LUT-kauppakorkeakoulu
Kauppatieteiden kandidaatintutkielma
Kansainvälinen liiketoiminta

Big datan hyödyntämisen vaikutus brändipääomaan

The Impact of the Utilization of Big Data on Brand Equity

6.1.2020
Tekijä: Saga Korkeaniemi
Ohjaaja: Pontus Huotari

TIIVISTELMÄ

Tekijä:	Saga Korkeaniemi
Tutkielman nimi:	Big datan hyödyntämisen vaikutus brändipääomaan
Akateeminen yksikkö:	LUT-kauppakorkeakoulu
Koulutusohjelma:	Kauppatieteet, Kansainvälinen liiketoiminta
Ohjaaja:	Pontus Huotari
Hakusanat:	big data, brändi, data-analytiikka, kuluttajatutkimus, paneelidata

Tässä kandidaatintutkielmassa pyritään selvittämään big datan hyödyntämisen vaikutusta yrityksen brändipääomaan. Aihetta lähestytään aikaisempien tutkimusten ja kirjallisuuden perusteella, joiden mukaan tehdään oletuksia näiden tekijöiden yhteydestä. Kirjallisuuskatsauksen avulla valitaan myös sopivat muuttujat tutkimukseen.

Tutkimus toteutetaan käyttämällä kvantitatiivista tutkimusmenetelmää. Tutkimusaineisto koostuu 121:sta pörssiyrityksestä, joista on kerätty havaintoja usealta eri ajankohdalta. Big dataa mittaava muuttuja on aineiston rajallisuuden vuoksi kerätty vain vuosilta 2013-2014, mutta loput muuttujat on kerätty ajalta 2013-2015.

Tutkimuksessa käytetään paneelidatan regressioanalyysia. Tutkimukseen sisällytetään kaksi eri selitettävää muuttujaa, joita molempia tarkastellaan omissa malleissaan. Osasta muuttujia otetaan logaritmiset ja viivästetyt arvot mallien parantamiseksi. Mallit toteutetaan käyttäen klusterirobusteja keskivirheitä.

Tutkimustulokset kertovat, että big datan hyödyntämisellä on positiivinen ja suuri vaikutus yrityksen brändipääomaan. Nämä tulokset ovat kuitenkin ristiriidassa myöhemmin esitellyissä parannetuissa malleissa, joten ne eivät ole konsistentteja. Tulokset eivät ole yleistettävissä, ja niihin tulee suhtautua kriittisesti.

ABSTRACT

Author:	Saga Korkeaniemi
Title:	The Impact of the Utilization of Big Data on Brand Equity
School:	School of Business and Management
Degree programme:	Business Administration, International Business
Supervisor:	Pontus Huotari
Keywords:	Big data, Brand, Data-analytics, Consumer research, Panel data

This bachelor's thesis is meant to explain the impact of the utilization of big data on brand equity. The subject will be approached on the basis of earlier research and literature that will help to make assumptions of the connection of these concepts. Also, appropriate variables will be chosen with the help of a literature review.

This research will be implemented using a quantitative approach. The research conducts of 121 companies listed on the stock market, from which observations have been gathered from several time periods. The variable measuring big data has been gathered from only years 2013-2014 due to limited data. Other variables, on the other hand, have been gathered from years 2013-2015.

The research is executed using panel data regression analysis. Two different dependent variables will be included in the research and both variables will be observed in their own models. Logarithmic transformations and the use of lagged values will be executed to a part of the variables to improve the models. Models will be also executed by using cluster-robust standard errors.

The findings of this research are that big data impacts brand equity positively. The effect is statistically significant and also large. However, these results are inconsistent compared to the improved models presented later. The results are not generalizable and need to be viewed critically.

SISÄLLYSLUETTELO

LIITTEET

1. JOHDANTO

”Voit omistaa dataa ilman informaatiota, mutta et voi omistaa informaatiota ilman dataa.” – Daniel Keys Moran (2012)

Yrityksen brändin jatkuva kehittäminen ja johtaminen ovat nykypäivän globaalissa markkinatilanteessa yksi suurimmista kilpailueduista. Vahva brändi antaa yritykselle sellaista aineetonta pääomaa, jota kilpailijat eivät pysty kopioimaan. Brändi on sen uniikkiuden ansiosta myös melko pysyvä kilpailuetu, jolla voi saada suuriakin markkinaetuja. Toimivan brändin kehittäminen ja johtaminen vaativat kuitenkin laaja-alaisen markkinatuntemuksen ja eri trendien tunnistamisen. Brändin markkinointi tulee olla sekä strategisella, että operatiivisella tasolla oikein suunnattua. (Wood, 2000) Tämän takia kuluttajien tutkiminen on yksi bränditutkimuksen kulmakivistä. Kuluttajatutkimusta on perinteisesti tehty muun muassa hyödyntäen kyselytutkimuksia, mutta digitalisaation ja sosiaalisen median ansiosta nykyaikana on tarjolla enemmän kuluttajadataa mitä pystytään edes generoimaan (Erevelles, Fukawa, Swayne, 2016).

Käsite ”big data” on ollut nyt muutaman vuoden pinnalla niin yritys- kuin tutkimusmaailmassa. Big datan tutkiminen on tärkeää, sillä teknologisen vallankumouksen ansiosta maapallolla oleva datan määrä kasvaa kumulatiivisella nopeudella. (Frankwick, Ramirez & Xu, 2015) Oikein käytettynä big data voi esimerkiksi ennustaa markkinatrendejä, parantaa asiakasymmärrystä sekä kehittää parempia tuotteita, palveluita ja liiketoimintamalleja (Grover, 2018). Sillä on siis suuria vaikutusmahdollisuuksia aivan uudenlaisen skaalan bränditutkimukseen.

1.1 Tutkimuksen aihe ja tavoitteet

Tämän tutkimuksen tavoitteena on selvittää big datan hyödyntämisen vaikutus yrityksen brändipääomaan. Työssä muodostetaan kysymys brändipääoman ja big datan hyödyntämisen yhteydestä, sekä vastataan tähän kysymykseen. Tutkimuskysymys on seuraava:

”Miten big datan hyödyntäminen vaikuttaa brändipääomaan?”

Tutkimuksessa on tarkoitus selvittää vaikutuksen suunnan lisäksi sen suuruutta, ja tutkimuskysymykseksi valitaan sellainen, joka pystyy tiivistämään nämä tekijät yhteen kysymykseen.

Brändipääoman ja big datan käytön yhteyttä on tutkittu hyvin vähän, sillä big data on terminä vielä melko uusi. Aiempi kirjallisuus kuitenkin osoittaa big datan hyödyntämisellä olevan positiivisia vaikutuksia muun muassa markkinointiin, tuoteinnovaatioihin ja prosessien parantamiseen. Oikein käytettynä big data tuo paljon arvoa yritysten päätöstentekoon päätöksen aihealueesta riippumatta. Tutkimukset ovat myös osoittaneet, että monet big data -projektit epäonnistuvat yrityksissä haastavan luonteensa vuoksi. Onnistuneeseen big datan implementointiin tarvitaan sitä tukeva infrastruktuuri, organisaatiokulttuuri ja osaaminen (Grover, Chiang, Liang & Zhang, 2018).

Tässä tutkimuksessa käydään läpi myös brändipääoman ja big datan arvon mittaamisen teorioita. Vaikka brändipääomaa onkin tutkittu jo kauan, sille ei ole tällä hetkellä olemassa yleisesti hyväksyttyä teoriaa, jonka avulla sitä voitaisiin mitata. Big datan tutkiminen on kasvanut termin suosion myötä, mutta sen suoria vaikutuksia organisaation arvoon on tutkittu melko vähän. Tutkielmassa käydään läpi eri näkökulmia sekä brändin, että big datan mittaamisesta, ja lopuksi valitaan sopivat mittarit perustuen aiempaan kirjallisuuteen ja teorioihin.

Tämän tutkielman aihe on tärkeä, sillä suoria vaikutuksia brändipääoman ja big datan hyödyntämisen välillä ei tietääkseni ole tutkittu. Sekä brändi, että big data tulevat olemaan tulevaisuudessa suuria kilpailuetuja yrityksille globaalin kilpailun kiristyessä.

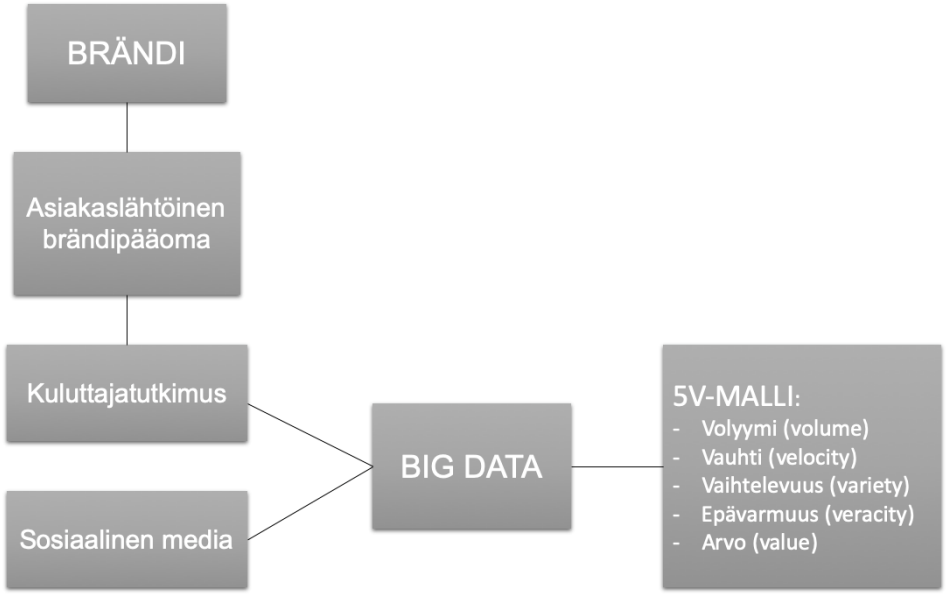
1.2 Teoreettinen viitekehys

Tämän tutkimuksen lähtökohtana on oletus, että big datalla on positiivisia vaikutuksia brändipääomaan. Oletuksen tekeminen perustellaan teoreettista taustaa esitellessä. Ydinajatuksena tutkielmassa on digitalisaation vaikutus bränditutkimukseen ja brändin johtamiseen.

Teoreettista viitekehystä lähdetään rakentamaan yhdistelemällä useita eri teorioita. Bränditutkijat ovat aikojen saatossa muodostaneet brändille lukuisia määritelmiä ja viitekehyksiä. Yksi tunnetuimmista on Kellerin (1993) luoma käsite asiakaslähtöisestä brändipääomasta. Tutkimus nojaa lähtökohtaisesti tähän käsitteeseen, jonka mukaan yrityksen brändi syntyy jokaisen yksilön omassa mielessä kognitiivisena ajatusrakennelmana. Asiakaslähtöisen brändipääoman avulla perustellaan kuinka tärkeää kuluttajatutkimus ja varsinkin kuluttajalähtöisen datan keruu ovat brändin johtamisessa. Etenkin sosiaalisen median nousevan aseman ansiosta nykypäivänä on enemmän yleisellä tasolla saatavissa olevaa kuluttajadataa kuin koskaan ennen. (Erevelles et al., 2016)

Näin ollen teoreettinen viitekehys siis yhdistää brändipääomaan käsitteen big dataan kuluttajatutkimuksen kautta. Big datan monimuotoisuus ja mittaamisperusteet määritellään 5V-mallin avulla, joka on myös tunnettu teoria asiantuntijoiden keskuudessa. 5V-malli määrittelee big dataa sen viiden V:n kautta, joita ovat volyymi *(volume)*, vauhti *(velocity)*, vaihtelevuus *(variety)*, epävarmuus *(veracity)* ja arvo *(value)* (IBM, 2012; Lycett, 2013; Oracle, 2012)

Kuva 1. Teoreettinen viitekehys



1.3 Tutkimuksen rajaus

Tutkimuksen rajaus tapahtuu aineiston saatavuuden rajoissa. Tutkimuksen luotettavuuden takaamiseksi aineistoa kerätään eri ajankohdilta. Datan saatavuutta kuitenkin rajoittaa erityisesti listat data-analyytikoista, joita on kerätty vain kahdelta vuodelta. Data-analyytikkolistat on kerätty vuodelta 2013-2014, kun taas muut muuttujat ovat ajalta 2013-2015. Tutkimus rajataan koskemaan vain pörssissä listattuja yrityksiä, jotta validia dataa olisi riittävästi saatavilla. Sen lisäksi aineistosta rajataan pois jo mukaan valittujen yhtiöiden tytäryhtiöt päällekkäisyyksien välttämiseksi. Maantieteelliset tai toimialakohtaiset rajoitteet päätetään jättää tekemättä, jotta aineisto ei suppenisi liikaa.

1.4 Tutkimusmenetelmät ja aineisto

Tutkimus toteutetaan kvantitatiivisella lähestymistavalla. Tutkimusmenetelmänä hyödynnetään paneelidatan regressioanalyysia. Aineistoa analysoidaan estimaattorin valintatesteillä, jotta sopivat estimaattorimenetelmät löytyvät. Mallien luotettavuutta analysoidaan myös regression yhteydessä ja sen jälkeen.

Aineisto on lukuisista eri lähteistä yhdistetty paneelidata. Paneelidata on aineistoa, jossa yhdistyy sekä poikkileikkausyksiköt, että aikasarjadata. Aineiston lähteitä ovat Forbesin Global 2000 -yrityslistat, Macrotrends.com sekä artikkelit yrityksissä työskentelevistä data-analyytikoista. Aineiston lähteet on kuvattu tarkemmin kappaleessa 3.4.

1.5 Tutkimuksen rakenne

Kandidaatintutkielma rakentuu viidestä eri pääluvusta. Johdannossa käsitellään tutkimuksen yleistä aihetta ja tavoitteita, sekä käydään tutkimusmenetelmä ja aineisto lyhyesti läpi. Kirjallisuuskatsaus ja teoreettinen aiheen tarkastelu tapahtuu toisessa luvussa. Luvussa käydään läpi tärkeimpiä käsitteitä, sekä aikaisempia tutkimuksia. Kolmas luku käsittelee tutkimusmenetelmää ja muuttujien valintaa. Osiossa käydään myös tutkimusaineisto läpi. Neljännessä luvussa esitellään regression tutkimustulokset. Luvussa arvioidaan myös tutkimuksen reliabiliteettia. Tutkielman päättää viides

luku, jossa kootaan johtopäätökset ja vastataan tutkimuskysymykseen. Luvussa pohditaan myös jatkotutkimuksen mahdollisuuksista.

2. TEOREETTINEN TAUSTA

Tässä luvussa käsitellään aiempaa kirjallisuutta ja tutkimuksia viitekehykseen liittyen. Osiossa määritellään oleellisia termejä, vertaillaan muuttujien mittausteorioita ja syvennytään brändipääoman ja big data-analytiikan yhteyteen.

2.1 Brändipääoman tutkiminen ja mittaaminen

Terminä brändi johtaa juurensa alun perin karjan polttomerkitsemisestä. Teollistumisen aikana ”brändäys” kuitenkin levisi myös yritysmaailmaan, kun yrityksen alkoivat käyttää personalisoituja logoja tuotteissaan. Nykypäivänä brändit ovat tärkeä osa yrityksen identiteettiä, mainetta ja imagoa. Systemaattinen brändin johtaminen voi auttaa yrityksiä saamaan sellaista aineetonta pääomaa, joka ei ole helposti kilpailijoiden kopioitavissa ja täten hyvin arvokasta. (Juuti, Laukkanen, Puusa & Reijonen, 2014, 228-229).

