Dehesa, Vaquera, Gonçalves, Mateus, Gomez-Ruano & Sampaio 2019).

Kirjallisuuskatsauksen perusteella voidaan todeta, että urheiluanalytiikka on erittäin merkittävä osa eri lajien seuraorganisaatioiden toimintaa ja päätöksentekoa. NBA:ssa datan keruu on edistyksellisellä tasolla, ja monet tutkimukset ovat tätä dataa hyödyntäneet. Nämä tutkimukset osoittavat hyökkäyksen ja puolustuksen olevan merkittäviä tekijöitä joukkueen suoriutumiseen liittyen. Tutkielman seuraavassa luvussa eli empiirisessä osuudessa tutkitaan aikaisempaan tutkimukseen pohjaten, miten hyvin historialliset ottelutilastot selittävät joukkueen suoriutumista. Tutkimusmenetelmänä hyödynnetään lineaarista regressioanalyysiä, jonka pohjalta myös testataan miten hyvin regressiomallilla pystytään ennustamaan voittoja tulevalla kaudella, koska tästä ei löytynyt aikaisempaa tutkimusta kuin yksittäisten pelien tasolla.

3 Tutkimusaineisto ja -menetelmät

Tässä luvussa käsitellään käytettävää tutkimusaineistoa, jonka lisäksi esitellään käytettävät tutkimusmenetelmät. Tutkimusaineisto sisältää NBA-joukkueiden runkosarjan ottelutilastot kausilta 2013/2014–2017/2018, jotka on valittu kausien peräkkäisyyksien sekä vertailukelpoisuuksien perusteella. Tutkimusaineisto on kerätty hyödyntäen NBA (2023a) ja Basketball-Reference (2023) tietokantoja. Datan käsittelyyn on käytetty Microsoft Excel-taulukkolaskentaohjelmistoa, jonka jälkeen tilastolliset testit on tehty Stata SE-ohjelmiston versiota 17.0 käyttäen. Tutkimusmenetelmistä käsitellään aluksi lineaarista regressiota yleisesti, josta syvennytään tarkemmin paneelidatan estimointimenetelmiin.

3.1 NBA-joukkueiden historialliset ottelutilastot

Tässä tutkimuksessa aineistona hyödynnetään kausien 2013/2014–2017/2018 perinteisiä ottelutilastoja, jotka kuvaavat joukkueiden hyökkäys- ja puolustuspään suoriutumista. Aikaisempaa tutkimusta mukaillen, ottelutilastoista on kerätty absoluuttisina lukuina joukkueiden pallonhallinnat, onnistuneet ja epäonnistuneet kahden ja kolmeen pisteen heitot, onnistuneet ja epäonnistuneet vapaaheitot eli yhden pisteen heitot, puolustus- ja hyökkäyslevypallot, syötöt, pallonmenetykset, riistot, torjunnat, tehdyt virheet sekä voittojen määrät runkosarjassa (Lorenzo et al. 2010; Sampaio & Janeira 2017; Thabtah et al. 2019; Cabarkapa et al 2022). Lisäksi, jotta saataisiin parempi kuva NBA-joukkueen puolustuksen merkityksestä, otetaan tutkimuksessa huomioon, kuinka paljon pisteitä joukkuetta vastaan on tehty, ja millä tehokkuudella. Täten tutkimukseen otettu mukaan joukkueiden vastustajan onnistuneet ja epäonnistuneet kahden sekä kolmen pisteen heitot ja onnistuneet sekä epäonnistuneet vapaaheitot (Baghal 2012; Kubatko et al. 2007).

Jotta tutkimuksessa pystyttäisiin luomaan toimiva malli joukkueen voittojen ennustamiseen, valittuja tunnuslukuja ei tutkimuksessa hyödynnetä sellaisenaan, vaan otetaan kirjallisuuskatsauksessa esitelty ottelutilastojen kritiikki huomioon. Perinteisen pelitilanneheittoprosentin kritiikkiin perustuen, tutkimuksessa käytetään Kubatko et al. 2007 tutkimuksessa esiteltyä tehokkaan pelitilanneheittoprosentin kaavaa, joka ottaa paremmin huomioon kolmen pisteen heitot verrattuna perinteiseen pelitilanneheittoprosenttiin. Tämä nostaa joukkueiden

pelitilanneheittoprosentteja, mutta ne kuvaavat paremmin kokonaisuutena pelitilanneheittojen tehokkuutta. Kaavan 1 avulla on laskettu jokaiselle joukkueelle tehokkaat pelitilanneheittoprosentit. Lisäksi joukkueiden vastustajille lasketaan tehokkaat pelitilanneheittoprosentit saman kaavan avulla.

$$EFG\% = \frac{FGM + 0{,}5 \times 3PM}{FGA} \tag{1}$$

Missä EFG% on joukkueen tehokas pelitilanneheittoprosentti, FGM on onnistuneet pelitilanneheitot, 3PM on onnistuneet 3-pisteen heitot ja FGA on pelitilanneheittojen yritykset

Tilastojen laskennassa on otettu huomioon myös joukkueen pallonhallintojen määrä, jotta eri joukkueiden pelitempon vaihtelun vaikutus vertailukelpoisuuteen voidaan eliminoida (Sampaio & Janeira 2017). Esimerkiksi levypallot on laskettu 100 pallonhallintaa kohti kaavan 2 mukaisesti (Oliver 2004, 20).

$$REB = \frac{REB}{POSS} \times 100 \tag{2}$$

Missä REB on levypallojen määrä ja POSS pallonhallintojen määrä

Taulukossa 1 on kuvattu tutkimuksessa käytettäviä muuttujia, sisältäen muuttujien keskiarvot, keskihajonnan sekä vaihteluvälin. Muuttujista esimerkiksi 3-pisteen heitoissa on paljon hajontaa, eli osa joukkueista ottaa enemmän heittoyrityksiä kauempaa, kun osa taas panostaa enemmän kahden pisteen heittoihin. Lisäksi myös vapaaheitoilla on paljon hajontaa, eli osa joukkueista pääsee useasti vapaaheittoviivalle, kun taas osa pääsee selkeästi vähemmän. Tämä linkittyy vahvasti myös joukkueen tekemien virheiden määrään, mikä vaihtelee myös suhteellisen paljon joukkueiden välillä. Suhteellisissa luvuissa, kuten heittoprosenteissa, ei luonnollisesti hajontaa ole paljoa, mutta kuitenkin eroja joukkueiden heittämisen tehokkuudessa löytyy.

Taulukko 1. Tutkimuksessa käytettävien muuttujien kuvailu

Tässä taulukossa on esiteltynä tutkimuksessa käytettävien muuttujien keskiarvot, keskihajonnat sekä vaihteluvälit koko tarkasteluperiodin ajalta. Laskennassa käytetty havaintojen lukumäärä on 150. Muuttujat on normalisoitu 100 pallonhallintaa kohti, pois lukien voittojen kokonaismäärät.

Muuttujan nimi	Selite	Keskiarvo	Keskihajonta	Vaihteluväli
W	Voittojen määrä runkosarjassa	41.00	12.60	[10, 73]
FGM	Onnistuneet pelitilanneheitot	39.65	1.31	[34.8, 42.7]
FGA	Pelitilanneheittojen yritykset	87.23	1.77	[82.1, 92.3]
FG%	Pelitilanneheittoprosentti	45.47	0.02	[40.8, 50.3]
EFG%	Tehokas pelitilanneheittoprosentti	50.70	0.02	[45.6, 56.9]
3PM	Onnistuneet 3-pisteen heitot	9.11	1.83	[5.2, 15.5]
3PA	3-pisteen heittojen yritykset	25.55	4.67	[15.3, 43]
3P%	3-pisteen heittoprosentti	35.59	0.02	[31.2, 41.6]
FTM	Onnistuneet vapaaheitot	17.98	1.80	[12.8, 23]
FTA	Vapaaheittojen yritykset	23.66	2.41	[17.4, 31.7]
FT%	Vapaaheittoprosentti	76.10	0.03	[66.8, 81.5]
OREB	Hyökkäyslevypallot	10.76	1.31	[8.2, 15.2]
DREB	Puolustuslevypallot	34.00	1.36	[30.6, 37]
REB	Levypallot	44.76	1.80	[39.7, 49.5]
AST	Koriinjohtaneet syötöt	23.15	1.85	[18.6, 30.1]
TOV	Pallonmenetykset	14.78	1.12	[11.9, 18.3]
STL	Riistot	7.99	0.83	[5.7, 10.1]
BLK	Torjunnat	4.96	0.77	[3.5, 7.4]
PF	Tehdyt virheet	20.84	1.52	[17.1, 24.5]
OPPFGM	Vastustajan onnistuneet pelitilanneheitot	39.65	1.20	[37.1, 42.8]
OPPFGA	Vastustajan pelitilanneheittojen yritykset	87.22	1.58	[83.2, 91.6]
OPPFG%	Vastustajan pelitilanneheittoprosentti	45.46	0.01	[42, 48.7]
OPPEFG%	Vastustajan tehokas pelitilanneheittoprosentti	50.69	0.02	[45.9, 54.2]
OPP3PM	Vastustajan onnistuneet 3-pisteen heitot	9.12	1.28	[6.7, 12.4]
OPP3PA	Vastustajan 3-pisteen heittojen yritykset	25.56	3.23	[19, 32.9]
OPP3P%	Vastustajan 3-pisteen heittoprosentti	35.63	0.01	[32.2, 38.2]
OPPFTM	Vastustajan onnistuneet vapaaheitot	17.98	1.64	[14.6, 22.5]
OPPFTA	Vastustajan vapaaheittojen yritykset	23.65	2.19	[18.3, 28.7]
OPPFT%	Vastustajan vapaaheittoprosentti	76.05	0.01	[72.7, 79.5]

3.2 Tutkimusmenetelmät

Koska tutkimuksessa on tarkoitus selittää ja ennustaa NBA-joukkueen voittojen määriä runkosarjassa useilla eri ottelutilastoilla, on pääasialliseksi tutkimusmenetelmäksi valittu lineaarinen regressioanalyysi, kuten Atkinson ja Nevill (2001) tutkimuksessaan suosittelevat. Lineaarisessa regressioanalyysissä pyritään selittämään ja ennustamaan yhtä selitettävää muuttujaa yhden tai useamman selittävän muuttujan avulla. Regressiomalli perustuu muuttujien väliseen vaihteluun, eli selitettävän muuttujan vaihtelua pyritään selittämään selittävien muuttujien vaihtelulla (Olive 2017, 2, 17). Usean selittäjän lineaarisen regression

yksinkertainen malli on kuvattu kaavassa 3, joka on esitelty (Hill, Griffiths & Lim 2018, 202) teoksessa.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e \quad (3)$$

Missä Y on selitettävä muuttuja, β_0 on vakiotermi, β_1 on x_1:n regressiokerroin, β_2 on x_2:n regressiokerroin, β_k on x_k:n regressiokerroin, x_1, x_2 sekä x_k ovat selittäviä muuttujia ja e on jäännöstermi/virhetermi.

