LUT-kauppakorkeakoulu

Kauppatieteiden kandidaatintutkielma

Talousjohtaminen

Big datan hyödyntämisen vaikutus pörssiyhtiöiden kannattavuuteen

The Impact of the Utilization of Big Data on the Profitability of Listed Companies

6.1.2020

Tekijä: Essi Rajala

Ohjaaja: Pontus Huotari

TIIVISTELMÄ

Tekijä:	Essi Rajala
Tutkielman nimi:	Big datan hyödyntämisen vaikutus pörssiyhtiöiden kannattavuuteen
Akateeminen yksikkö:	LUT-kauppakorkeakoulu
Koulutusohjelma:	Kauppatiede / Talousjohtaminen
Ohjaaja:	Pontus Huotari
Hakusanat:	Big data, kannattavuus, data-analytiikka, paneelidata, regressioanalyysi

Tämän kandidaatintutkielman tarkoituksena on selvittää, miten big datan hyödyntäminen vaikuttaa yrityksen kannattavuuteen. Teoriaosassa esitellään kannattavuuden, big datan ja data-analytiikan teoriaa. Teorian avulla valitaan tutkimuksen kannalta sopivat muuttujat: selittävänä muuttujana toimii data-analyytikoiden määrä ja selitettävinä muuttujina liiketulosprosentti ja sijoitetun pääoman tuottoprosentti.

Tutkimuksen aineisto on kerätty kolmen vuoden ajalta ja se koostuu noin 130 pörssiyrityksen kannattavuuteen liittyvistä tunnusluvuista sekä kyseisissä yrityksissä työskentelevien data-analyytikoiden määristä. Tutkimus toteutetaan kvantitatiivisena tutkimuksena, ja tutkimusmenetelmänä käytetään paneelidatan regressioanalyysiä. Tutkimuksen luotettavuutta parannetaan muuttujien logaritmi- sekä viivästysmuunnoksilla, ja saatujen tulosten reliabiliteettia arvioidaan tutkimalla muun muassa mallien residuaalien normaalijakautuneisuutta sekä autokorrelaatiota.

Tutkimuksen tuloksista voidaan huomata kummankin tehdyn mallin perusteella, että big datan hyödyntämisellä ei ole tilastollisesti merkitsevää yhteyttä yrityksen kannattavuuteen.

ABSTRACT

Author:	Essi Rajala
Title:	The impact of the utilization of big data on the profitability of listed companies
School:	School of Business and Management
Degree programme:	Business Administration / Financial Management
Supervisor:	Pontus Huotari
Keywords:	Big data, profitability, data analytics, panel data, regression analysis

The aim of this bachelor's thesis is to examine the effect of the utilization of big data on the profitability of listed companies. The theory of profitability, big data and data analytics will be explained in the theory section. The proper variables for the research are selected based on the theory section: the quantity of data analysts will be used as the independent variable, and operating margin and return on investment as the two dependent variables.

The research material covers three subsequent fiscal years and it consists of the financial data as well as the quantity of data analysts of roughly 130 listed companies. The research will be executed as statistical research, and panel data regression analysis is used as the research method. Logarithmic and lagged transformations will be performed in order to improve the reliability of the research, and the reliability will be evaluated by investigating the distributions of the estimated residuals and the autocorrelation.

Based on the results of the research, neither of the used models indicate that the utilization of big data has a statistically significant effect on the profitability of the company.

SISÄLLYSLUETTELO

LIITTEET

1. Johdanto

Viime vuosien aikana big data on luonut aivan uudenlaisen tavan analysoida yrityksen liiketoimintaan ja markkinoihin liittyviä tekijöitä. Big data tarjoaa tavan tehostaa yritysten liiketoimintaa ja ymmärtää sekä olemassa olevien että potentiaalisten asiakkaiden tarpeita entistä paremmin (Aggarwal, 2016). Viimeaikainen kiinnostus big dataan on saanut monet yritykset kehittämään kykyjään big data -analytiikan saralla parantaakseen suorituskykyään (Santhanam & Hartono, 2003). Hiljattain tehdyssä McKinseyn tutkimuksessa lähes jokainen kyselyyn osallistunut johtohenkilö kertoi yrityksensä tehneen huomattavia investointeja muun muassa analytiikkaohjelmiin sekä tietokantoihin. Big data -analytiikan hyödyntäminen lupailee yrityksille jopa 6% korkeampaa kannattavuutta. (Akter, 2016).

Vuonna 2013 lähes sadalle suuryrityksen johtajalle teetetystä kyselystä käy ilmi, että kyseisistä yrityksistä noin 91% investoi big data -projekteihin, kun samainen luku vuotta aiemmin oli 85% (Kiron, Prentice & Ferguson 2014). Etenkin suuret yritykset käyvät siis entistä kovempaa kilpailua siitä, kuka pystyy hyödyntämään big dataa liiketoiminnassaan ja päätöksenteossaan parhaiten.

Aiheena big data ja sen hyödyntäminen on siis ajankohtainen ja erittäin mielenkiintoinen. Aikaisempia tutkimuksia big datasta ja data-analytiikasta on paljonkin, mutta tutkimuksia big datan suorista hyödyistä yritysten kannattavuuteen on aiheen ajankohtaisuuteen verrattuna tehty melko vähän.

1.1 Tutkimuksen tavoitteet ja tutkimuskysymys

Tutkimuksen aiheena on big datan hyödyntämisen vaikutukset yrityksen kannattavuuteen. Tutkimuksen päätavoitteena on tutkia, onko big datalla ja data-analytiikalla vaikutuksia yrityksen kannattavuuteen ja jos on, niin minkä suuntaisia ja suuruisia vaikutukset ovat. Tutkimuksen päätavoitteen perusteella muotoiltu päätutkimuskysymys on täten seuraavanlainen:

”Miten big datan hyödyntäminen vaikuttaa yrityksen kannattavuuteen?”

Tutkimuksen toissijaisina tavoitteina on ensinnäkin selvittää, että mitkä tekijät kannattavuuteen vaikuttavat ja miten kannattavuutta voidaan mitata. Tämän lisäksi tutkimuksessa otetaan selvää siitä, millä keinoilla big data ja data-analytiikka voitaisiin valjastaa yrityksen käyttöön sellaisella tavalla, että se tuottaisi yritykselle mahdollisimman paljon taloudellista arvoa.

1.2 Tutkimuksen rajaukset

Tutkimus on rajattu koskemaan vain pörssiyhtiöitä, lähinnä sen takia että listatuista yhtiöistä on löydettävissä parempaa ja luotettavampaa dataa kuin listaamattomista yhtiöistä. Big datan hyödyntämisen osalta aineisto on kerätty vuosilta 2013 ja 2014 ja sen mittarina käytetään yrityksissä työskentelevien data-analyytikkojen määrällä. Kannattavuuteen liittyvä finanssidata on kerätty vuosilta 2013-2015.

Historia on osoittanut sen, että vie aikaa, että uusien teknologioiden käyttöönotto lopulta tuottaa yrityksen tavoittelemia konkreettisia hyötyjä. Myös big data -analytiikka noudattaa tätä samaa kaavaa, hyötyjen kasvaessa sitä enemmän, mitä enemmän aikaa kuluu. Bughinin (2016) tekemässä tutkimuksessa yritysten voittojen kasvu big da-

taan tehtyjen investointien ansiosta oli 6 prosenttia. Tämä kasvoi 9 prosenttiin kun kyseisiä voittoja tutkittiin viiden vuoden aikavälillä. Tämän takia kannattavuuteen liittyvää finanssidataa on kerätty pidemmältä aikaväliltä, jotta voitaisiin tutkia, miten data-analyytikkojen määrä vaikuttaa yritysten kannattavuuteen yli ajan.

1.3 Tutkimusaineisto ja -menetelmät

Tutkimus toteutetaan kvantitatiivisena tutkimuksena ja pääasiallisena analyysimenetelmänä käytetään paneelidatan regressioanalyysiä. Oikean estimointimenetelmän valitsemisen apuna toteutetaan kolme eri testiä. Tämän lisäksi tutkimuksen lopussa pohditaan tutkimuksen reliabiliteettiin vaikuttavia tekijöitä.

Tutkimusaineisto koostuu 132 pörssiyrityksen kannattavuuteen sekä kokoon liittyvästä datasta sekä vastaavissa yrityksissä töissä olevien data-analyytikkojen määrästä. Data-analytiikkojen määrän osalta data on kerätty vuosilta 2013 ja 2014, yritysten kannattavuuteen liittyvä data taas vuosilta 2013-2015. Kannattavuuteen liittyvää dataa on päätetty kerätä yhdeltä vuodelta enemmän, jotta voitaisiin tutkia, miten data-analyytikkojen määrä vaikuttaa yritysten kannattavuuteen yli ajan. Tutkimusmenetelmästä sekä -aineistosta kerrotaan lisää tutkielman kolmannessa luvussa.

1.4 Tutkimuksen rakenne

Tutkielma koostuu viidestä pääluvusta, joista ensimmäinen on johdanto. Johdantoluvun jälkeen siirrytään käsittelemään tutkielman teoreettista viitekehystä, joka koostuu aikaisempien tutkimusten läpikäymisestä sekä big dataan, data-analytiikkaan ja kannattavuuteen liittyvästä teoriasta. Kolmannessa luvussa esitellään tutkimuksessa käytettävä aineisto sekä paneelidatan tutkimusmenetelmien teoriaa. Neljäs luku on tutkielman empiirinen osuus, jossa käsitellään tehdyn tutkimuksen perusteella saatuja tuloksia. Lisäksi neljännessä luvussa pohditaan tutkimuksen reliabiliteettia sekä tutkimuksen kulkuun vaikuttaneita rajoitteita.

Tutkielma päättyy viidennen kappaleen yhteenvetoon, jossa kootaan yhteen tutkielman pääteemat, tiivistetään saadut tutkimustulokset sekä vastataan asetettuun tutkimuskysymykseen. Näiden lisäksi viimeisessä luvussa pohditaan mahdollisia aiheita jatkotutkimusta varten.

2. Kannattavuuden ja big datan teoriaa

Tässä luvussa käydään läpi tutkimuksen teoreettista viitekehystä aikaisempien tutkimusten, teorioiden sekä käsitteiden avulla. Ensimmäisessä alakappaleessa kerrotaan lyhyesti kannattavuuden määritelmästä sekä tunnusluvuista. Toisessa alakappaleessa siirrytään käsittelemään big dataa ja data-analytiikkaa tarkemmin, jonka jälkeen kolmannessa alakappaleessa tutkitaan sitä, miten yritykset voivat käyttää big data -analytiikkaa päätöksenteon tukena. Viimeinen alakappale käsittelee sitä, miten big data -analytiikkakyvykkyyttä voidaan mitata ja miksi sillä on tärkeä merkitys yrityksen kannattavuuden parantamisessa.

2.1 Kannattavuuden mittaaminen

Kannattavuus kuvaa liiketoiminnan taloudellista tulosta ja se on jatkuvan liiketoiminnan perusedellytys. Kannattavuutta voidaan mitata joko absoluuttisten lukujen avulla, tai vaihtoehtoisesti taseen tai tuloslaskelman eriin suhteutetuilla tuottomittareilla. Absoluuttisten kannattavuuden lukujen käyttö ei sellaisenaan ole kovinkaan hyödyllistä, koska niiden ajallinen ja yritysten välinen vertailukelpoisuus on heikko. (Ikäheimo, Malmi, Walden, 2016, 105) Tämän takia tutkimuksessa tullaan käyttämään vain suhteellisia kannattavuuden tunnuslukuja.

Liiketulosprosentti on erittäin yleisesti käytetty kannattavuuden mittari. Se kertoo, kuinka paljon varsinaisen liiketoiminnan tuotoista on jäänyt jäljelle juoksevien kulujen jälkeen ennen rahoituseriä ja veroja. Toisin sanoen liiketulosprosentti kuvaa sitä,

kuinka kustannustehokkaasti yritys toimii (Ikäheimo, Laitinen, Laitinen, Puttonen, 2014, 67) Liiketulosprosentti on monipuolinen tunnusluku, sillä se antaa yleiskattavan kuvan sekä yksittäisten yritysten kehityksestä, että sen kehityksestä verrattuna muihin yrityksiin. Liiketulosprosentin kaava on seuraavanlainen (kaava 1):

$$Liiketulos - \% = \frac{Liiketulos}{Liikevaihto} \times 100 \quad (1)$$

Sijoitetun pääoman tuottoprosentti (return on investment, ROI) kertoo nimensä mukaisesti sen, miten hyvin yritykseen sijoitettu pääoma on tuottanut. Se mittaa tuottoa, joka on saatu yritykseen sijoitetulle tuottoa vaativalle omalle ja vieraalle pääomalle. Voidaan myös sanoa, että sijoitetun pääoman tuottoprosentti ilmaisee, kuinka paljon yritys on saanut tuottoa suhteessa tämän tuoton saamiseksi tarvitulle pääomalle (Ikäheimo et al. 2014, 69). Sijoitetun pääoman tuottoprosentti voidaan laskea jakamalla tilikauden nettotulos, rahoituskulut ja verot keskimääräisellä sijoitetulla pääomalla (kaava 2):

$$Sijoitetun\ pääoman\ tuotto - \% = \frac{Nettotulos+rahoituskulut+verot}{Sijoitettu\ pääoma\ keskimäärin} \times 100 \quad (2)$$

Sekä liiketulosprosenttia että sijoitetun pääoman tuottoprosenttia voidaan yrityksen aiempien tilikausien vastaavien lukujen lisäksi verrata muihinkin yrityksiin, jolloin voidaan arvioida, kuinka kilpailukykyinen yritys on verrattuna toisiin yrityksiin. (Ikäheimo et al. 2016, 106) Kannattavuuteen vaikuttaa useat eri tekijät, niin yrityksen koosta ja kasvusta pienempiinkin seikkoihin. Seuraavissa alakappaleissa käydään läpi big datan ja data-analytiikan mahdollistamia hyötyjä, sekä sitä, miten niillä voidaan vaikuttaa yrityksen kannattavuuteen.