Kolmen viimeisen vuosikymmenen aikana yritysmaailmassa on tapahtunut paradigman muutos, jonka seurauksena brändistä on tullut tehokas työkalu kilpailuedun saavuttamiseen ja ylläpitämiseen. Sen kasvavan relevanssin ansiosta bränditutkimus on myös lisääntynyt. Brändi on käsitteenä hyvin moniulotteinen, ja sitä voidaan tutkia sekä psykologisesta, että liiketoiminnallisesta näkökulmasta. (Narayan, 2012). Brändin ydin on kuitenkin siinä ajatuksessa, että asiakas on valmis maksamaan lisähintaa tuotteesta, jonka brändistä hänelle tulee mieleen positiivisia mielleyhtymiä (Keller, 1993). Brändiin liittyvät assosiaatiot näin ollen tuovat yrityksille sellaista kilpailuetua, mikä tuo niille helposti tuottoja. Tässä osassa keskitytään erityisesti brändin mittareiden tunnistamiseen, eli yrityksen brändipääoman määrittämiseen nojaten aikaisempaan kirjallisuuteen.

Vaikka markkinointitutkijat ja alan ammattilaiset ovatkin yksimielisiä brändipääoman merkityksestä, yhtenäisesti hyväksyttyä mittaria tai arviointimallia ei ole kuitenkaan vielä syntynyt. Brändipääomaa voidaan tarkastella monesta eri näkökulmasta. (Narayan, 2012) Esimerkiksi de Oliveira, Silveira & Luce (2015) ovat jakaneet sen mallissaan erikseen käyttäytymispohjaiseksi sekä finanssipohjaiseksi. Finanssipohjaisessa mallissa pyritään määrittelemään brändin rahallinen arvo kirjanpidollisia tarkoituksia varten, kun taas markkinointinäkökulmaan nojaava käyttäytymispohjainen malli keskittyy enemmän yksilöllisiin kognitiivisiin mielikuviin brändistä ja tämän tuomasta kilpailuedusta. (de Oliveira et al., 2015) Kirjallisuuteen nojaten, voidaan todeta, että finanssipohjaisesta näkökulmasta brändipääomaa on helpompi määritellä ja mitata.

Brändipääomalla on ehdotettu olevan niin suuri merkitys yritysten arvossa, että se pitäisi erikseen huomioida yrityksen varoja kirjatessa (Narayan, 2012). Yksi taseellinen arvo, missä brändipääoman voidaan nähdä korostuvan, on yrityksen liikearvo, joka otetaan huomioon yrityskaupoissa tai fuusioissa. Jos yrityskauppa tehdään sellaisella summalla, joka ylittää ostettavan yrityksen kirjanpidollisen arvon, merkitään tällöin ylijäämä taseessa liikearvona. Sen siis voidaan katsoa heijastavan osakseen brändiarvoa. Puhtaan brändiarvon erotteleminen liikearvosta voi kuitenkin osoittautua melko monimutkaiseksi. (Johnson & Petrone, 1998). Yrityksen aineettomaan pääomaan voi kuulua myös paljon muita arvoa tuovia osuuksia, kuten tutkimus- ja kehitystoimintaa ja osaamispääomaa. Muun muassa Tollington (1998) onkin ehdottanut brändiarvon olevan niin tärkeä pääoma, että se pitäisi tunnustaa itsenäisesti taseessa.

Simon ja Sullivan (1993) ovat kehittäneet tekniikan brändipääoman mittaamiseen perustuen yrityksen arvoon markkinoilla. Tekniikka pohjautuu vahvasti tehokkaiden pääomamarkkinoiden hypoteesiin. Tämä hypoteesi ennustaa, että hyvin toimivissa pääomamarkkinoissa arvopapereiden arvostaminen tarjoaa parhaan saatavissa olevan sekä harhattoman arvion yhtiön varoista. Toisin sanoen, yhtiön osakekurssin voidaan ajatella heijastavan reaaliajassa kaikkia saatavilla olevia tietoja odotettavissa olevista kassavirroista. (Ross, 1983; Fama, 1970) Myös Schultzin (2002) tutkimus tukee osakepohjaista lähestymistapaa. Hän esittelee markkina-arvopohjaisen lähestymistavan, joka mittaa brändiä tulevien kassavirtojen kautta. Yeung ja Ramaswamy (2007) tutkivat

vuosina 2000 ja 2005 viittäkymmentä yhdysvaltalaista yritystä ja löysivät myös vahvan korrelaation yrityksen osakemarkkinoilla suoriutumisen ja brändipääoman välillä.

Aaker (1996) on luonut niin sanotun ”Brand Equity 10” -mallin, jossa kymmenen eri brändiarvon mittaria on jaettu viiteen eri kategoriaan: luotettavuuteen, bränditietoisuuteen, koettuun laatuun, mielleyhtymiin sekä markkinakäyttäytymiseen. Neljä ensimmäistä kategoriaa edustavat kuluttajalähtöisiä käsityksiä brändipääomasta ja niitä voidaan mitata muun muassa kyselytutkimuksilla. Viimeinen kategoria on mitattavissa ennen kaikkea markkinapohjaista informaatiota hyödyntäen. (Aaker, 1996, 105) De Oliveiran et al. (2015) teorian pohjalta myös Aakerin kategoriat voidaan jakaa finanssipohjaisiin ja käyttäytymispohjaisiin malleihin. Finanssipohjaista mallia edustaa markkinakäyttäytyminen, jota mitataan kahdella muuttujalla: markkinaosuudella ja markkina-arvolla. Markkinaosuus toimii heijastuksena brändien asemasta kuluttajien kannalta. Kun brändillä on suhteellinen etu asiakkaiden mielissä, sen markkinaosuuden tulisi kasvaa. Markkinaosuus voi kuitenkin osoittautua ongelmalliseksi mittausmielessä, sillä tarkan tuoteluokan ja kilpailusektorin määrittäminen voi olla vaikeaa.

Kerinin ja Sethuramanin (1998) tutkimus myötäilee myös markkina-arvon ja brändipääoman yhteyttä. Tutkimuksessa käytetään kuitenkin puhtaan markkina-arvon sijasta price-to-book -nimistä tunnuslukua. Tämä tunnusluku suhteuttaa yrityksen markkina-arvon yrityksen taseessa olevaan kirja-arvoon. Kun markkina-arvo on korkeampi kuin yrityksen oman pääoman arvo, voidaan ajatella yrityksen todellisen arvon olevan suurempi kuin taseessa esitetyt luvut. Tämä yleensä tarkoittaa, että yrityksellä on erikseen aineetonta pääomaa, usein brändipääomaa, joka tulisi huomioida erikseen. Tutkijat löysivät positiivisen ja merkittävän korrelaation brändiarvon ja P/B -luvun välillä, kun he käyttivät Financial Worldin arvioita yritysten brändiarvoista. Myös Little, Coffee ja Lirelyn (2005) mukaan yrityksillä, joilla on korkeat brändiarvot, on myös merkittävästi korkeammat P/B-luvut verrattuna yrityksiin pienellä tai jopa olemattomalla brändiarvolla.

P/B -tunnusluku voidaan laskea seuraavalla tavalla (Kailunki & Niemelä, 2004, 86):

$$P/B - luku = \frac{markkina-arvo}{oma\ pääoma} \quad (1)$$

2.2 Big data-analytiikan tutkiminen ja mittaaminen

”Big data” on terminä melko epäselvä, koska sille ei ole olemassa mitään tiettyjä volyymirajauksia. Big datalle onkin olemassa monta määritelmää. (Erevelles et al., 2016) Esimerkiksi Chen (2012) määrittelee sen valtavaksi datamassaksi, joka on jatkuvasti kasvava ja peräisin monista eri lähteistä. Yksi tapa määritellä big dataa, on 3V-malli, joissa V:t kuvaavat volyymia, vauhtia ja vaihtelevuutta (IBM, 2012; Lycett, 2013; Oracle, 2012). Big datan volyymia eli määrää mitataan tällä hetkellä petatavuina, exatavuina tai zettatavuina (Erevelles et al., 2016). Esimerkiksi yksi petatavu vastaa noin 20 miljoonaa perinteistä arkistokaappia tietoa (Mcafee & Brynjolffson, 2012). Big datalle ominaista on myös sen vauhti, joka kuvaa datavirtojen valtavaa syntymisnopeutta, kumulatiivisesti kasvavaa määrää ja ajankohtaisuutta. Kolmas V kuvaa datan vaihtelevaa luonnetta; se voi olla joko strukturoitua, epästrukturoitua tai jotain siltä väliltä. Strukturoitu data on helpommin luettavaa ja sitä voi löytää esimerkiksi tietokannoista. Epästrukturoituun dataan kuuluu erilaista järjestelemätöntä tekstidataa, videoita, kuvia ja äänitallenteita. Eniten epästrukturoitua dataa syntyy sosiaalisen median kautta, jossa yksilöt jakavat tietoa päivittäisestä elämästään. (Erevelles et al., 2016). Eri sosiaalisen median alustojen kasvulla on ollut suuri vaikutus epästrukturoidun datan syntymiselle, ja tällä hetkellä noin 95% big datasta on epästrukturoitua (Kaplan ja Haenlein, 2010; Grover et al., 2018). Jos yritys onnistuu valjastamaan sellaisen osaamisen tai ohjelmiston, joka järjestää epästrukturoidusta datasta strukturoitua, yritykselle käyttökelpoista tietoa, tuo se valtavasti aineetonta pääomaa yritykselle (Erevelles et al., 2016).

Muun muassa Ebner, Bühnen & Urbach (2014) sekä Lycett (2013) ovat muokanneet 3V-mallista 5V-mallin. Edellä mainittujen lisäksi big dataa voidaan kuvata sen epävarmuudella (*veracity*) ja arvolla (*value*). Epävarmuus täytyy ottaa huomioon, sillä kuluttajista kerätty data ei välttämättä ole virheetöntä. Epävarmuus nousee yhä suuremmaksi ongelmaksi, kun otetaan huomioon datan kasvava määrä, syntymisnopeus ja vaihtelevuus. Arvo taas kuvaa sen arvon määrää, joka datan hyödyntämisestä voi syntyä. (Erevelles et al., 2016)

Hitt, Lou ja Wu (2019) määrittelevät data-analytiikan kyvykkyydeksi prosessoida, analysoida ja muuntaa dataa päätöksenteon tueksi. Big data-analytiikka eli BDA on big

datasta arvokkaan tiedon keräystä ja hyödyntämistä käyttäen erilaisia data-analytiikan keinoja. BDA:n hyödyt perustuvat siihen, että suuri määrä epästrukturoitua tietoa useista eri lähteistä voi synnyttää oivalluksia, jotka auttavat yrityksiä muuttamaan liiketoimintaansa ja saavuttamaan etulyöntiaseman kilpailijoihin nähden (Chen, Chiang & Storey, 2012). Jotta tämä oivallusten tekeminen onnistuisi, yritys tarvitsee sopivan yhdistelmän dataa, teknologiaa, organisatorisia resursseja sekä ihmisosaamista (Vidgen, Shaw & Grant, 2017).

Big datan käyttöä on itsessään tärkeä tutkia, sillä se on hyvin kasvava ala. 2000-luvun alussa noin 25% varastoidusta tiedosta oli digitaalista, mutta jo vuonna 2015 98% maailman informaatiosta oli digitaalisessa muodossa. (Frankwick et al., 2015). PromptCloudin (2016) raportin mukaan big data teollisuudenalana on kasvanut kolmessa vuodessa 6.8 biljoonan dollarin arvosta 32 biljoonan dollarin arvoon. Kasvun ennustetaankin nousevan 23% vuodessa, joten big data tulee olemaan oleellinen voimavara tulevaisuudessa. Nykypäivänä organisaatiot, jotka aktiivisesti hyödyntävät big data-analytiikkaa erottuvat kilpailijoistaan, kun taas monet yritykset, jotka eivät ole implementoineet big datan käyttöä kamppailevat säilyttääkseen markkinaosuutensa. (IBM, 2014)

Vaikka big dataan kohdistuva tutkimustyö on käsitteen suosion myötä kasvanut, vaikutus liiketoiminnan arvoon on vielä melko vähän tutkittu aihe. Harvat tutkimukset ovat tutkineet big datan suoria vaikutuksia organisaation arvoon. (Forrester, 2011) Suurin ongelma on kuinka mitata big data -investointien vaikutusta yrityksen aineelliseen liiketoiminta-arvoon. Datan arvon mittaaminen on vaikeaa, sillä sen luonne on moneen muuhun pääomalähteeseen verrattuna melko uniikkia, kuten aikaisemmin esitelty 5V-malli demonstroi. Data on luonteeltaan uudelleenkäytettävää, integroitavaa eikä sitä voi niin sanotusti ”kuluttaa”. (Grover et al., 2018)

Suurin osa tähän mennessä tehdyistä raporteista koskien BDA:n luomaa arvoa yrityksissä tulevat valtamediasta tai case-tutkimuksista, jotka epäonnistuvat yleistettävien empiiristen tulosten esittelemisen laajemmassa skaalassa (Gupta & George, 2016). Yksi tapa lähteä mittaamaan big datan arvoa on yleisemmän yrityksen käyttämän tietotekniikan arvon näkökulmasta. Chau, Kuan ja Liang (2007) määrittelevät IT-arvon

tulokseksi, joka on seuraus tietotekniikan käytöstä. IT-arvoon liittyvässä kirjallisuudessa on tunnistettu, että se toimii merkittävänä tuottavuuden ajurina, mutta vain silloin, kun yritykset samalla implementoivat sitä täydentäviä organisatorisia käytäntöjä. (Melville, Kraemer & Gurbaxani, 2004) Nämä käytännöt auttavat yritystä keräämään ja tekemään tiedon pohjalta sisäisiä operationaalisia päätöksiä (Hitt, Jin & Wu, 2015).

McAfeen ja Brynjolfssonin (2012) tekemän tutkimuksen mukaan dataorientoituneisuus korreloi yrityksen performanssin kanssa. Mitä datasuuntautuneemmaksi yritykset itsensä kokivat, sitä paremmat operatiiviset ja taloudelliset performanssiluvut ne omistivat. Eniten data-analytiikkaa päätöksenteossa hyödyntävässä kolmasosassa olevat yritykset olivat myös noin 5 prosenttia tuottavampia ja 6 prosenttia kannattavampia kuin kilpailijansa. Myös Saunders ja Tamble (2013) toteavat tutkimuksessaan, että yritykset, jotka keskittyvät käyttämään dataa operationaalisella tasolla omistavat suuremmat tuottavuus- ja markkina-arviot. Täytyy kuitenkin ottaa huomioon, että datasta löytyvä tieto on vain lähtökohta päätöksentekoon, joka voi seurauksena tuottaa arvoa. Tiedon todellinen arvo paljastuu vasta, kun arvioidaan toimintaa, joka päätöksenteosta seuraa. (Grover et al., 2018)

2.3 Digitalisaation vaikutus bränditutkimukseen

Yhtenä brändikirjallisuuden lähtökohtana voidaan pitää Kellerin (1993) luomaa teoriaa asiakaslähtöisestä brändipääomasta. Sen mukaan organisaation sisällä syntyvän brändin sijasta, brändi syntyy jokaisen kuluttajan mielessä uniikkina kognitiivisena ajatusrakennelmana. Kuluttaja valitsee tietyn brändin tuotteen yli muiden tuotteiden, jos hän muistaa brändin ja omistaa positiivisia, vahvoja ja uniikkeja assosiaatioita siitä. Näin ollen asiakkaan ymmärtäminen on tärkein osa brändin luomista ja siihen liittyvää päätöstentekoa. (Keller, 1993)

Asiakkaita on perinteisesti tutkittu käyttämällä erilaisia kyselytutkimuksia ja kvalitatiivisia seurantatutkimuksia (Keller, 1993) Bränditutkimus on kuitenkin tällä hetkellä eräänlaisessa muutosaallossa, sillä digitalisaatio on mahdollistanut aivan uusia tapoja kerätä

kuluttajadataa. Big datan vaikutukset brändipääomaan perustuvat siihen, että asiakaslähtöistä brändipääomaa voidaan kasvattaa tutkimalla kuluttajia laajemmalla skaalalla kuin koskaan ennen. Kuluttajatutkimukseen big data on siis erittäin hyödyllinen keino.