Selitettävänä muuttujana on tutkimuksessa käytössä NBA-joukkueen voittojen määrä absoluuttisena lukuna, jotta mallilla pystytään ennustamaan voittojen määrää yksittäiselle NBA-joukkueelle. Selittävinä muuttujina, joilla pyritään selittämään voittomäärien vaihtelua, käytetään joukkueiden edellisessä alaluvussa tarkasteltuja hyökkäys- ja puolustuspään tilastoja. Vakiotermi kuvaa *Y*:n odotusarvoa tilanteessa, jossa kaikki selittävät muuttujat olisivat nollia, mikä ei tässä tutkimuksessa kuitenkaan ole käytännössä mahdollinen tilanne. Regressiokertoimet kuvaavat kuinka paljon selittävän muuttujan kasvaminen yhdellä yksiköllä kasvattaa tai vähentää *Y*:tä, eli voittojen määrää. (Hill et al. 2018, 199) Lisäksi mallissa on jäännöstermi *e*, joka kuvaa selittämättömän vaihtelun määrää, eli tässä tutkimuksessa tekijöitä, jotka vaikuttavat voittojen määrään, mutta jotka eivät sisälly malliin (Hill, Griffiths & Judge 2001, 4).

Yleisesti regressiomallin parametrien estimoimiseen käytetään pienimmän neliösumman menetelmää, jossa tarkoituksena on löytää sellaiset regressiokertoimet, jotka minimoivat jäännöstermin eli residuaalien neliösumman. Residuaali kuvaa havaintojen todellisen arvon ja mallin ennusteen välistä erotusta. Pienimmän neliösumman estimointimenetelmässä siis etsitään usean selittäjän tapauksessa sellainen pinta, jossa residuaalien neliöiden summa minimoituu. (Hill et al. 2001, 51–53)

Usean selittäjän lineaarisessa regressiossa on tiettyjä oletuksia, joiden tulisi täyttyä. Muuten estimoidussa mallissa kertoimet voivat olla harhaisia tai niiden keskivirheet voivat olla vääriä tai epäluotettavia. (Hill et al. 2001, 149–150) Ensimmäisenä oletuksena on selitettävän ja selittävien muuttujien välisen suhteen lineaarisuus, eli mallin muodostamisessa on tärkeää selittävän ja selittävien muuttujien välisen yhteyden oikea spesifiointi. Toisena oletuksena

on mallin harhattomuus, eli tällöin jäännöstermin odotusarvo on nolla. Kolmantena taustaoletuksena on mallin homoskedastisuus, joka tarkoittaa jäännöstermin varianssin vakioisuutta, eli jäännöstermi ei täten ole riippuvainen selittävän tai selittävien muuttujien arvoista tai ajasta. (Hill et al. 2018, 203)

Hill et al. (2018, 204) teoksessa neljännen taustaoletuksen mukaan mallissa ei saa olla autokorrelaatiota, eli havaintojen ja myös residuaalien tulisi olla riippumattomia toisistaan. Tällöin jäännöstermiparien kovarianssien tulisi olla nolla. Viidennessä taustaoletuksessa selittävät muuttujat eivät saisi olla multikollineaarisia, eli muuttujat eivät saisi korreloida liikaa keskenään. Kuudennessa taustaoletuksessa jäännöstermit noudattavat normaalijakaumaa silloin kun selitettävä muuttuja noudattaa normaalijakaumaa, mutta tämä ei kuitenkaan ole pakollista. Kun edellä läpikäydyt taustaoletukset lineaariselle regressiomallille pätevät, ovat pienimmän neliösumman menetelmän mukaiset estimaattorit Gauss-Markov teoreeman mukaisesti parhaat, koska niillä on pienin varianssi lineaaristen ja harhattomien estimaattoreiden joukossa (Hill et al. 2018, 72).

Ylempänä kuvattu perinteinen lineaarinen regressio on tutkimuksen kannalta tärkeää ymmärtää, koska käytettävä aineisto on todellisuudessa paneelidata. Se siis sisältää havaintoja useista havaintoyksiköistä useilta eri aikaperiodeilla, ja sen vuoksi syvennytään vielä tarkemmin paneelidataan soveltuvien estimointimenetelmien tarkasteluun. Tutkimuksessa vertaillaan pienimmän neliösumman regressiomenetelmän lisäksi kiinteiden- ja satunnaisten vaikutusten menetelmien soveltuvuutta käytettävään aineistoon joukkueiden ottelutilastoista. (Hill et al. 2001, 351) Kaavassa 4 on kuvattuna yksinkertainen paneeliregression malli, joka on esitelty Hill et al. (2018, 637) teoksessa.

$$Y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + v_{it} \quad (4)$$

Missä Y_{it} on selitettävä muuttuja, β_0 on vakiotermi, β_1 on x_{1it}:n regressiokerroin, β_2 on x_{2it}:n regressiokerroin, β_k on x_{kit}:n regressiokerroin, x_{1it}, x_{2it} sekä x_{kit} ovat selittäviä muuttujia ja v_{it} on jäännöstermi/virhetermi.

Paneeliregressiossa muuttujat vaihtelevat ajan ja yksiköiden välillä, mutta regressiokertoimet ovat vakioita kaikille muuttujille yli ajan. Mallissa voisi myös olla selittäviä muuttujia,

jotka säilyvät vakiona yli ajan, mutta koska tässä tutkimuksessa ei sellaisia ole, ei asiaa käsitellä tarkemmin. Kaavasta 4 nähdään myös, että jäännöstermi on muotoa v_{it} ja tämä johtuu siitä, että jäännöstermi koostuu paneeliregressiossa kahdesta osasta. Kaavassa 5 on esiteltynä termit u_i sekä v_{it}. jotka muodostavat jäännöstermin. Termi u_i sisältää ajasta riippumatonta yksikkökohtaista vaihtelua. Sillä tarkoitetaan havaitsematonta heterogeenisuutta eli yksikkökohtaisia kiinteitä vaikutuksia, joita ei ole sisällytettynä mallissa. (Hill et al. 2018, 637) Toinen termi e_{it} kuvaa Wooldridgen (2010, 285) mukaan idiosynkraattista jäännöstermiä, eli satunnaisia tekijöitä, jotka ovat riippuvaisia ajasta sekä yksiköstä ja joita ei ole sisällytetty malliin.

$$v_{it} = u_i + e_{it} \tag{5}$$

Missä v_{it} on kokonaisjäännöstermi, u_i on yksikkökohtaisten tekijöiden aiheuttama virhetermi ja e_{it} on yleinen idiosynkraattinen virhetermi

Kun pienimmän neliösumman menetelmää sovelletaan paneelidataan, puhutaan yhdistetystä pienimmän neliösumman menetelmästä, joka käsittelee paneelidatan havainnot yhtenä suurena yhdistettynä aineistona (Spada, Fiore & Galati 2023). Yhdistetty pienimmän neliösumman menetelmä ei kuitenkaan ota huomioon yksiköiden välisiä kiinteitä eroja eli havaitsematonta heterogeenisuutta, jonka lisäksi tärkeänä oletuksena on mallin eksogeenisuus eli nämä yksikkökohtaiset eroavaisuudet eivät saisi korreloida selittävien muuttujien kanssa (Hill et al. 2018, 635).

Kiinteiden vaikutusten menetelmää käytetään, kun havaitsematon heterogeenisuus eli yksilöiden väliset erot, jotka ovat vakioita ajassa, on korreloitunut selittävien muuttujien kanssa, mikä voi aiheuttaa endogeenisuutta. Kiinteiden vaikutusten estimaattori pyrkii poistamaan yksikkökohtaiset vaikutukset, ja hyödyntämään ainoastaan ajan suhteen tapahtuvaa vaihtelua yksiköiden sisällä. Kiinteiden vaikutusten mallissa perinteistä vakiotermiä ei ole sellaisenaan, sillä se sisältää yksikkökohtaiset kiinteät vaikutukset. Jokaiselle yksikölle on mallissa oma vakiotermi, joka huomioi yksikkökohtaiset kiinteät tekijät vähentämällä muuttujista niiden keskiarvot yli ajan, jolloin malli keskittyy siihen kuinka muuttujat muuttuvat ajan kuluessa yksiköiden sisällä. (Hill et al. 2018, 640-646)

Satunnaisten vaikutusten mallissa vakiotermi on satunnainen, ja se koostuu populaation keskiarvosta sekä yksikkökohtaisesta vaihtelusta. Satunnaisten vaikutusten malli kuitenkin

olettaa, että yksikkökohtainen vaihtelu on satunnaista, ja että se ei ole korreloitunut selittävien muuttujien kanssa. Malli hyödyntää täten yksiköiden sisäistä sekä yksiköiden välistä vaihtelua ja jos sen taustaoletukset pitävät paikkansa, se on paras estimointimenetelmä. (Hill et al. 2018, 651-654)

Paneeliregressiossa taustaoletukset vaihtelevat eri estimointimenetelmien välillä. Aiemmin kuvatun pienimmän neliösumman menetelmän taustaoletuksien lisäksi yhdistetyssä pienimmän neliösumman menetelmässä oletetaan, että havaitsematonta heterogeenisuutta ei ole ja että jäännöstermin ja selittävien muuttujien välillä ei ole korrelaatiota, eli malli on eksogeeninen. Kiinteiden vaikutusten menetelmässä oletuksena on, että havaitsematonta heterogeenisuutta on, ja että se korreloi selittävien muuttujien kanssa. Kuitenkin yleinen virhetermi ei saa olla korreloitunut selittävien muuttujien kanssa, koska se voi johtaa endogeenisuuteen. Lisäksi mallin tulee olla homoskedastinen, eli residuaalit eivät saa olla riippuvaisia selitettävästä tai selittävistä muuttujista eikä ajasta. Mallissa ei saa myöskään olla autokorrelaatiotaa eikä multikollineaarisuutta, kuten pienimmän neliösumman menetelmässä. Satunnaisten vaikutusten menetelmässä on muuten samat taustaoletukset kuin kiinteiden vaikutusten menetelmässä, mutta lisänä on oletus yksikkökohtaisten erojen satunnaisuudesta sekä niiden korreloimattomuudesta selittävien muuttujien kanssa.