2.2 Big data ja data-analytiikka

Big datalle on useita erilaisia määritelmiä, joista osa pyrkii vastaamaan siihen, mitä big data on ja osa taas siihen, mitä sen avulla voidaan tehdä. Tiivistetysti big data kuitenkin voidaan määritellä useista eri lähteistä generoiduiksi dataseteiksi, joiden avulla voidaan esimerkiksi tehostaa yrityksen päätöksentekoa sekä liiketoimintaa, tehdä tarkempia ennusteita tulevaisuudesta sekä parantaa taloudellista kannattavuutta. (Gandomi & Haider, 2015)

Yleisesti käytetty tapa kuvata big dataa on 3V-malli, jonka ulottuvuuksia ovat määrä, monimuotoisuus sekä nopeus (volume, variety & velocity). 3V-mallissa big data määritellään ”suureksi *määräksi monimuotoista* dataa, joka on luotu, tallennettu ja prosessoitu suurella *nopeudella*”. (Laney, 2001) Toinen määritelmä big datalle on seuraavanlainen: ”Big data on määrältään, nopeudeltaan ja monimuotoisuudeltaan suuri tietoresurssi, joka vaatii kustannustehokkaita ja innovatiivisia tapoja käsitellä tietoa. Datan analysointi antaa yritysjohdolle syvemmän ymmärryksen markkinoista, tehostaa päätöksentekoa sekä auttaa automatisoimaan yrityksen prosesseja.” (Gartner IT Glossary, n.d.)

Kuitenkaan pelkästään big datan keräämisestä ja varastoimisesta ei vielä itsessään koidu yritykselle minkäänlaista hyötyä. Big datan tuoma potentiaali vapautuu vasta silloin, kun sitä ruvetaan prosessoimaan ja analysoimaan oikeilla työkaluilla osaavien data-analyytikkojen toimesta. Data-analytiikka voidaan määritellä kyvyksi käsitellä, analysoida ja muokata dataa, jotta sen avulla voidaan havaita hyödyllisiä huomioita sekä tukea päätöksentekoa (Wu, Hitt & Lou, 2019). Big data -analytiikkaa pidetään suunnanmuuttajana, joka mahdollistaa liiketoiminnan tehostamisen ja kannattavuuden parantamisen sen suuren operatiivisen ja strategisen potentiaalinsa takia (Wamba, 2017).

Big data -prosessit voidaan jakaa kahteen eri luokkaan: datan hallintaan ja analytiikkaan. Datan hallintaan kuuluu sen hankkiminen ja tallentaminen, purkaminen ja siistiminen sekä integroiminen, kokoaminen ja esittäminen. Analytiikkaan sen sijaan kuuluu datan hallintaprosessin avulla hankitun valmiin datan mallintaminen ja analyysi sekä analyysiin pohjautuva tulkinta. (Gandomi & Haider, 2015) Tulee siis pitää mielessä, että datan hallinta ja itse data-analytiikka ovat kaksi erillistä vaihetta ja oikeastaan vasta analytiikkaprosessissa saadaan konkreettisia, kuten esimerkiksi taloudellisia hyötyjä. Tähän prosessien kahtiajakoon kuitenkin tekee poikkeuksen sellaiset yritykset, joiden liiketoimeen kuuluu datan kerääminen ja kokoaminen myyntitarkoituksessa muita yrityksiä varten, jolloin vain varsinainen analytiikkaprosessi jää datan ostavan yrityksen vastuulle.

Kyky muokata ja analysoida tätä valtavaa datamassaa yrityksen tarpeisiin sopivaan muotoon data-analytiikan työkalujen avulla antaa yritysjohdolle konkreettisia näkemyksiä toiminnan ja kannattavuuden parantamiseksi. Datan avulla voidaan esimerkiksi verrata yrityksen nykyistä toimintaa sen optimaaliseen tasoon. Tämän vertailun avulla voidaan tunnistaa sellaisia operationaalisia ongelmia, joita ei ilman analytiikkaa olisi välttämättä tunnistettu. Big data -analytiikka auttaa yritysjohtoa ymmärtämään syvemmin yrityksen toimintaa, joka puolestaan johtaa pysyvämpiin ja tehokkaampiin ratkaisuihin. (Ramsey, 2014)

2.3 Data-analytiikka päätöksenteon tukena

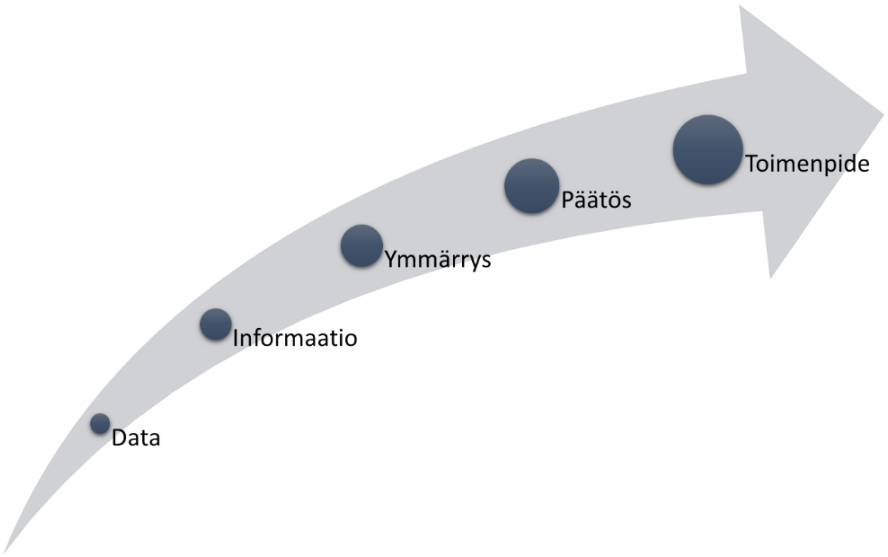
Big dataa on pidetty viime vuosina teknologisen kehityksen läpimurtona. Siitä huolimatta yrityksillä on edelleenkin rajallinen käsitys siitä, kuinka he voisivat muuttaa datan mahdollistaman potentiaalin oikeaksi taloudelliseksi arvoksi (Günther, 2017). Sellaisten organisaatioiden, jotka tukeutuvat big dataan organisaatiostrategioiden ja päivittäisen toiminnan ohjaamiseksi, odotetaan suoriutuvan paremmin taloudellisesti, kuin sellaisten organisaatioiden, jotka eivät hyödynnä big dataa (Lavalle, Lesser, Shockley, Hopkins & Kruschwitz, 2011)

Big datan vaikutukset yrityksen performanssiin tulevat esiin etenkin silloin, kun yritysjohto oppii hyödyntämään dataa tehokkaasti päätöksenteon tukena. Tietopohjaisen päätöksenteon (data driven decision making) arvoa koskeva tutkimus osoittaa, että analytiikalla voi olla huomattaviakin vaikutuksia yrityksen performanssiin (Müller, Fay & vom Brocke, 2018). McAfee ja Brynjolfsson (2012) havaitsivat tutkimuksessaan, että yritykset, jotka ovat toimialansa ylimmässä kolmanneksessa tietopohjaisen päätöksenteon käytössä, olivat keskimääräisesti 5 prosenttia tuottavampia ja 6 prosenttia kannattavampia kuin heidän kilpailijansa. Tämän perusteella tutkimukselle voidaan asettaa seuraavanlainen tutkimushypoteesi:

”Big datan hyödyntäminen vaikuttaa positiivisesti yrityksen kannattavuuteen”

Brynjolfsson, Hitt ja Kim (2011) ovat käyttäneet tietopohjaisen päätöksenteon mittareina kolmea eri tekijää. Ensimmäinen näistä tekijöistä on datan käyttö täysin uuden tuotteen tai palvelun luomisessa. Toisena tekijänä on datan käyttö päätöksenteossa koko yrityksen tasolla. Viimeisenä mittarina he ovat käyttäneet päätöksenteossa käytettävän datan määrää, joka käytännössä tarkoittaa koko yrityksellä hallussa olevan datan määrää, jota he käyttävät päätöksenteon apuna.

Big data -analytiikkaa voidaan hyödyntää hyvin laajalla skaalalla toimialasta riippumatta. Esimerkiksi pääomamarkkinoilla big dataa voidaan hyödyntää useiden tulevaisuusorientoituneiden päätösten, kuten riskianalyysien ja markkinoiden mahdollisten ongelmien suhteen (Singh, 2014). Tutkimuksista on käynyt myös ilmi, että esimerkiksi Yhdysvaltojen terveydenhuoltoala voisi vähentää kustannuksiaan jopa 8 prosentilla data-analytiikan mahdollistaman tehokkuuden kasvattamisen ja prosessien laadun parantamisen avulla. (Court, 2015)



Kuva 1 Datasta käytäntöön (Liu, 2015)

Yllä olevassa kuvassa on tiivistettynä tietopohjainen päätöksentekoprosessi. Prosessi lähtee liikkeelle raa'asta datasta, joka osaavien data-analyytikkojen ja data-analytiikan työkalujen avulla prosessoidaan yrityksen kannalta oleelliseksi informaatioksi. Informaatio mallinnetaan sellaiseen muotoon, että yritysjohto voi ymmärtää sen. Yritysjohdon ymmärrettyä informaation sisällön, voidaan viimein tehdä päätöksiä siitä, mitä aiotaan tehdä. Viimeinen vaihe on tehtävät toimenpiteet, jotka konkretisoivat päätöksentekoprosessin. Vasta viimeisen vaiheen jälkeen yritys voi saada konkreettisia hyötyjä datasta.

Big data -analytiikkaa ja sen perusteella tapahtuvaa päätöksentekoa pidetään merkittävänä erottavana tekijänä hyvin suoriutuvien ja heikosti suoriutuvien organisaatioiden välillä. Yrityksen suorituskyvyn ja data analytiikka -orientaation väliltä on löydetty positiivinen korrelaatio (Germann, Lilien, Fiedler, & Kraus, 2014). Analytiikka mahdollistaa yrityksen muutoksen proaktiiviseksi ja tulevaisuusorientoituneeksi organisaatioksi.

Dataorientoituneet yritykset suhtautuvat liiketoimintaan eri tavalla kuin heidän kilpailijansa. Tarkemmin sanottuna kyseiset yritykset hyödyntävät analytiikkaa päätöksente-

ossa mahdollisimman laajalla skaalalla. Laajalla skaalalla tarkoitetaan sitä, että yritysten tulisi tehdä päätöksiä tietopohjaisesti niin suurien kuin pienienkin päätösten kohdalla, sillä tällöin heillä on mahdollisuus kasvattaa voittomarginaaliaan jopa yli 60 prosentilla. (Court, 2015).