Kuluttajadata on yleensä vielä arvokkaampaa, kun eri tietoja integroidaan, ja tämä voi tarjota ennennäkemättömän mahdollisuuden vaikuttaa yrityspäätöksiin (Hitt et al., 2019). Nykypäivänä yritysten johtajat ymmärtävät sosiaalisen median tuomat mahdollisuudet, mutta heiltä kuitenkin usein puuttuu oikeat työkalut tällaisen prosessin johtamiseen (Mount, 2014).

Erityisesti ajankohtaisuus ja datan kumuloituva määrä osoittautuvat tärkeäksi, kun yritys yrittää ennustaa markkinatrendejä ja tehdä kuluttajatutkimusta (Erevelles et al., 2016). Perinteinen markkina-analyysi on keskittynyt pääosin tutkimaan ja parantamaan performanssiin liittyviä avainmittareita, kun taas BDA hyödyntää jatkuvaa informaatiovirtaa ja analysoi massiivisia datavolyymeja reaaliajassa kokonaisvaltaisen asiakasymmärryksen saavuttamiseksi. (Sathi, 2014) Big data -tutkimus on kyselytutkimuksiin verrattuna paljon kattavampi tapa tutkia kuluttajia, sillä kyselytutkimuksia tehdään verraten melko pienillä otoksilla (Wu & Brynjolfsson, 2009).

2.3.1 Brändin johtaminen sosiaalisen median kautta

Mount (2014) kuvaa yhdeksi onnistumisen esimerkiksi Nestle UK:n ”Kit Kat” -suklaapatukkabrändin, joka on kehittänyt vahvan kyvyn hallita sosiaalista ympäristöään ja löytänyt uusia oivalluksia hyödyntämällä laaja-alaista dataa sosiaalisesta mediasta. Nestle onnistui saamaan aikaan positiivisemman brändikuvan keräämällä ja kehittämällä dataa brändiä kannattavilta kuluttajilta. Yritys pystyi luomaan datan pohjalta *word-of-mouth* -markkinointistrategiansa, jolloin uskolliset kuluttajat myös markkinoivat brändiä eteenpäin. (Mount, 2014)

Nestlen tapauksen pohjalta kehitettiin SELI-viitekehys, joka auttaa yrityksiä hyödyntämään sosiaalisen median kaltaista big dataa kehittääkseen brändiään. Kehys koostuu neljästä vaiheesta; *scan, engage, learn, internalize*. Viitekehyksen mukaan yritysten tulisi aluksi aktiivisesti kartoittaa ja seurata käyttäjien luomaa dataa tunnistaakseen

nousevia trendejä. Esimerkiksi Nestle käytti kehittyneitä data-analytiikan työkaluja louhiakseen tietoa ja löytääkseen johtolankoja brändikiintymyksen parantamiseksi. Yksi oivallus oli, että erilaiset patukkamaut kiinnostivat kuluttajia ja olivat usein keskustelun aihe. (Mount, 2014)

Viitekehyksen mukaan organisaatioiden tulisi myös ottaa itse osaa keskusteluihin ja muodostaa dialogi asiakkaiden kanssa. Nykypäivänä kommunikoinnin keinot internetissä ovat hyvin monipuolisia, pelkän tekstin sijasta yritys voi käyttää muun muassa erilaisia audiovisuaalisia keinoja tai sosiaalisten medioiden integrointia. Näiden työkalujen avulla yritykset tavoittavat entistä enemmän erilaisia kuluttajaryhmiä tavalla, joka auttaa kehittämään ja määrittelemään brändikehitykseen tarvittavaa markkinatietoa. Esimerkiksi Nestle päätti järjestää asiakasäänestyksen uudesta patukkamausta. Yritys käytti integroituja äänestämistyökaluja, jolloin kuluttajat saivat helposti äänestää voittajan eri patukkamauista, ja aihe nostatti paljon lisäkiinnostusta brändiin liittyen. (Mount, 2014)

Kun yritys on onnistunut saavuttamaan toimivan dialogin kuluttajien kanssa, täytyy vielä koota yhteen kaikki kumuloitunut tieto ja jakaa se organisaation kesken. Mountin tutkimuksen mukaan osaamisen puuttuminen laajan datamäärän käsittelyssä nousee usein isoimmaksi esteeksi, kun yritys haluaa kerätä ja järjestää tietoa loppukäyttäjäyhteisöstä. Esimerkiksi Nestle on ratkaissut ongelman hyödyntäen erilaisia yrityskumppanuuksia ja konsultteja, kuten facebookia. Kun datasta on onnistuneesti saatu hyödyllisiä oivalluksia ja tietoa brändin kehitykseen, tämä tieto täytyy myös kommunikoida koko organisaatiolle, jotta uudet ratkaisut saadaan implementoitua yrityksen päivittäiseen toimintaan. Nestlen tapauksessa sosiaalisen median ja BDA:n hyödyntäminen mahdollisti bränditutkimuksen tekemisen laajemmalla skaalalla kuin koskaan ennen brändin historiassa. Yrityksen johtajat onnistuivat saamaan hyödyllistä markkinatietoa alhaisilla kustannuksilla ja nostamaan markkinapenetraatiotaan jopa kahdeksalla prosentilla. (Mount, 2014)

Nestlen bränditutkimus osoittaa ensinnäkin, että sisäisten BDA-kykyjen puuttuminen voi olla keskeinen kompastuskivi brändin johtamisessa sosiaalisen median kautta.

Nestlen yrityskumppanit ja konsultit olivat avainasemassa Kit Kat -brändin menestymisessä. Toinen huomioonotettava asia on yrityksen hierarkia. Mountin mukaan organisatoriset rakenteet voivat olla esteitä ulkopuolisen tiedon pääsyyn yrityksen sisälle. Vähemmän byrokraattiset rakenteet osoittautuvat melkein yhtä tärkeäksi sosiaalisen median onnistuneelle hyödyntämiselle, kuin korkealaatuinen data-analyysi. (Mount, 2014)

2.3.2 Data-analyytikoiden merkitys

Kun otetaan huomioon 5V-mallin esittämä datan monimuotoisuus ja muuttuva ympäristö, voidaan perustella, että big dataan investoiminen poikkeaa merkittävästi perinteisimmistä investoinneista. Investointi big dataan ja analytiikkaan vaatii ainakin sopivan infrastruktuurin, organisaatiokulttuurin ja osaamisen. Grover et al. (2018) ehdottavat, että tärkein tekijä näistä olisi ihmisosaaminen. Big datan implementoinnin onnistuminen vaatii ammattiosaamista sekä kokemusta, jota data-analyytikot voivat tarjota. Analyytikkojen tarve on kriittisintä etenkin, kun suunnitellaan toteutettavaa tiedonkeruun strategiaa ja sen lisäksi myöhemmin tulosten tulkinnassa. (Grover et al., 2018)

McAfee ja Brynjolfsson (2012) ovat osoittaneet, että erilaiset dataa täydentävät organisatoriset rakenteet nousevat yhä tärkeämpään rooliin datan määrän kasvaessa. He myös ehdottavat data-analyytikoiden olevan yksi tärkeimmistä voimavaroista BDA-projektissa. Mitä tulee big dataan, perinteisten tilastotieteilijöiden taidot voivat olla riittämättömiä tiedon käsittelyyn tarvittavien tekniikoiden käyttöön. Tutkijat määrittelevät tärkeimmiksi data-analyytikoiden taidoiksi suurten datamassojen erottelemisen ja järjestelemisen.

Aiemman kirjallisuuden perusteella organisaatiossa työskentelevät data-analyytikot voivat toimia myös indikaattorina, kun mitataan yrityksen datan käyttöä. Muun muassa Hitt et al. (2019) käyttävät dataorientoituneisuuden mittaamiseen yrityksissä työskentelevien data-analyytikkojen määrää. He laskevat kokonaismäärän työntekijöistä, joilla on relevantit tietojen analysointitaidot yli kuudesta miljoonasta ansioluettelosta. Tässä

tapauksessa analysointitaidot voivat kattaa monenlaisia taitoja, kuten tiedon puhdistamisen, tietojen muuntamisen ohjelmistotyökaluja hyödyntäen, sekä edistyneen koneoppimisen ja tekoälyn taitoja.

3. TUTKIMUSMENETELMÄ JA -AINEISTO

Tässä osiossa esitellään käytettävä aineisto, ja käydään läpi tutkimukseen valitut muuttujat. Muuttujien valintaa perustellaan myös tässä osiossa, sekä käytetään aiemmin läpikäydyn teorian perusteita. Lisäksi luvussa esitellään tutkimusmenetelmä, jolla tutkimus tehdään, ja tehdään tähän nojaten tarvittavat muuttujamuunnokset.

3.1 Muuttujien kuvailu

Aiemmin esitellyn kirjallisuuden pohjalta selittäjämuuttujaa, eli big datan hyödyntämistä, on perusteltua mitata käyttämällä yrityksessä työskentelevien data-analytiikkojen määrää. Muuttujavalintaa voidaan perustella myös sillä, että big datan hyödyntämisen mittaaminen muilla keinoilla olisi tämän tutkimuksen laajuuden ulkopuolella.

Jotta tutkimus ottaisi huomioon mahdollisimman hyvin vain selittäjämuuttujan vaikutuksen, on tarpeen ottaa mukaan myös kontrollimuuttujia, jotka voivat selittää pois yleisiä selitettävään muuttujaan vaikuttavia tekijöitä. Tässä tutkimuksessa kontrollimuuttujien osalta osakseen nojataan Littlen ja Coffeen (2000) malliin brändiarvon mittaamisesta. Tutkijat löysivät merkittävän korrelaation price-to-book -tunnusluvun sekä yrityksen koon, kasvun, ja pääomaintensiteetin välillä. Korrelaatiota P/B-luvun ja kontrollimuuttujien välillä myötäilevät myös osakseen Faman ja Frenchin (1992) tutkimus, jonka mukaan suuremmilla yrityksillä on korkeammat P/B-luvut.

Yrityksen koko on yleinen toiminnan laajuuden indikaattori. Laajuutta voidaan mitata monella eri tavalla, mutta useimmin käytettyjä mittareita ovat liikevaihto, taseen loppusumma ja henkilöstömäärä. Yleensä koon mittarit kuitenkin korreloivat voimakkaasti

toistensa kanssa, joten tässä tutkimuksessa on perusteltua käyttää vain yhtä koon mittaria. (Ikäheimo, Malmi & Walden, 2016) Tutkielmaan koon mittariksi valitaan liikevaihto, sillä se on kohtuullisen yleismaailmallinen ja helposti löydettävissä oleva taloudellinen arvo. Esimerkiksi henkilöstömäärää ei perustellusti käytetä, sillä siihen vaikuttavat monet muutkin tekijät, kuten osa-aikaiset työntekijät tai henkilöstöpalveluiden käyttö.

Yrityksen kasvu voidaan määritellä toiminnan laajenemiseksi, ja käytännössä se ilmenee erilaisten toimintaa kuvaavien tunnuslukujen kasvuna. Yleensä kasvun mittareina käytetään samoja mittareita kuin kokoon, mutta ne suhteutetaan toisiinsa yli ajan. Kasvua kuvaava tunnusluku kasvu-% voidaan kuvata seuraavalla tavalla (Ikäheimo et al., 2016):

$$Kasvu - \% = \frac{koko_t - koko_{t-1}}{koko_{t-1}} * 100 \quad (2)$$

Tutkielmassa käytetään edellisiin perusteisiin nojaten liikevaihtoa mittaamaan kasvua siten, että vuoden 2012 ja 2013 liikevaihdot suhteutetaan toisiinsa mittaamaan vuoden 2013 kasvua ja niin edelleen.

Kolmantena kontrollimuuttujana käytetään yrityksen pääomaintensiteettiä. Tätä tunnuslukua käytetään mittaamaan, kuinka hyvin yritys pystyy hyödyntämään pääomaansa. Käytännössä pääomaintensiteetti indikoi kuinka paljon pääomaa tulee sijoittaa, jotta saadaan yhden rahayksikön tuotot. Tämä tunnusluku voidaan laskea täten jakamalla yrityksen taseen loppusumma liikevaihdolla seuraavalla kaavalla (Malik & Shaheen, 2012):

$$Pääomaintensiteetti = \frac{taseen\ loppusumma}{liikevaihto} \quad (3)$$

3.2 Paneelidata

Paneelidata on aineistoa, joka koostuu erilaisista poikkileikkauksellisista ryhmistä, joita mitataan yli ajan. Paneelidatassa siis yhdistyvät sekä poikkileikkaus-, että aikasarja-aineiston ominaisuudet. On olemassa kolmentyyppisiä paneelidata-aineistoja: Pitkiä ja

kapeita, lyhyitä ja leveitä, sekä pitkiä ja leveitä. Tässä tutkimuksessa aineisto on lyhyttä ja leveää, eli erilaisia ryhmiä on monta, mutta aikaulottuvuus on kapea. Lisäksi aineisto voi olla tasapainoista tai epätasapainoista. Tasapainoisessa aineistossa on sama määrä ajanjaksoja jokaiselle tarkasteltavalle yksikölle. (Hill, Griffiths & Lim, 2018, 9, 634-636) Tässä tapauksessa aineisto on epätasapainoinen, sillä havaintoja on eri määrä yksiköiden kesken.

Paneelidata mahdollistaa yksikkökohtaisen heterogeenisyyden kontrolloinnin ja muutoksen tarkastelun yli ajan. Lisäksi vapausaste on korkeampi ja selitettävien muuttujien välillä on vähemmän kollineaarisuutta. Paneelidatan antamat tulokset ovat siis informatiivisimpia ja luotettavampia verrattuna pelkkään aikasarja- tai poikkileikkausdataan. (Hill et al., 2018, 635) Tässä tutkimuksessa paneelidatasta on hyötyä juuri sen takia, että yritysten väliset erot tulevat huomioitua paremmin, kun otetaan aikaulottuvuus mukaan.