Sopivimman estimointimenetelmän valintaan on olemassa tilastollisia testejä, joiden perusteella paras estimaattori pystytään valitsemaan. Taulukossa 2 on kuvattuna Park (2010) ja Hill et al. (2018, 640-656) teoksissa esiteltyjä tilastollisia testejä, joiden perusteella sopivin estimointimenetelmä voidaan valita. Kiinteiden vaikutusten F-testin avulla pystytään tutkimaan, onko mallissa kiinteitä vaikutuksia, jotka tulisi ottaa huomioon. Testi vertaa mallien sopivuutta aineistolle, ja jos molemmat toimivat yhtä hyvin, eli nollahypoteesi jää voimaan, kiinteitä vaikutuksia ei ole ja täten kiinteiden vaikutusten menetelmää ei ole järkevää käyttää. Tällöin yhdistetty pienimmän neliösumman menetelmä olisi parempi vaihtoehto.

Breusch-Pagan testillä tutkitaan satunnaisten vaikutusten olemassaoloa mallissa. Testissä testataan, onko satunnaisten vaikutusten varianssi yksiköiden välillä nolla. Jos nollahypoteesi jää voimaan, ei mallissa satunnaisia vaikutuksia löydy ja täten satunnaisten vaikutusten menetelmän käyttämisestä ei ole hyötyä.

Taulukko 2. Paneelidatan estimointiin käytettävän menetelmän valinta

Tässä taulukossa on kuvattuna Park (2010) ja Hill et al. (2018, 640-656) teoksissa esiteltyjen tilastollisten testien hyödyntäminen sopivimman estimointimenetelmän valitsemiseksi.

F-testi*	Breusch-Pagan testi**	Käytettävä estimointimenetelmä
H_0 jää voimaan	H_0 jää voimaan	Yhdistetty pienimmän neliösumman menetelmä
H_0 hylätään	H_0 jää voimaan	Kiinteiden vaikutusten menetelmä
H_0 jää voimaan	H_0 hylätään	Satunnaisten vaikutusten menetelmä
H_0 hylätään	H_0 hylätään	1) Jos Hausman−testin*** H_0 jää voimaan, satunnaisten vaikutusten menetelmää voidaan käyttää. Muuten kiinteiden vaikutusten malli, tai 2) Sekä kiinteiden- että satunnaisten vaikutusten menetelmä

* F-testin nollahypoteesi on että kiinteitä vaikutuksia ei ole

** Breusch-Pagan testin nollahypoteesi on että satunnaisia vaikutuksia ei ole

*** Hausman-testin nollahypoteesi on että kiinteiden ja satunnaisten vaikutusten mallien kertoimissa ei ole eroja

Vaikka kiinteiden vaikutusten F-testin ja Breusch-Pagan testin nollahypoteesit hylätään, eli kiinteiden- ja satunnaisten vaikutusten menetelmät ovat parempia yhdistetyn pienimmän neliösumman menetelmään verrattuna, on satunnaisten vaikutusten menetelmä kiinteiden vaikutusten menetelmää parempi sen hyödyntäessä enemmän informaatiota. Satunnaisten vaikutusten menetelmän käyttämiseksi tulee kuitenkin testata, onko se konsistentti, eli jäännöstermi ei saa korreloida minkään selittävän muuttujan kanssa. Tätä ongelmaa ei kiinteiden vaikutusten menetelmässä ole, joten Hausman-testin avulla testataan, onko kiinteiden ja satunnaisten vaikutusten mallien kertoimissa eroa. Jos nollahypoteesi hylätään, eli kertoimissa on eroa, ei satunnaisten vaikutusten menetelmää saa käyttää. Jos kertoimissa ei ole eroa, voidaan satunnaisten vaikutusten menetelmää käyttää, tai vaihtoehtoisesti voidaan raportoida molempien menetelmien tulokset.

4 Tutkimustulokset

Tässä luvussa käsitellään aluksi regressiomallin luomiseen valittuja muuttujia sekä parhaiten soveltuvaa parametrien estimointimenetelmää. Tämän jälkeen käydään läpi lopullisen mallin taustaoletukset sekä tutkimustulokset. Lisäksi ennustemallin toimivuutta testataan mallin luomiseen käytetyn tarkasteluperiodin jälkeisellä periodilla, eli NBA-kaudella 2018/2019, ja tarkastellaan ennustemallin tuloksia.

4.1 Regressiomallin muodostaminen

Lineaarisen regressiomallin muodostamiseen oli käytettävissä suuri määrä muuttujia, jonka vuoksi aluksi tarkasteltiin selittävien muuttujien yhteyksiä ja erityisesti korrelaatioita, pitäen mielessä lineaarisen regression multikollineaarisuuden taustaoletuksen. Korrelaatiomatriisin perusteella esimerkiksi tehokas pelitilanneheittoprosentti korreloi voimakkaasti perinteisen pelitilanneheittoprosentin sekä onnistuneiden pelitilanneheittojen ja kolmen pisteen heittojen kanssa, mikä johtuu siitä, että pelitilanneheittoprosentit perustuvat absoluuttisiin onnistuneiden heittojen määrään suhteessa heittoyrityksiin. Tämän takia malliin valikoitui joukkueiden pisteiden tekemisen tehokkuutta kuvaava tehokas pelitilanneheittoprosentti, ja absoluuttiset heitot sekä heittoyritykset jätettiin pois mallista. Lisäksi vastustajan vapaaheitot korreloivat vahvasti joukkueen tekemien virheiden kanssa, mikä on luonnollista koska koripallossa virheet johtavat usein vastustajan pääsyyn vapaaheittoviivalle. Täten mallista jätettiin myös joukkueen tekemien virheiden määrä pois selittäjien joukosta. Lisäksi puolustuslevypallot korreloivat voimakkaasti vastustajan pelitilanneheittojen ja vapaaheittojen kanssa, joten malliin yhdistettiin hyökkäys- ja puolustuslevypallot kuvaamaan joukkueen kykyä saada levypalloja kokonaisuutena.

Lopulliseen malliin valittiin korrelaatioiden tarkastelun jälkeen selitettäväksi muuttujaksi joukkueen voittojen määrä runkosarjassa absoluuttisena lukuna, ja lisäksi selittäviksi muuttujiksi valikoitui tehokas pelitilanneheittoprosentti, joukkueen onnistuneiden vapaaheittojen määrä, levypallot, pallonmenetykset, riistot, torjunnat sekä vastustajan tehokas pelitilanneheittoprosentti ja vastustajan onnistuneiden vapaaheittojen määrä. Vaikka joitain muuttujia

jouduttiin heti aluksi pudottamaan tarkastelusta pois, nämä tilastolliset muuttujat kuvaavat joukkueen hyökkäys- ja puolustuspään suoriutumista kuitenkin kattavasti.

Sopivaa estimointimenetelmää tutkimuksen paneelidatalle tarkasteltiin kappaleessa 3.2 esiteltyjen tilastollisten testien avulla. Ensimmäiseksi vertailtiin kiinteiden vaikutusten sekä yhdistetyn pienimmän neliösumman menetelmää kiinteiden vaikutusten F-testin avulla. Testin nollahypoteesi hylättiin, eli testin mukaan kiinteitä vaikutuksia on olemassa. Täten kiinteiden vaikutusten menetelmä sopisi yhdistetyn pienimmän neliösumman menetelmää paremmin mallin parametrien estimoimiseen. (Liite 1) Satunnaisten vaikutusten menetelmän soveltuvuutta verrattuna yhdistetyn pienimmän neliösumman menetelmään testattiin Breusch-Paganin-testillä, jonka nollahypoteesi myös hylättiin, eli testin perusteella satunnaisia vaikutuksia löytyy. Tällöin kannattaisi mieluummin käyttää satunnaisten vaikutusten estimointimenetelmää. (Liite 2) Lopulta testattiin vielä, että onko satunnaisten vaikutusten malli konsistentti hyödyntämällä Hausman-testiä. Testin perusteella kiinteiden ja satunnaisten vaikutusten mallien kertoimissa ei ollut eroja, eli satunnaisten vaikutusten menetelmä on konsistentti, minkä seurauksena parametrien estimoimiseen valittiin satunnaisten vaikutusten menetelmä (Liite 3).

4.2 Lineaarisen regression taustaoletukset

Mallin tarkastelussa lähdettiin liikkeelle taustaoletusten täyttymistä, jotta mallia voisi pitää hyvänä. Ensimmäiseksi lähdettiin liikkeelle mallin spesifioinnista, ja selittävän ja selittävien muuttujien lineaarisen suhteen tarkastelusta sirontakuvioiden avulla. Useilla muuttujilla lineaarinen suhde on nähtävissä selkeästi, kuten tehokkaassa heittotilanneprosentissa ja puolustuslevypalloissa. Kuitenkin muilla muuttujilla lineaarinen suhde ei ole erityisen vahvan näköinen voittojen määrän kanssa, mikä voi johtua koripallossa yksittäisen tilastorivin pienestä vaikutuksesta voittojen määrään.

Mallin heteroskedastisuutta tarkasteltiin aluksi residuaalikuvien avulla erikseen idiosynkraattiselle virheelle sekä yksikkökohtaiselle virheelle suhteessa mallin ennustettuihin eli sovitettuihin arvoihin. Residuaalikuvaajien perusteella mallissa saattaisi esiintyä heteroskedastisuutta, koska molemmissa kuvaajissa virheiden varianssit eivät vaikuta olevan täysin vakioita. Esimerkiksi idiosynkraattisen virheen varianssi on suurempaa voittojen pienemmillä arvoilla verrattuna suurempiin arvoihin. (Liite 4)

Mallin autokorrelaatiota tutkittiin Wooldridgen (2010, 282-283) testillä, jossa virhetermien viivästetyillä arvoilla testataan korreloivatko virhetermit keskenään eri aikaperiodeilla. Testin nollahypoteesina on, että mallissa ei ole ensimmäisen asteen autokorrelaatiota. Testin nollahypoteesi jäi voimaan, eli sen mukaan tutkimuksen malli ei sisällä ensimmäisen asteen autokorrelaatiota (Liite 5). Multikollineaarisuutta tarkasteltiin Pearsonin korrelaatiotestien avulla, jotka suoritettiin kaikkien selittävien muuttujien kesken. Suurin korrelaatiokerroin (-0.4994) löytyy levypallojen ja vastustajan tehokkaan pelitilanneheittoprosentin väliltä, mutta muuten selittävien muuttujien väliset korrelaatiokertoimet ovat kaikki matalia, eli suurta ongelmaa multikollineaarisuuden kanssa ei ole (Liite 6).