Dataorientoituneet yritykset käyttävät analytiikkaa apuvälineenä sekä tulevaisuuden strategioiden muodostamiseen että päivittäisen toiminnan ohjaamiseen kaksi kertaa todennäköisemmin kuin heidän kilpailijansa. He myös tekevät päätökset tilastojen ja analyysin perusteella yli kaksi kertaa useammin kuin huonommin menestyvät yritykset. Analytiikkaan perustuvan johtamisen sekä yrityksen suorituskyvyn välillä on löydetty selkeä korrelaatio, ja siitä on hyötyä etsivätpä yritykset sitten kasvua, toiminnan tehostamista tai kilpailuetua. (Lavalle, Lesser, Shockley, Hopkins & Kruschwitz, 2011)

Tietopohjaisen päätöksenteon ja paremman tuottavuuden sekä markkina-arvon väliltä on löydetty positiivinen korrelaatio. Sen lisäksi, että tietopohjainen päätöksenteko vaikuttaa suoraan yritysten tuottavuuteen ja markkina-arvoon, on sillä havaittu olevan vaikutuksia myös kannattavuuteen. (Brynjolfsson, Hitt & Kim, 2011)

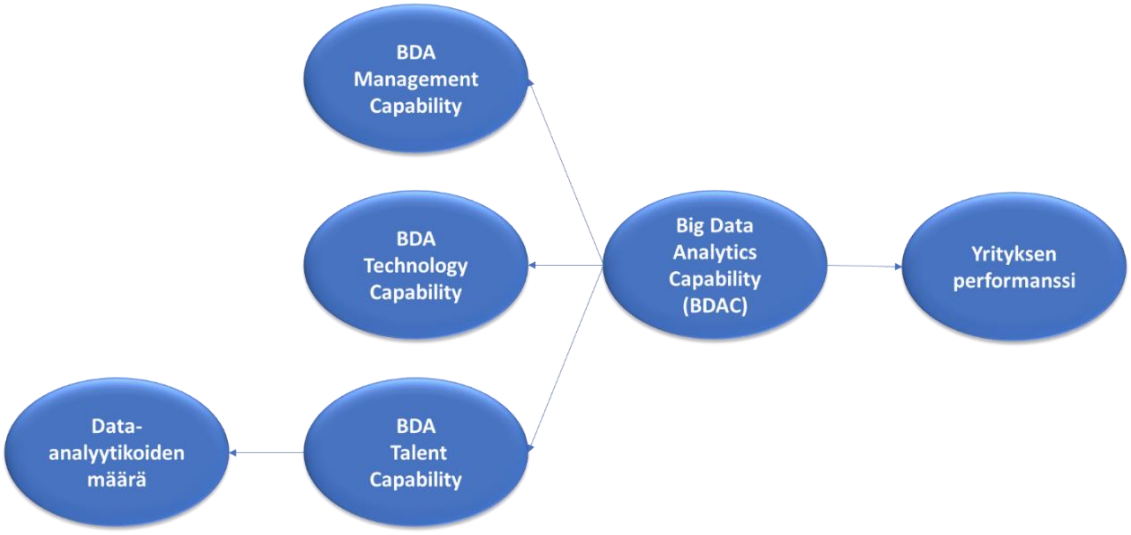
2.4 Big data -analytiikkakyvykkyys

Viimevuosien aikana yhä useammat yhtiöt ovat tavoitelleet big datasta ja data-analytiikasta saatavia hyötyjä. Data-analytiikkaan liittyvien investointien kannattavuudesta ja tuottoasteesta on kuitenkin herännyt kysymyksiä. Tutkimustulosten mukaan merkittävien tuottojen saamiseksi yritysten tulee investoida ennen kaikkea osaajiin big data -analytiikan saralla. (Bughin, 2016)

Big data -analytiikkakyvykkyydellä (Big Data Analytics Capability, BDAC) on sanottu olevan potentiaalia muuttaa täysin johtamisen teoria (George, Haas & Pentland, 2014) ja sitä on luonnehdittu jopa ”tieteen neljänneksi paradigmaksi” (Strawn, 2012). Resurssipohjaisen teorian perusteella BDAC määritellään laajasti yritysten ainutlaatuiseksi

kyvykkyydeksi muun muassa asettaa tuotteille optimaalinen hinta, havaita laatuongelmat, päättää alin mahdollinen varastotaso sekä tunnistaa lojaalit ja kannattavat asiakkaat (Davenport & Harris, 2007). Nämä kaikki tekijät myös tehostavat yrityksen toimintaa, kasvattavat liiketoiminnasta saatavia voittoja ja näin ollen tekevät yrityksestä taloudellisesti kannattavamman. BDAC:in mittarina voidaan käyttää muun muassa data-analytiikkatyöntekijöiden määrää, kuten Wu, Hitt ja Lou (2019) tutkimuksessaan tekevät.

Organisaatiossa työskentelevien data-analyytikkojen ja liikevoiton kasvamisen väliltä on löydetty positiivista korrelaatiota. 15% big data -analytiikkainvestoinneista johtuneesta liikevoiton kasvusta selittyi data-analytiikkaosaajien palkkaamisella (Court, 2015). Organisaation koon lisäksi myös organisaatiossa töissä olevien data-analyytikkojen määrä ennustaa vahvasti tietopohjaista päätöksentekoa, joka puolestaan parantaa yrityksen suorituskykyä (Brynjolfsson & McElheran, 2016).



Kuva 1 BDAC:n osatekijät (Akter 2016)

Yllä olevassa kuvassa on tiivistetty BDAC:n osatekijät, jotka liittyvät yritysjohtoon, teknologiaan sekä osaamiseen. Akterin (2016) tekemän tutkimuksen perusteella näistä kolmesta tekijästä merkittävimmäksi osoittautui osaamiseen liittyvä kyvykkyys, jonka mittarina voidaan käyttää esimerkiksi data-analyytikkojen määrää. BDAC:lla taas on suora vaikutus yrityksen performanssiin

Data-analyytikkojen määrän lisäksi myös sillä on merkitystä, kuinka laajalla skaalalla analyytikkoja palkataan organisaation palvelukseen. Kun yritykset sovittavat IT-pääomaan tehdyt sijoitukset ammattitaitoisiin työntekijöihin tehtyihin sijoituksiin, suorituskyky paranee huomattavasti. Osaajia tulisi palkata jokaiselta analytiikan osa-alueelta, jotta yrityksellä olisi resursseja sekä liiketoiminnan nykyisiin haasteisiin vastaamiseen että uusien innovaatioiden kehittämiseen. (Court, 2015) Analytiikkataidot käsittävät useita eri taitoja, kuten esimerkiksi datan louhimisen ja siistimisen sekä esittämisen tilastointi- ja ohjelmointityökaluilla (Wu, Hitt & Lou, 2019). Tällä hetkellä yritykset ovat kuitenkin sellaisen ongelman äärellä, että data-analyytikkojen tarve kehittyy eksponentiaalisesti nopeammin kuin osaavien data-analyytikkojen määrä.

3V-mallia mukaillen big datan nopeus tarkoittaa tiedon luomisen, tallentamisen, analysoimisen, visualisoinnin ja manipuloinnin vauhdin. Käytännössä nopeudella tarkoitetaan myös vauhtia, joilla nämä toimenpiteet tulisi suorittaa (Gandomi & Haider, 2015). Suurimpien organisaatioiden haaste nykypäivänä on selviytyä tästä valtavasta data-nopeudesta, kun datan luomisen ja käytön tulisi olla reaaliaikaista (Aggarwal, 2016). Big datan luonteeseen liittyy myös olennaisesti se, että dataa sekä luodaan että varastoidaan valtavan suuria määriä. Yrityksillä hallussa olevan big datan määrä keskimäärin kaksinkertaistuu kahden vuoden välein, jonka takia sen varastointiin ja prosessointiin kuluu kasvavissa määrin enemmän työtä (Aggarwal, 2016).

Big dataa on käytännössä kolmea eri tyyppiä: strukturoitua, ei-strukturoitua sekä puolistrukturoitua. Strukturoitu data tarkoittaa taulukkomuotoista dataa, jossa tieto on helposti luettavissa ja analysoitavissa. Epästrukturoitu data sen sijaan ei noudata mitään tiettyä kaavaa, ja vain yhden tietotyypin sijaan se koostuu useista eri tietotyypeistä,

kuten kuva-, teksti- ja audiotiedostoista. Puolistrukturoitu data taas on jotain strukturoidun ja epästrukturoidun datan välillä eli siinä esiintyy kummallekin tyypille ominaisia puolia. (Agrawal, 2017)

Kullakin datatyypillä on omat ominaisuutensa, ja näin ollen kuhunkin liittyy omat ongelmansa ja lähestymistapansa. Epästrukturoidun datan kokoaminen ja analysoiminen vaatii enemmän vaivaa ja ammattitaitoa kuin strukturoitu data. Epästrukturoidun datan määrä kasvaa 15 kertaa nopeammin kuin strukturoidun ja tälläkin hetkellä 95% big datasta on epästrukturoitua (Aggarwal, 2016). Näin ollen yrityksen tulee panostaa päteviin työntekijöihin, joilla on hyvät data-analytiikkataidot, jotta he pystyvät suoriutumaan epästrukturoidun datan luomista haasteista.

Yritykset voivat vastata big datan nopeuden, määrän sekä monimuotoisuuden luomiin ongelmiin kasvattamalla big data -resurssejaan optimaaliselle tasolle. Resurssiperusteisen teorian mukaan yrityksen performanssi riippuu sen kyvystä hallita kriittisiä resursseja (kuten henkilöstöä) tehokkaasti, saavuttaakseen parhaan mahdollisen suorituskyvyn (Akter, 2016).

Kehityksen aallonharjalla pysyminen, alan osaajiin ja uusimpiin teknologioihin tehtävien investointien tasapainottaminen sekä huipputason kykyjen palkkaaminen ovat ensisijaisia näkökohtia johtajille, jotka haluavat muuttaa data-analytiikan myötä saatavat hyödyt laajemmiksi ja merkittävimmiksi (Court, 2015).

3. Tutkimusaineisto ja -menetelmät

Tässä luvussa käydään läpi tutkimuksessa käytettävät selittävät sekä selitettävät muuttujat, aineiston rakenne sekä käytettävän tutkimusmenetelmän teoriaa. Tämän lisäksi esitellään mahdolliset estimointimenetelmät sekä se, miten mallille valitaan oikea estimointimenetelmä. Tutkimuksen muuttujat on valittu sekä aineiston saatavuuden, että teoriakappaleesta tehtyjen havaintojen perusteella.

3.1 Aineiston ja muuttujien kuvailu

Tutkimuksessa käytettävä aineisto on haettu ja kerätty kahdesta eri avoimesta tilastolähteestä. Data Science Central on tehnyt julkaisut yrityksissä töissä olevien data-analyytikkojen määristä. Listat on julkaistu joulukuussa 2013 ja tammikuussa 2015, joten voidaan olettaa, että data-analyytikkojen määrät on kerätty vuosilta 2013 ja 2014. Vuoden 2013 listassa on mukana 6000 yritystä, kun taas vuoden 2015 listassa yritysten määrä on noussut 25 prosentilla, havaintojen määrän ollessa 7500.

Muiden muuttujien arvot on kerätty Macrotrends nimisestä lähteestä manuaalisesti. Macrotrends on sivusto, johon on koottu finanssidataa tuhansista pörssiyrityksistä. Sieltä kerätyt arvot ovat liikevaihto (*Revenue*), nettotulos (*Net Revenue*), velkaantuneisuusaste (*Debt/Equity Margin*), liiketulosprosentti (*EBIT*) sekä sijoitetun pääoman tuottoprosentti (*ROI*). Kasvuprosentti (*Growth)* on laskettu kerätyn datan perusteella jakamalla kunkin vuoden liikevaihto edellisen vuoden liikevaihdolla.

Yrityksiä on otettu tutkimukseen mukaan ilman sen suurempia rajauksia, sillä Data Science Centralin listassa oli vain rajallinen määrä pörssiyrityksiä, joilta löytyi halutut tunnusluvut. Myös suuri osa pankeista ja finanssilaitoksista jouduttiin karsimaan, sillä niiden liiketoimintamallit usein poikkeavat merkittävällä tasolla tavallisten yritysten liiketoimintamalleista, ja näin ollen myös niiden tunnusluvut olivat poikkeavalla tasolla muihin verrattuna. Tämän lisäksi karsittiin pois sellaiset yritykset, joiden tunnusluvut olivat niin huomattavasti poikkeavalla tasolla, että ne olisivat selkeästi olleet outlier-havaintoja.

Data-analytiikan hyödyntämisen mittarina käytetään tässä tutkimuksessa data-analyytikoiden määrää (*DA)*. Mittarin valintaan vaikutti ensinnäkin se, että Wu, Hitt ja Lou (2019) käyttivät tutkimuksessaan juurikin data-analyytikoiden määrää mittaamaan big data -analytiikkakyvykkyyttä. Kuten teoriakappaleessa kävi ilmi, data-analytiikkakyvykkyydellä on havaittu olevan suora vaikutus yrityksen performanssiin. Tämän lisäksi

kunnollisen datan saatavuus oli melko heikkoa, ja data-analyytikoiden määrä oli tutkimuksen kannalta sopivin löydetty mittari.

Tutkimukseen on valittu kannattavuuden mittareiksi kaksi selitettävää muuttujaa: liiketulosprosentti (*EBIT*) sekä sijoitetun pääoman tuottoprosentti (*ROI*). Kummatkin ovat suhteellisia kannattavuuden tunnuslukuja, joten niiden avulla on mahdollista tehdä yritysten välistä vertailua pelkän yrityskohtaisen vertailun lisäksi. Suhteellisten tunnuslukujen käyttö on tärkeää tämänkaltaista tutkimusta tehdessä, sillä esimerkiksi pelkkä liiketulos tai sijoitetun pääoman tuotto ei vielä kerro siitä, kuinka hyvin yritys on suoriutunut sekä muihin yrityksiin että yrityksen aikaisempiin tilikausiin verrattuna. Sekä liiketulosprosentin että sijoitetun pääoman tuottoprosentin kaavat on esitelty kappaleessa 2.1.