3.2.1 Analysointimenetelmät

Paneelidatan analysointiin on olemassa erilaisia malleja. Ensimmäisenä mallina kuvataan Pooled OLS (*pooled ordinary least squares*). Tämä malli ei huomioi aineiston paneelimaisuutta, vaan kohtelee muuttujia niin kuin tavallisessa usean selittävän muuttujan regressioanalyysissa. Tämän takia mallia pidetään usein epäkäytännöllisenä paneelidatan analysoinnissa. Mallia voidaan kuitenkin käyttää, jos aineiston yksiköiden välillä ei ole yksilöllisiä eroja. Pooled OLS -mallin kaava voidaan esittää muodossa (Koop, 2008, 256):

$$Y_{it} = \propto + \beta x_{it} + \varepsilon_{it} \tag{4}$$

Paneelidataa analysoitaessa käytetäänkin useammin joko kiinteiden vaikutusten mallia (FE) tai satunnaisten vaikutusten mallia (RE), sillä ne huomioivat aineiston yksikkökohtaiset erot (Koop, 2008, 255). FE-mallilla on kaksi eri mallinnustapaa riippuen yksiköiden määrästä. Kun yksiköitä on hyvin pieni määrä, eroja voidaan selittää dummy muuttuja -mallia hyödyntäen. Tässä tutkimuksessa tapa osoittautuu kuitenkin vaja-

vaiseksi, sillä yksiköitä on enemmän. Suuremman aineiston analysointiin sopii paremmin kiinteiden vaikutusten estimaattori, jossa käytetään yli ajan otettuja yksikkökohtaisia keskiarvoja. Näin ollen muuttujat ovat lukujen poikkeamia keskiarvosta. Oletuksena mallissa on, että erot yksiköiden välillä kuuluvat vakiotermiin. Kiinteiden vaikutusten menetelmä hyödyntää pelkästään yksiköiden sisäistä vaihtelua ja sitä kutsutaankin usein *within* -estimaattoriksi. Vakiotermi β_{1i} on ainoa termi, joka vaihtelee yksiköiden välillä. Tämän estimointimenetelmän kaava voidaan esittää muodossa (Hill et al., 2018, 642):

$$Y_{it} = \beta_{1i} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \varepsilon_{it}, \quad (5)$$

jossa ε_{it} riippumattomia, $E(\varepsilon_{it}) = 0$, $var(\varepsilon_{it}) = \sigma_e^2$

Estimaattori eliminoi endogeenisuusongelman, joka johtuu selittävien muuttujien ja mallin residuaalin korreloituneisuudesta. Mallin heikkoutena pidetään kuitenkin sitä, että siihen ei voi sisällyttää muuttujia, jotka säilyvät vakioina yli ajan. (Koop, 2008, 262)

Satunnaisten vaikutusten malli eli RE eroaa kiinteistä vaikutuksista, sillä yksiköiden väliset erot arvioidaan sekä yksikköjen sisällä, että yksikköjen välillä. RE-mallin mukaan aineisto on satunnaisotos tietystä populaatiosta, ja näin ollen se kohdistaa tutkimuksen koko populaatioon. Kiinteiden vaikutusten malli taas keskittyy tutkimaan vain yksiköitä, jotka ovat aineistossa. Satunnaisten vaikutusten yhtälö voidaan kuvata seuraavalla tavalla (Hill et al. 2018, 651; Koop, 2008, 263):

$$Y_{it} = \beta_{1i} + \beta_2 x_{2it} + \beta_3 x_{3it} + v_{it} \text{ , i=1, …, N,} \quad (6)$$

jossa $v_{it} = u_i + \varepsilon_{it}$

Estimaattorissa oletetaan FE-mallin tapaan yksiköiden välisten erojen kuuluvan vakiotermiin, mutta RE-malli jakaa virhetermin yleiseen virhetermiin ε_{it}, sekä satunnaiseen yksilövaikutukseen u_i,. Avainoletuksena mallissa on, että saman yksikön virhetermit korreloivat toistensa kanssa yli ajan. (Koop, 2008, 263-264)

3.2.2 Klusterirobustit keskivirheet ja interaktiomuuttujat

Vaikka läpikäydyillä estimaattoreilla on erilaisia hyöty- ja haittapuolia, voidaan niistä yleensä havaita yleiset sarjakorrelaation ja heteroskedastisuuden ongelmat. Jos aineiston tarkasteltavien yksiköiden lukumäärä N on paljon suurempi kuin aikaulottuvuus T, nämä ongelmat ratkeavat hyödyntämällä klusteroituja robusteja keskivirheitä. Klusterirobustit keskivirheet huomioivat samoja yksikköjä koskevien havaintojen riippuvuutta. Tässä menetelmässä huomio keskittyy näiden ryhmien eli klustereiden sisäiseen eli *within* -vaihteluun. (Hill et al., 2018, 650, 662)

Regressiomallissa voidaan kontrollimuuttujien lisäksi käyttää niin sanottuja interaktiomuuttujia. Interaktiomuuttujan avulla voidaan tarkastella, onko arvolla Z vaikutusta selittävän (X) ja selitettävän (Y) muuttujan väliseen suhteeseen. Interaktiomuuttujan vaikutus voidaan nähdä joko muutoksena yhtälön kulmakertoimessa tai sen siirtymisen koordinaatistossa eli vakiotermin muuttumisena. (Hill et al., 2018, 320) Interaktion vaikutus regressiomallissa voidaan esittää seuraavalla kaavalla (Hill et al., 2018, 320):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma(x_1 x_2) + \varepsilon \tag{7}$$

Interaktiomuuttuja on siis tulomuuttuja, jossa nähdään kahden eri selittäjän yhteisvaikutus. Kun interaktiossa hyödynnetään dummy-muuttujaa, joka saa vain arvoja 0 tai 1, sen vaikutus voidaan esittää tarkemmin (Hill et al., 2018, 320):

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma(x_1 x_2) = \begin{cases} \beta_0 + (\beta_1 + \gamma)x_1 & kun\ x_2 = 1 \\ \beta_0 + \beta_1 x_1 & kun\ x_2 = 0 \end{cases} \tag{8}$$

3.3 Paneelidatan analysointi

Seuraavaksi on syytä käydä läpi, millä perusteella oikea tutkimusmenetelmä tulisi valita. Aluksi tutkitaan, onko aineistossa yksikkökohtaisia eroja. Tämä tapahtuu kiinteiden vaikutusten F-testin avulla. Nollahypoteesina on, että vakiot ovat yhtä suuria. Jos nollahypoteesi jää voimaan, tällöin on järkevämpää käyttää Pooled OLS -mallia, sillä kiinteiden vaikutusten estimaattorista ei tässä tapauksessa ole mitään hyötyä. Jos taas

nollahypoteesi hylätään, kiinteiden vaikutusten menetelmä on järkevämpi vaihtoehto. (Hill et al., 2018, 661)

Breusch-Pagan testillä testataan mallin heteroskedastisuutta eli satunnaisia eroja. Testin nollahypoteesina on, että satunnaisia eroja ei ole. Jos nollahypoteesi jää voimaan, satunnaisten vaikutusten mallin käytöstä ei ole tällöin hyötyä ja voidaan käyttää Pooled OLS -estimaattoria. Tilanteessa, jossa nollahypoteesi hylätään ja mallin ollessa konsistentti, voidaan käyttää satunnaisten vaikutusten mallia. (Hill et al., 2018, 653)

Tilanteessa, jossa sekä F-testin, että Breusch-Pagan testin nollahypoteesit hylätään, tulee vertailla keskenään FE- ja RE-estimaattoreita. Tämä tehdään yleensä Hausman -testin avulla, jossa nollahypoteesina on, että malli ei ole endogeeninen. Testi testaa siis havaitsemattoman heterogeenisuuden ja selittävien muuttujien korreloituneisuutta. Jos nollahypoteesi hyväksytään, endogeenisuutta ei ole, ja näin ollen voidaan käyttää satunnaisten vaikutusten estimaattoria. Jos taas nollahypoteesi hylätään, satunnaisten vaikutusten malli ei ole konsistentti, joten on käytettävä kiinteiden vaikutusten mallia. (Hill et al., 2018, 658-659)

Tässä tutkielmassa Hausman testi korvataan Sargan-Hansen testillä, sillä robustien keskivirheiden mallissa Hausman testiä ei ole mahdollista käyttää. Sargan-Hansen testi on käytännössä testi rajoitusten tunnistamiseksi, ja sillä voidaan vertailla Hausman testin tapaan FE- ja RE-mallia. Nollahypoteesi antaa saman lopputuleman kuin Hausman testissä: Jos nollahypoteesi hyväksytään, RE-malli on parempi estimaattori, mutta jos hypoteesi hylätään, tulisi käyttää FE-mallia. Taulukko 1 esittää yhteenvedon testeistä ja estimaattorimenetelmän valinnasta.

Taulukko 1. Estimaattorin valintaperusteet

F-testi	**Breusch-Pagan testi**	**Käytettävä malli**
H_0 jää voimaan	H_0 jää voimaan	Pooled OLS
H_0 jää voimaan	H_0 hylätään	RE-malli
H_0 hylätään	H_0 jää voimaan	FE-malli
H_0 hylätään	H_0 hylätään	Hausman/Sargan-Hansen testi: 1) H0 hyväksytään: **RE-malli** 2) H0 hylätään: **FE-malli**

3.4 Aineiston kuvailu ja muokkaus

Tutkimusaineisto perustuu Granvillen (2013; 2015) listaan yrityksistä, joilla on eniten data-analyytikko -kontakteja. Vuoden 2013 lista on julkaistu joulukuussa 2013 ja siinä on listattu 6000 yritystä, jotka palkkaavat eniten data-analyytikkoja. Vuoden 2015 lista on julkaistu tammikuussa 2015 ja se kattaa 7500 yritystä. Julkaisuajankohtien perusteella voidaan olettaa, että listat kuvaavat vuosien 2013 ja 2014 tilanteita.

Selittävästä muuttujasta poiketen muut muuttujat kerätään vuodelle 2013, 2014 ja 2015, jotta tutkimusaineisto olisi laajempi. Yritysten markkina-arvot kerätään osakseen vuoden 2013 ja 2014 Forbes Global 2000 -listalta ja osakseen manuaalisesti Macrotrends -nimiseltä sivustolta. Forbesin vuoden 2014 markkina-arvot on kerätty 1. huhtikuuta 2014 pörssitilanteesta (Murphy, 2014) ja vuoden 2013 listan markkina-arvot ovat kerätty 15. maaliskuuta 2013 tilanteesta (DeCarlo, 2013). Yritykset, joille markkina-arvoja ei löytynyt Forbesin listalta, kerätään Macrotrends -sivustolta mahdollisimman läheltä markkinatilannetta 15.3.2013, 1.4.2014 ja 1.4.2015. P/B-luvut, liikevaihdot, sekä taseen loppusummat kerätään myös samalta sivustolta mahdollisimman läheltä näitä markkinatilanteita.

Listasta karsitaan alustavasti yritykset, jotka eivät ole pörssissä noteerattuja, jotta etsittävät selitettävät ja kontrollimuuttujat olisivat mahdollisimman valideja ja helppoja löytää. Myös jo valittujen yhtiöiden tytäryhtiöt jätetään pois konsernitason päällekkäisyyksien välttämiseksi. Tämän lisäksi yritykset jaotellaan sekä IT-sektoriin, että ei-IT-sektoriin. Monien IT-alan yritysten arvo pörssissä on viime vuosina ollut korkea, sillä itse IT-sektorille on odotettavissa paljon kasvua digitalisaation ja teknologisen vallankumouksen ansiosta. Sektorikohtainen jaottelu auttaa kontrolloimaan itse toimialan kasvun odotuksia. Myöhemmin tätä jaottelua hyödynnetään myös osana interaktiomuuttujaa. Jaottelu tapahtuu manuaalisesti tutkimalla erinäisiä julkisia lähteitä, kuten yritysten nettisivuja. Kaiken kaikkiaan aineiston keräys tapahtuu datan saatavuuden rajoissa, ja aineisto on epätasapainoinen. Tutkimukseen valikoituu listasta kaiken kaikkiaan 121 yritystä, jotka voidaan nähdä liitteestä 1.

Taulukko 2. Tutkimuksessa käytettävien muuttujien perustiedot

Muuttuja	N	Keskiarvo	Keskihajonta	Minimi	Maksimi	Yksikkö
MA	361	57.89	79.40	0.07	714.09	miljardi dollaria
PB	360	5.36	9.68	-50.54	85.83	suhdeluku
data	242	7.91	14.89	1	110	kpl
liikevaihto	361	33.51	54.83	0.1	473.44	miljardi dollaria
po_int	361	4.43	7.37	0.29	44.60	suhdeluku
kasvu	356	8.74	27.74	-50.88	282.97	prosentti
IT	363	0.33	0.47	0	1	kategorinen (IT=1 muut=0)

Markkina-arvo (MA) ja liikevaihto on esitetty miljardeina dollareina. P/B-luku (PB) ja pääomaintensiteetti (po_int) ovat suhdelukuja, kun taas kasvuprosentti (kasvu) on esitetty prosenttilukuna. Data-analyytikot (data) ovat henkilömääräisiä ja IT-sektori (IT) on kategorisena muuttujana IT-sektori = 1 ja ei-IT-sektori = 0.

Muuttujien jakaumista (liite 2) voidaan huomata, että muuttujat eivät noudata normaalijakaumaa. Tämä tarkoittaa, että monet niistä ovat vinoja ja huipukkaita. Muuttujia voitaisiin saada normaalijakautuneemmaksi erilaisten muuttujamuunnosten, kuten logaritmien ja neliöjuurimuunnosten avulla. Muuttujamuunnokset kuitenkin vaikeuttavat tulsoten tulkintaa ja voivat vaikuttaa negatiivisesti havaintojen lukumäärään. Lineaarisen regressiossa oletuksena ei ole muuttujien normaalijakautuneisuus, vaan mallin virhetermien eli residuaalien normaalijakautuneisuus. Jos malli sisältää poikkeavia havaintoja, ja residuaali on kovin vino tai huipukas, se voi vääristää analyysin tuloksia (Melin, 2006, 339). Tässä tilanteessa muuttujamuunnoksia ei vielä tehdä pelkästään muuttujien jakauman vinouden takia, sillä se poistaisi osan aineiston havainnoista niiden negatiivisen arvon vuoksi. Residuaaleja tarkastellaan itse mallien valinnan jälkeen.

Viivästetyillä arvoilla voidaan vähentää malleissa esiintyvää autokorrelaatiota. Autokorrelaatio kuvaa aikasarjan havaintojen välistä riippuvuutta ja sitä, että aikasarjan havainnot eivät ole satunnaisia. Esimerkiksi tässä tutkimuksessa taloudelliset arvot, kuten markkina-arvo riippuvat luonnollisesti myös edellisten vuosien arvoista. Yleensä

myös taloudellisia muuttujia hyödyntävissä aikasarjatutkimuksissa on havaittavissa tietynlainen muuttujan kehitystaso yli ajan. Tätä kehitystä kutsutaan trendiksi, ja esimerkiksi hintatasoissa, teollisessa tuotannossa, kulutuksessa ja osakemarkkinaindekseissä on havaittavissa trendikasvua. Viivästetyt arvot voivat näin ollen kontrolloida myös trendin vaikutuksia. (Koop, 2008, 174, 178) Tässä tutkimuksessa hyödynnetään yhden vuoden viivästettyä arvoa molemmista selitettävistä muuttujista, ja ne sisällytetään erikseen molempiin malleihin kontrollimuuttujina.

Monesti myös selittävän muuttujan vaikutus selitettävään muuttujaan näkyy vasta tietyn ajan kuluttua (Koop, 2008, 174). Tässä tapauksessa esimerkiksi data-analyytikon palkkaaminen ja tätä kautta yrityksen BDA:n hyödyntäminen ei välttämättä heti näy yrityksen suoriutumisessa. Vaikutus etenkin markkina-arvoon ja tätä kautta aineettomaan pääomaan, kuten brändipääomaan ei oletuksena voi näkyä heti. Tässä tutkimuksessa on kiinnostavaa vertailla data-analyytikoiden vaikutuksen merkittävyyttä, jos muuttujaa viivästytetään vuodella. Taulukossa 3 on esitetty sekä selitettäville muuttujille, että selittävälle muuttujalle data viivästetyt arvot.

Taulukko 3. Viivästettyjen muuttujien perustiedot

Muuttuja	N	Keskiarvo	Keskihajonta	Minimi	Maksimi
MAlag	242	54.08	72.09	0.07	456.81
PBlag	242	4.92	7.60	-3.21	85.83
datalag	242	7.91	14.89	1	110

4. TULOKSET

Tässä osiossa käydään läpi estimointimenetelmän valinnan tulokset. Luvussa esitellään myös itse regressiomallien tulokset ja sen jälkeen paranneltujen mallien tulokset. Lopuksi arvioidaan tutkimuksen reliabiliteettia.

4.1 Korrelaation tarkastelu ja estimointimenetelmän valinta

Seuraavaksi raportoidaan valitut estimointimenetelmät perusteluineen. Tutkimuksessa on tarkoitus tarkastella markkina-arvoa ja P/B-lukua erikseen omissa malleissaan. Näin ollen molempiin malleihin sisällytetään selittäväksi muuttujaksi data-analyytikoiden määrä, sekä aiemmin perustellut kontrollimuuttujat. Kahden selitettävän muuttujan hyödyntäminen lisää tutkimuksen monipuolisuutta ja laajentaa tutkimustuloksia.

Ennen varsinaisten analyysimenetelmien valintaa ja tulosten raportointia tarkastellaan hieman muuttujien korrelaatiota. Regressioanalyysissa selitettävän ja selittävän muuttujan korrelaatio on positiivinen asia, mutta toistensa kanssa korreloituneet selittävät muuttujat saattavat heikentää tutkimuksen tulosten tarkkuutta. Liitteen 3 hajontakuvioiden avulla voidaan tehdä alustavaa tarkastelua selitettävien ja selittäjämuuttujan välillä. Kirjallisuuskatsauksen hypoteesista poiketen suoraa korrelaatiota ei ole silmämääräisesti havaittavissa. Tämä voi johtua osakseen kuitenkin rajallisesta aineistosta ja outliereista. Korrelaatiota on syytä tutkia tarkemmin hyödyntäen regressioanalyysia.