Selitettävän muuttujan ja residuaalien normaalisuutta testattiin graafisen tarkastelun ja Shapiro-Wilkin normaalijakautuneisuustestien perusteella. Graafisen tarkastelun perusteella voittojen määrä sekä idiosynkraattinen virhe ja yksikkökohtainen virhe vaikuttavat noudattavan normaalijakaumaa. Shapiro-Wilkin testin perusteella ainoastaan yksikkökohtaisella virheellä nollahypoteesi ei jää voimaan, eli testin mukaan se ei olisi normaalijakautunut vaikkakin p-arvo 0.04970 miltei riittää viiden prosentin riskitasolla. Vaikka taustaoletukset vaikuttavat yleisesti mallissa toteutuvan, on kuitenkin otettava mahdollinen heteroskedastisuus huomioon. (Liite 7; Liite 8; Liite 9) Tämän vuoksi lopullisessa mallissa hyödynnetään Whiten korjattuja keskivirheitä, jotka tarjoavat luotettavimpia p-arvoja ja luottamusvälejä, vaikka jäännöstermin varianssi ei olisi vakio (Hill et al. 2018, 374–375).

4.3 Regressioanalyysin tulokset

Muodostetun regressiomallin pääasialliset tulokset ovat esiteltynä taulukossa 3. Mallin muodostamiseen käytettiin 30 joukkueesta havaintoja viiden kauden ajalta, eli havaintojen lukumäärä oli 150 kappaletta. Regressiomallin selitysaste on 0.9166, mikä tarkoittaa, että voittomäärän vaihtelusta pystytään selittämään 91.66 prosenttia pelitilastojen vaihtelun avulla. Lisäksi yhden prosentin riskitasolla malli sekä kaikki selittävät muuttujat ovat tilastollisesti merkitseviä, suurimman p-arvon ollessa torjunnoilla 0.003. Vakiotermi ei kuitenkaan ole tilastollisesti merkitsevä, mikä ei ole ongelma, koska tilanne, jossa kaikki selittävät muuttujat olisivat nolla, on epärealistinen.

Taulukko 3. Lineaarisen regression tulokset

Tässä taulukossa on esiteltynä satunnaisten vaikutusten menetelmällä estimoidun regressiomallin tulokset sekä havaintojen ja ryhmien eli joukkueiden lukumäärät.

Havaintojen lukumäärä	150	**Joukkueiden lukumäärä**	30	**Selitysaste**	0.9166
	Kerroin	**Korjattu keskivirhe**	**z**	**P-arvo**	**Luottamusväli**
EFG%	3.789347	13.74477	27.57	0.000	[3.519954, 4.058739]
FTM	0.9564867	0.2273	4.2100	0.000	[0.5109207, 1.402053]
OPPFTM	-0.8145725	0.1705	-4.78	0.000	[-1.148661, -0.4804839]
REB	1.637704	0.2350125	6.97	0.000	[1.177088, 2.09832]
TOV	-2.612514	0.3665466	-7.13	0.000	[-3.330932, -1.894096]
STL	3.908891	0.5353335	7.3	0.000	[2.859657, 4.958126]
BLK	-1.425966	0.4734939	-3.01	0.003	[-2.353997, -0.497935]
OPPEFG	-3.541206	26.45607	-13.39	0.000	[-4.059735, -3.022676]
Vakiotermi	-32.98483	28.27958	-1.17	0.243	[-88.41179, 22.44212]

Mallin kertoimien mukaan joukkueen voittomäärien selittämisessä merkittävin tekijä on joukkueen tehokas pelitilanneheittoprosentti. Mallia tulkittaessa on kuitenkin tärkeää ymmärtää, että luvut on normalisoitu 100 pallonhallintaa kohti, ja malli tarkastelee ainoastaan runkosarjan tilastoja. Eli kun joukkueen pelitilanneheittoprosentti kasvaa runkosarjassa yhdellä prosenttiyksiköllä 100 pallonhallintaa kohti, pitäisi runkosarjan voittomäärien nousta 3.789347 yksiköllä. Vastustajan tehokkaalla pelitilanneprosentilla on miltei yhtä suuri vaikutus, mutta vastakkaiseen suuntaan. Eli kun vastustajan tehokas pelitilanneheittoprosentti nousee yhdellä prosenttiyksiköllä 100 pallonhallintaa kohti runkosarjassa, laskee joukkueen runkosarjan voittomäärä 3.541206 yksiköllä. Koska molemmat muuttujat ovat mitattuna samalla skaalalla, tästä voidaan vetää johtopäätös, että joukkueen onnistuminen korinteossa on merkittävämpää kuin kyky vaikeuttaa vastustajan korintekokykyä.

Mallin absoluuttisissa luvuissa on myös tärkeää ottaa huomioon muuttujien normalisointi sekä ainoastaan runkosarjan tarkastelu. Esimerkiksi riistojen tulkinta mallin mukaisesti on, että kun joukkue saa yhden riiston 100 pallonhallintaa kohti enemmän runkosarjassa, nousee joukkueen runkosarjan voittomäärä 3.908891 yksiköllä. Vaikka riistojen kerroin on suurempi kuin tehokkaiden pelitilanneheittoprosenttien, on myös ymmärrettävä muuttujien mitta-asteikoiden erilaisuus. Esimerkiksi aiemmin taulukossa 1 esiteltyjen muuttujien

kuvailun perusteella nähdään, että joukkueen tehokkaan pelitilanneheittoprosentin keskiarvo on 50.70 prosenttia, kun taas riistojen keskiarvo on 7.99. Tällöin vaikka riistojen kerroin mallissa on suuri, se ei yksiselitteisesti tarkoita, että riistot olisivat merkittävin tekijä voittojen takana.

Mielenkiintoista tuloksissa on joukkueen onnistuneiden vapaaheittojen sekä vastustajan onnistuneiden vapaaheittojen merkitys voittojen määrälle. Mallin kertoimien perusteella joukkueen onnistuneiden vapaaheittojen määrä ei ole erityisen merkittävä joukkueen voittomäärien takana, mutta sillä on kuitenkin suurempi vaikutus kuin vastustajan tekemien vapaaheittojen määrällä. Mallissa tehokkaiden pelitilanneheittoprosenttien jälkeen merkittävimpänä muuttujana voidaan pitää levypalloja, koska vaikka sen kerroin ei ole yhtä suuri esimerkiksi pallonmenetysten ja riistojen kanssa, levypalloja tulee aikaisemmin esitellyn taulukon 1 perusteella joukkueella paljon enemmän kuin pallonmenetyksiä tai riistoja.

Mallin perusteella torjuntojen määrä vaikuttaa negatiivisesti joukkueen voittojen määriin, mitä voidaan pitää yllättävänä, koska torjunnat tarkoittavat vastustajan heiton torjumista puhtaasti, jolloin joukkue usein saa pallon itselleen. Toisaalta suurempi torjuntojen määrä voi kertoa aggressiivisemmasta pelistä, jossa vastustaja ajaa enemmän korille, mikä voi johtaa joukkueen torjuntoihin mutta myös esimerkiksi virheisiin. Tämä johtuu siitä, että usein torjuntoja tulee kontaktitilanteissa korin alla, jolloin myös virheitä voi tulla joukkueelle enemmän mikä vaikuttaa vastustajan pallon takaisin saamiseen ja vapaaheittoviivalle pääsemiseen.

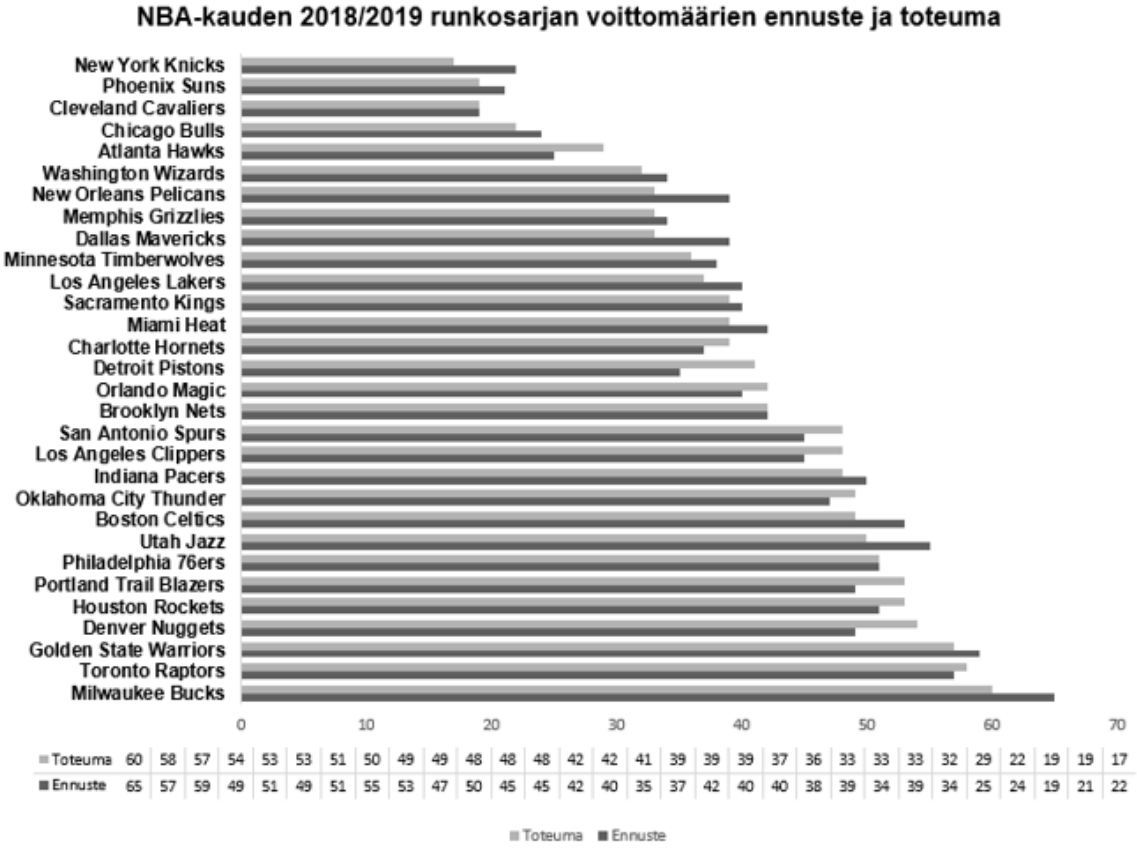
4.4 Regressiomallin toimivuus voittojen ennustamisessa kaudella 2018/2019