Tämän lisäksi mukaan on valittu kolme kontrollimuuttujaa: liikevaihto (*Revenue*), kasvuprosentti (*Growth)* sekä velkaantuneisuusaste (*DE*). Kontrollimuuttujien tarkoituksena on estää vääristymiä mallin selittävien muuttujien selitysasteissa selitettävään muuttujaan. (Hill, Griffiths & Lim, 2018, 278-280) Käytännössä kontrollimuuttujien olisi siis hyvä olla sellaisia muuttujia, jotka voisivat myös mahdollisesti vaikuttaa selitettävään muuttujaan, eli tässä tapauksessa kahteen kannattavuuden tunnuslukuun. Kontrollimuuttujat eivät kuitenkaan saisi olla liian samankaltaisia selitettävien muuttujien kanssa. Tämän takia yhdeksi kontrollimuuttujaksi suunniteltu nettotulosprosentti (*NRP*) jätettiin pois tutkimuksesta.

3.2 Paneelidata

Paneelidatassa yhdistyy poikkileikkausaineistolle sekä aikasarja-aineistolle ominaiset elementit. Paneelidatan tarkoituksena on siis tarkastella useaa eri poikkileikkausyksikköä, kuten esimerkiksi yritystä, useana eri ajanhetkenä (Hill, Griffiths & Lim, 2018, 9). Poikkileikkausyksiköitä eli tämän tutkimuksen tapauksessa yrityksiä on *N* kappaletta, ja näiltä yrityksiltä kerätään *T* kappaletta havaintoja. Tällöin otoskooksi muodostuu *NT* kappaletta. Paneelidataa on useaa eri tyyppiä, mutta kyse on kuitenkin aina suuresta

ja monipuolisesta datasta, johon liittyy useiden haasteiden lisäksi myös useita etuja. (Hill et al., 2018, 635)

Paneelidataa on pääsääntöisesti kolmea eri tyyppiä: pitkää ja kapeaa, pitkää ja leveää sekä lyhyttä ja leveää. Pitkä ja kapea data tarkoittaa sitä, että tarkasteluajanjakso on pitkä, mutta tarkasteltavien yksiköiden määrä sen sijaan on pieni. Pitkä ja leveä aineisto taas sisältää sekä pitkän tarkasteluajanjakson ja sen lisäksi paljon tarkasteltavia yksiköitä. Tässä tutkimuksessa käytettävä aineisto on tyypiltään lyhyt ja leveä, joka tarkoittaa sitä, että aineisto sisältää suhteessa useita poikkileikkausyksiköitä, mutta tarkasteluajanjakso on melko lyhyt. (Hill et al, 2018, 635)

Tasapainoisessa paneelissa datasta ei puutu yhtäkään arvoa. Sen sijaan, jos datasta puuttuu arvoja, on paneeli epätasapainoinen. (Yaffee, 2003) Useimmiten etenkin talouteen liittyvät paneelidatat ovat epätasapainoisia, sillä kaikkien vuosien arvoja voi olla hankalaa kerätä (Croissant & Millo, 2008). Tässäkin tutkimuksessa käytettävä paneelidata on epätasapainoinen, sillä sen lisäksi, että data-analyytikoiden määrät puuttuvat kokonaan yhdeltä vuodelta niin myös yksittäisiä tunnuslukuja puuttuu tietyiltä yrityksiltä.

Paneelidatassa on useita hyviä puolia, jotka tekevät tutkimuksesta luotettavamman ja kattavamman kuin esimerkiksi tavallisesta lineaarisesta regressioanalyysistä. Datapisteiden määrä kasvaa poikkileikkaus- ja aikasarjadatoihin verrattuna, joten vapausasteita on enemmän. Paneelidatan avulla on myös mahdollista tutkia dynaamisia muutoksia samasta havaintokohteesta eri ajankohtina, jolloin monimutkaisempien yhteyksien analyysi mahdollistuu. (Wooldridge, 2002)

3.3 Paneelidatan regressioanalyysi

Paneelidatan regressioanalyysin tarkoituksena on selvittää selittävien muuttujien (*X*) vaikutus selitettävään muuttujaan (*Y*) (Wooldridge, 2013, 20) Etenkin paneelidatan kohdalla oikean estimointimenetelmän valinta on tärkeää, jotta voidaan varmistua mahdollisimman merkityksellisistä ja valideista tuloksista. Oikea estimointimenetelmän

valinta voidaan tehdä tarkastelemalla kolmea erilaista paneelidatan regressioanalyysimallia: yhdistetty OLS:ia, kiinteiden vaikutusten mallia sekä satunnaisten vaikutusten mallia. Seuraavaksi esitellään nämä kolme menetelmää, sekä perusteet, joilla oikea estimointimenetelmä voidaan valita.

Aluksi on tärkeää tarkastella, onko aineistossa olevien yksiköiden välillä eroja tai korrelaatiota. Jos yksiköiden välillä ei ole eroja, eli ne eivät ole heterogeenisia, voidaan mallina käyttää yhdistettyä OLS:ia (pooled OLS). Malli ei käytännössä hyödynnä paneelidatan ominaisuuksia, ja tällöin voidaan suorittaa tavallinen usean selittävän muuttujan regressioanalyysi. Koop (2008, 256) esittää yhdistetyn OLS:in kaavan seuraavasti (kaava 3):

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it} \quad (3)$$

Tässä tutkimuksessa tutkitaan eri yrityksen välisiä eroja, ja koska yhdistetty OLS perustuu sellaiseen olettamukseen, että parametrit ovat samanlaisia eri yksiköiden kesken, on melko epätodennäköistä, että päädyttäisiin käyttämään yhdistettyä OLS:ia. Tätä olettamusta voidaan tutkia Breusch-Pagan -testillä, jonka nollahypoteesi on, että malli ei ole heterogeeninen, eli yksiköt ovat samanlaisia. Jos testin nollahypoteesi jää voimaan, voidaan käyttää yhdistettyä OLS:ia.

Jos taas yksiköiden välillä on eroja, on käytettävälle mallille kaksi vaihtoehtoa: kiinteiden vaikutusten malli sekä satunnaisten vaikutusten malli. Kiinteiden vaikutusten mallissa (fixed effects model) kaikki erot yksiköiden välillä oletetaan kuuluvan vakiotermiin, jolloin mallissa vakiotermi on ainut vaihteleva asia. Vakiotermi siis ilmentää niitä eroja, joita yksiköiden välillä on (Koop, 2008, 263). Koop (2008, 261) esittää kiinteiden vaikutusten mallin kaavan seuraavasti (kaava 4):

$$Y_{it} = \alpha_i + \beta X_{it} + \varepsilon_{it} \quad (4)$$

Kiinteiden vaikutusten mallin heikkoutena voidaan pitää sitä, että siihen ei voida sisällyttää sellaisia muuttujia, jotka pysyvät yli ajan vakioina. Tämänkaltaisia muuttujia ei kuitenkaan ole käytettävässä aineistossa, joten tämä ei tule muodostumaan ongelmaksi tutkimuksen kannalta. Kiinteiden vaikutusten mallissa hyvä puoli on se, että se yleensä eliminoi heterogeenisuuden, eli täten se eliminoi myös endogeenisuusongelman.

Kun kiinteiden vaikutusten mallissa tutkitaan vain niitä yksiköitä, joita tutkittava aineisto koskee, niin satunnaisten vaikutusten malli kohdistuu koko populaatioon, jonka yksi osa malli olettaa aineiston olevan. Koop (2008, 263) esittää satunnaisten vaikutusten mallin kaavan seuraavassa muodossa:

$$Y_{it} = \alpha_i + \beta X_{it} + u_{it} \quad (5)$$

Satunnaisten vaikutusten mallissa eroavaisuudet yksiköiden välillä oletetaan myös kuuluvan vakiotermiin. Toisin kuin kiinteiden vaikutusten mallissa, satunnaiset vaikutukset huomioidaan jakamalla vakiotermi kahteen eri osaan: koko populaation keskiarvoon sekä yksikkökohtaiseen satunnaiseen vaihteluun.

Oikean estimointimenetelmän valinta tapahtuu käyttämällä kolmea eri testiä. Taulukossa 1 esitellään estimointimenetelmän valinnassa apuna käytettävä testirunko. Käytännössä oikea malli valitaan sen perusteella, hylätäänkö testien nollahypoteesit, vai jäävätkö ne voimaan.

Taulukko 1. Estimointimenetelmän valinta (Park 2010)

F-testi	***Breusch-Pagan testi***	***Käytettävä malli***
H_0 hyväksytään	H_0 hyväksytään	Yhdistetty OLS
H_0 hylätään	H_0 hyväksytään	Kiinteiden vaikutusten malli
H_0 hyväksytään	H_0 hylätään	Satunnaisten vaikutusten malli
H_0 hylätään	H_0 hylätään	Valitaan kiinteiden tai satunnaisten vaikutusten malli riippuen Sargan-Hansenin testin tuloksesta

Estimointimenetelmän valinta aloitetaan tekemällä kiinteiden vaikutusten F-testi, jonka tarkoituksena on vertailla, onko mallissa kiinteitä vaikutuksia vai ei. Siinä siis vertaillaan, onko hyödyllisempää käyttää yhdistettyä OLS:ia vai kiinteiden vaikutusten mallia. F-testin nollahypoteesina on, että kaikki vakiot ovat yhtä suuria. Jos nollahypoteesi jää voimaan, yksiköiden vakioissa ei ole eroja ja tällöin on mahdollista käyttää yhdistettyä OLS:ia. Jos taas nollahypoteesi hylätään, kaikki vakiot eivät ole yhtä suuria. Tämä tarkoittaa sitä, että tulee estimointimenetelmänä käyttää kiinteiden vaikutusten mallia, joka ottaa huomioon yksiköiden väliset erot.

Mallin satunnaisten vaikutusten olemassaoloa testataan Breusch-Pagan -testillä. Testin tarkoituksena on siis vertailla, onko parempi käyttää yhdistettyä OLS:ia vai satunnaisten vaikutusten mallia. Testin nollahypoteesina on, että satunnaisten vaikutusten varianssi yksiköiden välillä on nolla, eli että yksiköt ovat samanlaisia. Mikäli nollahypoteesi jää voimaan, satunnaisia vaikutuksia ei ole, eikä satunnaisten vaikutusten menetelmän käytöstä ole hyötyä. Tällöin voidaan estimointimenetelmänä käyttää yhdistettyä OLS:ia. Sen sijaan, jos nollahypoteesi hylätään, yksiköiden välillä on satunnaisia eroja, jolloin satunnaisten vaikutusten malli on parempi estimointimenetelmä.

Jos sekä F-testin että Breusch-Pagan testin nollahypoteesit hylätään, vertaillaan kiinteiden vaikutusten mallia satunnaisten vaikutusten malliin joko Hausmanin tai Sargan-Hansenin testillä. Koska tutkimuksessa käytetään klusteroituja robusteja keskivirheitä, Hausmanin testiä ei voida käyttää, eli testaus pitää toteuttaa Sargan-Hansenin testillä.

Testin nollahypoteesi on, että satunnaisten ja kiinteiden mallien kertoimet eivät eroa. Jos nollahypoteesi hyväksytään, on mahdollista raportoida tulokset kummankin mallin perusteella. Jos taas nollahypoteesi hylätään, kertoimissa on eroja ja tulee käyttää kiinteiden vaikutusten menetelmää.

4. Tutkimustulokset ja analyysi

Tässä luvussa käsitellään tutkielman empiriaa. Luku alkaa aineiston kuvailulla ja muuttujien logaritmi- sekä viivästysmuokkauksilla. Tämän jälkeen tutkitaan kummallekin mallille oikeaa estimointimenetelmää, ja toteutetaan tutkimus valituilla menetelmillä. Viimeisessä alakappaleessa arvioidaan tutkimuksen reliabiliteettia, sekä pohditaan tutkimukseen mahdollisesti vaikuttaneita rajoitteita.

4.1 Aineiston kuvailu ja muokkaus

Tutkimuksessa käytettävien muuttujien havaintojen määrät, keskiarvot, keskihajonnat ja minimi- sekä maksimiarvot esitetään taulukossa 2. Data-analyytikoiden määrät vaihtelevat 1 ja 110 välillä, keskiarvon ollessa alle 8. Tämä osoittaa sen, että vaikka analyytikkojen määrä on joissain yrityksissä todella suuri, on niiden keskimääräinen määrä melko alhainen suurimmassa osassa yrityksistä. Liikevaihdonkin perusteella voidaan todeta, että aineistossa on todella eri kokoisia yrityksiä, sillä pienin liikevaihto on vain 0,08 miljoonaa dollaria, suurimman ollessa 485,65 miljoonaa dollaria. Kasvuprosenttikin vaihtelee melko paljon negatiivisista arvoista positiivisiin arvoihin, mutta keskimäärin tarkasteltavat yritykset ovat kuitenkin kasvaneet maltillisesti. Muuttujat ovat jakautuneet melko epätasaisesti, sillä jokaisen muuttujan keskiarvot ovat niiden keskihajontoja pienempiä.

Taulukko 2. Muuttujien DA, EBIT, ROI, Revenue, Growth ja DE tunnuslukuja.