Korrelaatiomatriisi kuvaa selittävien muuttujien korreloituneisuutta. Suuri korrelaatio eli multikollineaarisuus vaikeuttaa yksittäisten selittäjien vaikutuksen erottamista. Jos selittävien muuttujien korrelaatio nousee yli 0.9, voi kyseessä olla multikollineaarisuusongelma. (KvantiMOTV, 2003) Taulukosta 4 huomataan, että selittävillä muuttujilla ei ole havaittavissa keskenään multikollineaarisuutta. Perusteltua on siis olla poistamatta mallista mitään muuttujia. Suurin korrelaatio on muuttujien data ja IT välillä. Tämä tarkoittaa, että IT-alan yritykset todennäköisemmin palkkaavat enemmän data-analyytikoita, mikä oli odotettavissakin. Korrelaatio on kuitenkin vain 20%, joten itse tutkimuksessa se ei ole haittatekijä.

Taulukko 4. Selittävien muuttujien korrelaatiomatriisi

	data	liikevaihto	po_int	kasvu	IT
data	1				
liikevaihto	0.1856	1			
po_int	0.001	0.0663	1		
kasvu	-0.0204	-0.1248	-0.1186	1	
IT	0.2316	-0.2194	-0.2608	0.1651	1

Klusterirobustien keskivirheiden käyttö vaikuttaa tehtäviin estimaattorin valintatesteihin. StataSE 16 -ohjelmisto ei pysty laskemaan kiinteiden vaikutusten F-testiä robusteille keskivirheille sen monimutkaisuuden takia. Näin ollen F-testi testataan käyttämällä normaaleja keskivirheitä, ja sen tulokset laajennetaan koskemaan myös robustien keskivirheiden mallia. Myös Hausman testi joudutaan korvaamaan käyttämällä Sargan-Hansen testiä, kuten aiemmin mainittu.

Taulukko 5 kuvaa yhteenvedon tutkimusmenetelmien valinnasta. Aluksi tutkitaan markkina-arvoa selitettävänä muuttujana. F-testi hylätään viiden prosentin riskiarvolla, joten mallissa on kiinteitä vaikutuksia. Breusch-Pagan testin nollahypoteesi hyväksytään kyseisellä riskitasolla, joten mallissa ei ole satunnaisia vaikutuksia. Näin ollen satunnaisten vaikutusten estimaattorin hyödyntäminen ei ole perusteltua. Myös Sargan-Hansen testi tukee kiinteiden vaikutusten mallin käyttämistä. P-arvo ollessa alle 0.0001 nollahypoteesi hylätään kaikilla yleisesti käytetyillä riskitasoilla. Mallissa esiintyy siis endogeenisuutta, jonka kiinteiden vaikutusten estimaattori ottaa huomioon.

P/B-luku selitettävänä muuttujana on seuraavaksi tutkimusmenetelmän valinnan testien kohteena. F-testin nollahypoteesi jää voimaan viiden prosentin riskitasolla p-arvon ollessa 0.0543. Näin ollen mallin vakiot ovat yhtä suuria, joten tässä kohtaa on järkevämpää käyttää Pooled OLS -mallia kuin kiinteiden vaikutusten mallia. Myös Breusch-Pagan testin nollahypoteesi hyväksytään kaikilla yleisesti käytetyillä riskitasoilla. Mallissa ei siis myöskään ole satunnaisia vaikutuksia, joten satunnaisten vaikutusten estimaattorin käyttö ei ole perusteltua. Näin ollen parhaat ja konsisteimmat tulokset antava malli on Pooled OLS, joka toimii tavallisen lineaarisen regression tavoin. Tämä

tulee yllätyksenä, sillä oletuksena olisi, että yritysten välillä olisi havaittavissa yksikkökohtaisia eroja. Aineiston huipukkuudella, ajallisesti pienellä koolla ja klusterirobustien keskivirheiden käytöllä toisaalta saattaa olla jonkunlainen vaikutus lopputulokseen.

Taulukko 5. Estimointimenetelmän valintatestit

	MA	PB
F-testi	H_1 (0.0000)	H_0 (0.0543)
Breusch-Pagan LM	H_0 (0.3002)	H_0 (0.3404)
Hausman / Sargan-Hansen	H_1 (0.0000)	H_1 (0.0000)
Valittu estimaattori	FE-malli	Pooled OLS

4.2 Mallien tulokset

Ennen valittujen mallien raportointia tarkastellaan mallien residuaalien normaalijakautuneisuutta. Liitteen 4 jakaumakuviosta huomataan, että mallien residuaalit eivät kumpikaan noudata normaalijakaumaa. Jotta mallin estimaatit olisivat harhattomia, tulisi residuaalien olla normaalijakautuneita. Selitettävistä muuttujista otetaan logaritmit, jonka jälkeen mallien residuaalien jakaumakuviot (liite 5) ovat tasaisemmat. Taulukosta 6 voidaan tarkastella paranneltujen muuttujien tunnuslukuja. Huomataan, että muuttujan lgPB havaintojen määrä väheni. Negatiiviset arvot pudotettiin aineistosta, sillä niistä ei ole mahdollista ottaa logaritmia. Arvoja tippui kuitenkin sen verran vähän, että uutta havaintomäärää voidaan käyttää tutkimuksessa. Täten selitettävät muuttujat korvataan molemmissa malleissa niiden logaritmimuunnoksilla.

Taulukko 6. Logaritmisten muuttujien perustiedot

Muuttuja	N	Keskiarvo	Keskihajonta	Minimi	Maksimi
lgMA	361	3.10	1.65	-2.66	6.57
lgPB	354	1.13	1.02	-0.80	4.45

Taulukko 7 esittää markkina-arvon kiinteiden vaikutusten mallin parametriestimaatit, keskivirheet ja niiden tilastolliset merkitsevyydet. Kiinteiden vaikutusten mallista joudutaan jättämään toimialavertailu pois, sillä se on yli ajan muuttumaton kategorinen muuttuja. Koko malli on tilastollisesti merkitsevä yhden prosentin riskitasolla, sillä Prob > F

= 0.002. Muuttuja data on tilastollisesti merkitsevä kymmenen prosentin riskitasolla. Selitettävän muuttujan ollessa logaritmimuunnos, selityskertoimien tulkinta tapahtuu osittaisjoustona. Jos selittävä muuttuja kasvaa yhdellä yksiköllä, kasvaa selitettävä muuttuja selityskertoimella kerrottuna sadalla prosentilla. Näin ollen datan kasvu yhdellä yksiköllä kasvattaa markkina-arvoa 100 x 0.0071 eli 0.71%. Vaikutus on siis mallin mukaan olemassa, ja ottaen huomioon muuttujien yksiköt, myös yllättävän suuri. Jo yhden data-analyytikon palkkaaminen lisää kasvattaa yrityksen markkina-arvoa 0.71 prosenttia, joten tämän regression perusteella data-analyytikoiden palkkaaminen yritykseen olisi erittäin perusteltua. Markkina-arvon ollessa esimerkiksi 1 miljardi dollaria (1 mrd), data-analyytikon palkkaaminen kasvattaa markkina-arvoa 1 mrd dollaria x 0.0071 = 7.1 miljoonaa dollaria. Viiden prosentin riskitasolla tilastollisesti merkitseviä muuttujia ovat kontrollit liikevaihto ja vuodella viivästetty arvo markkina-arvosta. Markkina-arvon viivästettyä arvoa voidaan pitää tärkeänä mallissa, sillä se kontrolloi mallin autokorrelaatiota. Molempien muuttujien vaikutukset ovat positiivisia. Liikevaihdon yhden yksikön nousu kasvattaa markkina-arvoa 0.3%, kun taas edellisen vuoden markkina-arvolla on 0.2% vaikutus. Kumpikaan merkittävistä kontrolleista ei kuitenkaan yletä muuttujan data vaikutuksen tasolle, kun otetaan huomioon, että muuttuja data on henkilömääräinen ja kontrollit ovat esitetty miljardeissa dollareissa. Pääomaintensiteetti ja kasvuprosentti eivät kumpikaan ole tilastollisesti merkitseviä, mikä on ristiriidassa Littlen ja Coffeen (2000) mallin kanssa. Tutkijoiden mallissa käytettiin kuitenkin P/B-lukua markkina-arvon sijasta, joka voi selittää tämän ristiriidan. Lisäksi robustien keskivirheiden käyttö myös nostaa p-arvoa, eli laskee tilastollista merkitsevyyttä.

Mallin kokonaisselitysaste on 47%, eli selittävät muuttujat selittävät noin puolet mallin kokonaisvaihtelusta. Tässä tapauksessa selitysaste jää melko matalaksi, vaikka kontrollimuuttujat on valittu aiemman validin tutkimuksen perusteella. Matalaa selityskerrointa voi selittää kontrollimuuttujien pieni lukumäärä, sillä markkina-arvoon vaikuttaa varmasti lukemattoman moni tekijä. Myös selittäjien pienet selitysasteet voivat alentaa mallin kokonaista selityskerrointa.

Taulukko 7. Markkina-arvo selitettävänä muuttujana FE-mallissa

Muuttuja	Kerroin	Robust keskivirhe	P-arvo
data	0.0071	0.0038	0.062
MAlag	0.0020	0.0010	0.050
liikevaihto	0.0031	0.0015	0.043
po_int	0.0229	0.0291	0.433
kasvu	0.0020	0.0018	0.275
_cons	2.8073	0.1396	<0.0001

R-squared	
within	0.14
between	0.48
overall	0.47

Vaikka estimaattorien valintatestit puolsivatkin FE-mallia, on kiinnostavaa tutkia myös antavatko muut mallit samankaltaisia tuloksia. Pooled OLS -mallissa data muuttuja ei ole tilastollisesti merkitsevä, mutta malli on harhainen, sillä F-testin nollahypoteesi hylättiin. RE-malli taas antaa melko saman selityskertoimen viiden prosentin riskitasolla, kuin FE-malli. RE-malli ei tosin ota huomioon mallissa esiintyvää endogeenisuutta, joten näihinkin tuloksiin tulee suhtautua kriittisesti.

Taulukossa 8 esitetään Pooled OLS -mallin tulokset, jossa P/B-luku on selitettävänä muuttujana. Mallissa 1 on otettu kontrollien lisäksi mukaan sekä data, että yhden vuoden viivästetty arvo datalag. Kumpikaan muuttuja ei kuitenkaan ole tilastollisesti merkitsevä, joten niiden selitysastetta ei voida arvioida. Nämä muuttujat, ollen suoraan johdannaisia toisistaan, ovat kuitenkin vahvasti korreloituneita toistensa kanssa, mikä voi vaikuttaa malliin tehden siitä harhaisen. Yhden prosentin riskitasolla tilastollisesti merkitseviä ovat kontrollimuuttujat liikevaihto ja po_int. Kun liikevaihto kasvaa yhden yksikön, se mallin mukaan vähentää P/B-lukua 0.28%. Tämä on hieman yllättävä tulos, sillä oletuksena olisi, että liikevaihdon kasvulla ei olisi vaikutusta P/B-luvun vähenemiseen. Voidaan tässä kohtaa olettaa, että aineiston pienellä koolla ja laajuudella on osansa tutkimustulosten merkittävän suuriin arvoihin. Toisaalta vaikutuksen voi osaksi selittää myös P/B-tunnusluvun kaava, jossa suuri liikevaihto voi kasvattaa loppujen lopuksi yrityksen tulosta, joka lisätään taseeseen, eli kaavassa olevaan jakajaan. Pääomaintensiteetillä on negatiivinen vaikutus P/B-lukuun, mutta se voi osakseen olla jälleen selitettävissä tunnuslukujen kaavojen avulla. Oman pääoman nouseva arvo vaikuttaa osittain myös taseen nousevaan arvoon, ja kun oma pääoma jaettavana kasvaa, kasvaa myös taseen arvo jakajana. Näin ollen arvojen kasvun vaikutukset ovat

päinvastaiset. Viiden prosentin riskitasolla merkitseväksi nousee myös viivästetty arvo selitettävästä muuttujasta. Mallin kokonaisselityskerroin on 52%, joten voimme sanoa, että selittävät muuttujat selittävät hieman yli puolet P/B-luvun vaihtelusta.

Taulukko 8. P/B-luku selitettävänä muuttujana Pooled OLS-mallissa

Malli 1	R-squared	0.52	
Muuttuja	Kerroin	Robust keskivirhe	P-arvo
data	-0.0028	0.0112	0.805
datalag	0.0114	0.0121	0.344
PBlag	0.0439	0.0211	0.04
liikevaihto	-0.0028	0.0008	0.001
kasvu	0.6192	0.4281	0.151
po_int	-0.049	0.0071	<0.001
1.IT	0.0021	0.1396	0.988
_cons	0.5193	0.4253	0.225

Jotta estimaattorin avulla saataisiin tilastollisesti merkitseviä tuloksia data muuttujasta, testataan mallia vielä käyttämällä muuttujia data ja datalag erikseen. Näin saadaan selville, onko pelkällä normaalilla tai viivästetyllä arvolla tilastollisesti merkitsevää vaikutusta. Taulukossa 9 on esitelty parametriestimaattien tulokset. Mallissa 2 käytetään muuttujaa data ja mallissa 3 viivästettyä arvoa muuttujasta. Molemmissa malleissa muuttujat nousevat tilastollisesti merkitseviksi: muuttuja data 5% riskitasolla ja datalag 1% riskitasolla. Mallissa 2 datan yhden yksikön nousu nostaa P/B-lukua 0.76% ja mallissa 3 0.72%. Voidaan siis sanoa, kuten markkina-arvon kohdalla, vaikutusten olevan taas merkitsevän suuria. Vaikutusten erot ovat kuitenkin hyvin pieniä, ja oletus, jossa viivästetyllä arvolla olisi mahdollisesti suurempi vaikutus markkina-arvoon, ei pidä tämän estimaation mukaan paikkaansa. Mallin 2 ja 3 kokonaisselityskertoimet ovat hyvin samaa luokkaa mallin 1 kanssa, eli ne selittävät kohtuullisen hyvin selitettävän muuttujan vaihtelua. Malleissa liikevaihto ja pääomaintensiteetti nousevat tilastollisesti merkitseviksi yhden prosentin riskitasolla. Molemmilla on samankaltainen vaikutus, kuten mallissa 1: liikevaihto vaikuttaa -0.3% ja pääomaintensiteetti noin -5%. Muuttuja IT ei

sen sijaan ole noussut missään mallissa tilastollisesti merkittäväksi tekijäksi, joten sen toimivuuden kontrollina näissä malleissa voi kyseenalaistaa.

Taulukko 9. Pooled OLS, jossa data ja datalag erikseen

Malli 2	R-squared	0.52	
Muuttuja	Kerroin	Robust keskivirhe	P-arvo
data	0.0076	0.0032	0.019
PBlag	0.0442	0.0211	0.038
liikevaihto	-0.0029	0.0008	0.001
kasvu	0.618	0.4194	0.143
po_int	-0.0491	0.0072	<0.001
1.IT	-0.013	0.1371	0.924
_cons	0.5128	0.4174	0.222

Malli 3	R-squared	0.54	
Muuttuja	Kerroin	Robust keskivirhe	P-arvo
datalag	0.0072	0.0024	0.003
PBlag	0.0641	0.0294	0.03
liikevaihto	-0.0031	0.0008	<0.001
kasvu	0.3779	0.2139	0.079
po_int	-0.0548	0.007	<0.001
1.IT	-0.0569	0.0937	0.545
_cons	0.7937	0.2237	<0.001

Vaikka Pooled OLS valittiinkin testien perusteella oikeaksi menetelmäksi, mallit ajetaan vielä käyttäen estimaattoreita, jotka ottavat aineiston paneelimaisuuden huomioon. Käytetään sekä kiinteiden vaikutusten mallia, että satunnaisten vaikutusten mallia, jotta voitaisiin huomioida mahdollisia yritysten yksikkökohtaisia eroja. FE-mallissa arvo data ei ole tilastollisesti merkitsevä ja koko mallin selityskertoimeksi jää vain 5%. Tämä voi kertoa aineiston mahdollisesta vajavaisuudesta, sillä FE-mallin tulokset eivät ole konsistentteja Pooled OLS mallin kanssa. RE-mallissa data nousee merkitseväksi selittäjäksi viiden prosentin riskitasolla. Malli antaa samankaltaiset tulokset Pooled OLS-malliin verrattuna; datan nousu vaikuttaa positiivisesti 0.78% P/B-lukuun.