Regressiomallin ennustekykyä on tarkoitus tarkastella hyödyntäen mallin muodostamiseen käytetystä tarkasteluperiodista seuraavaa aikaperiodia, eli NBA:n kautta 2018/2019. Regressiomallin myötä saatujen kertoimien avulla testataan, onnistuuko malli ennustamaan oikean voittojen määrän joukkueille. Regressioanalyysin myötä estimoitu regressiomalli on esitelty kaavassa 6.

$$W = - 32.98483 + 3.789347 * EFG\% + 0.9564867 * FTM – 0.8145725 * OPPFTM + 1.637704 * REB – 2.612514 * TOV + 3.908891 * STL – 1.425966 * BLK – 3.541206 * OPPEFG\% \quad (6)$$

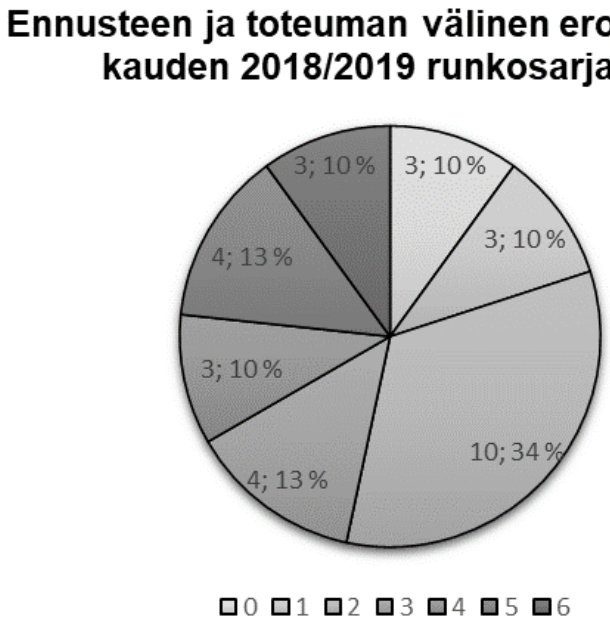
Missä W on voittojen määrä runkosarjassa, EFG% on tehokas pelitilanneheittoprosentti, FTM on joukkueen onnistuneet vapaaheitot, OPPFTM on vastustajan onnistuneet vapaaheitot, REB on levypallot, TOV on pallonmenetykset, STL on riistot, BLK on torjunnat ja OPPEFG% on vastustajan tehokas pelitilanneheittoprosentti

Tämän mallin avulla on laskettu kauden 2018/2019 jokaisen joukkueen mallin ennustamat voittojen määrät, hyödyntämällä kauden 2018/2019 runkosarjan dataa. Kauden 2018/2019 data on myös normalisoitu 100 pallonhallintaa kohti, kuten mallin muodostamiseen käytetyssä datassa oli menetelty. Kuvassa 3 on näkyvissä ennustusten tulokset, sekä oikeat kauden 2018/2019 voittojen määrien toteumat.



Kuva 3. NBA-kauden 2018/2019 runkosarjan voittomäärien ennuste ja toteuma

Kuvaajan alapuolella on taulukoitu toteutuneet ja ennustetut arvot, jotka ovat järjestyksessä suurimmasta pienimpään toteutumien perusteella, eli ensimmäinen sarake kuvaa Milwaukee Bucksin voittomääriä ja viimeinen sarake New York Knicksin voittomääriä. Kuvaajassa haaleampi palkki kuvaa toteutumaa, ja tummempi palkki ennustetta. Kuvaajasta nähdään, että regressiomallilla pystytään ennustamaan suhteellisen lähelle toteutuneita voittomääriä. Kuvassa 4 on esiteltynä mallin ennustuksen ja toteuman välinen erotus, eli kuinka monen voiton päähän malli on onnistunut ennustamaan joukkueiden voittojen määriä. Malli onnistui ennustamaan kolmelle joukkueelle eli kymmenelle prosentille oikean voittojen määrän. Kuitenkin enimmillään ennusteet olivat kolmelle joukkueelle kuuden voiton päässä toteutuneesta. 16 joukkueen ennustettu määrä oli kuitenkin korkeintaan kahden voiton päässä toteutuneesta, mitä voi pitää suhteellisen hyvänä tuloksena.



Kuva 4. Ennusteen ja toteuman välinen erotus NBA-kauden 2018/2019 runkosarjassa

Tämän mallin perusteella voisi tutkia tarkemmin, mikä yhdistää joukkueita, joiden ennuste on todella lähellä toteutumaa, ja mikä joukkueita kauempana toteutumasta. Tällöin voisi miettiä painottaako malli jotain tilastoriviä liikaa, jonka vuoksi ennuste ei toimi yhtä hyvin kaikkiin joukkueisiin, vaikka mallin rakentamiseen on käytetty jokaisen joukkueen tilastoja tarkasteluperiodin ajalta. Ennustemallin toimivuutta olisi myös hyvä kokeilla muissa otoksissa, ja tutkia kuinka hyvin ennuste toimii muilla kausilla.

5 Johtopäätökset

Tämän tutkimuksen tavoitteena oli selvittää, miten hyvin historiallisilla ottelutilastoilla voidaan selittää NBA-joukkueen voittomääriä runkosarjassa. Näiden ottelutilastojen avulla oli myös tarkoitus luoda ennustemalli, jonka avulla pyrittiin ennustamaan runkosarjan voittojen määriä eri otoksessa. Tässä luvussa esitellään johdannossa muodostetut tutkimuskysymykset uudelleen ja vastataan niihin tutkimuksen perusteella. Lisäksi arvioidaan tutkimuksen luotettavuutta ja mahdollisia jatkotutkimusehdotuksia.

5.1 Tutkimustulokset sekä yhteenveto

Tutkimusongelma kiteytettiin yhteen päätutkimuskysymykseen sekä kahteen alatutkimuskysymykseen.

Päätutkimuskysymys:

> *Kuinka historiallisen otteludatan avulla voidaan selittää NBA-joukkueen suoriutumista?*

Alatutkimuskysymykset:

> *Mitkä ovat merkittävimpiä tilastollisia muuttujia, jotka vaikuttavat NBA-joukkueen suoriutumiseen*?
>
> *Kuinka hyvin historialliseen dataan perustuva tilastollinen malli pystyy ennustamaan joukkueiden suoriutumista tulevilla kausilla*?

Tutkimuksen perusteella päätutkimuskysymykseen pystytään toteamaan, että historiallisen otteludatan avulla NBA-joukkueen suoriutumista voidaan selittää tilastollisen analyysin perusteella hyvin. Vaikka NBA:ssa tilastoidaan erilaisia tilastoja todella suuri määrä, jo suhteellisen pieni hyökkäystä ja puolustusta kuvaavien muuttujien määrä onnistuu selittämään voittomäärien vaihtelua runkosarjassa hyvin. Luodun regressiomallin selitysasteeksi saatiin 91,66 prosenttia, joka kertoo siitä, että ottelutilastoilla pystytään selittämään joukkueen voittojen määrää runkosarjassa hyvin. Tämä on myös linjassa esimerkiksi Li et al. (2021)

tutkimuksen kanssa, jossa ottelutilastojen avulla saatiin hyviä tuloksia logistisen regressioanalyysin avulla.

Ensimmäiseen alatutkimuskysymykseen vastataan tutkimustulosten perusteella. Kuten Ottenin ja Millerin (2015) tutkimuksessa, myös tämän tutkimuksen perusteella joukkueen tehokasta pelitilanneheittoprosenttia sekä vastustajan tehokasta pelitilanneheittoprosenttia voidaan pitää merkittävimpänä tekijöinä, jotka vaikuttavat joukkueen suoriutumiseen. Cabarkapa et al. (2022) nostivat tutkimuksessaan puolustuslevypallot erittäin merkittäväksi tekijäksi omassa tutkimuksessaan, mikä voidaan myös nähdä tämän tutkimuksen tuloksissa, vaikka tarkastelun kohteena oli levypallot kokonaisuutena ilman erittelyä hyökkäys- ja puolustuslevypalloihin. Myös Teramoton ja Crossin (2010) tutkimuksessa merkittäväksi tekijäksi noussut pallonmenetysten määrä on suhteellisen merkittävä tässä tutkimuksessa. Vapaaheittojen merkitys jäi kuitenkin yllättävän pieneksi, mikä tosin näkyy myös Li et al. (2021) tutkimuksessa.

Toiseen alatutkimuskysymykseen vastataan ennustemallin perusteella. Regressiomallin kertoimilla pystyttiin suhteellisen hyvin ennustamaan NBA-kauden 2018/2019 joukkueiden voittojen määriä. Kolmelle eli kymmenelle prosentille joukkueista malli osasi ennustaa toteutuneen arvon, ja 16 eli 53.33 prosenttia ennustuksista oli korkeintaan kahden voiton päässä toteutuneesta, mitä voidaan pitää hyvänä tuloksena. Toisaalta myös kymmenen prosenttia oli kuuden voiton päässä, mikä on jo kaukana toteutuneesta.

Tämä tutkimus vahvistaa aikaisempia tutkimustuloksia merkittävimmistä tekijöistä, jotka vaikuttavat NBA-joukkueen suoriutumiseen runkosarjassa, ja tämä tieto on tärkeää esimerkiksi valmentajille ja muille seuraorganisaation päättäjille, jotka pystyvät vaikuttamaan NBA-joukkueen kokoonpanoon. Joukkueet pystyvät arvioimaan omien ottelutilastojensa perusteella missä ne suoriutuvat hyvin ja missä eivät, jolloin esimerkiksi uusien pelaajien hankintaa voidaan suhteuttaa tähän tietoon, esimerkiksi tarvitaanko joukkueeseen parempia puolustuspelaajia, vai suurempaa hyökkäysvoimaa. Lisäksi tutkimuksessa luotu ennustemalli voi antaa päättäjille kuvaa siitä, millaista voittojen määrää joukkueelle voisi olla ennustettavissa ottelutilastojen perusteella kauden aikana. Ylipäätään ymmärtämällä paremmin joukkueen suoriutumiseen vaikuttavia tekijöitä, voivat joukkueet pärjätä paremmin kilpailussa muita vastaan ja täten menestyä paremmin.

5.2 Tutkimuksen luotettavuus ja jatkotutkimusehdotukset

Tutkimuksen luotettavuutta arvioidessa tulee ottaa huomioon tarkasteluperiodin pituus, eli tutkimuksessa hyödynnettiin ottelutilastoja ainoastaan viiden kauden ajalta, joten otos ei ole erityisen suuri. Lisäksi on hyvä huomioida rajoitukset, joita on tehty, eli tutkimuksen kohteena on ollut NBA ja sen runkosarja, eli mallin tuloksia ei voida yleistää muihin koripalloliigoihin, koska eri koripallosarjoissa säännöt, esimerkiksi peliaika, ovat erilaisia.