Muuttuja	***Havainnot***	***Keskiarvo***	***Keskihajonta***	***Minimi***	***Maksimi***
DA	264	7,72	14,27	1	110
EBIT	387	0,15	0,39	-0,96	7,02
ROI	396	0,09	0,13	-1,12	0,65
Revenue	396	34,68	56,49	0,08	485,65
Growth	396	0,10	0,28	-0,78	2,85
DE	376	1,31	4,25	-10,15	59,38

Liitteissä 8 ja 9 on esitelty mallien 1 ja 2 residuaalien jakaumat ja histogrammit. Residuaaleja on tutkittu kummankin selitettävän muuttujan kanssa ilman muokattuja muuttujia, viivästysmuutosten kanssa sekä viivästysmuutosten ja logaritmimuutosten kanssa. Viivästysmuutosten tarkoituksena on kontrolloida autokorrelaatiota. Autokorrelaatiolla tarkoitetaan sitä, että aikasarjan havaintojen arvoon vaikuttaa aikaisemmat havaintojen arvot ja tämä taas saattaa vääristää saatuja tuloksia. (Hill et al., 2018, 424) Residuaalien normaalijakautuneisuuden voidaan huomata parantuvan huomattavasti, kun kumpaankin malliin lisätään selitettävän muuttujan *Y* viivästetty arvo Y_{t-1} (mallissa 1 *EBITlag* ja mallissa 2 *ROIlag*).

Viivästysmuuttujien lisäksi tarkastellaan, miten kunkin mallin y-muuttujien logaritmimuunnokset vaikuttavat residuaaleihin. Logaritmimuunnokset voivat joissain tapauksissa vaikuttaa residuaalien normaalijakautuneisuuteen positiivisesti. Kuten jakaumista ja histogrammeista voidaan huomata, myös tässäkin tapauksessa y-muuttujan logaritmimuunnos auttaa, joten kummankin mallin selitettävä muuttuja muunnetaan logaritmimuotoon (mallissa 1 *lgEBIT* ja mallissa 2 *lgROI*). Logaritmimuunnoksia kokeiltiin myös mallin kaikkiin muihin muuttujiin, mutta todettiin, että se ei tuota tarpeeksi positiivisia vaikutuksia verrattuna sen aiheuttamiin negatiivisiin vaikutuksiin. Kaikkien muuttujien logaritmimuunnokset aiheuttaisivat epäluotettavuutta ja vääristymiä tutkimuksen tuloksissa, sillä aineistosta häviäisi kaikki negatiivisen arvon saaneet havainnot.

4.2 Estimointimenetelmän valinta ja tulokset

Muuttujiin tehtyjen muutosten jälkeen tutkimuksen ensimmäisessä mallissa käytettävä selitettävä muuttuja on lg*EBIT*, selittävä muuttuja *DA* ja kontrollimuuttujat *lgEBITlag, Revenue, Growth sekä DE*. Toisessa mallissa selitettävänä muuttujana on *lgROI*, selittävänä muuttujana jälleen *DA* ja kontrollimuuttujina *lgROIlag, Revenue, Growth* sekä *DE*.

Taulukko 3. Pearsonin korrelaatiokertoimet

	lgEBIT	**lgROI**	**DA**	**Revenue**	**Growth**	**DE**
lgEBIT	1,0000					
lgROI	0,4342	1,0000				
DA	0,0624	0,1616	1,0000			
Revenue	-0,1558	0,0673	0,1430	1,0000		
Growth	-0,0397	0,0111	-0,0476	-0,1277	1,0000	
DE	0,0682	0,0234	-0,0360	-0,0107	-0,0283	1,0000

Taulukossa 3 käydään läpi vielä muuttujien korrelaatiokertoimet. Regressioanalyysissä on yleistä, että selittävät muuttujat korreloivat keskenään. Jos korrelaatio on kuitenkin liian suurta, tulee ongelmaksi multikollineaarisuus. Multikollineaarisuuden seurauksena regressioanalyysin tuloksissa saattaa esiintyä tarkkuusongelmia, koska tällöin voi olla hankalaa erottaa selittävän muuttujan vaikutusta selitettävään (Koop, 2008, 49). Korrelaatiomatriisin perusteella muuttujien välillä ei ole havaittavissa kovinkaan suuria korrelaatioita.

Seuraavissa alakappaleissa käydään läpi sekä mallin 1 että mallin 2 estimointimenetelmän valintaprosessi tarkemmin. Parhaaksi osoittautuneen estimointimenetelmän perusteella voidaan tehdä tutkimusanalyysi sekä muodostaa johtopäätökset.

4.2.1 Big datan hyödyntämisen vaikutus liiketulosprosenttiin

Mallin 1 selitettävä muuttuja on *EBIT*, selittävä muuttuja *DA* ja kontrollimuuttujat*, lgEBITlag, Revenue, Growth ja DE*. Tämän lisäksi malliin on lisätty vuosimuuttujasta tehty dummy-muuttuja *Year kontrolloimaan ajan vaikutusta.* F-testin p-arvo on 0,0002, eli nollahypoteesi hylätään. Tämä tarkoittaa sitä, että vakiot eivät ole yhtä suuria, eli kiinteiden vaikutusten malli on parempi vaihtoehto kuin yhdistetty OLS. Tehdään lisäksi Breusch-Pagan -testi, jolla selvitetään, onko järkevämpää käyttää satunnaisten vaikutusten mallia, kuin yhdistettyä OLS:ia. Testin p-arvo on 0,0039, joka viittaa siihen, että satunnaisten vaikutusten malli olisi myös parempi vaihtoehto kuin yhdistetty OLS. Sargan-Hansenin testillä voidaan varmistua siitä, onko satunnaisten vai kiinteiden vaikutusten malli parempi. Koska testin p-arvo on 0,0000, on kiinteiden vaikutusten malli paras estimointimenetelmä. Estimointimenetelmän valintaprosessissa käytettyjen testien p-arvot tiivistetään vielä yhteen taulukossa 4.

Taulukko 4. Estimointimenetelmän valinnassa käytettyjen testien p-arvot

	EBIT (Malli 1)
F-testi	→ H1 (0.0002)
Breusch-Pagan	→ H1 (0.0039)
Sargan-Hansen	→ H1 (0.0000)

Taulukosta 5 käy ilmi, että mallin overall-selitysaste on 24,05%. Mitä suurempi mallin selitysaste on, sitä suuremman osan malli selittää selitettävän muuttujan kokonaisvaihtelusta. Lisäksi on hyvä huomioida mallin p-arvon olevan 0,0000, mikä tarkoittaa sitä, että malli on tilastollisesti merkitsevä.

Taulukko 5. Valittu estimointimenetelmä mallille 1

	EBIT
Valittu estimointimenetelmä	Kiinteiden vaikutusten malli
Havaintojen lukumäärä	220
Mallin p-arvo	0,0000
Mallin selitysaste (overall)	0,2405

Teoriaan ja aikaisempiin tutkimuksiin pohjautuva nollahypoteesi mallin mahdollisista tuloksista on seuraavanlainen:

H0: Big datan hyödyntäminen ei vaikuta liiketulosprosenttiin

Taulukossa 6 on listattu kiinteiden vaikutusten mallin muuttujien ja vakiotermin kertoimet, klusteroidut robustit keskivirheet sekä tilastolliset merkitsevyydet. Koska selitettävä muuttuja on logaritmimuodossa, tulee ottaa huomioon, että estimaatin tulkinta tapahtuu eri tavalla kuin normaalisti. Kun vertaillaan tavallisen muuttujan vaikutusta logaritmimuotoiseen muuttujaan, tulee vertailu tehdä osittaisjoustona. Tämä tarkoittaa sitä, että kun x-muuttuja muuttuu yhden yksikön, niin y-muuttuja muuttaa x-muuttujan arvoa 100 prosentilla kerrottuna estimaatin arvolla.

Muuttujista tilastollisesti merkitsevät ovat *lgEBITlag* ja Revenue viiden prosentin riskitasolla. Kun liikevaihto kasvaa yhdellä yksiköllä, niin liiketulosprosentti kasvaa noin 0,02 prosentilla. Kun taas viivästetty arvo liiketulosprosentista muuttuu yhdellä yksiköllä, niin liiketulosprosentti vähenee noin 0,82 prosentilla.

Taulukko 6. Kiinteiden vaikutusten mallin selitysasteet, muuttujien kertoimet, kluste-roidut robustit keskivirheet ja p-arvot. Selitettävänä muuttujana lgEBIT.

Muuttuja	Kerroin	Robust keskivirhe	P-arvo
DA	0,0055	0,0111	0,618
lgEBITlag	-0,8187	0,2567	0,002***
Revenue	0,0197	0,0098	0,048**
Growth	-0,1052	0,0976	0,284
DE	-0,0030	0,0027	0,376
Year	-0,0352	0,0777	0,651
_cons	-4,4417	0,6115	0,000

R-squared: within	0,2194
R-squared: between	0,3194
R-squared: overall	0,2405

Muuttujan *DA* p-arvo on 0,618, eli se ei ole tilastollisesti merkitsevä. Voidaan siis sanoa, että selittävällä muuttujalla ei ole tilastollisesti merkitsevää yhteyttä selitettävään muuttujaan. Näin ollen tehty nollahypoteesi pysyy voimassa, eli big datan hyödyntämisellä ei ole vaikutusta yrityksen liiketulosprosenttiin. Tulos on sinänsä yllättävä, sillä teorian perusteella odotettavissa oleva tulos olisi ollut se, että big datan hyödyntämisen ja yrityksen kannattavuuden välillä olisi ollut positiivinen yhteys. Myöskään kasvuprosentilla, velkaantuneisuusasteella tai vuoden dummy-muunnoksella ei ole vaikutusta liiketulosprosenttiin tilastollisesti merkitsevällä tasolla. Liitteessä 3 on esitelty mallin 1 kiinteiden vaikutusten malli kokonaisuudessaan.

4.2.2 Big datan hyödyntämisen vaikutus sijoitetun pääoman tuottoprosenttiin

Mallin 2 selitettävä muuttuja on *ROI*, selittävä muuttuja *DA* ja kontrollimuuttujat *lgROI-lag, Revenue, Growth ja DE.* Tämän lisäksi myös tähän malliin on lisätty kontrolliksi vuosimuuttujan dummy-muuttuja *Year kontrolloimaan ajan vaikutusta.* Estimointimenetelmän valintaan liittyvien testien tulokset on tiivistetty taulukkoon 7. F-testin p-arvo on 0,0000, eli nollahypoteesi hylätään. Tämä tarkoittaa sitä, että vakiot eivät ole yhtä

suuria, eli kiinteiden vaikutusten malli on parempi vaihtoehto kuin yhdistetty OLS. Tehdään lisäksi Breusch-Pagan -testi, jolla selvitetään, onko järkevämpää käyttää satunnaisten vaikutusten mallia, kuin yhdistettyä OLS:ia. Testin p-arvo on 0,0229, joka viittaa siihen, että satunnaisten vaikutusten malli olisi parempi vaihtoehto verrattuna yhdistetty OLS:iin. Sargan-Hansenin testillä voidaan varmistua siitä, onko satunnaisten vai kiinteiden vaikutusten malli parempi. Koska testin p-arvo on 0,0000, on kiinteiden vaikutusten malli paras estimointimenetelmä.

Taulukko 7. Estimointimenetelmän valinnassa käytettyjen testien p-arvot

	ROI (Malli 2)
F-testi	→ H1 (0.0000)
Breusch-Pagan	→ H1 (0.0229)
Sargan-Hansen	→ H1 (0.0000)

Taulukossa 8 on esitelty mallille 2 valittu estimointimenetelmä, mallissa mukanaolevien havaintojen määrä, mallin p-arvo sekä sen overall-selitysaste:

Taulukko 8. Valittu estimointimenetelmä selitettävälle muuttujalle ROI.

	ROI
Valittu estimointimenetelmä	Kiinteiden vaikutusten malli
Havaintojen lukumäärä	219
P-arvo	0,0001
Mallin selitysaste (overall)	0,0718

Teorian ja aikaisempien tutkimusten pohjalta voidaan mallille 2 asettaa seuraavanlainen nollahypoteesi:

H0: Big datan hyödyntämisellä ei ole vaikutusta sijoitetun pääoman tuottoprosenttiin.

Taulukkoon 9 on listattu kiinteiden vaikutusten mallin muuttujien ja vakiotermin kertoimet, robustit keskivirheet sekä tilastolliset merkitsevyydet. Tulee jälleen huomioida, että selitettävä muuttuja on logaritmimuodossa, joten tulosten tulkinta tehdään osittaisjoustoina. Voidaan huomata, että *DA* ei tässäkään mallissa ole tilastollisesti merkitsevä edes kymmenen prosentin riskitasolla, sillä sen p-arvo on 0,711. Nollahypoteesi jää siis voimaan tässäkin mallissa, eli big datan hyödyntämisellä ei ole tilastollisesti merkitsevää vaikutusta sijoitetun pääoman tuottoprosenttiin. Muuttujan kerroin on kuitenkin positiivinen, eli jos tilastollista merkitsevyyttä olisi, olisi vaikutus positiivinen.

Taulukko 9. Kiinteiden vaikutusten mallin selitysasteet sekä muuttujien kertoimet, klusteroidut robustit keskivirheet ja p-arvot. Selitettävänä muuttujana lgROI.