4.3 Parannellut mallit ja tulokset

Seuraavaksi pyritään parantamaan mallien selitysastetta sekä luotettavuutta muokkaamalla malleja. Paranneltujen mallien avulla voidaan tarkastella pysyvätkö tulokset merkittävän suurina, ja näin arvioida aikaisempien tulosten luotettavuutta. Aikaisemmin IT-kontrollimuuttuja ei noussut tilastollisesti merkitseväksi missään mallissa. Nyt muuttujan roolia muutetaan hyödyntämällä sitä interaktiomuuttujana. Näin ollen malliin otetaan mukaan muuttujien data ja IT välinen interaktio. Datan ja IT:n interaktio ottaa huomioon oletuksen, että IT-alan yritykset palkkaavat suhteessa enemmän data-analyytikoita, kuin muiden alojen yritykset. Mallien luotettavuutta parannetaan myös ottamalla huomioon ajan vaikutus paremmin. Malleihin otetaan mukaan kategorinen vuosimuuttuja t, joka kontrolloi ajan vaikutusta. Vuosimuuttujaa on hyvä käyttää kontrollina etenkin absoluuttisia lukuja tarkastellessa, tässä tapauksessa markkina-arvon kohdalla.

Paranneltujen mallien estimointimenetelmien valintaperusteet voidaan nähdä taulukosta 10. Sopivaksi estimaattoriksi markkina-arvon mallille valikoituu jälleen kiinteiden vaikutusten malli. Jokaisen testin nollahypoteesi hylätään kaikilla yleisesti käytetyillä riskitasoilla p-arvon ollessa alle 0.0001. Testien perusteella tällä kertaa myös P/B-luvun malleista valitaan kiinteiden vaikutusten malli, ja tässäkin tapauksessa testien nollahypoteesit hylätään kaikilla yleisesti käytetyillä riskitasoilla. Tämä saattaa viestiä mallin parantumisesta, sillä kiinteiden vaikutusten menetelmää voidaan pitää Pooled OLS -menetelmää todennäköisempänä tämänkaltaisessa paneelidatassa. Liitteitä 6 ja 7 vertailemalla voidaan huomata, että molemmissa malleissa logaritminen muoto selitettävästä muuttujasta parantaa jälleen residuaalien normaalijakautuneisuutta, joten malleissa käytetään muuttujia lgMA ja lgPB. Tästä johtuen parametriestimaatteja tulkitaan jälleen osittaisjoustoina.

Taulukko 10. Paranneltujen mallien estimointimenetelmän valintatestit

	MA	PB
F-testi	H_1 (<0.0001)	H_1 (<0.0001)
Breusch-Pagan LM	H_1 (<0.0001)	H_1 (<0.0001)
Hausman / Sargan-Hansen	H_1 (<0.0001)	H_1 (<0.0001)
Valittu estimaattori	FE-malli	FE-malli

Taulukossa 11 esitellään tulokset, jossa markkina-arvo toimii selitettävänä muuttujana. Parannellussa mallissa data ei ole tilastollisesti merkitsevä, joka on ristiriidassa aikaisempien tulosten kanssa. Tämä vahvistaa sitä käsitystä, että aikaisemmat mallit saattavat olla epäluotettavia. Sekä interaktiomuuttuja, että vuosidummy nousevat tilastollisesti merkitseviksi viiden prosentin riskitasolla. Molemmilla on positiivinen vaikutus markkina-arvoon. Yhden vuoden ajallinen vaikutus nostaa markkina-arvoa 100 x 0.096 eli 9.6%. Interaktiomuuttujan vaikutus nostaa markkina-arvoa 1.6%. Muista kontrolleista ainoastaan liikevaihto nousee tilastollisesti merkitseväksi 10% riskitasolla. Vaikutuksen suuruus on sama, kuin aikaisemmassa mallissa (0.3%). Aikaisempaan malliin verratessa täytyy kuitenkin huomioida, että kokonaisselityskerroin jää parannellussa mallissa jopa 20 yksikköä pienemmäksi (0.27%). Tämä on myös yksi tekijä, joka vahvistaa käsitystä aineiston puutteellisuudesta ja kannustaa suhtautumaan kriittisesti tutkimuksen tuloksiin.

Taulukko 11. Paranneltu malli, markkina-arvo selitettävänä muuttujana FE-mallissa

Muuttuja	Kerroin	Robust keskivirhe	P-arvo
data	-0.0115	0.0078	0.143
MAlag	0.0005	0.0012	0.663
liikevaihto	0.0032	0.0018	0.083
po_int	0.0335	0.0306	0.277
kasvu	0.0019	0.0019	0.329
t(2015)	0.0961	0.0318	0.003
IT#c.data(1)	0.0163	0.0078	0.038
_cons	2.8713	0.1448	<0.001

R-squared	
within	0.23
between	0.28
overall	0.27

Taulukosta 12 nähdään yhteenveto P/B-luvusta selitettävänä muuttujana parannellussa mallissa. Voidaan huomata, että 10% riskitasolla datan selityskerroin on negatiivinen, mikä on ristiriidassa aikaisempien mallien kanssa. Täytyy kuitenkin ottaa huomioon myös mallin erityisen alhainen kokonaisselitysaste, joka on vain 2%. Myös tämän mallin tuloksiin tulee suhtautua kriittisesti. Tällä kertaa uusista kontrolleista tilastollisesti merkitseväksi nousee ainoastaan vuosidummy. Vuoden ajallinen lisäys kasvattaa P/B-lukua jopa 15.9% Muita tilastollisesti merkitseväksi nousevia selittäjiä ovat 5% riskitasolla PBlag ja 10% riskitasolla liikevaihto. Eroten edellisistä P/B-luvun estimaattoreista, tässä mallissa liikevaihdon vaikutus on positiivinen ja samaa luokkaa kuin malleissa, jossa markkina-arvo toimii selitettävänä muuttujana.

Taulukko 12. Paranneltu malli, P/B-luku selitettävänä muuttujana FE-mallissa

Muuttuja	Kerroin	Robust keskivirhe	P-arvo
data	-0.0185	0.0099	0.065
PBlag	0.0105	0.0046	0.023
liikevaihto	0.0045	0.0027	0.094
po_int	-0.0430	0.0434	0.323
kasvu	0.0008	0.0016	0.599
vuosi(2015)	0.1588	0.0462	0.001
IT#c.data(1)	0.0071	0.0093	0.447
_cons	1.3283	0.2094	<0.001

R-squared	
within	0.18
between	0.02
overall	0.02

Kaiken kaikkiaan parannelluista malleista voidaan todeta, että ne eivät antaneet tukea edellisten tutkimustulosten luotettavuuteen. Pikemminkin heikot kokonaisselityskertoimet sekä edellisiin malleihin verratut vaihtelevat datan parametriestimaattien etumerkit antavat informaatiota aineiston puutteellisuudesta ja siitä, että tulokset eivät ole yleistettävissä koko populaatiolle. Jos tulokset olisivat pysyneet konsistentteina, voitaisiin niistä olla vakuuttuneimpia. Parannetut mallit kuitenkin osoittivat, että etenkin vuosidummyn hyödyntäminen kontrollina on perusteltua. Myös interaktiomuuttujan tilastollisesti merkitsevä vaikutus ensimmäisessä mallissa oli oletuksen mukainen.

4.4 Tutkimuksen luotettavuuden arviointi

Seuraavaksi on tarkoitus arvioida tutkimuksen luotettavuutta ja mahdollisia epäkohtia, jotka voivat esiintyä aineistossa, sekä itse tutkimuksessa. Rajoitteet ja tutkimuksen luotettavuus on tarkoitus huomioida myös johtopäätöksiä tehtäessä.

Vaikka malleissa korjattiinkin residuaalien normaalijakautuneisuutta muuttujamuunnosten avulla, residuaalit eivät silti ole täysin normaalijakautuneita. Tutkimuksesta on myös rajattu tiettyjä menetelmiä pois, jotta aihe pysyisi kandidaatin tutkielman laajuuden rajoissa. Mallien stationaarisuutta ei esimerkiksi testattu eikä outlier -havaintoja lähdetty karsimaan. Nämä tekijät täytyy ottaa huomioon, kun arvioidaan tutkimuksen luotettavuutta.

Vaikka klusterirobustit keskivirheet soveltuvat yleensä lyhyeen ja leveään paneeliin, voi olla mahdollista, että poikkileikkausaineiston eli yksiköiden määrän suppeuden takia robustit keskivirheet eivät sovellu tähän tutkimukseen yhtä hyvin. Jotta tutkimus robusteilla keskivirheillä olisi luotettava, yksiköiden määrä tulee olla suuri verrattuna aikasarjojen määrään. Esimerkkinä on annettu tuhat yksikköä kolmea ajankohtaa vasten, johon verrattuna tämä tutkimus jää hyvin vajaaksi. (Hill et al., 2018, 650)

Aineisto on kooltaan juuri ja juuri paneelidatan rajoissa. Ajallisesti se on varsinkin hyvin pieni, sillä aineistoa on vain kolmelta vuodelta. Selittäjämuuttujan osalta aineisto on vielä rajallisempi, sillä data-analyytikoita on mitattu vain kahdelta vuodelta. Kyseenalaiseksi nousee myös aineiston alkuperä, ja sen kyky mitata big datan hyödyntämistä. Aineisto perustuu artikkelin julkaisseen tahon yli 10000 LinkedIn -kontaktiin yrityksissä työskenteleviin data-analyytikkoihin (Granville, 2013). Aineiston siis voidaan sanoa olevan todella pieni otos populaatiosta, joka ei ole yleistettävissä. Aineisto on sen lisäksi hyvin epätasaisesti jakautunut. Alkuperäisestä 6000 ja 7500 yrityksen listoista yritysten lukumäärä on karsiutunut rajallisten yritystietojen takia. Tutkimuksessa käytettäviä yrityksiä oli loppujen lopuksi vain 121. Vaikka ohjelmisto millä tutkimus tehdään, pystyy käsittelemään epätasapainoista aineistoa, täytyy ottaa huomioon tutkimuksen luotettavuuden arvioinnissa, että aineistossa oli paljon puuttuvia arvoja.

Estimaattoreiden valintatestit etenkin Pooled OLS -mallin kohdalla eivät välttämättä pitäisi paikkaansa, jos aineisto laajennettaisiin suuremmalla poikkileikkausaineistolla ja varsinkin aikasarja-aineistolla. Eri yrityksillä on luonnollisesti olemassa yksikkökohtaisia eroja, joita käytettävät kontrollit eivät saa kontrolloitua pois, joten Pooled OLS oli hieman yllättävä tulos estimaattoriksi. Mallit on kuitenkin tästä syystä ajettu myös hyödyntämällä datan paneelimaisuuden ja yksikkökohtaisten erojen huomioonottavia malleja, sekä niitä on myöhemmin pyritty parantamaan lisäämällä vuosidummy ja interaktiomuuttuja.

Varsinkin taloudellisia arvoja tutkittaessa on yleistä, että yksikköjen sisäisten arvojen välillä on korrelaatiota. On mahdollista, että tutkimuksessa arvojen välillä on havaittavissa kausaliteettia tai käänteistä kausaliteettia. Esimerkiksi liikevaihdon, eli koko -muuttujan kasvusta seuraa myös pääomaintensiteetin pieneneminen, sillä liikevaihto on kyseisessä tunnusluvussa jakajana.

5. JOHTOPÄÄTÖKSET

Tutkimuksessa analysoitiin big datan hyödyntämisen vaikutusta brändipääomaan. Tarkastelu tehtiin hyödyntämällä paneelidatan regressioanalyysia kahdella eri selitettävällä muuttujalla omissa regressiomalleissaan. Brändipääoman ja big datan yhteyttä lähestyttiin asiakaslähtöisen brändipääoman, kuluttajatutkimuksen ja etenkin sosiaaliseen mediaan kertyvän kuluttajadatan näkökulmista. Kirjallisuuskatsauksen avulla perusteltiin, miksi näiden kahden tekijän yhteisvaikutuksen tutkiminen on tärkeää yrityksille nyt ja tulevaisuudessa.

Muuttujat valittiin teorian ja aikaisempien tutkimusten pohjalta. Brändipääoman mittareina käytettiin lukuja, joiden avulla voidaan mitata yrityksen aineetonta pääomaa: markkina-arvoa sekä P/B-lukua. Big datan mittarina käytettiin yrityksessä työskentelevien data-analyytikoiden määrää. Malleihin valittiin yhteensä viisi kontrollimuuttujaa. Kolme niistä ovat yritystä kuvaavia taloudellisia tunnuslukuja, jotka valittiin aiemman tutkimuksen perusteella: koko (liikevaihto), kasvuprosentti ja pääomaintensiteetti. Tutkimukseen valittiin myös kategorinen muuttuja IT kontrolloimaan oletusta siitä, että IT-

alan yrityksillä on keskimäärin suuremmat kasvuodotukset ja tätä kautta myös suuremmat markkina-arvot suhteessa muiden alojen yrityksiin. Parannelluissa malleissa IT -muuttujaa hyödynnettiin datan ja IT:n interaktiossa, jolla pyritään kontrolloimaan oletusta, että IT-alan yritykset todennäköisemmin palkkaavat enemmän data-analyytikkoja. Näiden lisäksi valittiin myös viivästetty arvo selitettävästä muuttujasta kontrolloimaan autokorrelaatiota ja trendikehitystä.

Tutkimustulokset osoittivat osakseen, että big datan hyödyntämisellä on positiivinen vaikutus brändipääomaan ja vaikutus on todella suuri. Jos mallien luotettavuutta ei otettaisi huomioon, tulokset kannustaisivat yrityksiä palkkaamaan data-analyytikkoja erittäin perustellusti. Mallit, joissa vaikutus oli tilastollisesti merkitsevä, jo yhden data-analyytikon palkkaaminen nosti brändipääomaa melkein prosentin. Näiden tutkimustulosten perusteella tutkimuskysymykseen saadaan seuraava vastaus:

”Miten big datan hyödyntäminen vaikuttaa brändipääomaan?”

”Vaikutus on positiivinen ja suuri.”

Ottaen huomioon paranneltujen mallien ristiriitaiset tulokset, emme kuitenkaan voida luottaa tämän tutkimuksen tuloksiin. Aineiston rajallisuus on yksi pääsyy epäkonsistentteihin tuloksiin, jotka eivät ole yleistettävissä.

Big data on käsitteenä melko abstrakti. Big datan rajoja on vaikea määrittää teorioista huolimatta, sillä se on loppujen lopuksi vain suuri määrä dataa. Näin ollen data-analyytikoiden määrä ei välttämättä ole paras mittari sille, vaikka tässä tutkimuksessa sen käyttö onkin perusteltua. Se, että yrityksessä työskentelee data-analyytikkoja ei välttämättä tarkoita, että yritys hyödyntäisi big data -analytiikkaa. Big datan käyttöä voitaisiin mitata paremmin esimerkiksi yritysten johdolle tai IT-osastolle kohdistuvilla kyselytutkimuksilla, mutta menetelmä oli tämän tutkielman resurssien ulkopuolella. Myös brändipääoma on itsessään todella monimuotoinen käsite. Tässä tutkimuksessa hyödynnettiin finanssipohjaista lähestymistapaa tutkimalla taloudellisia aineettoman pääoman

osoittavia arvoja, mutta itse brändiarvoa niistä ei saatu eroteltua. Pelkän finanssipohjaisen muuttujan käytön voidaan ajatella olevan myös hieman ristiriidassa Kellerin asiakaslähtöisen brändipääoman teorian kanssa.