Aineiston luotettavuudessa täytyy arvioida sen oikeellisuutta, mikä tässä tutkimuksessa toteutuu. Regressiomallin luotettavuutta arvioidessa tulee ottaa huomioon mallin yleistettävyys ja sen taustaoletukset. Regressiomallin taustaoletusten paikkaansa pitävyyttä testattiin, ja epäiltiin mahdollisuutta heteroskedastisuutta. Tämän vuoksi mallissa käytettiin korjattuja keskivirheitä. Malli ja sen selittävät muuttujat olivat tilastollisesti merkitseviä, mutta mallia ei validoitu vertailemalla tuloksia uudessa otoksessa. Tulokset kuitenkin mukailivat aikaisempaa tutkimusta, joten sen puolesta niitä voidaan pitää suhteellisen luotettavana. Lisäksi ennustemallin toimivuutta testattiin ainoastaan yhteen NBA-kauteen, eli ei ole varmuutta mallin toimivuudesta myös muilla NBA-kausilla.

Mielenkiintoinen jatkotutkimus olisi esimerkiksi merkittävimpien tekijöiden vertailu runkosarjan ja pudotuspelien välillä, jossa ottelut ovat tasaisempia ja enemmän paineistettuja. Tämän lisäksi joukkueen suoriutumista voisi tutkia myös yksittäisen pelin tasolla. Nämä voisivat antaa päätöksentekijöille tietoa siitä, mitä joukkueelta vaatii suoriutua hyvin runkosarjan lisäksi pudotuspeleissä, tai miten yksittäiseen peliin vaikuttaviin tekijöihin pystyisi mahdollisesti valmistautumaan ennen pelejä. Lisäksi yksittäisten pelaajien vaikutusta joukkueen suoriutumiseen olisi mielenkiintoista tutkia. Tämä tarjoaisi arvokasta tietoa pelaajien hankintaan, jos yksittäisiä pelaajia pystyttäisiin etukäteen datan avulla sovittamaan nykyiseen kokoonpanoon, ja tutkimaan mahdollisia vaikutuksia siihen.

Lähteet

Atkinson, G. & Nevill, A. M. (2001) Selected issues in the design and analysis of sport performance research. Journal of Sports Sciences. 19, 811–827.

Abeza, G., O'Reilly, N., Nadeau, J. & Abdourazakou, Y. (2022) Big data in professional sport: the perspective of practitioners in the NFL, MLB, NBA, and NHL. Journal of Strategic Marketing. 1–21.

Ali, A., Qadir, J., Rasool, R. Sathiaseelan, A., Zwitter, A. & Crowcroft, J. (2016) Big data for development: applications and techniques. Big Data Analytics. 1 (1), 2.

Angelov, P., Gu, X. & Kangin, D. (2017) Empirical data analytics. International Journal of Intelligent Systems. 32 (12), 1261–1284.

Baghal, T. (2012) Are the "four factors" indicators of one factor? An application of structural equation odeling methodology to NBA data in prediction of winning percentage. Journal of Quantitative Analysis in Sports. 8 (1).

Basketball-Reference. (2023) [verkkodokumentti]. [Viitattu 17.11.2023]. Saatavilla: https://www.basketball-reference.com/

Baumer, B. S., Matthews, G. J. & Nguyen, Q. (2023) Big ideas in sports analytics and statistical tools for their investigation. Wiley Interdisciplinary Reviews: Computational Statistics.

Bradbury, J. C. (2019) Determinants of revenue in sports leagues: an empirical assessment. Economic Inquiry. 57 (1), 121–140.

Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R. & Ramanujam, K. (1997) Advanced scout: data mining and knowledge discovery in NBA data. Data Mining and Knowledge Discovery. 1 (1), 121–125.

Cabarkapa, D., Deane, M. A., Fry, A., Jones, G. T., Cabarkapa, D. V., Philipp, N. M., Yu, D. & Abián-Vicén, J. (2022) Game statistics that discriminate winning and losing at the NBA level of basketball competition. Plos One. 17 (8), 0273427

Caporale, T. & Collier, T. C. (2015) To three or not to three? Shot selection and managerial performance in the national basketball association. Journal of Labor Research. 36 (1), 1–8.

Chartier, T. (2017) Valuing data. The Journal of Corporate Accounting and Finance. 28 (2), 88–89.

Chen, C. L. P. & Zhang, C.-Y. (2014) Data-intensive applications, challenges, techniques and technologies: a survey on big data. Information Sciences. 275, 314–347.

Colás, S. (2020) Numbers don't lie : new adventures in counting and what counts in basketball analytics. Lincoln: University of Nebraska Press.

Davenport, T. H. (2014) What businesses can learn from sports analytics. MIT Sloan Management Review. 55 (4), 10–13.

Davenport, T. H. & Harris, J. G. (2017) Competing on analytics: the new science of winning. Boston: Harvard Business Review Press.

Davenport, T. H. & Kim, J. (2013) Keeping up with the quants: your guide to understanding and using analytics. Boston: Harvard Business Review Press.

Dehesa, R., Vaquera, A., Gonçalves, B., Mateus, N., Gomez-Ruano, M.-Á. & Sampaio, J. (2019) Key game indicators in NBA players' performance profiles. Kinesiology. 51 (1), 92–101.

Dubey, R., Gunasekaran, A., Childe, S. J., Blome, C. & Papadopoulos, T. (2019) Big data and predictive analytics and manufacturing performance: integrating institutional theory, resource-based view and big data culture. British Journal of Management. 30, 341–361.

Duquette, C. M., Cebula, R. J. & Mixon, F. G. (2019) Major League Baseball's Moneyball at age 15: a re-appraisal. Applied Economics. 51 (52), 5694–5700.

Elitzur, R. (2020) Data analytics effects in Major League Baseball. Omega. 90, 102001.

ESPN. (2023) NBA. Stats. [verkkodokumentti]. [Viitattu 17.11.2023]. Saatavilla: https://www.espn.com/nba/stats

Gong, H. (2022) The effect of the crowd on home bias: evidence from NBA games during the COVID-19 pandemic. Journal of Sports Economics. 23 (7), 950–975.

Groll, A., Manisera, M., Schauberger, G. & Zuccolotto, P. (2018) Guest editorial 'statistical modelling for sports analytics'. Statistical Modelling. 18 (5–6), 385–387.

Hill, R. C., Griffiths, W. E. & Judge, G. G. (2001) Undergraduate econometrics. 2. p. Hoboken: John Wiley & Sons.

Hill, R. C., Griffiths, W. E. & Lim, G. C. (2018) Principles of econometrics. 5. p. Hoboken: John Wiley & Sons.

Horvat, T., Job, J., Logozar, R. & Livada, Č. (2023) A data-driven machine learning algorithm for predicting the outcomes of NBA games. Symmetry. 15 (4), 798.

Houghton, D. M., Nowlin, E. L. & Walker, D. (2019) From fantasy to reality: the role of fantasy sports in sports betting and online gambling. Journal of Public Policy & Marketing. 38 (3), 332-353.

Hurwitz, J. Nugent, A., Halper, F. & Kaufman, M. (2013) Big data for dummies. Hoboken: John Wiley & Sons.

Kubatko, J., Oliver, D., Pelton, K. & Rosenbaum, D. T. (2007) A starting point for analyzing basketball statistics. Journal of Quantitative Analysis in Sports. 3 (3), 1.

Lewis, M. (2003) Moneyball: the art of winning an unfair game. New York: W. W. Norton.

Li, Y., Wang, L. & Li, F. (2021) A data-driven prediction approach for sports team performance and its application to National Basketball Association. Omega. 98, 102123.

Liberatore, M. & Luo, W. (2010) The analytics movement: implications for operations. Interfaces. 40 (4), 313–324.

Lorenzo, A., Gómez, M. Á., Ortega, E., Ibáñez, S. J. & Sampaio, J. (2010) Game related statistics which discriminate between winning and losing under-16 male basketball games. Journal of Sports Science & Medicine. 9 (4), 664–668.

Mandić, R., Jakovljević, S., Erčulj, F. & Štrumbelj, E. (2019) Trends in NBA and Euroleague basketball: analysis and comparison of statistical data from 2000 to 2017. Plos One. 14 (10), 0223524.

Mohbey, K. K., Pandey, A. & Rajput, D. S. (2020) Predictive analytics using statistics and big data: concepts and modeling. Singapore: Bentham Science Publishers.

Mondello, M. & Kamke, C. (2014) Management Whitepaper: The introduction and application of sports analytics in professional sport organizations: A case study of the Tampa Bay Lightning. Journal of Applied Sport Management. 6 (2).

Morgulev, E., Azar, O. H. & Lidor, R. (2018) Sports analytics and the big-data era. International Journal of Data Science and Analytics. 5 (4), 213–222.

National Basketball Association. (2016) NBA announces multiyear partnership with Sportradar and Second Spectrum. [verkkodokumentti]. [Viitattu 17.11.2023]. Saatavilla: https://pr.nba.com/nba-announces-multiyear-partnership-sportradar-second-spectrum/

National Basketball Association. (2020a) Everything you need to know about the 2019-20 NBA season restart. [verkkodokumentti]. [Viitattu 4.10.2023]. Saatavilla: https://www.nba.com/nba-returns-2020-faq

National Basketball Association. (2020b) NBA announces structure and format for 2020-21 season. [verkkodokumentti]. [Viitattu 4.10.2023]. Saatavilla: https://www.nba.com/news/nba-announces-structure-and-format-for-2020-21-season

National Basketball Association. (2023a) Stats. [verkkodokumentti]. [Viitattu 4.10.2023]. Saatavilla: https://www.nba.com/stats/help/faq

National Basketball Association. (2023b) NBA and Genius Sports/Second Spectrum expand partnership to deepen NBA league pass innovations with enhanced basketball analytics and develop new next gen platform. [verkkodokumentti]. [Viitattu 17.11.2023]. Saatavilla: https://pr.nba.com/nba-genius-sports-second-spectrum-expanded-partnership/

NBAstuffer. (2023) How the NBA schedule is made. [verkkodokumentti]. [Viitattu 4.10.2023]. Saatavilla: https://www.nbastuffer.com/analytics101/how-the-nba-schedule-is-made/

Olive, D. J. (2017) Linear regression. Cham: Springer International Publishing.

Oliver, D. (2004) Basketball on paper: rules and tools for performance analysis. 1. p. Lincoln: Potomac Books.