Muuttuja	Kerroin	Robust keskivirhe	P-arvo
DA	0,0067	0,0182	0,711
lgROIlag	-0,3972	0,2192	0,073
Revenue	0,0115	0,0051	0,027**
Growth	0,2004	0,1142	0,042**
DE	-0,0005	0,0036	0,089
Year	0,0056	0,1068	0,958
_cons	-3,7613	0,5300	0,000

R-squared: within	0,2308
R-squared: between	0,1356
R-squared: overall	0,0718

Tilastollisesti merkitseviä muuttujia 5% riskitasolla ovat *Revenue* sekä *Growth.* Liikevaihdolla sekä kasvuprosentilla on kummallakin positiivinen vaikutus yrityksen sijoitetun pääoman tuottoprosenttiin. Muuttujat *lgROIlag*, *DE* sekä *Year* eivät ole tilastollisesti merkitseviä. Liitteessä 4 on esitelty mallin 2 kiinteiden vaikutusten malli kokonaisuudessaan.

Big datan hyödyntämisellä ei voida todeta olevan vaikutusta yritysten kannattavuuteen kummankaan mallin perusteella. Kappaleessa 5 pohditaan enemmän mahdollisia syitä siihen, minkä takia vastoin teorian ja aikaisempien tutkimusten perusteella tehtyjä oletuksia kummankaan mallin kohdalla ei löytynyt positiivista vaikutusta big datan hyödyntämisen ja kannattavuuden välillä.

4.2.3 Yrityksen toimialan vaikutus

Mielenkiinnon vuoksi tutkimus toteutettiin myös lisäämällä kumpaankin malliin interaktiotermi toimialan ja data-analyytikoiden välillä, jolloin voidaan tutkia sitä, vaikuttaako data-analyytikoiden määrän lisäksi yrityksen toimiala selitettävän muuttujan arvoihin. Interaktiotermi tarkoittaa selittävää muuttujaa, joka muodostuu kahdesta muusta selittävästä muuttujasta (Wooldridge, 2013, 844). Yritykset jaoteltiin IT-toimialaan sekä ei-IT-toimialaan sen takia, että IT-alan yrityksissä voidaan olettaa työskentelevän enemmän data-analyytikoita kuin muiden alojen yrityksissä. Tarkoituksena on siis moderoida toimialan vaikutusta tutkimuksen tuloksiin. Interaktiotermien arvoja tulkitaan kuten muidenkin muuttujien arvoja niiden erilaisesta luonteesta riippumatta (Hill et al., 2018, 320).

Taulukossa 10 on esitelty mallin 1 tulokset interaktiomuuttujan kanssa. Tuloksista voidaan huomata, että interaktiomuuttujalla *IT#DA* ei ole tilastollisesti merkitsevää vaikutusta selitettävään muuttujaan *EBIT*. Näin ollen voidaan sanoa, että toimialan lisäämisellä malliin ei ole merkitystä yrityksen kannattavuuteen.

Taulukko 10. Interaktiomuuttuja mallissa 1, selitettävänä muuttujana EBIT.

Muuttuja	Kerroin	Robust keskivirhe	P-arvo
DA	0,0283	0,0387	0,466
lgEBITlag	-0,8509	0,2591	0,001
Revenue	0,0201	0,0099	0,043
Growth	-0,1026	0,9731	0,294
DE	-0,0023	0,0028	0,408
Year	-0,0599	0,0743	0,421
IT#DA	-0,0409	0,0411	0,322
_cons	-4,5822	0,6795	0,000

Sama interaktiomuuttuja lisättiin myös malliin 2, jonka selitettävänä muuttujana on *ROI*. Taulukossa 11 on esitelty mallin tulokset, joista voidaan jälleen huomata, että interaktiomuuttujalla *IT#DA* ei ole tilastollisesti merkitsevää vaikutusta selitettävään muuttujaan.

Taulukko 11. Interaktiomuuttuja mallissa 2, selitettävänä muuttujana ROI.

Muuttuja	Kerroin	Robust keskivirhe	P-arvo
DA	0,0307	0,0342	0,371
lgROIlag	-0,3978	0,2193	0,072
Revenue	0,1173	0,0048	0,017
Growth	0,2031	0,1142	0,078
DE	-0,0004	0,0037	0,909
Year	-0,0248	0,1061	0,816
IT#DA	-0,0480	0,0312	0,127
_cons	-3,8174	0,5354	0,000

Kummankin mallin tulosten perusteella voidaan tehdä johtopäätös siitä, että vaikka yrityksen toimialan vaikutusta moderoitaisiinkin interaktiotermin avulla, ei data-analyytikoiden määrällä siltikään ole vaikutusta yrityksen kannattavuuteen.

4.3 Tulosten reliabiliteetin arviointi ja rajoitteet

Tutkimuksen tulosten luotettavuuden arviointi on tärkeä osa tutkimuksen tekoa. Seuraavaksi käydään läpi regressioanalyysiin liittyviä taustaoletuksia, ja arvioidaan kuinka luotettavia tutkimuksesta saadut tulokset ovat. Lisäksi käydään läpi tutkimuksen tuloksiin mahdollisesti vaikuttaneita rajoitteita.

Mallien residuaalien normaalijakautuneisuus on yksi luotettavan mallin ominaisuuksista. Jos malli ei ole normaalijakautunut, ei mallille voida tehdä kovinkaan luotettavia hypoteeseja. Kappaleessa 4.2 yritetty parannella residuaalien normaalijakautuneisuutta muuttujamuunnoksille, ja liitteiden 8 ja 9 jakaumakuvioista voidaan tulkita, miten hyvin residuaalit ovat normaalijakautuneita. Jakaumista nähdään, että vaikka parannuksia on tehty, ei kummankaan mallin residuaalit silti ole aivan normaalijakautuneita, sillä mallin 1 residuaalit ovat melko vinot ja mallin 2 hieman huipukkaat.

Multikollineaarisuutta käsiteltiin jo aiemmin kappaleessa 4.2, ja taulukkoon 3 on listattu muuttujien Pearsonin korrelaatiokertoimet. Liitteessä 5 on esitelty samat korrelaatiokertoimet p-arvojen kanssa. Muuttujat eivät korreloi keskenään tilastollisesti merkitsevillä tasoilla. Muuttujien lineaarista korrelaatiota voidaan testata lisäksi VIF-testillä, jonka tulokset ovat liitteissä 6 ja 7. VIF-testin perusteella multikollineaarisuutta ei ole, sillä minkään muuttujan VIF-arvo ei ole yli kymmentä.

Autokorrelaatiota on pyritty poistamaan tekemällä selitettävistä muuttujista *EBIT* ja *ROI* viivästysmuunnokset *lgEBITlag* ja *ROIlag*, jotka on asetettu mallien kontrolleiksi. Toimenpiteistä huolimatta mallissa voidaan silti olettaa olevan autokorrelaatiota, sillä käytännössä kaikkien mallissa käytettävien muuttujien arvot ovat jollain tavalla riippuvaisia aikaisempien vuosien arvoista.

Tutkimuksen luotettavuutta horjuttaa myös aineistoon liittyvät rajoitteet. Tutkimukseen sopivaa dataa big datan hyödyntämiseen liittyen oli hieman huonosti saatavilla, mutta käytetty data oli kuitenkin paras löydetty vaihtoehto. Data-analyytikoiden määrät olivat saatavilla vain kahdelta vuodelta, joka on melko lyhyt aika tämänkaltaisessa aikasarjoihin perustuvassa paneelitutkimuksessa. Tunnuslukudataa päädyttiinkin keräämään

vielä kolmannelta vuodelta, mutta kolmenkaan vuoden tarkasteluvälistä ei oikeastaan pysty tekemään kovinkaan luotettavia ja yleistettäviä havaintoja. Tämä lisäksi aiheutti datan epätasapainoisuuden, koska data-analyytikot puuttuivat kokonaan yhdeltä vuodelta.

5. Yhteenveto ja johtopäätökset

Tutkimuksen tarkoituksena oli ottaa selvää siitä, vaikuttaako big datan hyödyntäminen yrityksen kannattavuuteen, ja jos vaikuttaa, minkä suuntainen vaikutus on. Aineisto rajattiin koskemaan vain pörssiyrityksiä, eikä tarkempia rajauksia tehty. Vaikutusta tutkittiin paneelidatan regressioanalyysillä kahdella eri mallilla. Kummankin mallin analyysiin käytettiin kiinteiden vaikutusten menetelmää klusteroiduilla robusteilla keskivirheillä. Kuten johdannossa kävi ilmi, tutkimukselle asetettu päätutkimuskysymys oli seuraavanlainen:

”Miten big datan hyödyntäminen vaikuttaa yrityksen kannattavuuteen?”

Teoriaosuudessa käytiin aluksi lyhyesti läpi kannattavuuden teoriaa ja valittuja selitettäviä muuttujia. Sen jälkeen siirryttiin tutkimaan big datan ja data-analytiikan teoriaa ja niiden vaikutuksia yrityksen kannattavuuteen. Teorian pohjalta tutkimukseen valittiin kaksi selitettävää muuttujaa, selittävä muuttuja ja sopivat kontrollit. Kummankin selitettävän muuttujan pohjalta tehtiin hypoteesit, miten selittävän muuttujan niihin tulisi vaikuttaa.

Tutkimus toteutettiin kahden mallin avulla, ensimmäisessä selitettävänä muuttujana oli liiketulosprosentti ja toisessa sijoitetun pääoman tuottoprosentti. Kummastakaan mallista ei löytynyt tilastollisesti merkitsevää yhteyttä big datan hyödyntämisen ja yrityksen kannattavuuden väliltä. Tutkimustulosten perusteella tutkimuskysymykseen voidaan siis vastata seuraavasti:

”Big datan hyödyntämisellä ei ole vaikutusta yrityksen kannattavuuteen.”

Tutkimustulosten perusteella saatu lopputulos on hieman ristiriidassa aikaisempien tutkimusten ja teorian kanssa. Aiemmissa tutkimuksissa on löydetty vahva positiivinen vaikutus big datan hyödyntämisen ja yrityksen kannattavuuden väliltä. Yksi positiivinen asia on kuitenkin se, että hypoteesit kumoutuivat kummankin selitettävän muuttujan kohdalla, koska tämä viestii siitä, että aineisto ei ole ristiriidassa itsensä kanssa.

Se, että etsittyä yhteyttä ei tässä tutkimuksessa löytynyt johtuu varmasti useammasta eri tekijästä. Yksi syy on ainakin varmasti se, että aineisto oli melko pieni eli havaintoja oli vähän ja tutkittava aikaväli lyhyt. Lisäksi, jos big datan hyödyntämisen mittarina olisi käytetty useampia eri selittäviä muuttujia, olisi myös voitu saada parempia tutkimustuloksia.

Saatujen tutkimustulosten perusteella yritysten ei välttämättä olisi kannattavaa palkata enempää data-analyytikoita paremman kannattavuuden toivossa. Kuitenkin teoria ja aikaisemmat tutkimukset ovat osoittaneet sen, että big datan hyödyntämisellä on positiivisia vaikutuksia yrityksen kannattavuuteen. Näin ollen tästä tutkimuksesta ei voida tehdä kovinkaan yleistettäviä johtopäätöksiä, joten jatkotutkimusehdotukset ovat paikallaan. Ensinnäkin aihetta olisi mielenkiintoista tutkia suuremmalla aineistolla useamman vuoden ajalta. Jos havaintoja olisi enemmän, olisi mahdollista tehdä laajempaa vertailua eri toimialojen välillä. Big datan hyödyntämisen mittarina olisi myöskin hyvä käyttää useampia eri selittäviä muuttujia, jolloin pystyttäisiin tekemään syvempää analyysiä tutkimusten tuloksista.

Lähdeluettelo

Aggarwal, A. 2016. Identification of quality parameters associated with 3V's of Big Data.

Agrawal, A. 2017. Big Data: An Introduction. International Journal of Advanced Research in Computer Science, 8(3).

Akter, S. 2016. How to improve firm performance using big data analytics capability and business strategy alignment? International Journal of Production Economics, 182(C), pp. 113-131.

Brynjolfsson, E., Hitt, L. & Kim, H. 2011. Strength in numbers: How does data-driven decision-making affect firm performance?

Brynjolfsson, E. & McElheran, K. 2016. The rapid adoption of data-driven decision-making. American Economic Review, 106(5), pp. 133-139.

Bughin, J. 2016. BIG DATA: GETTING A BETTER READ ON PERFORMANCE. The McKinsey Quarterly, 1, p. 8.

Court, D. 2015. Getting big impact from big data. The McKinsey Quarterly, 1, p. 52.

Croissant, Y. & Millo, G. 2008. Panel data econometrics in R: The plm package. Journal of statistical software, vol. 27, no. 2, p. 1-43.

Davenport, T.H., Harris, J.G., 2007. Competing on Analytics: The New Science of Winning. Harvard Business School Press, Brighton, Boston.

Gartner IT Glossary (n.d.). Retrieved from https://www.gartner.com/en/information-technology/glossary/big-data

Gandomi, A. & Haider, M. 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), pp.137-144.

George, G., Haas, M. & Pentland, A. 2014. From the editors: Big data and management. Academy of Management Journal, 57(2), pp. 321-326.