Tämä tutkimus on relevantti etenkin yrityksille, jotka harkitsevat BDA:n käyttöönottoa ja tätä kautta data-analyytikoiden palkkaamista. Jos tulokset olisivat luotettavia, tutkimus kertoisi, että yritysten ehdottomasti pitäisi palkata data-analyytikoita. Emme kuitenkaan voi myöskään perustella, ettei BDA:n hyödyntämisellä ja data-analyytikoiden palkkaamisella olisi mitään vaikutusta yrityksen brändipääomaan, vaikka tulokset olivatkin ristiriitaisia. Teorian pohjalta on järkevää olettaa, että BDA:n oikeaoppinen valjastaminen kasvattaa yrityksen markkina-arvoa ja tätä kautta brändipääomaa.

Aihe on melko eksploratiivinen tutkimuksena, sillä big datan hyödyntämisen suoria vaikutuksia brändipääomaan ei ole juurikaan tutkittu aiemmin. Tulevaisuuden jatkotutkimuksessa voitaisiin tutkia enemmän ja syvällisemmin mistä brändipääoma koostuu. Mahdollisuus olisi käyttää tutkimuksessa myös kvalitatiivisia menetelmiä, sillä loppujen lopuksi brändiä voidaan pitää kognitiivisena ajatusrakenteena kuluttajien mielissä. Myös syvällisempi big datan mittareiden selvittäminen olisi hyödyllistä tulevaisuuden tutkimuksessa. Kontrollimuuttujien määrää sekä aineiston poikkileikkaus- ja aikasarjamäärää voitaisiin laajentaa, jotta tutkimus antaisi luotettavampia tuloksia. Jatkotutkimus aiheesta on tärkeää yrityksille, sillä brändi ja big data nousevat tulevaisuudessa yhä tärkeämmiksi kilpailueduiksi globaalissa ja digitaalisessa maailmassa.

LÄHDELUETTELO

Chau, P.Y., Kuan, K.K & Liang, T. (2007). Research on IT value: What we have done in Asia and Europe. *European Journal of Information Systems*, vol. 16(3), s.196-201.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, vol. 36(4), s. 1165-1188.

DeCarlo, S. (2013) Global 2000: How We Crunch the Numbers. Forbes. [verkkodokumentti] [viitattu 7.10.2019] Saatavilla: https://www.forbes.com/sites/scottdecarlo/2013/04/17/global-2000-methodology-how-we-crunch-the-numbers/#75e57e54208d

de Oliveira, M., Rovedder, M.O., Silveira, C. & Luce, F. (2015), Brand equity estimation model, *Journal of Business Research,* vol. 68, no. 12, s. 2560-2568.

de Vries, A.; Chituc, C.-M.; and Pommeé, F. (2016) Towards identifying the business value of big data in a digital business ecosystem: A case study from the financial services industry. *Lecture Notes in Business Information Processing*, vol. 255, s. 28–40.

Ebner, K., Bühnen, T. & Urbach, N. (2014). Think big with Big Data: Identifying suitable big Data strategies in corporate environments. *Proceedings of the 47th International Conference on System Science*, *Washington, IEEE Computing Society,* s. 3748-3757

Erevelles, S., Fukawa, N. & Swayne, L. (2016) Big Data consumer analytics and the transformation of marketing. *Journal of Business Research* vol. 69, s.897-904.

Fama, E. (1970) Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance* vol. 25, s. 383-417.

Fama, E. & French, K. (1995) The effect of the set of comparable firms on the accuracy of the price earning valuation method, *Journal of Accounting Research, Spring*, s. 94-108

Ferrando-Llopis, R., Lopez-Berzosa, D & Mulligan, C. (2013) Advancing value creation and value capture in data-intensive contexts. *Proceeding conference on big Data, IEEE, October 6-9*, s. 5-9.

Forrester (2011) Expand your digital horizon with big data. [verkkodokumentti] [Viitattu 3.11.2019] Saatavilla: http:// www.asterdata.com/newsletter-images/30-04-2012/resources/Forrester_Expand_ Your_Digital_Horiz.pdf

Frankwick, G., Ramirez, E. & Zhenning, X. (2015) Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective, *Journal of Business Research* vol. 69, s. 1562-1566.

Gartner Report. (2016) Survey Analysis: Big Data Investments Begin Tapering in 2016. [Verkkodokumentti]. [Viitattu 5.11.2019]. Saatavilla: https://www.gartner.com/doc/3446724/survey-analysis-big-data- investments

Granville, V. (2015) 7500 companies hiring data scientists. Data Science Central. [verkkodokumentti] [viitattu 6.10.2019] Saatavilla: https://www.datasciencecentral.com/profiles/blogs/7500-companies-hiring-data-scientists

Granville, V. (2013) 6000 Companies Hiring Data Scientists. Data Science Central. [verkkodokumentti] [viitattu 6.10.2019] Saatavilla: https://www.datasciencecentral.com/profiles/blogs/6000-companies-hiring-data-scientists

Grover, V., Chiang, R., Liang, T-P. & Zhang, D. (2018) Creating Strategic Business Value from Big Data Analytics: A Research Framework, *Journal of Management Information Systems,* vol. 35, no. 2, s. 388-423.

Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information & Management*, vol. 53(8), 1049–1064.

Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A. & Khan, S.U. (2015) The rise of “big data” on cloud computing: Review and open research issues. *Inf. Syst*., vol. 47, s. 98–115.

Hill, R., Griffiths, W. & Lim, G. (2018) Principles of Econometrics, Wiley Custom, 5th edition, s.635

Hitt, L., Jin, F., Wu, L. (2015) Data analytics and corporate value of social media. *Working paper, University of Pennsylvania, Philadelphia*

Hitt, L. Jin, F. Wu, L. (2019) Data Analytics, Innovation and Firm Productivity, *INFORMS: Institute for Operations Research and the Management Sciences*, s.1-23.

IBM. (2014) Better business outcomes with IBM Big Data & Analytics. [Verkkodokumentti]. [Viitattu 3.11.2019]. Saatavilla: http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/59898_Better%20Business%20Outcomes_White%20Paper_Final_NIW03048-USEN-00_Final_Jan21_14.pdf

IBM. (2012) What is big data?. [Verkkodokumentti]. [Viitattu 2.11.2019]. Saatavilla: http://www-01.ibm.com/software/data/bigdata/

Ikäheimo, S., Malmi, T. & Walden, R. (2016) Yrityksen laskentatoimi, Helsinki: Talentum Pro, s. 104-105.

Johnson, L.T. & Petrone, K.R. (1998) Commentary: is goodwill an asset?, *Accounting horizons,* vol. 12, no. 3.

Kallunki, J. & Niemelä, J. (2004) Uusi yrityksen arvonmääritys *Helsinki: Talentum*. s. 86

Kaplan, A.M. and Haenlein, M. (2010), Users of the World, unite! The challenges and opportunities of social media, *Business Horizons*, Vol. 53(1), s. 59-68.

Keller, K.L. (1993) Conceptualizing, Measuring, and Managing Customer-Based Brand Equity, *Journal of Marketing,* vol. 57(1), s. 1-22.

Kiron, D. (2017). Lessons from becoming a data-driven organization. *MIT Sloan Management Review*, vol. 58(2).

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

KvantiMOTV. 2003b. Regressioanalyysin rajoitteet. Kvantitatiivisten menetelmien tietovaranto. [Viitattu: 20.11.2019] Saatavilla: https://www.fsd.uta.fi/menetelmaopetus/regressio/rajoitteet.html

Lycett, M. (2013). 'Datafication': Making sense of (big) data in a complex world. *European Journal of Information Systems*, vol. 22(4), s. 381–386.

Malik, Q.A., & Shaheen, S. (2012) The Impact of Capital Intensity, Size of Firm and Profitability on Debt Financing in Textile Industry of Pakistan, *Interdisciplinary Journal of Contemporary Research in Business,* vol. 3(10), s.1061-1066.

Mcafee, A. (2012) Big Data: The Management Revolution, *Harvard business review,* vol. 90(10), s. 60-68.

Mellin, I. (2006) Tilastolliset menetelmät: Lineaarinen regressioanalyysi. [verkkodokumentti] [viitattu 30.11.2019] Saatavilla: https://math.aalto.fi/opetus/sovtoda/oppikirja/Regranal.pdf

Melville, N., Kraemer, K. & Gurbaxani, V. (2004). Review: Information technology and organizational performance: An integrative model of IT business value. *MIS Quarterly*, s. 283-322.

Mount, M. (2014) Rejuvenating a Brand Through Social Media, *MIT Sloan Management Review,* vol. 55(4), s. 14-16.

Murphy, A. (2014) Global 2000: How We Crunch the Numbers. Forbes. [verkkodokumentti] [viitattu 7.10.2019] Saatavilla: https://www.forbes.com/sites/andreamurphy/2014/05/07/global-2000-how-we-crunch-the-numbers/#164dda204c6e

Oracle (2012). Big data for the enterprise. Oracle White Paper [verkkodokumentti] [Viitattu 30.10.2019] Saatavilla: http://www.oracle.com/us/products/database/big-data-for- enterprise-519135.pdf

Popovič, A., Hackney, R., Tassabehji, R., & Castelli, M. (2018). The impact of big data analytics on firms' high value business performance. *Information Systems Frontiers*, vol. 20(2), s. 209–222.

Prompt Cloud. (2016) Big Data Industry Report – 2016. [Verkkodokumentti]. [Viitattu 30.9.2019]. Saatavilla: https://www.promptcloud.com/big-data-industry-report-2016/

Puusa, A., Reijonen, H., Juuti, P. & Laukkanen, T. (2014) Akatemiasta markkinapaikalle: johtaminen ja markkinointi aikansa kuvina, 4. uud. painos, *Talentum, Helsinki*.

Ross, S. A. (1983) Accounting and Economics. *The Accounting Review,* vol. 58(2), s. 375-380.

Sathi, A. (2014). Engaging customers using big data: how Marketing analytics are transforming business. *New York: Palgrave Macmillan*.

Tollington T (1998), Separating the Brand Asset from the Goodwill Asset, *Journal of Product & Brand Management*, Vol. 7(4), s. 291-304.

Vidgen, R., Shaw, S. & Grant, D. B. (2017) Management challenges in creating value from business analytics. *European Journal of Operational Research,* vol.261(2), s. 626-639

Wood, L. (2000) Brands and Brand Equity: Definition and Management. *Management Decision*, vol. 38(9), s. 662-669.

Wu, L, Brynjolfsson, E (2009) The future of prediction: How Google searches foreshadow housing prices and sales. *SSRN Electronic Journal,* s. 89–118.

LIITTEET

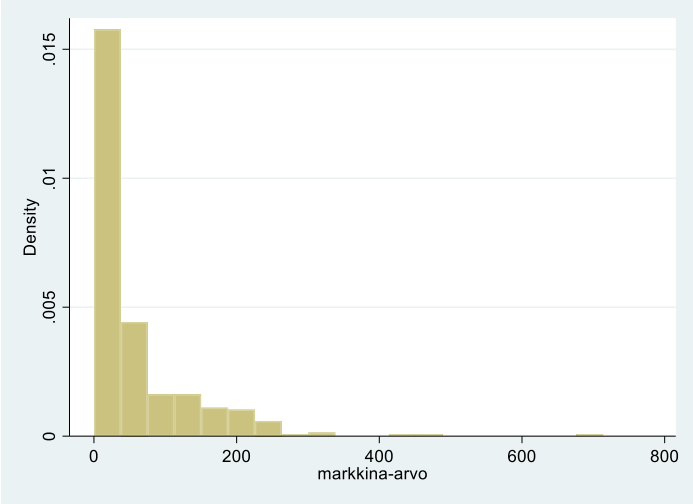
Liite 1. Tutkimuksessa käytettävät yritykset

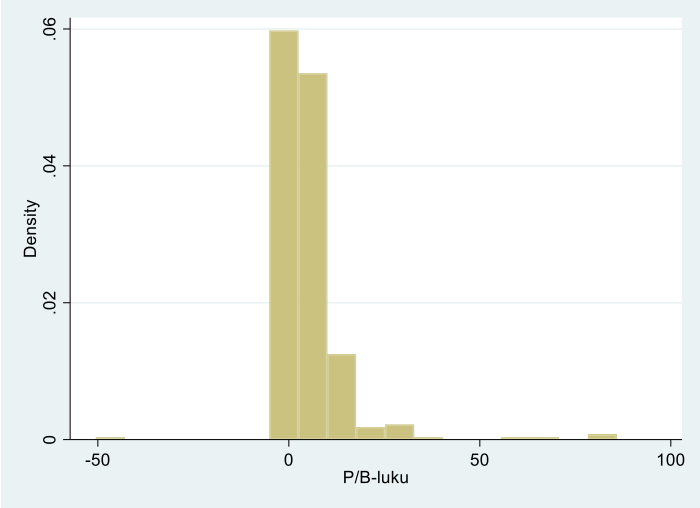
AbbVie	Koninklijke Philips
Accenture	Leidos Holdings
Adobe	Lockheed Martin
Aflac	ManpowerGroup
Akamai Technologies	Marchex
Allstate	MasterCard
Amazon	Medtronic
Amdocs	Merck
American Airlines	MetLife
American Express	Microsoft
Amgen	Morgan Stanley
Apple	NetApp
AstraZeneca	Netflix
Autodesk	Nielsen Holdings
Bank of America	Nike
BlackBerry	Nordstrom
BlackRock	Northrop Grumman
Blucora	Novartis AG
Boeing	Office Depot
Broadcom	Oracle
Broadridge Financial Solutions	PepsiCo
Capital One Financial	Pfizer
Charles Schwab	Pitney Bowes
Chevron	PNC Financial Services
Chubb	Progressive
Cigna	Proofpoint
Cisco	Providence Health & Services
Citigroup	Prudential Financial

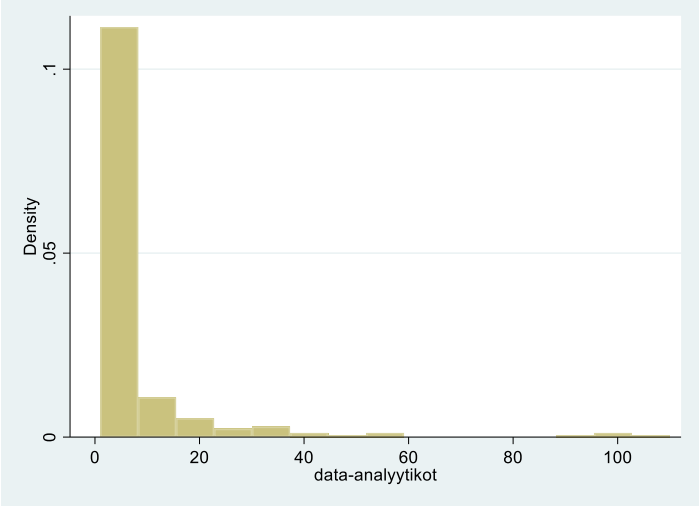
Citrix Systems	Raytheon
Cognizant Technology Solutions	Red Hat
Comcast	Robert Half
CommVault Systems	Royal Dutch Shell
CoreLogic	Salesforce
Cray Inc	Shutterfly
Credit Suisse Group	Splunk
Deutsche Bank Aktiengesellschaft	Starbucks
Disney	SunTrust Banks
Dun and Bradstreet	Symantec
Eaton	Syntel
Electronic Arts	Tableau Software
Enphase Energy	Target
Equifax	TELUS
Facebook	Teradata
Fifth Third Bancorp	The Home Depot
Fiserv	Thermo Fisher Scientific
Ford Motor Company	Thomson Reuters
Gartner	Tremor Video
General Motors	TripAdvisor
Goldman Sachs	UnitedHealth Group
Groupon	Verint Systems
Hartford Financial Services	Verisk Analytics
Humana	Verizon
IBM	Viacom
Infosys	Visa
Intel Corporation	VMware
Intuit	Vodafone Group
Johnson Controls	Wells Fargo
JPMorgan Chase	Western Union
Juniper Networks	Workday
Kforce	Xerox