Otten, M. P. & Miller, T. J. (2015) A balanced team wins championships: 66 years of data from the National Basketball Association and the National Football League. Perceptual and Motor Skills. 121 (3), 654–665.

Park, H. M. (2010) Practical guides to panel data analysis. [verkkodokumentti]. [Viitattu 22.12.2023]. Saatavilla: http://www.iuj.ac.jp/faculty/kucc625/writing/panel_guidelines.pdf

Patel, D., Shah, D. & Shah, M. (2020) The intertwine of brain and body: a quantitative analysis on how big data influences the system of sports. Annals of data science. 7(1), 1–16.

Qin, S. J. (2014) Process data analytics in the big data era. AIChE Journal. 60 (9), 3092–3100.

Rashedi, J. (2022) The data-driven organization: using data for the success of your company. 1. p. Cham: Springer Cham.

Runkler, T. A. (2016) Data analytics models and algorithms for intelligent data analysis. 2. p. Wiesbaden: Springer Vieweg Wiesbaden.

Sampaio, J. & Janeira, M. (2003) Statistical analyses of basketball team performance: understanding teams' wins and losses according to a different index of ball possessions. International Journal of Performance Analysis in Sport. 3 (1), 40–49.

Sarlis, V. & Tjortjis, C. (2020) Sports analytics — evaluation of basketball players and team performance. Information Systems. 93, 101562.

Schumaker, R. P., Solieman, O. K. & Chen, H. (2010) Sports data mining. In: Sharda, R & Voß, S. (toim.) Integrated Series in Information Systems. Boston: Springer US.

Spada, A., Fiore, M. & Galati, A. (2023) The impact of education and culture on poverty reduction: evidence from panel data of European countries. Social Indicators Research.

Tang, L. & Meng, Y. (2021) Data analytics and optimization for smart industry. Frontiers of Engineering Management. 8 (2), 157–171.

Teramoto, M. & Cross, C. L. (2010) Relative importance of performance factors in winning NBA games in regular season versus playoffs. Journal of Quantitative Analysis in Sports. 6 (3).

Thabtah, F., Zhang, L. & Abdelhamid, N. (2019) NBA game result prediction using feature analysis and machine learning. Annals of Data Science. 6 (1), 103–116.

Tsai, C.-W., Lai, C.-F., Chao, H.-C. & Vasilakos, A. V. (2015) Big data analytics: a survey. Journal of Big Data. 2 (1), 1–32.

Tukey, J. W. (1962) The future of data analysis. The Annals of Mathematical Statistics. 33 (1), 1–67.

Watanabe, N. M., Shapiro, S. & Drayer, J. (2021) Big data and analytics in sport management. Journal of Sport Management. 35, 197–202.

Wooldridge, J. M. (2010) Econometric analysis of cross section and panel data. 2. p. Cambridge: MIT Press.

Wu, S. & Bornn, L. (2018) Modeling offensive player movement in professional basketball. The American Statistician. 72 (1), 72–79.

Yahoo Sports. (2023) NBA. Stats. [verkkodokumentti]. [Viitattu 17.11.2023]. Saatavilla: https://sports.yahoo.com/nba/stats/

Yan, X. & Su, X. (2009) Linear regression analysis theory and computing. Singapore: World Scientific Publishing.

Liitteet

Liite 1. Kiinteiden vaikutusten F-testin tulokset

Tässä taulukossa on esiteltynä kiinteiden vaikutusten F-testin tulokset, jolla verrataan kiinteiden vaikutusten menetelmän soveltuvuutta verrattuna yhdistetyn pienimmän neliösumman menetelmään.

H0: Ei kiinteitä vaikutuksia

H1: On kiinteitä vaikutuksia

F	Prob > F
2.08	0.0035

Liite 2. Breusch-Paganin satunnaisten vaikutusten testin tulokset

Tässä taulukossa on esiteltynä satunnaisten vaikutusten Breusch-Pagan testin tulokset, jolla verrataan satunnaisten vaikutusten menetelmän soveltuvuutta verrattuna yhdistetyn pienimmän neliösumman menetelmään.

H0: Ei satunnaisia vaikutuksia

H1: On satunnaisia vaikutuksia

$\chi 2$	Prob > $\chi 2$
7.67	0.0028

Liite 3. Hausman-testin tulokset

Tässä taulukossa on esiteltynä Hausman-testin tulokset joilla on tutkittu, onko satunnaisten vaikutusten malli konsistentti.

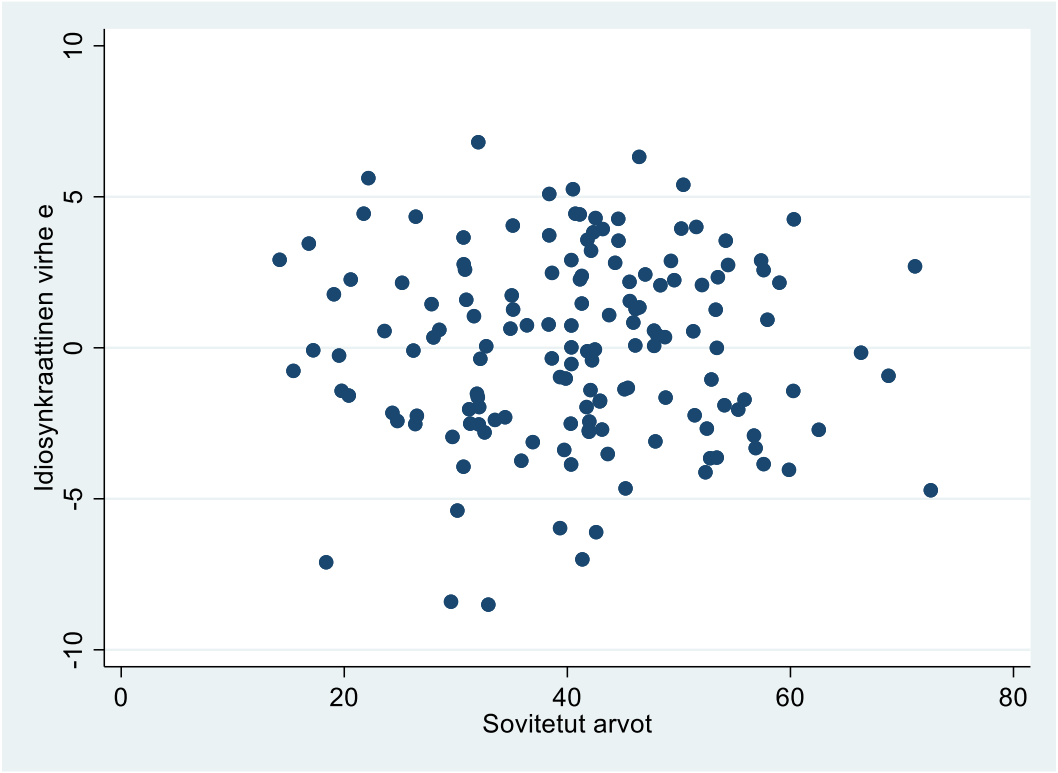
H0: Satunnaisten ja kiinteiden vaikutusten mallien kertoimissa ei ole eroa

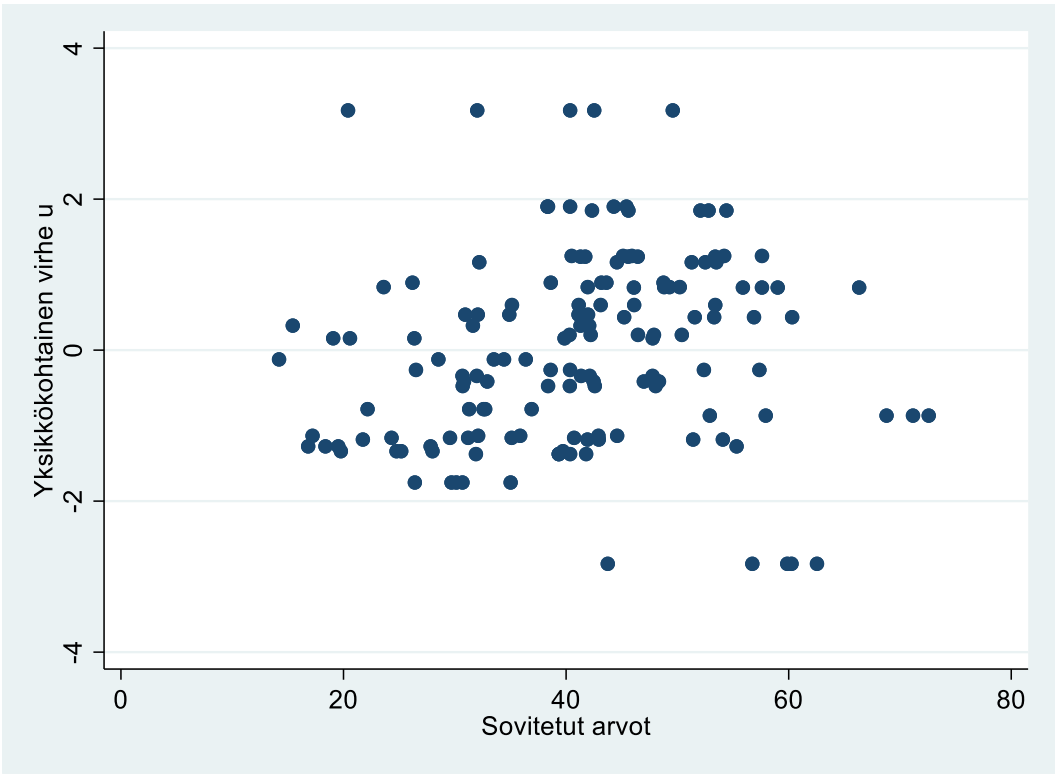
H1: Satunnaisten ja kiinteiden vaikutusten mallien kertoimissa on eroa

$\chi 2$	Prob > $\chi 2$
5.38	0.7159

Liite 4. Residuaalikuvaajat

Tässä on esiteltynä ensimmäisessä kuvassa idiosynkraattinen jäännöstermi suhteessa mallin ennustamiin eli sovitettuihin arvoihin, ja toisessa kuvassa yksikkökohtaisten erojen aiheuttama virhetermi suhtessaa sovitettuihin arvoihin





Liite 5. Wooldridgen autokorrelaatiotestin tulokset

Tässä taulukossa on esiteltynä Wooldridgen autokorrelaatiotestin tulokset, joilla on testattu onko mallissa ensimmäisen asteen autokorrelaatiota.