Germann, F., Lilien, G. L., Fiedler, L. & Kraus, M. 2014. Do Retailers Benefit from Deploying Customer Analytics? Journal of Retailing, 90(4), pp. 587-593.

Günther, W. A. 2017. Debating big data: A literature review on realizing value from big data. Journal of Strategic Information Systems, 26(3), pp. 191-209.

Hill, C.R., Griffiths, W.E. & Lim. G.C. 2018. Principles of Econometrics. Fifth Edition. John Wiley & Sons.

Ikäheimo, S., Malmi, T. & Walden, R. 2016. Yrityksen laskentatoimi. 6., uudistettu painos. Helsinki: Talentum Pro.

Ikäheimo, S., Laitinen, E., Laitinen, T. & Puttonen, V. 2014. Yrityksen taloushallinto tänään. Vaasan Yritysinformaatio Oy.

Kiron, D., Prentice, P. & Ferguson, R. B. 2014. The Analytics Mandate. MIT Sloan Management Review, pp. 1-25.

Kiron, D. & Shockley, R. 2011. Creating Business Value with Analytics. MIT Sloan Management Review, 53(1), pp. 57-63.

Koop, G. 2008. Introduction to Econometrics. Wiley.

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. & Kruschwitz, N. 2011. Big Data, Analytics and the Path from Insights to Value. MIT Sloan Management Review, 52(2), pp. 21-32.

Laney, D. 2001. 3D data management: Controlling data volume, velocity and variety, META Group Research Note, 6(70).

Liu, Y. 2014. Big Data and Predictive Business Analytics. The Journal of Business Forecasting, 33(4), pp. 40-42.

McAfee, A., Brynjolfsson, E. 2012. Big data: the management revolution. Harvard business review, 90(10), pp.60-68.

Müller, O., Fay, M. & vom Brocke, J. 2018. The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. Journal of Management Information Systems, 35(2), pp. 488-509.

Park, H. (2010) Practical Guides to Panel Data Analysis [verkkodokumentti] [viitattu 15.11.2019] Saatavilla: http://www.iuj.ac.jp/faculty/kucc625/writing/panel_guidelines.pdf

Ramsey, M. 2014. Using 'Big Data' to deliver a competitive advantage. Plant Engineering.

Santhanam, R. & Hartono, E. 2003. Issues in Linking Information Technology Capability to Firm Performance. MIS Quarterly, 27(1), pp. 125-153.

Singh, M. 2014. Big Data in Capital Markets. International Journal of Computer Applications, 107(5).

Strawn, G. O. 2012. Scientific Research: How Many Paradigms? EDUCAUSE Review, 47(3), p. 26.

Wamba, S. F. 2017. Big data analytics and firm performance: Effects of dynamic capabilities. Journal of Business Research, 70(C), pp. 356-365.

Wooldridge, J. M. (2002) Econometric analysis of cross section and panel data. MIT Press.

Wooldridge, J. M. (2013) Introductory econometrics: a modern approach. 5th ed. SouthWestern Cengage Learning.

Wu, L., Hitt, L. & Lou, B. 2019. Data Analytics, Innovation, and Firm Productivity. Management Science.

Yaffee, R. 2003. A primer for panel data analysis. Connect: Information Technology at NYU, pp. 1-11.

Liitteet

Liite 1. Aineistossa mukana olevat yritykset

1. Accenture
2. Adobe
3. ADP
4. AIG
5. Alexion Pharmeceuticals
6. Alibaba.com
7. Allstate
8. Amazon
9. American Airlines
10. American Express
11. AmerisourceBergen
12. Amgen
13. Aon
14. Apple
15. AstraZeneca
16. Autodesk
17. BHP Billiton
18. BlackBerry
19. Blucora
20. Boeing
21. Booz Allen Hamilton
22. British American Tobacco
23. Broadcom
24. Broadridge Financial Solutions
25. CGI
26. Charles Schwab
27. Chegg, Inc.
28. Chevron
29. Chubb Insurance
30. Cigna
31. Cisco
32. Citrix
33. Cognizant Technology Solutions
34. Comcast
35. CommVault
36. CoreLogic
37. Costco Wholesale
38. Criteo
39. Dell
40. Diageo
41. Eaton
42. Electronic Arts
43. Equifax
44. Ericsson
45. Etsy
46. Expedia, Inc.
47. Facebook
48. FedEx Services
49. FICO
50. Fiserv
51. Ford Motor Company
52. Franklin Templeton Investments
53. Gartner
54. GE
55. General Motors
56. Genpact
57. Goldman Sachs
58. Groupon
59. Heidrick & Struggles
60. Hewlett-Packard
61. Honeywell
62. HSBC
63. Huron Consulting Group
64. IBM

65. IHS
66. Infosys
67. Intel Corporation
68. Intuit
69. Iron Mountain
70. Johnson & Johnson
71. Johnson Controls
72. Juniper Networks
73. Leidos
74. Lockheed Martin
75. Magellan Health Services
76. MasterCard
77. McKesson
78. Medidata Solutions
79. Merck
80. MetLife
81. Microsoft
82. Moody's Analytics
83. Morgan Stanley
84. NetApp
85. Netflix
86. Nielsen
87. Nike
88. Nordstrom
89. Novartis
90. Office Depot
91. Oracle
92. Orange
93. PayPal
94. PepsiCo
95. Perficient
96. Pfizer
97. Pitney Bowes
98. Proofpoint
99. Providence Health & Services
100. Qualcomm
101. IQVIA
102. Raytheon
103. Rubicon Project
104. Salesforce.com
105. Sanofi
106. SAP
107. Splunk
108. Starbucks
109. Suncor Energy
110. Symantec
111. Target
112. TELUS
113. Teradata
114. The Coca-Cola Company
115. The Hartford
116. The Home Depot
117. The Walt Disney Company
118. Thermo Fisher Scientific
119. Thomson Reuters
120. T-Mobile
121. TransUnion
122. Twitter
123. UnitedHealth Group
124. Walmart
125. Wells Fargo
126. Verisk Analytics
127. Verizon
128. Wipro BPO
129. Visa
130. VMWare
131. Workday
132. Xerox
133. Zillow

Liite 2. Muuttujien selitteet ja käytetyt yksiköt

Muuttuja	**Selite**	**Yksikkö**
DA	Data-analyytikoiden määrä	kpl
EBIT	Liiketulos-%	prosentti
ROI	Sijoitetun pääoman tuotto-%	prosentti
Revenue	Liikevaihto	miljoonaa dollaria
Growth	Kasvu-%	prosentti
DE	Velkaantuneisuusaste	suhdeluku

Muokattu muuttuja	**Selite**	**Yksikkö**
lgEBIT	Liiketulos-% logaritmimuunnettu muuttuja	prosentti
lgROI	Sijoitetun pääoman tuotto-% logaritmi-muunnettu muuttuja	prosentti
lgEBITlag	Liiketulos-% logaritmimuunnettu ja viivästetty muuttuja	miljoonaa dollaria
lgROIlag	Sijoitetun pääoman tuotto-% logaritmi-muunnettu ja viivästetty muuttuja	prosentti

Liite 3. Kiinteiden vaikutusten malli 1

Dependent Variable	lgEBIT			
Variable	**Coef.**	**Robust Std. Err.**	**T**	**P > |t|**
DA	-.0055347	.0110564	-0.50	0.681
lgEBITlag	-.8187012	.2566717	-3.19	0.002
Revenue	.0197057	.0098487	2.00	0.048
Growth	-.1051665	.0976257	-1.08	0.284
DE	-.0023979	.0026993	-0.89	0.376
Year	-.0352148	.077663	-0.45	0.651
_cons	-4.44166	.6114868	-7.26	0.000

sigma_u	2.1084093
sigma_e	.50689122
rho	.94535926

Liite 4. Kiinteiden vaikutusten malli 2

Dependent Variable	lgROI			
Variable	**Coef.**	**Robust Std. Err.**	**T**	**P > |t|**
DA	.0067853	.0182479	-0.37	0.711
lgROIlag	-.397198	.219255	-1.81	0.073
Revenue	.0115161	.0051366	2.24	0.027
Growth	.2003692	.1142425	1.75	0.042
DE	-.0005051	.0036411	-0.14	0.089
Year	.0055865	.1067826	0.05	0.958
_cons	-3.761326	.5300027	-7.10	0.000

sigma_u	1.27961
sigma_e	.52407614
rho	.8563559

Liite 5. Pearsonin korrelaatiokertoimet ja p-arvot

	lgEBIT	lgROI	DA	Revenue	Growth	DE
lgEBIT	1,0000					
lgROI	0,4342	1,0000				
	0,0620					
DA	0,0624	0,1616	1,0000			
	0,4476	0,0573				
Revenue	-0,1558	0,0673	0,1430	1,0000		
	0,0735	0,2096	0,5148			
Growth	-0,0397	0,0111	-0,0476	-0,1277	1,0000	
	0,4593	0,8367	0,5413	0,0510		
DE	0,0682	0,0234	-0,0360	-0,0107	-0,0283	1,0000
	0,2117	0,6676	0,5756	0,8365	0,5850	

Liite 6. VIF-arvot malli 1

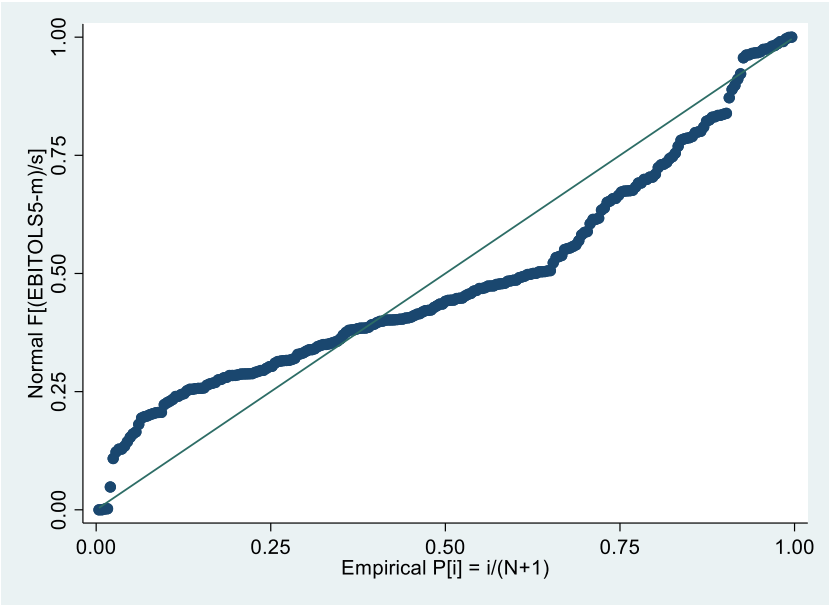
Muuttuja	**VIF-arvo**
DA	1,03
lgEBITlag	1,08
Revenue	1,09
Growth	1,03
DE	1,01