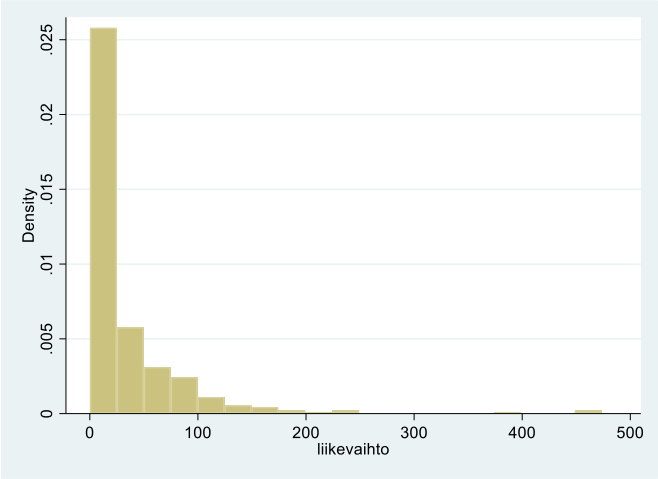
Kimberly-Clark	

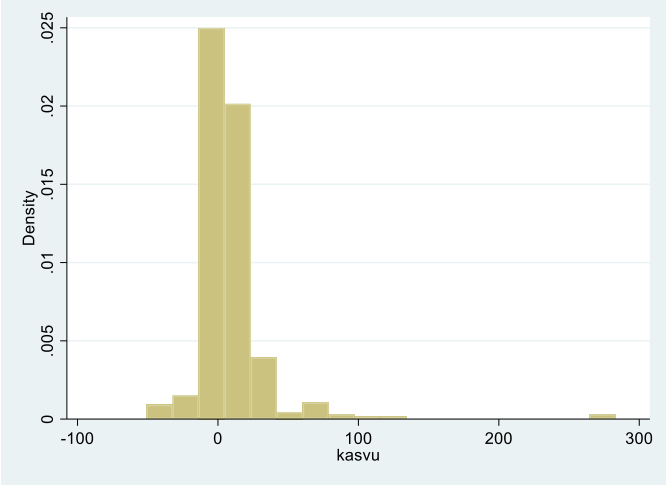
Liite 2. Muuttujien jakaumat

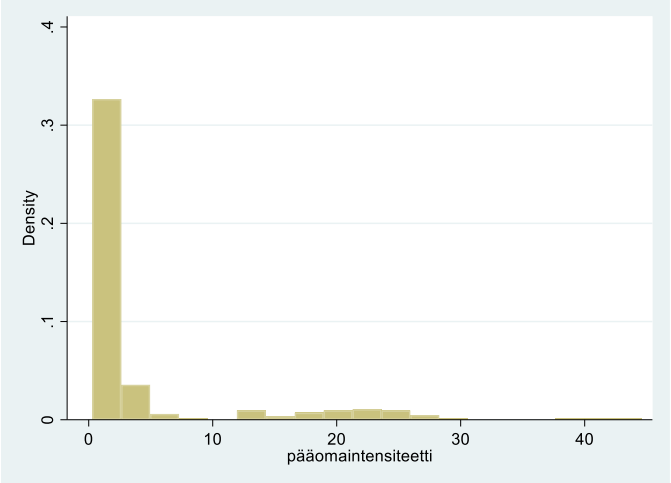


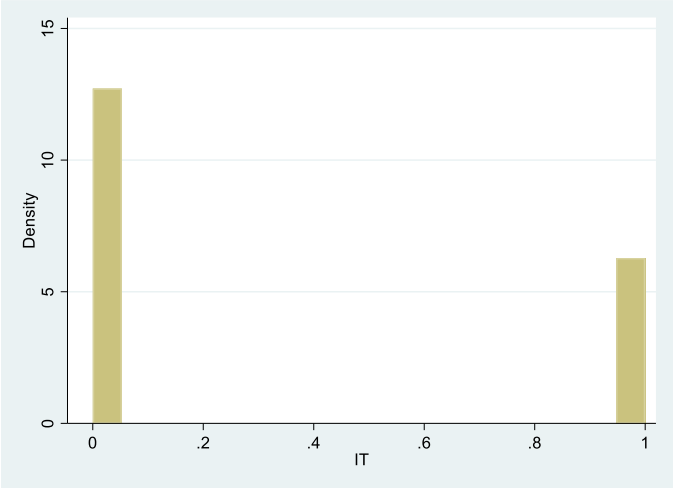




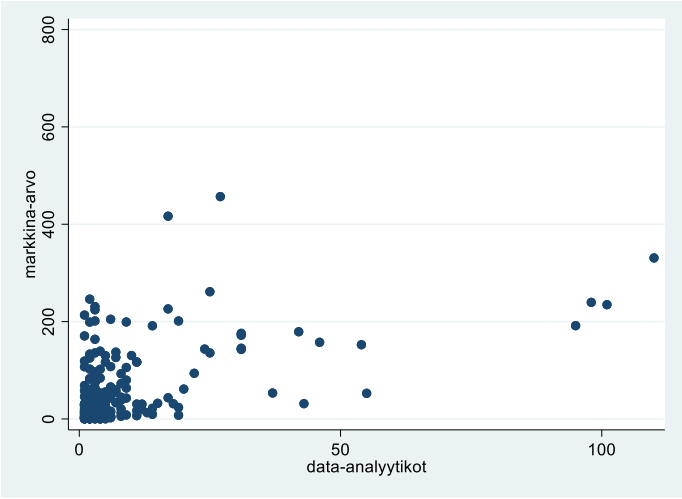

Density
0
.005
.01
.015
.02
.025
0
100
200
300
400
500
liikevaihto

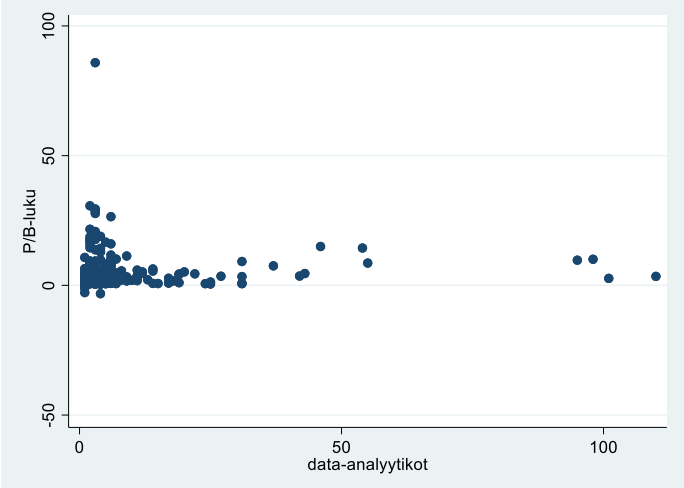

Density
0
.005
.01
.015
.02
.025
-100
0
100
200
300
kasvu


Density
0
.1
.2
.3
.4
0
10
20
30
40
pääomaintensiteetti



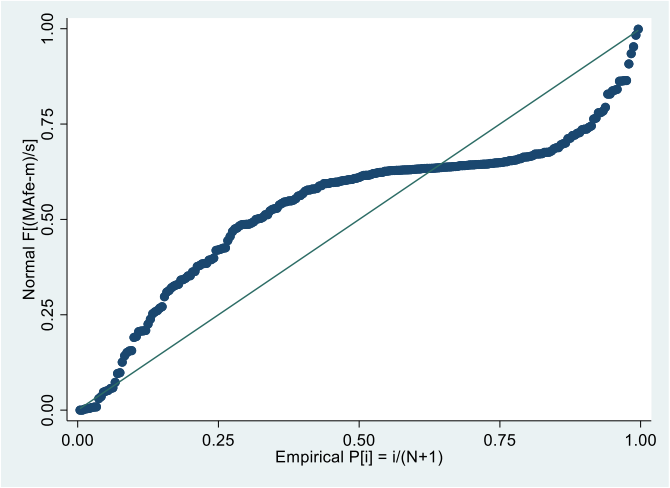
Liite 3. Selitettävien ja selittäjän korrelaatiot



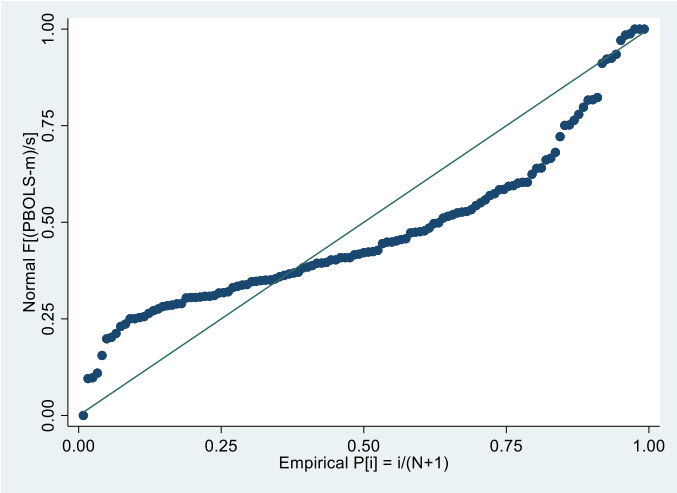


Liite 4. Mallien residuaalikuviot

Markkina-arvo selitettävänä muuttujana

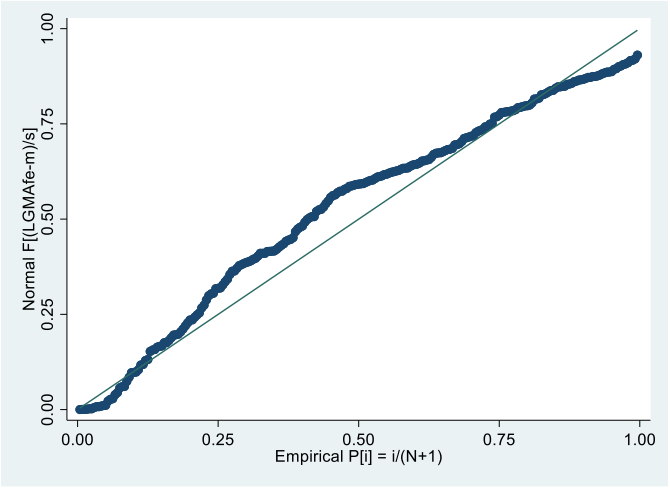


P/B-luku selitettävänä muuttujana

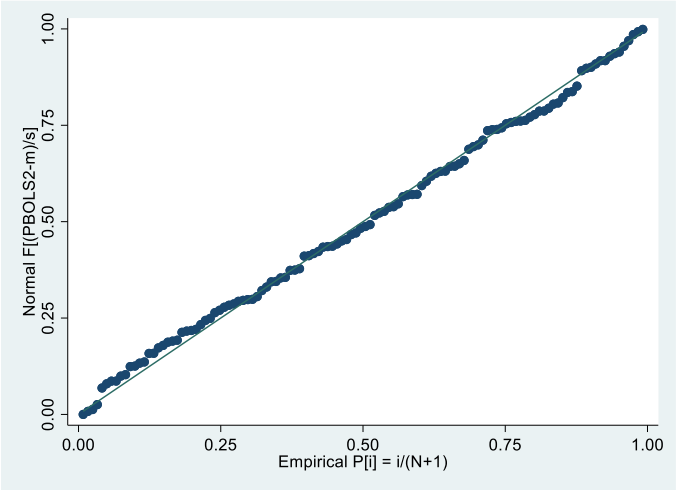


Liite 5. Mallien residuaalikuviot logaritmimuunnosten jälkeen

Markkina-arvo selitettävänä muuttujana

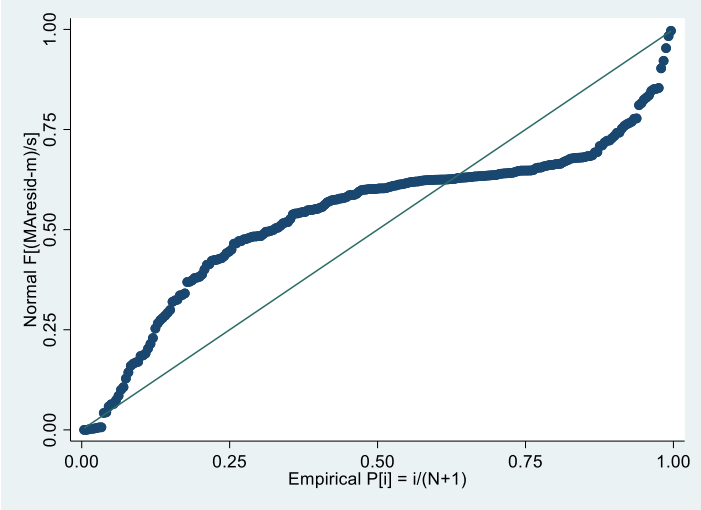


P/B-luku selitettävänä muuttujana

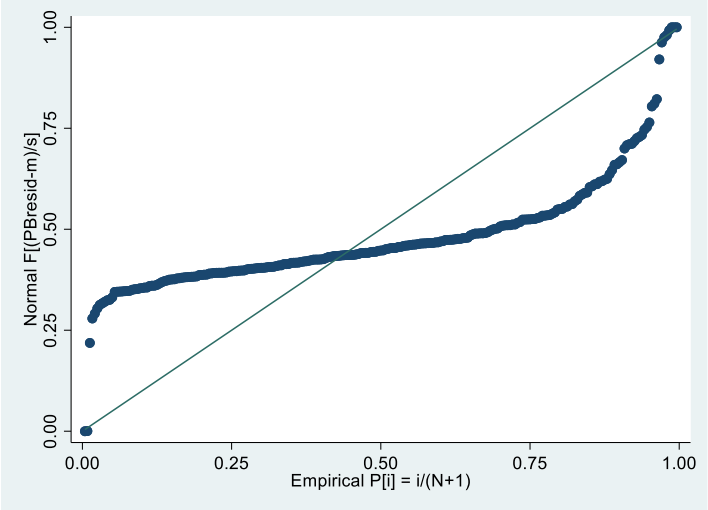


Liite 6. Paranneltujen mallien residuaalikuviot

Markkina-arvo selitettävänä muuttujana parannellussa mallissa

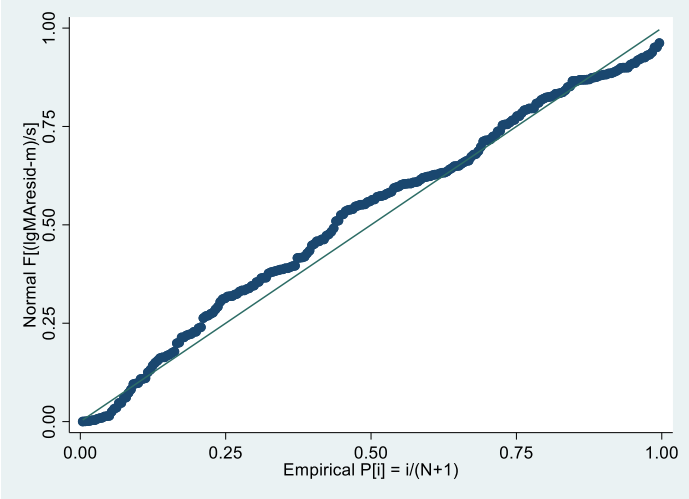


P/B-luku selitettävänä muuttujana parannellussa mallissa



Liite 7. Paranneltujen mallien residuaalikuviot logaritmimuunnosten jälkeen

Markkina-arvo selitettävänä muuttujana parannellussa mallissa



P/B-luku selitettävänä muuttujana parannellussa mallissa