H0: Ei ensimmäisen asteen autokorrelaatiota

H1: On ensimmäisen asteen autokorrelaatiota

F	Prob > F
1.35	0.2554

Liite 6. Pearsonin korrelaatiotestien tulokset

Tässä taulukossa on esiteltynä korrelaatiomatriisi regressiomallin selittävistä muuttujista, joille on laskettu korrelaatiot hyödyntäen Pearsonin testiä. Korrelaatiokertoimen alapuolella on sulkuihin merkitty testin p-arvo.

	EFG%	FTM	OPPFTM	REB	TOV	STL	BLK	OPPEFG%
EFG%	1.0000							
FTM	-0.0447	1.0000						
	(0.5869)							
OPPFTM	-0.2210	0.1133	1.0000					
	(0.0066)	(0.1674)						
REB	-0.2454	0.1619	-0.1556	1.0000				
	(0.0025)	(0.0478)	(0.0573)					
TOV	-0.1425	0.0449	0.2927	-0.0386	1.0000			
	(0.0820)	(0.5856)	(0.0003)	(0.6387)				
STL	0.1880	-0.0018	0.1417	-0.4266	0.2494	1.0000		
	(0.0212)	(0.9828)	(0.0838)	(0.0000)	(0.0021)			
BLK	0.1716	0.0318	0.0368	0.2370	0.1069	0.0422	1.0000	
	(0.0357)	(0.6997)	(0.6545)	(0.0035)	(0.1928)	(0.6077)		
OPPEFG%	-0.0269	-0.1054	-0.0524	-0.4994	-0.1121	0.0043	-0.5259	1.0000
	(0.7441)	(0.1993)	(0.5245)	(0.0000)	(0.1720)	(0.9580)	(0.0000)	

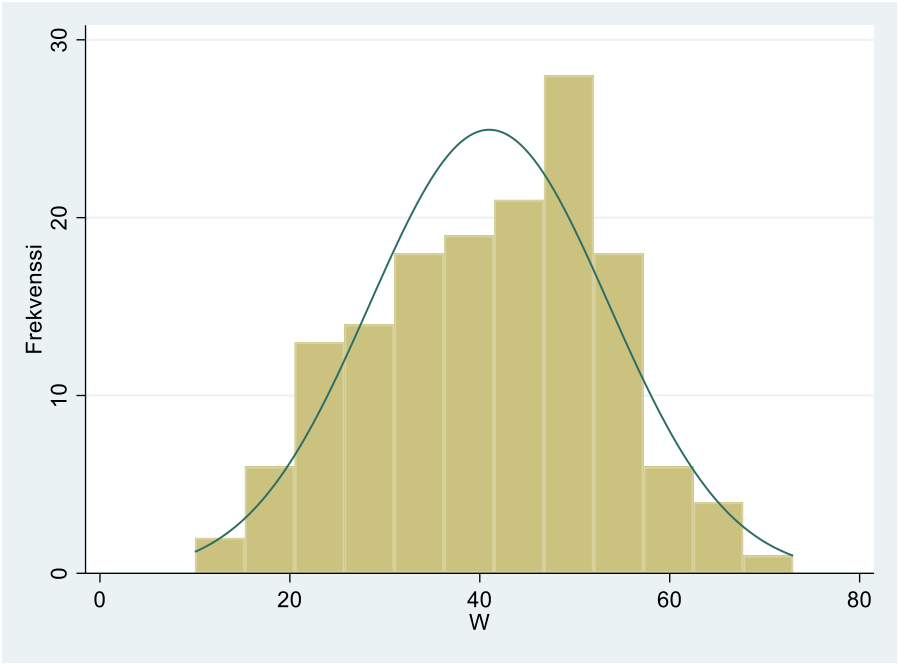
Liite 7. Selitettävän muuttujan normaalijakautuneisuuden tulokset

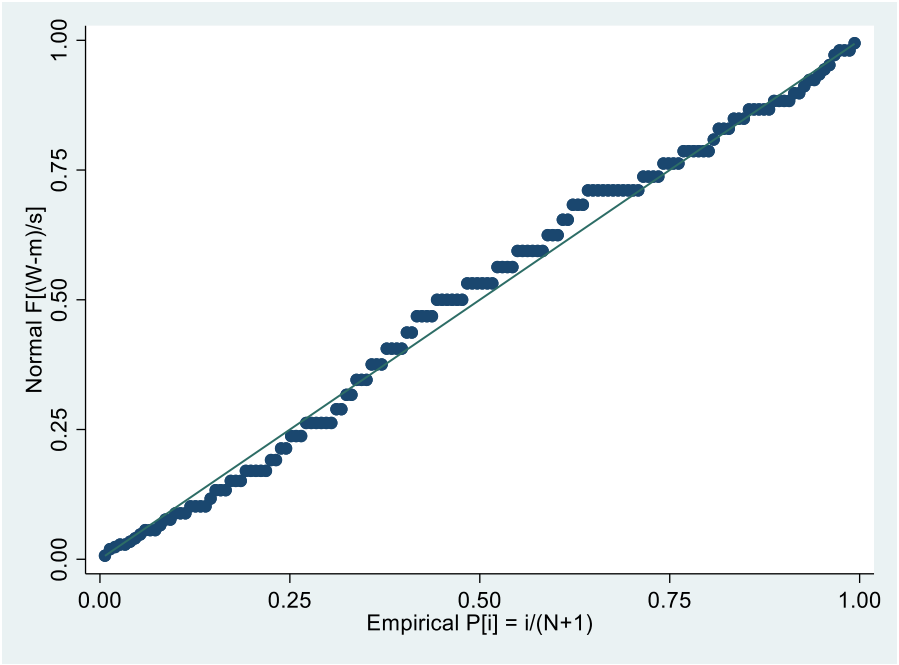
Tässä taulukossa on esiteltynä Shapiro-Wilk normaalijakaumatestin tulokset sekä kuvaajat normaalijakautuneisuudesta.

H0: Noudattaa normaalijakaumaa

H1: Ei noudata normaalijakaumaa

	W	Prob > F
	0.9914	0.5023





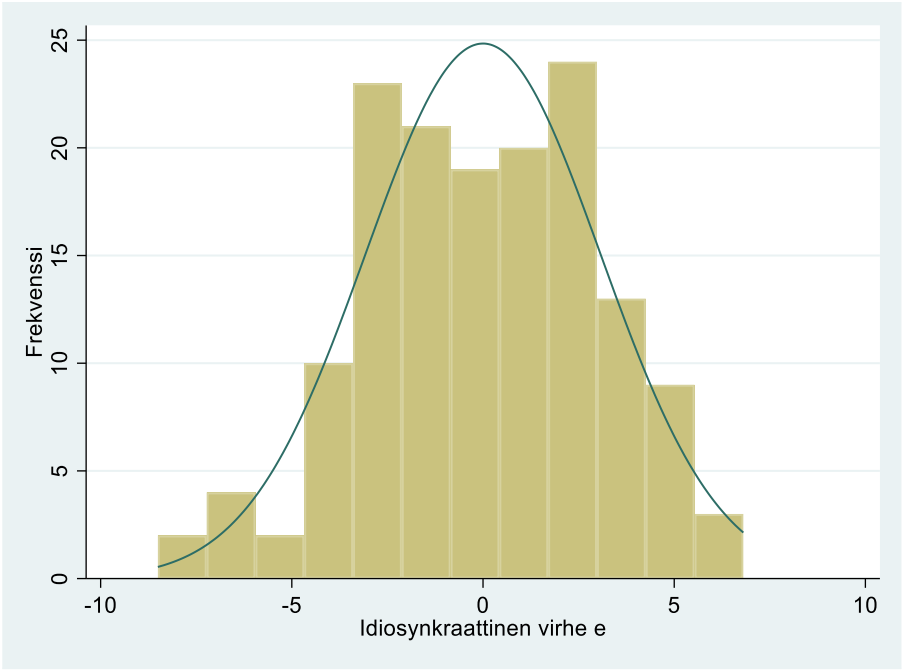
Liite 8. Idiosynkraattisen jäännöstermin normaalijakautuneisuuden tulokset

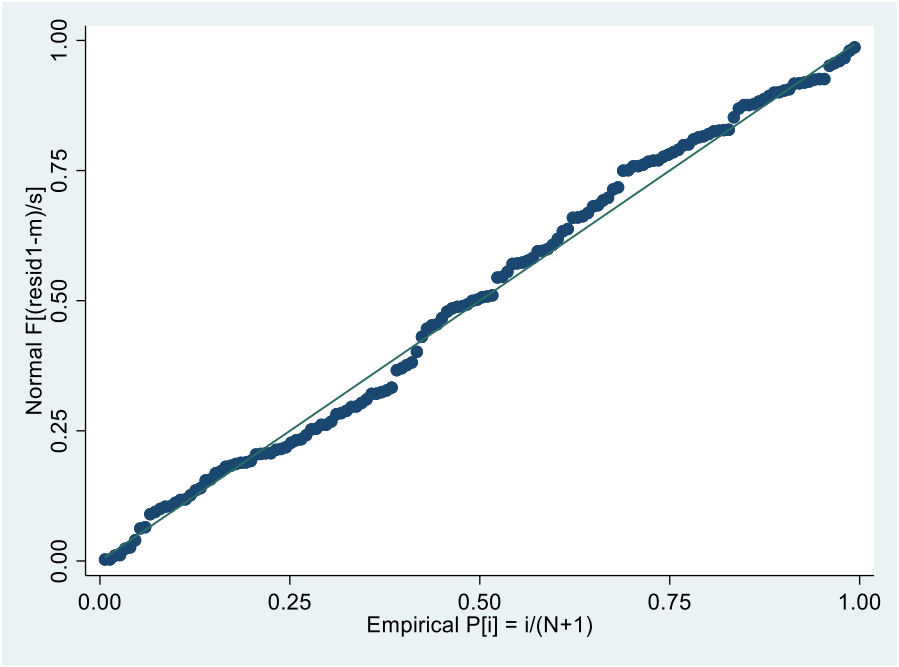
Tässä taulukossa on esiteltynä Shapiro-Wilk normaalijakaumatestin tulokset sekä kuvaajat normaalijakautuneisuudesta.

H0: Noudattaa normaalijakaumaa

H1: Ei noudata normaalijakaumaa

	W	Prob > F
	0.9866	0.1551





Liite 9. Yksikkökohtaisen jäännöstermin normaalijakautuneisuuden tulokset

Tässä taulukossa on esiteltynä Shapiro-Wilk normaalijakaumatestin tulokset sekä kuvaajat normaalijakautuneisuudesta.

H0: Noudattaa normaalijakaumaa

H1: Ei noudata normaalijakaumaa

	W	Prob > F
	0.9822	0.0497