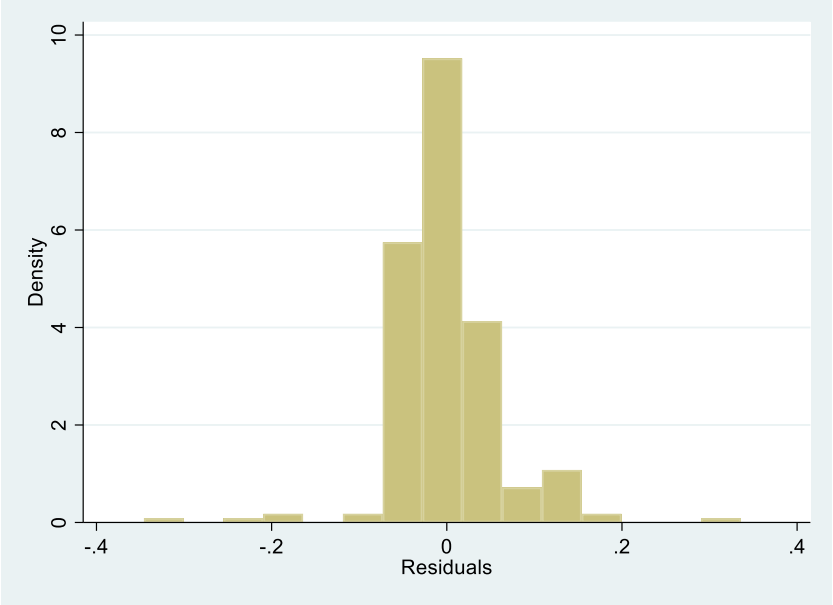
Liite 7. VIF-arvot malli 2

Muuttuja	**VIF-arvo**
DA	1,05
lgROIlag	1,04
Revenue	1,02
Growth	1,02
DE	1,03

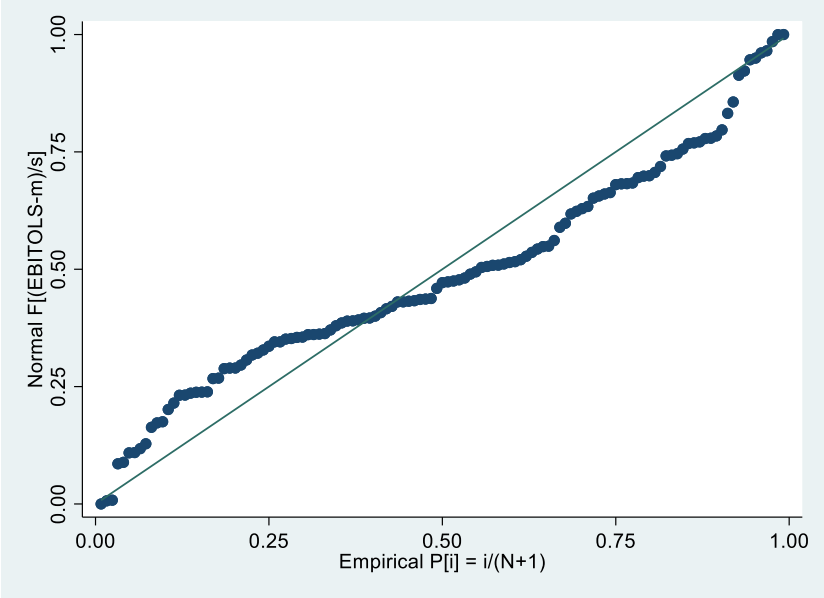
Liite 8. Mallin 1 residuaalikuvaajat

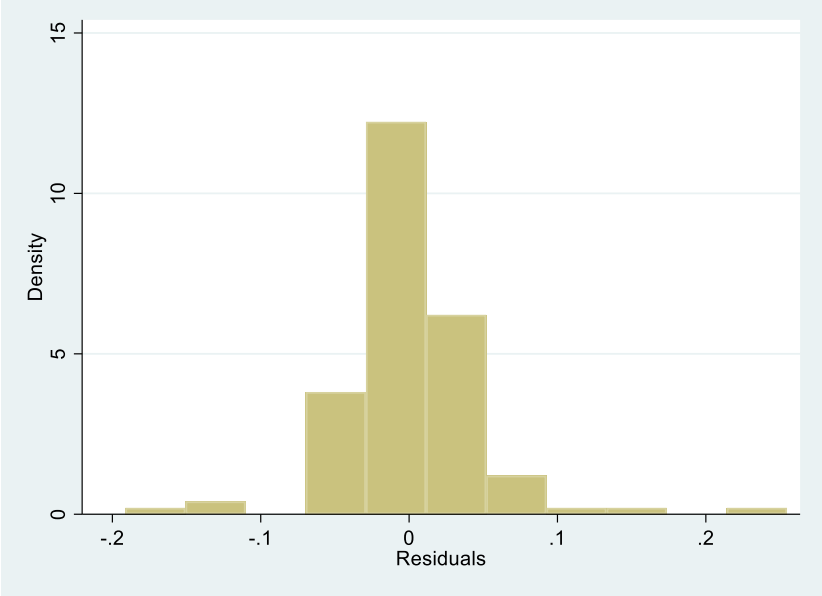
Ei muuttujamuunnoksia:



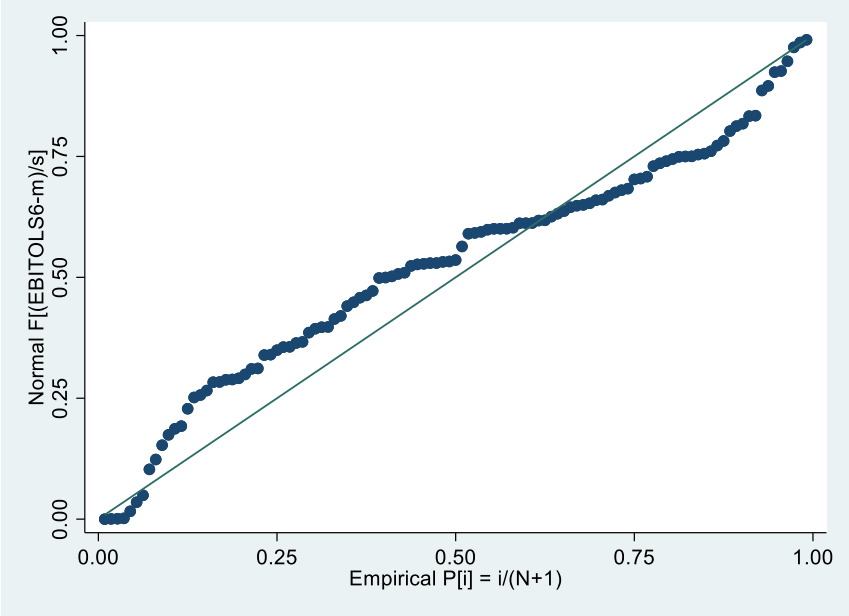


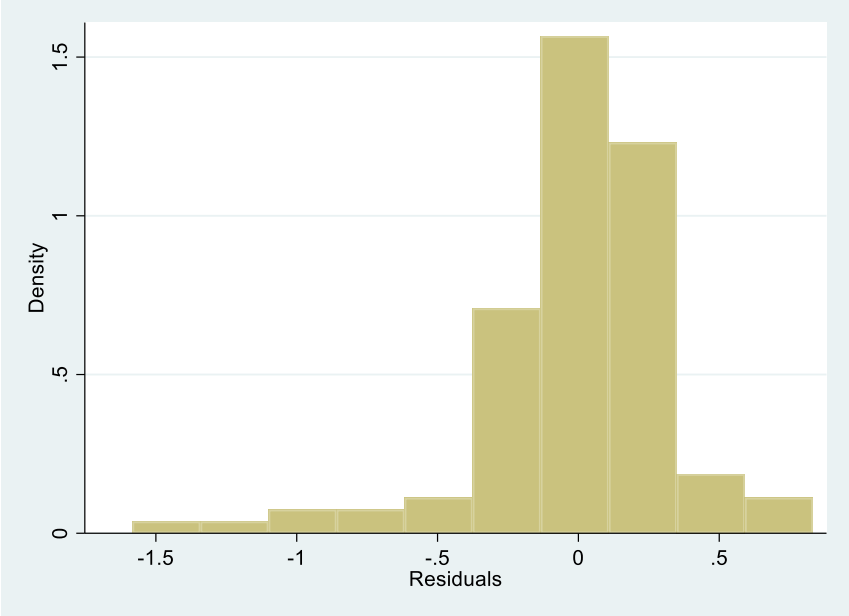
Viivästysmuutokset kontrolleina:





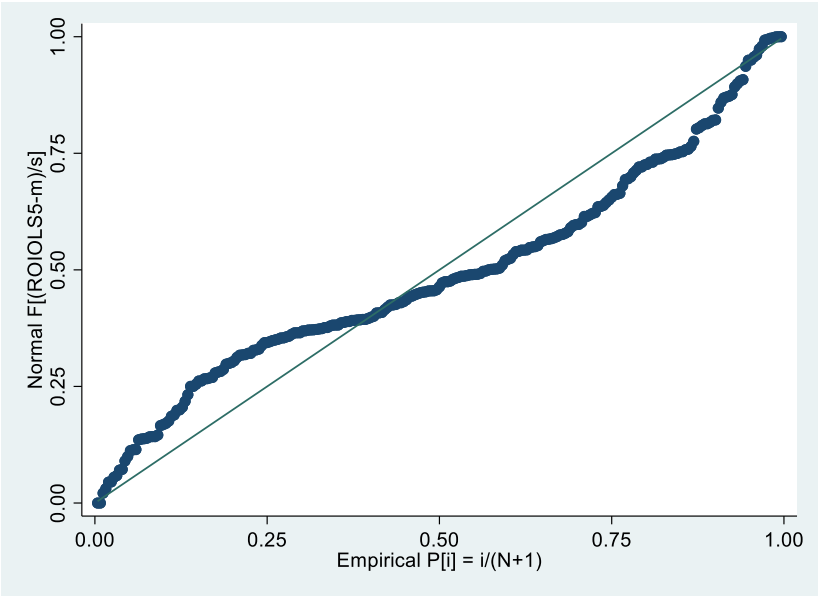
Y-muuttujan logaritmimuunnos:

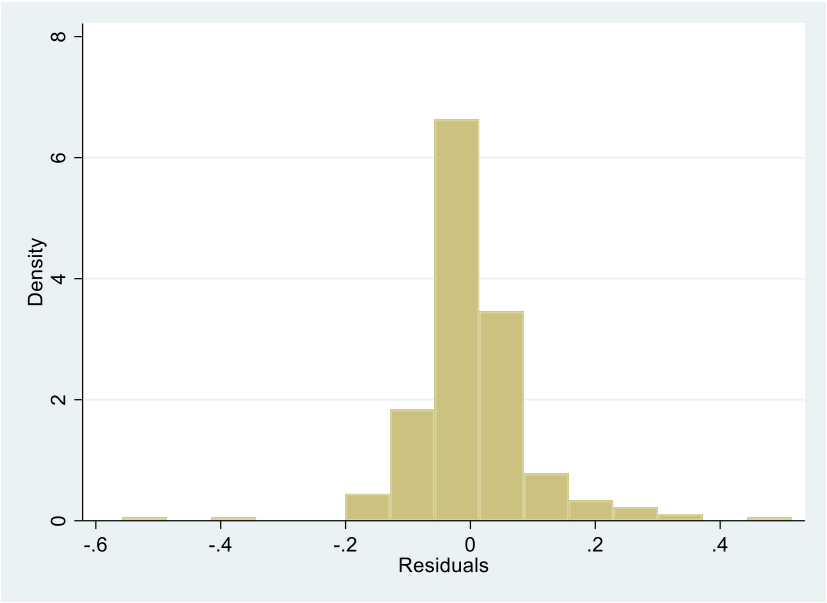




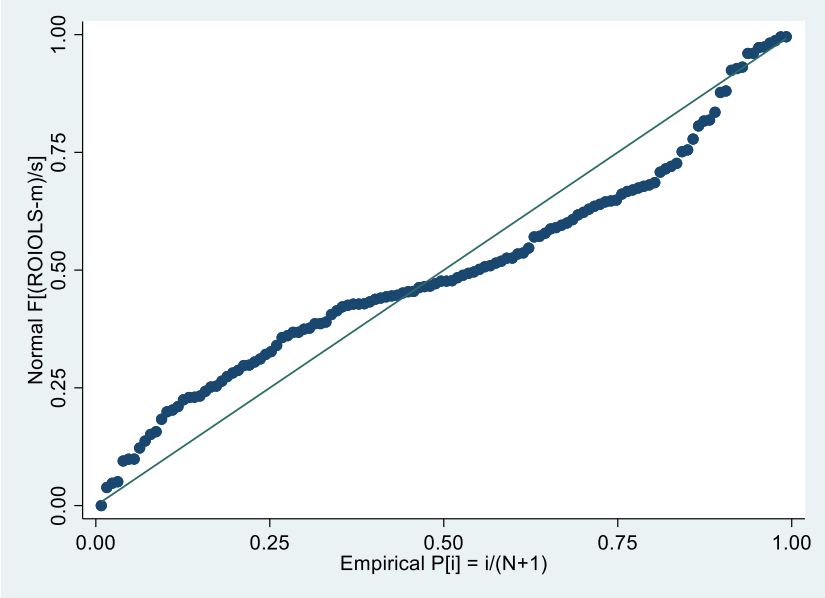
Liite 9. Mallin 2 residuaalikuvaajat

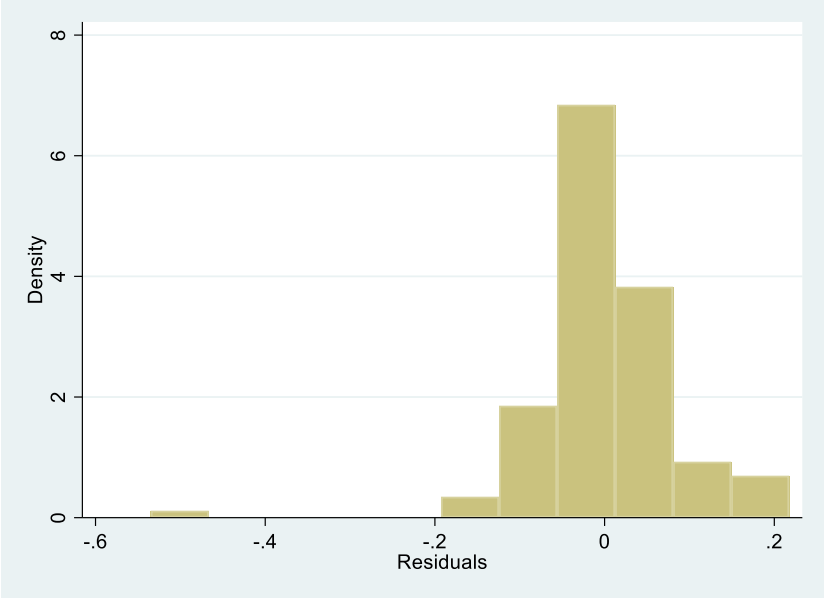
Ei muuttujamuunnoksia:





Viivästysmuutokset kontrolleina:





Y-muuttujan logaritmimuunnos: