PRO GRADU -TUTKIELMA

Niina Matikainen

Auton arvon aleneminen iän ja käytön myötä

TAMPEREEN YLIOPISTO
Luonnontieteiden tiedekunta
Tilastotiede
Toukokuu 2017

Tampereen yliopisto
Luonnontieteiden tiedekunta
Matikainen, Niina: Auton arvon aleneminen iän ja käytön myötä
Pro gradu -tutkielma, 40 s., 1 liites.
Tilastotiede
Toukokuu 2017

Tiivistelmä

Suomalaiset liikkuvat paljon autoilla. Monessa taloudessa onkin vähintään yksi auto. Auto ei kuitenkaan ole ihan mikä tahansa kotitalouden kulutustuote, vaan se voi olla yksi talouden isoimmista hankinnoista esimerkiksi asunnon ostamisen jälkeen. Autoa ei välttämättä osteta täysin uutena, vaan usein se voidaan hankkia myös käytettynä.

Käytettyjen autojen kauppaa käydäänkin vilkkaasti, ja Suomessa tapahtuu vuositasolla satoja tuhansia käytettyjen autojen kauppoja. Kauppaa käydään nykyään esimerkiksi internet-sivustoilla, kuten tämän tutkielman aineiston lähteenä toimivalla sivustolla.

Kuten kaikkeen kaupankäyntiin, myös automarkkinoilla hinnan määrittely on tärkeää. Liian suuri pyyntihinta ei houkuttele potentiaalisia ostajia. Sopivaksi tai jopa alakanttiin määritelty hinta saa ostajan kiinnostumaan, ja auto voi mennä kaupaksi hyvinkin nopeasti. Pyyntihinta pitää suhteuttaa muun muassa auton ominaisuuksiin, ikään ja ajettuihin kilometreihin.

Tämän tutkielman tavoitteena on määritellä, miten auton arvon laskee sen ikääntyessä ja mittarilukeman kilometrimäärän kasvaessa. Keskimääräisen arvon alenemisen lisäksi on mielenkiintoista verrata tähän lopputyöhön valittujen eri automerkkien välisiä eroja. Tuloksia voidaan hyödyntää käytetyn auton arvon määrittämisessä tai auton omistamisen kokonaiskustannusten laskemisessa.

Tilastolliseksi menetelmäksi analyysiä tehdessä valittiin polynominen regressiomalli. Tavallinen lineaarinen malli ei sovi aineistoon, sillä auton arvo ei alene samalla kulmakertoimella eri vaiheessa auton elinkaarta. Polynomisen mallin avulla on mahdollista mallintaa selittävien muuttujien epälineaarista vaikutusta selitettävään muuttujaan.

Tutkimuskysymystä lähdettiin tarkastelemaan keräämällä aineistoksi vuoden 2016 aikana Autotalli.com-sivustolle jätettyjä autojen myynti-ilmoituksia. Aineiston tiedot, kuten mittarilukemasta luettu ajettujen kilometrien lukumäärä, saadaan ilmoitukseen merkityistä tietokentistä. Autojen myynti-ilmoituksia ovat jättäneet sekä autoliikkeet että yksityiset henkilöt. Ilmoituksista kerättyyn aineistoon liitettiin toisesta tietokannasta arvio siitä, kuinka paljon auto on maksanut uutena. Tämä tieto on olennainen, kun lasketaan auton arvon alenemista kyseiseen hetkeen asti, mutta sitä tietoa ei ollut alkuperäisessä ilmoitusten datassa.

Tällöin analyysin aineisto on siis yhdistelmä kahta eri lähdettä, joten lopullinen aineisto muodostui niistä autojen myynti-ilmoituksista, joissa tarpeelliset tiedot löy-

tyivät molemmista aineistoista. Aineistoa rajattiin luotettavuuden takia poistamalla hyvin erikoiset havainnot. Lopulta analyysiin jäi 34 068 havaintoa.

Tulosten perusteella voi päätellä, että auton arvo alenee keskimäärin eniten auton ollessa lähes uusi. Kilometrit vievät auton jälleenmyyntihinnasta pois vähiten silloin, kun autolla on ajettu jo satoja tuhansia kilometrejä. Auton ensimmäiset kilometrit koituvat siis autoilun kustannuksia laskettaessa kalleimmiksi, jos kilometrin hintaa mitataan auton arvon aleneman mukaan. Autoilun kokonaiskustannuksina pitää huomioida kuitenkin myös korjauskustannukset, joita tässä työssä ei tarkasteltu.

Tässä lopputyössä on laskukaava sille, kuinka paljon auton arvo alenee keskimäärin iän ja kilometrien myötä. Sen avulla yksityinen autonomistaja voi havainnollistaa autonsa mahdollista jälleenmyyntihintaa. Yksityisen kuluttajan mielenkiinto voi olla myös siinä, miten auton arvo alenee käytön myötä ja mitkä ovat autoilun todelliset kustannukset. Työn loppupuolella on myös esimerkkitaulukkoja auton arvon keskimääräisestä alenemasta, kun tehdään oletus keskimääräisestä vuosittaisesta ajokilometrimäärästä.

Asiasanat: polynominen regressiomalli, mallintaminen, R-ohjelmisto, autojen hinnat, auton hinnoittelu

Sisältö

1 Johdanto

Suomessa oli vuoden 2015 lopulla 2 635 643 liikennekäytössä olevaa henkilöautoa, ja kaikkiaan rekisterissä oli 3 257 581 henkilöautoa (Tilastokeskuksen katsaus 2016). Autot ovatkin Suomen toiseksi yleisin kulkuneuvo polkupyörän jälkeen (Liimatainen et al. 2015, sivu 55).

Autoalan tiedotuskeskuksen mukaan Suomessa tehdään vuosittain noin 550 000 - 600 000 käytetyn auton kauppaa. Näistä puolet tapahtuu merkkiliikkeissä, neljäsosa järjestäytymättömissä autoliikkeissä ja neljäsosa vaihtaa omistajaa kuluttajien välisessä kaupassa. (Autoalan tiedotuskeskus 2017)

Autojen myynti-ilmoituksia on internetissä, ja verkkopalvelut ovatkin hyviä markkinapaikkoja autojen myymiseen ja ostamiseen. Ilmoituksia laittavat sekä yksityiset auton myyjät että autokauppaan erikoistuneet yritykset. Internetin käyttö on yhä suositumpaa, mikä käy ilmi Tilastokeskuksen vuonna 2016 julkistamassa väestön tietoja viestintätekniikan käyttöä kartoittaneessa tutkimuksessa. Siinä osoitettiin, että jopa 92 prosenttia 16-74-vuotiaasta väestöstä käyttää internetiä. (Suomen virallinen tilasto 2015b)

Autojen ostajaehdokkaat liikkuvat siis yhä enemmän internetissä ja auton markkinoinnissa on tärkeää laittaa ilmoitus internetin markkinapaikalle. Ostettavien autojen vertailu on helppoa, kun samassa palvelussa on tarjolla useita myynnissä olevia autoja. Valinnanvaraa on usein runsaasti, ja esimerkiksi Autotalli.com-palvelussa onkin sekä yksityisten että autoliikkeiden jättämiä autojen myynti-ilmoituksia. Myynti-ilmoituksen jättämiseen liittyy auton pyyntihinnan määrittely, mikä kannattaa tehdä huolellisesti, jotta auton myyntiaika ei veny liian korkean pyyntihinnan vuoksi.

Autojen pyyntihinnan määrittelyssä voidaan hyödyntää vastaavien myynnissä olevien autojen pyyntihintoja. Tällöin internetin markkinapaikat voivat toimia vastaavien autojen myynti-ilmoitusten selailusivustoina, ja auton hinnan määrittämiseen vaikuttavat ne ilmoitukset, jotka ovat juuri silloin esillä. Uuden ilmoituksen tekemisessä ja auton hinnoittelussa saattaa olla ylikorostununeina ne ilmoitukset, joissa hinta on ostajakunnan mielestä liian korkea, jolloin auton myyntiaika on pidempi, ja niitä illmoituksia on runsaammin esillä kuin lyhyen markkinointiajan ilmoituksia. Sopivasti tai alhaisesti hinnoitellut autot myydään nopeasti. Tähän tutkimukseen otetut havainnot ovat Autotalli.comin ilmoitusten tietokannasta, joten kohteen markkinointiaika ei vaikuta sen todennäköisyyteen olla mukana aineistossa.

Auton myynti-ilmoitukseen laitettavan hinnan määrittämisessä voidaan myös käyttää apuna sitä varten tehtyjä laskureita, joita voi löytää autoiluun keskittyviltä sivustoilta. Tämän pro gradu -työn tuloksia voidaan joiltain osin soveltaa auton hinnoitteluun, sillä tuloksien avulla voidaan selvittää, kuinka paljon auton alkuperäisestä hinnasta on keskimäärin jäljellä sen hetkisellä mittarilukemalla ja iällä. Tässä työssä olevan laskukaavan avulla saadaan keskimääräinen hinta sen ikäiselle ja niin paljon ajetulle autolle, mikä voi toimia lähtökohtana hinnoittelulle. Auton muiden ominaisuuksien, kuten vaikkapa mukana tulevien renkaiden tai viimeisen huoltoajankohdan mukaan tätä keskimääräistä hintaa voi muokata oman tiedon perusteella

sopivammaksi. Hinnoittelussa kannattaa myös huomioida haluttu ja odotettu myyntiaika.

Myös vakuutusyhtiöt joutuvat pohtimaan auton arvoa lunastustilanteissa, kun pitää määritellä asiakkaalle korvattavaa summaa. Auton käypä arvo on oleellinen tieto, kun pohditaan auton korjaamisen kustannuksia verrattaen sen lunastushintaan. Hyttisen (2016) insinöörityössä käsitellään auton lunastustoiminnan perusteita tarkemmin.

Autotalli.com-palvelusta löytyy kymmeniä tuhansia ilmoituksia, joita sinne ovat jättäneet sekä yksityiset myyjät että autoalan ammattilaiset. Tämän lopputyön aineistona on vuoden 2016 aikana palveluun jätetyt autojen myynti-ilmoitukset. Ilmoituksessa olleiden tietojen lisäksi aineistoon lisättiin arvio auton alkuperäisestä hinnasta uutena, mikä tehtiin täsmäämällä autojen ominaisuuksia toisessa tietokannassa saatavilla oleviin uusien autojen hintatietoihin. Tällöin päästiin lähelle sitä tietoa, kuinka paljon auton hinta on ollut uutena ja kuinka paljon sen hinta on tippunut myyntihetkeen mennessä.

Mielenkiinnon kohteena on tutkia, kuinka suuri osuus auton alkuperäisestä hinnasta on jäljellä riippuen auton iästä ja mittarilukeman kilometreistä. Tähän lopputyöhön on otettu myös vertailu muutaman yleisimmän automerkkien välillä.

Lopputyön rakenne muodostuu niin, että johdannon jälkeen tutustutaan analyysimenetelmään ja käydään läpi sen teoriaa. Sen jälkeen esitellään aineistoa ja siihen tehtyjä muuttujien johdannaisia sekä rajauksia. Nelosluvussa käydään läpi analyysiä ja tuloksia. Lopuksi on johtopäätelmiä, joissa työn tuloksia tiivistetään ja pohditaan jatkotutkimuksen ideoita.

1.1 Aiheeseen liittyvät aiemmat tutkimukset

Autoiluun liittyviä tutkimuksia on tehty Suomessa aiemminkin. Auton hintalaskureista kertovia tutkimuksia ei ole julkisesti saatavilla, mutta autoilun kokonaiskustannuksia selvittäneessä opinnäytetyössä pohdittiin myös auton arvon alenemista (Leppikangas 2015). Kyseisessä työssä ei kuitenkaan paneuduttu kovin syvällisesti auton arvon alenemaan, ja lähteenä auton arvon laskemiselle käytettiin Taloussanomien artikkelia (Kanniainen, 2014).

Autoilun kokonaiskustannuksia selvitettiin myös Keurulaisen opinnäytetyössä vuonna 2010. Työssä keskityttiin erityisesti pääkaupunkiseutuun, mutta työssä on lisäksi valtakunnallista tietoa. Työssä mainittiin myös auton arvon aleneminen osana kokonaisukustannuksia, ja lähteenä tähän asiaan oli Ilta-Sanomien artikkeli (Nurmela, 2009).

Aivan erilainen tutkimus vuodelta 2016 käsittelee autoja sijoituskohteena (Pulkkinen, 2016). Opinnäytetyössä tarkastellaan auton arvon kehittymistä ja lasketaan myös auton pitämiseen liittyviä kustannuksia. Pulkkisen opinnäytetyö koskee harrasteja klassikkoautoja, joten sen aineistona on erilainen ryhmä autoja kuin tässä lopputyössä. Vaikka aihe sinällään oli hyvin läheinen, oli tutkimukseen valittu näkökulma ja aineisto hyvin poikkeava tähän tutkimukseen verrattuna.

Käytetyn auton hankintaan liittyvien ostopäätösten tekemistä on pohdittu vuonna

2016 julkaistussa opinnäytetyössä. Kyseisessä työssä käsitellään myös hintaa ostoprosessin tärkeänä alueena, mutta siinä ei oteta kantaa käytetyn auton hinnoitteluun tai arvon laskemiseen (Kangosjärvi & Sassi 2016).

Toivasen (2005) pro gradu -tutkielmassa selvitetään käytetyn auton arvon määrittelyn prosessia. Tutkielmassa selvitetään auton hinnan määrittelyssä käytettäviä apuvälineitä, joita autoalan ammattilaiset hyödyntävät työssään. Tutkielmassa kerrotaan Grey-Hen Oy:n tekemästä auton arvon määrittämiseen tehdystä työkalusta, ja käsitellään siihen käytetyn tilastollisen mallin teoriaa. Työssä ei kuitenkaan julkaista laskukaavaa siitä, miten auton arvo alenee käytön ja iän myötä. Tutkielman päätelmät-osiossa todetaan, että kyseisen työn tekemisen ajanhetkellä oli vaikea löytää tutkimuksia autojen hinnoitteluprosessista. Tutkielman lähteenä on käytetty paljon Grey-Hen Oy:n tuottamaa tietoa, kuten haastatteluja ja sisäisiä raportteja, jotka eivät ole julkisesti saatavilla.

Kuluttajatutkimuskeskus julkaisi vuonna 2003 tutkimuksen käytettyjen ajoneuvojen markkinahinnoista. Tutkimuksessa selvitetään esimerkiksi ajoneuvojen pyyntihintojen hajontaa. Siinäkin tutkimuksessa, kuten edellisessä kappaleessa mainitussa, käsitellään Grey-Hen Oy:n tuottamaa tilastollista mallia, mutta vain teoriatasolla. Tutkimuksessa ei kuitenkaan kerrota esimerkiksi regressiomallin kertoimia, joten siitä ei käy ilmi, miten ikä ja ajetut kilometrit vaikuttavat auton arvoon. (Aalto-Setälä & Halonen, 2003)

Autoiluun liittyviä tutkimuksia on siis tehty aiemmin Suomessa, mutta viime vuosina ei ole julkaistu tämän työn sisältöä vastaavia julkaisuja. Tämän lopputyön kaltaista lopputulemaa auton arvon keskimääräisestä alenemasta ei ole julkisesti saatavilla olevaa tietoa. Myöskään Autotalli.comin tietokannan dataa ei ole aiemmin hyödynnetty tällaiseen tutkimukseen.

2 Analyysimenetelmä

Tässä luvussa esitellään aineiston analyysiin ja tulosten saamiseen käytettyä menetelmää. Analyysimenetelmän esittely aloitetaan tavallisella lineaarisella regressiolla. Sen jälkeen esitetään polynomisen regressiomallintamisen teoriaa. Lisäksi käydään läpi regressiomallin sovittamisen teoriaa ja lopuksi nostetaan esiin muutamia regressioanalyysin haasteita.

Analyysimenetelmän teorian lähteenä on käytetty teoksia Linear Regression Analysis (Lee & Seber, 2003), Introduction to Linear Regression Analysis (Montgomery, Peck & Vining, 2006) ja Applied Linear Regression (Weisberg, 2005). Suomenkieliset termit liittyen regressioanalyysiin olen oppinut alunperin Jukka Nyblomilta Jyväskylän yliopistossa, ja käytin tässäkin työssä lähteenä hänen luentomonistettaan (Nyblom, 2015).

2.1 Lineaarinen malli

Lineaarisella regressiomallilla voidaan selvittää kahden tai useamman muuttujan välistä yhteyttä. Vastemuuttujan arvojen oletetaan olevan riippuvaisia yhdestä tai useammasta prediktorimuuttujasta, ja tämän riippuvuuden oletetaan olevan lineaarista. Toisella tapaa sanottuna, lineaarisessa regressioanalyysissä yritetään selittää vastemuuttujan vaihtelua käyttäen selittäviä muuttujia.

Saatua regressiomallia voidaan käyttää vasteen arvon ennustamiseen, kun sille annetaan prediktorimuuttujien arvot. Regressiomallin kertoimia voidaan käyttää ennusteiden eron tulkitsemiseen tai selvittämään tiettyjen keskiarvojen eroja. Regressioyhtälön avulla voidaan laskea ennustettuja eroja prediktorien eri arvoilla.

Tässä työssä käydään läpi ensin yhden selittävän muuttujan lineaarinen regressiomalli, ja sen jälkeen useamman selittävän muuttujan lineaarinen regressiomalli. Selitettävää muuttujaa merkitään y:llä ja sen oletetaan olevan jatkuva. Selittäviä muuttujia merkitään x_i:llä, $i = 1, 2, \ldots, k$, ja selittävä muuttuja voi olla jatkuva tai ns. dummy-muuttuja, jolloin sen arvo on joko 0 tai 1. Dummy-muuttuja voi olla siis kvalitatiivinen muuttuja, jos sillä on vain kaksi luokkaa.

2.1.1 Yksinkertainen lineaarinen regressiomalli

Yksinkertainen, yhden selittävän muuttujan lineaarinen regressiomalli on muotoa

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

missä y on ennustettava muuttuja eli vastemuuttuja, β_0 on vakiotermi, β_1 on kulmakerroin, x on selittävä muuttuja eli prediktori ja ϵ virhetermi. Vakiotermi β_0 kuvaa kohtaa, missä regressiosuora leikkaa y-akselin, eli mikä on ennusteen arvo, kun $x = 0$. Kulmakerroin β_1 kuvaa ennustettavan arvon muutosta x-muuttujan arvon muuttuessa.

Regressiomallin viimeiselle termille ϵ pätee

$$\epsilon = y - E(y|x) \tag{2.2}$$

ϵ on mallin virhetermi, johon sisältyy mallin jäännöstermit eli residuaalit, jotka johtuvat aineiston satunnaisvaihtelusta. Jäännöstermit ovat erotuksia havaitun arvon y ja ennustetun arvon välillä, eli jäännöstermiin sisältyy se vaihtelu, jota regressiomalli ei selitä. Kun tarkastellaan yhtälöitä (2.2) ja (2.3), voidaan todeta jäännöstermien olevan erotuksia havaitun arvon ja suoran $\beta_0 + \beta_1 x$ arvon välillä.

ϵ on satunnaismuuttuja jakaumalla $N(0, \sigma^2)$. Lisäksi virhetermien pitää olla keskenään tilastollisesti riippumattomat eli ne eivät saa korreloida keskenään, eivätkä ne saa myöskään riippua prediktorien arvoista.

Jos tehdään oletus, että selittävä muuttuja x on kiinteä, voimme laskea muuttujan y odotusarvon. Koska virhetermin ϵ odotusarvo on nolla, vastemuuttujan y odotusarvo ehdolla x on $\beta_0 + \beta_1 x$, eli seuraava yhtälö pätee

$$E(y|x) = \mu_{y|x} = E(\beta_0 + \beta_1 x) = \beta_0 + \beta_1 x \tag{2.3}$$

Yhtälön (2.3) avulla voimme siis laskea ennusteen y:lle, mikäli tiedämme muuttujan x arvon.

Muuttujan y varianssi on

$$Var(y|x) = \sigma^2_{y|x} = Var(\beta_0 + \beta_1 x + \epsilon) = \sigma^2 \tag{2.4}$$

Regressiosuora muodostuu siis muuttujan y odotusarvoista muuttujan x eri arvoilla. Kulmakerrointa β_1 voidaan tulkita vastemuuttujan y keskiarvon muuttumisena, kun selittävän muuttujan x arvo muuttuu yhden yksikön. Mikäli regressiomallin kulmakerroin on nolla, selittävän muuttujan x ja selitettävän muuttujan y välillä ei ole lineaarista yhteyttä, jolloin ennustettu arvo muuttujalle y on aineiston keskiarvo. Mikäli kulmakerroin β_1 on positiivinen, muuttujien x ja y välillä on positiivinen korrelaatio. Vastaavasti, jos kulmakerroin β_1 on negatiivinen, vastemuuttujan y ennustettu arvo pienenee, kun x kasvaa. Muuttujan y varianssi ei kuitenkaan saa oletuksen mukaan olla riippuvainen muuttujan x arvosta, vaan se on aina virhetermin varianssi σ^2.

Jos aineisto sisältää muuttujan x arvoja välillä $x_1 \leq x \leq x_2$, regressiomallia ei saisi käyttää ennustamaan vastemuuttujan y arvoja muuttujan x arvoilla $x \leq x_1$ tai $x \geq x_2$. Regressiosuoran ei siis oleteta toimivan hyvin havaintoalueen ulkopuolella, vaan sitä tulisi käyttää vain havaittujen arvojen alueen sisäpuolella.

2.1.2 Useamman selittävän muuttujan malli

Yhtälö (2.1) sisältää vain yhden selittävän muuttujan, mutta regressiomallia on mahdollista laajentaa niin, että vastemuuttujan y arvoja voidaan ennustaa useamman selittävän muuttujan avulla.

Vastemuuttuja y voi olla riippuvainen k selittävästä muuttujasta, $x_1, x_2, \ldots, x_k$ niin, että

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \tag{2.5}$$

Tätä kutsutaan useamman selittävän muuttujan lineaariseksi regressiomalliksi, koska muuttujan y arvoja ennustetaan useammalla kuin yhdellä selittävällä muuttujalla.

Oletukset ja määritelmät pätevät kuten edellä, eli esimerkiksi kulmakertoimia $\beta_1, \beta_2, \ldots, \beta_k$ tulkitaan ennustettavan arvon keskimääräisinä eroina eri selittävien muuttujien $x_1, x_2, \ldots, x_k$ arvoilla. Esimerkiksi parametri β_1 indikoi vastemuuttujan y odotettua muutosta, kun muuttujan x_1 arvo muuttuu yhden yksikön verran, kun muiden muuttujien $x_2, x_3, \ldots x_k$ arvot pidetään vakioina. Myös oletukset virhetermille ϵ pätevät kuten edellä.

Kaava (2.5) voidaan kirjoittaa myös matriisimuodossa. Tämä mahdollistaa kompaktimman tavan näyttää mallin ja tuloksia. Kaava (2.5) on matriisimuodossa

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.6}$$

missä

(2.7)

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Kaavassa (2.7) $\mathbf{y}$ on $n \times 1$ -vektori, joka sisältää havainnot, $\mathbf{X}$ on $n \times p$ -matriisi, joka sisältää selittävien muuttujien tasot, $\boldsymbol{\beta}$ on $p \times 1$ -vektori, joka sisältää regressiokertoimet ja $\boldsymbol{\epsilon}$ on satunnaisvirheet sisältävä $n \times 1$ -vektori.

2.2 Polynominen regressiomalli

Regressioanalyysissä mallinnetaan selitettävää muuttujaa yhdellä tai useammalla taustamuuttujalla. Tämän työn analyysissä on useampi selittävä muuttuja, joilla on selkeästi epälineaarinen yhteys selitettävään muuttujaan. Tällöin ei voida käyttää tavallista lineaarista mallia, vaan mallinnuksessa käytetään polynomista regressiomallia, sillä tavallinen lineaarinen malli ei selittäisi muuttujien välistä yhteyttä tarpeeksi hyvin.

Polynomisessa regressiossa vastemuuttujan saamia arvoja mallinnetaan prediktorilla, joka on polynomimuodossa. Tällöin siis selitettävän muuttujan ei oleteta riippuvan lineaarisesti prediktorista, vaan sille sallitaan epälineaarinen yhteys. Muuttujien välinen yhteys voi olla suoran sijaan käyrän muotoinen. Mallin selittävän muuttujan x ja selitettävän muuttujan y välistä yhteyttä mallinnetaan k:nnen asteen polynomin avulla.

Toisen asteen polynominen malli yhdellä selittävällä muuttujalla on muotoa

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \tag{2.8}$$

ja vastaavasti toisen asteen polynominen malli kahdella selittävällä muuttujalla voidaan kirjoittaa muotoon

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon \tag{2.9}$$

Tällöin ei puhuta enää regressiosuorasta, vaan vastemuuttujan ja selittävien muuttujien välinen yhteys on käyrän muotoinen.

Yhtälön (2.8) mukaan muuttujan y odotusarvo on

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{2.10}$$

joka on toisen asteen käyrän yhtälö. Parametri β_0 on muuttujan y keskiarvo, jos $x = 0$, eli β_0 on myös muuttujan y ennustettu arvo kiinnitetyllä arvolla $x = 0$, mikäli aineiston vaihteluväli sisältää kohdan $x = 0$. Muuten vakiotermillä β_0 ei ole tulkintaa. Parametrilla β_1 ei ole samanlaista tulkintaa kuin yksinkertaisessa lineaarisessa regressiossa, sillä muuttujan x arvon muuttuessa, ennustettuun muuttujan y arvoon vaikuttavat molemmat parametrit β_1 ja β_2.

Yhtälöä (2.8) voidaan yhä yleistää k:nnen asteen polynomiksi, jolloin saadaan yhtälö

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \epsilon \tag{2.11}$$

Kuten yhtälön (2.8) mukaan muuttujan y odotusarvo saadaan yhtälöstä (2.10), myös k:nnen asteen polynomin regressiomallissa muuttujan y odotusarvo saadaan vastaavasti kaavasta

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k. \tag{2.12}$$

Jos asetetaan $x_j = x^j$, $j = 1, 2, \ldots, k$, niin yhtälöstä (2.11) tulee monimuuttujainen regressiomalli, jossa on k kappaletta selittäviä muuttujia $x_1, x_2, \ldots, x_k$.

Mikäli polynominen regressiomalli sisältää ylimmän asteen regressiotermin lisäksi kaikki sitä alempien asteiden regressiotermit, mallia voidaan kutsua hierarkiseksi malliksi. Esimerkiksi mallia

$$y = \beta_0 + \beta_1 x + \beta_3 x^3 + \epsilon \tag{2.13}$$

ei voi kutsua hierarkiseksi malliksi, sillä siitä puuttuu muuttujan x toisen asteen regressiotermi.

2.3 Regressiomallin sovittaminen

Regressiomallin sovittaminen tapahtuu pienimmän neliösumman avulla[1]. Esimerkiksi yhtälön (2.1) parametrit β_0 ja β_1 ovat tuntemattomat ja ne pitää estimoida havaintoaineiston perusteella. β_0 ja β_1 estimoidaan niin, että minimoidaan havaintojen y_i ja regressiosuoran erotuksien neliösumma.

Yksittäisen otoksen yhtälö voidaan kirjoittaa yhtälön (2.1) mukaisesti

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{2.14}$$

Tällöin pienimmän neliösumman kriteeri on

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2 \tag{2.15}$$

Pienimmän neliösumman estimaattorien eli $\hat{\beta}_0$ ja $\hat{\beta}_1$ täytyy toteuttaa yhtälöt

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0 \tag{2.16}$$

ja

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) x_i = 0 \tag{2.17}$$

Yksinkertaistamalla yhtälöt (2.16) ja (2.17) saadaan yhtälöt

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{2.18}$$

ja

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \tag{2.19}$$

Ratkaisemalla yhtälöt (2.18) ja (2.19) saadaan

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.20}$$

[1] engl. ordinary least squares, OLS

ja

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{2.21}$$

missä

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{ja} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.22}$$

ovat muuttujien y_i ja x_i keskiarvot.

Tällöin β_0 ja β_1 yhtälöissä (2.20) ja (2.21) ovat vakion ja kulmakertoimen pienimmän neliösumman estimaattorit.

Sovitettu yksinkertainen lineaarinen regressiomalli on silloin muotoa

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.23}$$

Yhtälö (2.23) antaa piste-estimaatin muuttujan y keskiarvolle tietyllä muuttujan x arvolla.

Koska yhtälön (2.21) nimittäjä on muuttujan x_i korjattu neliösumma ja osoittaja on muuttujien x_i ja y_i ristitulon neliösumma, nämä voidaan kirjoittaa kompaktimmassa esityksessä seuraavasti

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2.24}$$

ja

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x}) \tag{2.25}$$

Tällöin voimme kirjoittaa yhtälön (2.21) lyhemmässä muodossa

(2.26) $$\hat{\beta}_1 = \frac{S_{xy}}{S_{xy}}$$

Havaitun arvon y_i ja sitä vastaavan sovitetun arvon $\hat{y}_i$ välistä erotusta kutsutaan residuaaliksi. Matemaattisesti kirjoitettuna i:nnes residuaali on

(2.27) $$\epsilon_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, 2, \ldots, n$$

Residuaaleja voidaan käyttää mallin sopivuutta tarkasteltaessa. Niiden avulla voidaan selvittää, toteutuuko mallin oletukset, kuten se, että residuaalit eivät saa olla riippuvaisia prediktorin arvoista ja että niiden keskiarvo on nolla.

Usean selittävän muuttujan regressiomallin sovittaminen tapahtuu hyvin vastaavalla tavalla. Matriisimuodon yhtälön (2.6) kirjoitustapaa mukaillen mallin sovittaminen tapahtuu niin, että pyritään löytämään pienimmän neliösumman estimaattorien vektori $\hat{\boldsymbol{\beta}}$, joka minimoi yhtälön

(2.28) $$S(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$

Koska $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$ on 1×1 -matriisi eli skalaari, ja matriisien laskusääntöjen mukaan sen transpoosi $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{X}'\boldsymbol{\beta}$ on sama skalaari, yhtälön (2.28) voi kirjoittaa myös muodossa

(2.29) $$S(\hat{\boldsymbol{\beta}}) = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}$$

Pienimmän neliösumman estimaattorien täytyy toteuttaa yhtälö

(2.30) $$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0}$$

ja tämä yksinkertaistuu lyhempään muotoon

(2.31) $$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

Yhtälön (2.31) ratkaisu löytyy, kun sen molemmat puolet kerrotaan matriisin $\boldsymbol{X}'\boldsymbol{X}$ käänteismatriisilla. Tällöin mallin sovittamisen kannalta oleelliset $\boldsymbol{\beta}$:n pienimmän neliösumman estimaattorit saadaan yhtälöstä

(2.32) $$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

edellyttäen, että käänteismatriisi $(\boldsymbol{X'X})^{-1}$ on olemassa. Matriisi $(\boldsymbol{X'X})^{-1}$ on aina olemassa, jos selittävät muuttujat ovat lineaarisesti riippumattomat, mikä toteutuu silloin, kun yksikään matriisin $\boldsymbol{X}$ sarake ei ole lineaarikombinaatio muista sarakkeista.

Sovitettujen arvojen $\hat{y}_i$ vektori on muotoa

$$\boldsymbol{X\hat{\beta}} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = \boldsymbol{Hy} \tag{2.33}$$

jossa $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$ on $n \times n$ -matriisi, jota kutsutaan myös nimellä tasoittajamatriisi tai hattumatriisi[2]. Tämä matriisi on oleellinen regressioanalyysissä, sillä se projektoi havaittujen arvojen vektorin sovitettujen arvojen vektoriksi.

Kuten aiemmin on määritelty, residuaali eli jäännöstermi ϵ_i on havaitun arvon y_i ja sitä vastaavan sovitetun arvon $\hat{y}_i$ välinen erotus, eli $\epsilon_i = y_i - \hat{y}_i$. Tällöin voidaan kirjoittaa matriisimuodossa n residuaalia seuraavasti

$$\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{\hat{y}} \tag{2.34}$$

Yhdistämällä aiemmin esiteltyjä yhtälöitä, voidaan residuaalivektoria $\boldsymbol{\epsilon}$ esittää myös muilla tavoin, kuten

$$\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X\hat{\beta}} = \boldsymbol{y} - \boldsymbol{Hy} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \tag{2.35}$$

2.3.1 Regressiomallin merkitsevyydestä

Sovitetun mallin tarkastelussa täytyy ottaa huomioon myös regressiomallin merkitsevyys. Tilastollisella merkitsevyydellä tarkoitetaan yksinkertaisen regressiomallin tilanteessa sitä, että mallin antama kulmakerroin β_1 selittävälle muuttujalle x ei ole nolla valitulla luottamustasolla. Yksinkertaisessa lineaarisessa regressiossa se tarkoittaa kahden hypoteesin, H_0 ja H_1, testaamista.

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0 \tag{2.36}$$

Testaaminen aloitetaan oletuksella, että H_0 pitää paikkansa, eli muuttujilla x ja y ei ole lineaarista yhteyttä. Tämä tarkoittaa sitä, että muuttujan x arvon tunteminen ei anna lisäarvoa muuttujan y ennusteen muodostamiseen, vaan paras estimaatti muuttujalle y jokaisella x on y:n keskiarvo eli $\hat{y} = \bar{y}$, tai että muuttujien x ja y mahdollinen riippuvuus ei ole ainakaan lineaarista. Jos nollahypoteesia $H_0 : \beta_1 = 0$ ei voida hylätä, voidaan siis sanoa, että muuttuja x ei selitä muuttujan y vaihtelua.

Vaihtoehtoisesti, jos $H_0 : \beta_1 = 0$ hylätään ja siis päädytään hypoteesiin $H_1 : \beta_1 \neq 0$, voidaan sanoa, että x selittää muuttujan y vaihtelua. Tällöin niiden välillä

[2]engl. hat matrix

on lineaarinen riippuvuus. Vaikka päädytään hylkäämään nollahypoteesi ja toteamaan muuttujien välinen lineaarinen riippuvuus, se ei vielä tarkoita sitä, etteikö regressiomalli toimisi vielä paremmin, mikäli malliin lisättäisiin polynomitermejä.

Useamman selittävän muuttujan tilanteessa hypoteesit ovat

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{ja} \quad H_1 : \beta_j \neq 0 \text{ ainakin yhdelle } j \geq 1 \tag{2.37}$$

Nollahypoteesin hylkääminen tarkoittaa sitä, että ainakin jokin regressiomuuttujista $x_1, x_2, \ldots, x_k$ on tilastollisesti merkitsevä ja tällöin selittää vastemuuttujan y vaihtelua merkitsevästi.

Testisuureen laskeminen tapahtuu varianssianalyysin tavoin. Tämä tarkoittaa lineaarisessa regressiossa sitä, että vastemuuttujan y vaihtelu jaetaan osiin. Testisuureen laskemista varten havaintojen kokonaisneliösumma[3] SS_T jaetaan kahteen osaan, regressioneliösummaan[4] SS_R ja jäännösneliösummaan[5] SS_{Res}.

Edellä mainitut SS_T, SS_R ja SS_{Res} määritellään seuraavien yhtälöiden avulla

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i^2 - n\bar{y})^2 \tag{2.38}$$

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \tag{2.39}$$

ja

$$SS_{Res} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2. \tag{2.40}$$

Varianssianalyysihajotelma on siis muotoa

$$SS_T = SS_R + SS_{Res}. \tag{2.41}$$

Kuten yhtälöt (2.38), (2.39) ja (2.40) osoittavat, varianssianalyysihajotelmassa SS_T mittaa selitettävän muuttujan y kokonaisvahtelua. Se koostuu kahdesta osasta, jossa SS_R kuvaa sitä osuutta, jonka regressiomalli selittää eli se mittaa estimoidun muuttujan $\hat{y}$ arvojen vaihtelua, ja SS_{Res} kuvaa sitä osuutta, jota malli ei pysty selittämään eli jäännösten ϵ_i vaihtelua.

Kokonaisneliösumman regressioneliösumman ja jäännösneliösumman yhtälöt voidaan kirjoittaa myös matriisimuotoa apuna käyttäen. Tällöin ne voidaan kirjoittaa niin, että kokonaisneliösumma SS_T on muodossa

[3] engl. total sum of squares

[4] engl. regression sum of squares

[5] engl. residual sum of squares

$$SS_T = \boldsymbol{y'y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}, \tag{2.42}$$

regressioneliösumma SS_R on muodossa

$$SS_R = \hat{\boldsymbol{\beta}}'\boldsymbol{X'y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{2.43}$$

ja jäännösneliösumma SS_{Res} on muodossa

$$SS_{Res} = \boldsymbol{y'y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X'y}. \tag{2.44}$$

Testisuure F_0 saadaan $F_{k,n-k-1}$ -jakaumasta yhtälön

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}} \tag{2.45}$$

avulla, missä k on selittävien muuttujien lukumäärä ja n on havaintojen lukumäärä.

Testisuureen F_0 tulisi olla suuri, mikäli ainakin yksi regressiokerroin $\beta_j \neq 0$, $j = 1, 2, \ldots, k$, eli toisin sanoen nollahypoteesi voidaan hylätä, kun F_0 on tarpeeksi suuri.

Nollahypoteesia $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ voidaan testata F-testillä, ja se voidaan hylätä, mikäli

$$F_0 > F_{\alpha,k,n-k-1}, \tag{2.46}$$

missä α on riskitaso.

2.3.2 Regressiomallin selitysasteesta R^2

Regressiomallin hyvyyttä voidaan tarkastella myös selitysasteen R^2 avulla. Selitysasteella mitataan, kuinka suuri osuus muuttujan y vaihtelusta voidaan selittää regressiomuuttujien avulla, eli se saadaan kaavasta

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}, \tag{2.47}$$

missä SS_R, SS_T ja SS_{Res} määritellään aiemmin esiteltyjen yhtälöiden (2.39), (2.38) ja (2.40) avulla.

SS_T mittaa muuttujan y vaihtelua ilman regressiotermien huomioimista, kun taas virhetermien neliösumma SS_{Res} mittaa aineistoon jäävää vaihtelua regressiotermien vaikutuksen huomioon ottamisen jälkeen.

Koska $0 \leq SS_{Res} \leq SS_T$, siitä seuraa, että $0 \leq R^2 \leq 1$. Selitysasteen R^2 arvot lähellä arvoa 1 viittaavat siihen, että regressiomalli selittää suurimman osan muuttujan y vaihtelusta. Toisin sanoen tällöin regressiomallia voidaan käyttää ennustamaan muuttujan y arvoja. Selitysaste voidaan antaa myös prosenttilukuna, jolloin selitysasteen tulkinta siitä, kuinka suuren osan mallin selittävät muuttujat selittävät muuttujan y kokonaisvaihtelusta, on intuitiivisempi.

Selitysasteen R^2 käyttämisessä pitää kuitenkin olla varovainen, sillä regressiomalli ei aina silti ole sitä parempi, mitä suuremman arvon R^2 saa. Selitysastetta R^2 on nimittäin mahdollista kasvattaa lisäämällä regressiomalliin selittäviä tekijöitä, vaikka samaan aikaan mallin ennuste ei välttämättä parane aineiston ulkopuolisille havainnoille. Esimerkiksi jos aineistossa on vain yksi havaittu arvo muuttujalle y kutakin muuttujan x arvoa vastaavasti, $n-1$-asteen polynomi antaa "täydellisen"yhteensopivuuden ($R^2 = 1$) n:lle datapisteelle. Polynominen regressiomalli on siis mahdollista sovittaa niin, että ennustekäyrä sovittuu jokaisen pisteen kautta ja tällöin selitysaste on suurin mahdollinen, mutta mallin antamat ennusteet mallinnuksessa käytetyn aineiston ulkopuolisille havainnoille voivat olla huonoja.

Vaikka R^2 ei ikinä pienene, kun malliin lisätään termejä, se ei silti tarkoita, että uusi malli olisi parempi kuin yksinkertaisempi malli. Mitä useampi regressiotermi mallissa on, sitä enemmän mallintamisessa pitää laskea regressiokertoimia, mikä vähentää lopullisten vapausasteiden määrää. R-ohjelmiston lm-funktio laskee selitysasteen myös niin, että jäljelle jäävät vapausasteet huomioidaan tavallisen selityasteen lisäksi, ja tämä termi on tulosteessa nimellä "Adjusted R-squared"eli korjattu selitysaste. Korjattu selitysaste sopii mallien vertaamiseen erityisesti silloin, kun pyritään välttämään ylisovittamista[6]. Ylisovittamisen tilanteessa estimoitavien parametrien lukumäärä on liian suuri havaintoihin verrattuna, ja tilastollinen regressiomalli selittääkin aineiston satunnaisvaihtelua eikä prediktorien todellista yhteyttä vastemuuttujaan.

Teoksen Linear Regression Analysis (Lee & Seber, 2003) kappaleessa 12 korjattu selitysaste määritellään kaavalla

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n}{n-p}, \tag{2.48}$$

missä R^2 on tavallinen selitysaste, n on havaintojen lukumäärä ja p on estimoitavien parametrien eli samalla myös regressiokertoimien lukumäärä. Myös korjattu selitysaste on välillä [0,1].

Malli voidaan siis valita niin, että pyritään löytämään suurin korjattu selitysaste $\overline{R}^2$. Sen avulla voidaan verrata "täyttä mallia", jossa on kaikki mahdolliset regressiotermit, yksinkertaisempaan malliin. Vaikka regressiotermien lisääminen kasvattaa

[6]engl. overfitting

normaalia selitysastetta R^2, se voi pienentää korjattua selitysastetta $\overline{R}^2$, koska estimoitavien regressiokertoimien lukumäärä kasvaa ja regressiomallin vapausasteet vähenevät.

Teoksessa Introduction to Linear Regression Analysis (Montgomery, Peck & Vining, 2006) huomautetaan myös joistain väärinkäsityksistä, jotka liittyvät selitysasteeseen R^2. R^2:n arvo ei tarkoita regressiosuoran jyrkkyyttä, eli suuri selitysaste ei tarkoita, että regressiomallissa olisi jyrkkä kulmakerroin. Suuri R^2 ei myöskään tarkoita, että valittu lineaarinen malli on hyvä ennustemalli, sillä se voi olla suuri, vaikka yhteys onkin epälineaarinen. Regressiomalli ei välttämättä anna hyviä ennusteita, vaikka selitysaste olisikin suuri.

Kirjassa huomautetaan myös kausaalisuuden tulkinnasta, eli syy-seuraussuhteen pohtimisesta. Kausaalisuudesta seuraa aina jonkinlainen korrelaatio, mutta korrelaatiosta ei aina seuraa kausaalisuutta. Sen vuoksi, vaikka regressioanalyysin tuloksena muuttujien välillä löydetään yhteys, se ei silti anna vahvaa näyttöä niiden kausaalisuudesta. Muuttujien luonne tai niiden välinen aikajana voi tarkoittaa, että muuttujien välillä on kausaalisuutta, mutta regressioanalyysi ei suoranaisesti anna aihetta puhua syystä ja seurauksesta. Usein kuitenkin analyysin tekijä tuntee aiheen niin hyvin, että osaa laittaa selitettäväksi tekijäksi sen, jonka oletetaan riippuvan muista, riippumattomista tekijöistä eli selittäjistä.

Yksi haaste regressiomallinnuksessa on myös se, että selittävät muuttujat voivat olla yhteydessä toisiinsa. Erityisesti polynomisessa regressiossa vaarana on se, että selittävät muuttujat korreloivat keskenään, eli puhutaan tällöin multikollineaarisuudesta. Aiemmin määriteltiin, että parametrin β pienimmän neliösumman estimaattorit saadaan kaavalla (2.32) edellyttäen, että käänteismatriisi $(\boldsymbol{X'X})^{-1}$ on olemassa. Tämä edellyttää, että yksikään matriisin $\boldsymbol{X}$ sarake ei ole lineaarikombinaatio muista sarakkeista. Multikollineaarisuus voi kuitenkin aiheuttaa matriisin $\boldsymbol{X}$ huonovointisuutta[7]. Tämä voidaan välttää ortogonaalisella polynomiregressiolla. Teoksessa Linear Regression Analysis (Lee & Seber, 2003) kuitenkin huomautetaan, että regressiomatriisin $\boldsymbol{X}$ huonovointisuutta ilmenee yleensä vasta silloin, kun yritetään sovittaa kuudennen tai suuremman asteen polynomifunktiota. Polynomiregression sovittamista k:nnen asteen polynomille, kun $k < 6$, ei siis tämän huomion perusteella ole syytä välttää sen takia, että regressiomallin $\boldsymbol{X}$-matriisin huonovointisuus estäisi parametrin β pienimmän neliösumman estimaattorien laskemista.

[7]engl. ill-conditioning

3 Aineiston kuvaus

3.1 Myynti-ilmoitukset ja niihin liitetty tieto auton alkuperäisestä hinnasta

Autotalli.com-palvelusta saadussa datassa on 1.1.-31.12.2016 julkaistujen autojen myynti-ilmoitusten tietoja. Tähän työhön tarpeellisimmat tiedot olivat auton vuosimalli, jonka perusteella pääteltiin auton ikä, sekä mittarilukeman kilometrimäärä. Tietenkin myös mallin vasteena käytetty auton pyyntihinta on tärkeä muuttuja aineistossa.

Lisäksi ilmoituksesta poimittiin muita tietoja, joiden avulla selvitettiin auton alkuperäistä hintaa uutena. Tätä tietoa ei ilmoituksessa kerrottu suoraan, joten sen selvittämiseen tarvittiin Autotalli.comin käytössä olevaa uusien autojen hintoja sisältävää tietokantaa. Työssä yhdistettiin siis kahta eri tietokantaa, joista muodostettiin lopullinen aineisto.

Analyysissä käytetyssä aineistossa oli 34 068 havaintoa, ja tilastoyksikkönä toimii yksittäinen auton myynti-ilmoitus. Tähän tietoon on yhdistetty arvio kyseisen auton hinnasta uutena.

3.2 Aineiston rajaaminen ja muokkaus

Mittarilukeman kilometrimäärä on myyjän itse ilmoittama, joten siinä oli myös jonkin verran poikkeavuuksia, jotka olivat mahdollisesti kirjoitusvirheitä tai ehkä jopa tahallisesti virheellisiksi asetettuja. Aineistoon hyväksyttyjen ilmoitusten tiedoista rajattiin joitain kohteita pois poikkeavien ajokilometrien vuoksi. Koska tämä lopputyö koskee erityisesti käytettyjä autoja, kilometrimäärää rajoitettiin alhaalta niin, että mukaan hyväksyttiin vasta alkaen 500 kilometriä ajetut autot.

Toisaalta kilometrimäärän toisesta ääripäästä löytyi hyvin harvakseltaan luotettavia havaintoja enää 500 tuhannen kilometrin jälkeen. Ennustemallin käytettävyyden kannalta se yläraja riittää tässä lopputyössä, joten myös se rajasi joitain havaintoja pois mallintamisesta. Tämän rajapyykin jälkeen auton pyyntihintaan voi vaikuttaa niin yksilölliset tekijät, että havaintojen harvalukuisuuden takia tilastollinen malli voi olla epäluotettava. Autojen huolloilla voi olla vaikutusta sen hinnoittelussa, ja erityisesti siis vanhojen autojen kohdalla. Tarpeeksi paljon käytetyistä autoista myös osa poistetaan käytöstä, kuten vaikkapa erityisen kuluneet tai huonosti toimivat autot, joten jäljelle jäävät ja myyntiin laitetut autot eivät enää edusta sellaista satunnaisuutta, että niiden avulla laskettu auton arvon kehittyminen voitaisiin yleistää suurimpaan osaan autoista.

Koska auton alkuperäistä hintaa uutena ei tiedetä tarkasti, sitä arvioitiin täsmäämällä ilmoituksen auton tietoja toiseen tietokantaan, jolloin osa havainnoista tippui pois. Tämä tapahtui siitä syystä, että kaikkiin ilmoituksiin ei saatu täsmättyä tietoa uuden auton hinnasta. Autoja täsmättiin toisen tietokannan dataan käyttämällä ilmoi-

tuksen tietoja auton merkistä, mallista, vuosimallista, moottorin tilavuudesta, auton tehosta sekä vaihteiston ja polttoaineen tyypeistä.

Joihinkin ilmoituksiin saatiin täsmällinen tieto auton hinnasta uutena, mutta useampaan löydettiin useampi hinta-arvio. Näistä arvioista tämän työn aineistoon otettiin omiksi sarakkeiksi minimi ja maksimi. Lopulliseksi arvioksi auton hinnasta uutena otettiin näiden keskiarvo. Tällöin arvion virheen maksimaalinen matka molemmille puolille arviota on yhtä pitkä, eikä toisaalta ole syytä tehdä oletusta siitä, mihin kohtaan hinta-arvion haitaria auton hinta oikeasti sijoittui täysin uutena.

Auton alkuperäisen hinnan arvion maksimaalista potentiaalista virhettä käytettiin myös aineiston rajaamiseen. Sen avulla poistettiin niitä havaintoja, joissa arvion minimin ja maksimin väli on suuri. Mallintamiseen käytettyyn aineistoon hyväksyttiin vain ne havainnot, joissa maksimaalinen potentiaalinen virhe voi olla enintään 5 prosenttia hinta-arviosta. Näin voidaan varmistua siitä, että hinta-arvio auton hinnasta uutena on hyvin lähellä todellista. Kun rajaus on prosentuaalinen osuus, se sallii euroissa mitattuna suuremman virheen kalliille autoille, mutta pienemmän vaihteluvälin edullisemmissa autoissa.

Lisäksi aineistosta poistettiin myös selkeästi virheelliset pyyntihinnat, ja lopulliseen aineistoon jäivät pyyntihinnaltaan 190 eurosta 99 000 euroon olevien autojen ilmoitukset. Tämä väli on riittävä sen kannalta, minkä arvoisiin käytettyihin autoihin lopputyön tuloksia halutaan soveltaa. Pois jäivät siis sellaiset autot, jotka on ilmoitettu esimerkiksi nollahinnalla, tai toisaalta harvinaisen suurella pyyntihinnalla ilmoitetut. Aineiston rajaaminen ei siis aiheuta ongelmia sen suhteen, onko tulokset yhä yleistettävissä suurimpaan osaan käytettyjä autoja.

3.3 Lopullisen aineiston kuvailu

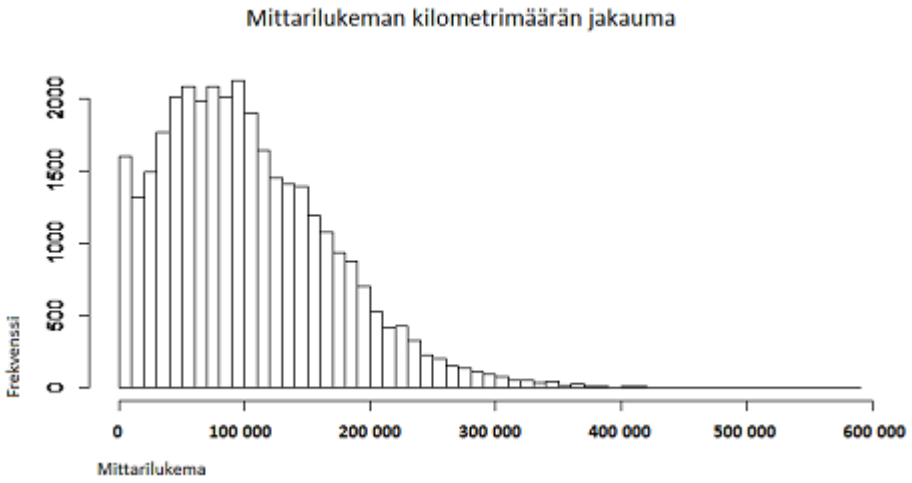
Lopulliseen aineistoon jäi rajaamisen jälkeen 34 068 auton myynti-ilmoitusta. Havaintoja on siis hyvin paljon ja niissä on mukana monta erilaista automerkkiä ja -mallia.

Tämän työn lopun liitteenä olevassa kuvassa on hajontakuvio, josta näkee analyysissä käytettävien muuttujien keskinäistä hajontaa. Mallinnuksessa käytetyt muuttujat ovat auton ikä vuosissa ja auton mittarissa oleva ajettu kilometrimäärä. Iän jakauma on alla olevassa taulukossa ja mittarilukeman jakauma on histogrammina.

Taulukko 3.1. Havaintojen lukumäärä, eri ikäiset autot

0 v	1 v	2 v	3 v	4 v	5 v	6 v	7 v	8 v	9 v	10 v	11 v
1224	2205	3367	4388	4947	5950	3448	2830	2242	1557	1471	439

Yleisimmät automerkit tässä aineistossa ovat järjestyksessään Volkswagen (6 276 havaintoyksikköä), Audi (4 657 havaintoyksikköä), Mercedes-Benz (2 801 havaintoyksikköä), Kia (2 507 havaintoyksikköä), Ford (2 329 havaintoyksikköä), Toyota (1 864 havaintoyksikköä) ja Volvo (1 839 havaintoyksikköä). Mukana oli myös harvinaisimpia automerkkejä, joista alle kymmenen havaintoa oli automerkeistä Abarth, Lancia ja Lada. Kaiken kaikkiaan eri automerkkejä oli yhteensä 37.



Kuvio 3.1. Auton mittarilukeman jakauma histogrammina

Yksittäisistä automalleista eniten tässä aineistossa oli järjestyksessään seuraavia: Audi A6 (1 555 havaintoyksikköä), Audi A4 (1 467 havaintoyksikköä), Skoda Octavia (1 145 havaintoyksikköä), Volkswagen Polo (1 041 havaintoyksikköä) sekä Volkswagen Golf (997 havaintoyksikköä). Kaiken kaikkiaan aineistossa oli 338 eri automallia, joista osaa oli vain yksi tai kaksi havaintoyksikköä.

Moottorin tilavuuden vaihteluväli oli välillä 875 cm^3 - 5 654 cm^3. Sen keskiarvo oli 1816 cm^3, ja useampi kuin neljä viidestä havainnosta oli alle 2 000 cm^3.

Aineistossa oli mukana 16 934 automaattivaihteista autoa ja 17 134 manuaalivaihteista autoa. Polttoaineen mukaan aineisto jakautui niin, että 17 855 oli bensakäyttöistä autoa, ja 16 213 aineiston autoista kulkee bensalla.

4 Aineiston analyysi

4.1 Analyysin vaiheet

Analyysin ensimmäisenä vaiheena on selvittää, voiko aineistoon käyttää polynomista regressiota. Myös tilastollisesti merkitsevien muuttujien selvittäminen on tärkeää, mikä tarkoittaa tässä työssä sitä, että pyritään selvittämään polynomien asteet. Mallien vertailun avulla selvitetään, mikä tilastollinen malli lopulta on paras auton arvon laskemisessa käytettävään kaavaan.

Mallintamisen jälkeen regressiomallista selvitetään selittävien muuttujien kulmakertoimet. Näiden avulla voidaan muodostaa auton arvon alenemisen laskemiseen tarvittavaa laskukaava, jolla voidaan laskea auton arvon alenemaa käytön ja iän myötä sen alkuperäisestä hinnasta uutena. Koska analyysimenetelmäksi valikoitui parametrinen menetelmä, saatu tulos toimii hyvin yleistyksenä auton arvon laskukaavalle.

Regressiomallin jäännöstermien tutkiminen automerkeittäin auttaa tutkimaan sitä, onko automerkkien välillä eroa auton arvon alenemisessa. Yleisen kaavan löytämisen jälkeen mallinnetaan myös muutamille yleisimmille automerkeille omat kaavansa. Niiden avulla voidaan tutkia, miten eri automerkkien välillä auton arvon aleneminen vaihtelee. Taulukon esimerkin avulla voi arvioida eri automerkkien välisiä eroja auton arvon aleneman suuruudessa.

Aineiston analyysissä käytettiin R-ohjelmistoa. R-ohjelmistossa on lm-funktio, jonka avulla voidaan sovittaa lineaarinen malli. Funktio sallii myös polynomisen prediktorin sovittamisen. Funktion avulla tilastollisesta mallista selviää esimerkiksi prediktorien kertoimet, niiden luottamusvälit ja tilastollinen merkitsevyys. Lisäksi mallille lasketaan sen selitysaste, eli se, kuinka hyvin malli sopii aineistoon ja kuinka hyvin prediktorit selittävät vastemuuttujan saamaa vaihtelua.

R-ohjelmistolla voidaan myös piirtää tilastollisen mallin sopivuutta tarkastelevia kuvia, kuten jäännöskuvioita. Näiden avulla voidaan varmistua siitä, että malli todella sopii aineistoon ja oletukset täyttyvät. Esimerkiksi jäännökset eli residuaalit eivät saa olla riippuvaisia sovitteen tai selittävien muuttujien arvoista, eli kuvassa ei pitäisi näkyä mitään systemaattista kuviota. Lisäksi eri malleja voidaan verrata ANOVA-testillä, jolla testataan eri mallien välistä tilastollista merkitsevyyttä. Sen avulla voidaan pohtia, mikä malli on paras ilmiön selittämiseen, mutta toisaalta tarpeeksi yksinkertainen.

4.1.1 Polynomisen mallin sovittaminen

Regressiomallin selittäviksi tekijöiksi laitettiin auton ikä ja mittarilukeman mukaan autolla ajetut kilometrit. Vastemuuttuja oli prosenttiluku siitä, kuinka suuri osuus auton arvosta uutena oli jäljellä myyntihetkellä. Tämä oli saatu vertaamalla auton myynti-ilmoituksessa ilmoitettua hintaa siihen hintaan, joka oli ilmoituksen tiedoilla täsmäytetty toisen tietokannan tietoihin autojen hinnoista uutena.

Aineiston hajontakuvio on tämän työn liitteenä. Hajontakuviosta näkee silmä-

määräisesti, että aineistoon ei sovi ainakaan tavallinen lineaarinen regressiosuora. Lähdetään siis selvittämään, sopiiko aineistoon polynominen regressio, ja monennenko asteen polynomi prediktoriksi sopii.

R-ohjelmiston lm-funktion sisälle voidaan prediktoriksi laittaa toinen funktio, poly, jonka avulla selittävän tekijän vaikutusta selittävään tekijään tutkitaan polynomin avulla. Tässä työssä molemmat selittävät tekijät, eli auton ikä ja mittarilukema, laitetaan poly-funktion sisälle, joten tutkitaan niiden molempien vaikutusta vastemuuttujaan polynomisena.

Poly-funktioon merkitään, monennenko asteen polynomina selittävän muuttujan vaikutusta tarkastellaan. Ensimmäisen asteen polynomi on suora, eikä siis eroa tavallisesta lineaarisesta mallista. Toisen asteen polynomi on paraabeli, joten selittävän muuttujan vaikutusta vastemuuttujaan tarkastellaan käyrän avulla, ja tässä käyrässä voi olla enintään yksi derivaatan nollakohta.

Vastaavasti kolmannen asteen polynomilla voidaan tarkastella monimutkaisempaa muuttujien välistä yhteyttä käyrällä, jossa voi olla enintään kaksi derivaatan nollakohtaa. Tämä mahdollistaa jo jonkin verran erilaisen riippuvuuden tarkastelua, kuten vaikkapa sitä, että kilometrien vaikutus auton arvoon on aluksi nopeasti sitä vähentävää, sitten auton mittariin kertyvät kilometrit vähentävät auton arvoa taas hieman hitaammin ja lopulta vaikutus auton arvoon on taas entistä suurempaa eli sen vaikutus auton arvoon laskevasti kiihtyy taas.

Mallintaminen aloitettiin sovittamalla aineistoon useampi regressiomalli, joiden erot olivat siinä, monennenko asteen polynomi selittäväksi tekijäksi sallittiin. Malleja vertailtiin regressiokertoimien tilastollisen merkitsevyyden perusteella sekä testaamalla eri mallien tilastollista merkitsevyyttä ANOVA-testin avulla, eli hierarkisia malleja vertailtiin keskenään. Näin pyrittiin löytämään sellainen malli, jossa jokainen mukana oleva termi on tilastollisesti merkitsevä, eikä polynomifunktion astelukua kasvateta turhan suureksi. Analyysin edetessä huomioitiin myös mallin selitysaste, mikä auttaa arvioimaan sitä, kuinka hyvin eri mallit selittävät auton arvon alenemaa. Korjatun selityasteen avulla arvioitiin myös sitä, ettei polynomisen regressiomallin kohdalla päädytä valitsemaan liian suuren asteluvun polynomifunktiota selittäväksi tekijäksi.

Lopulta parhaaksi malliksi valikoitui sellainen polynominen regressiomalli, jossa auton ikä oli prediktorina neljännen asteen polynomin muodossa, ja mittarilukema oli mukana kolmannen asteen polynomin muodossa. Ero esimerkiksi hieman yksinkertaisempaan malliin, jossa molemmista selittävistä tekijöistä oli prediktorina vain kolmannen asteen polynomi, oli ANOVA-testin mukaan tilastollisesti merkitsevä ($p<0.01$), eli ottamalla mukaan iän neljännen asteen polynomifunktio saavutetaan tilastollisesti merkitsevä parannus kolmannen asteen polynomifunktioon verrattuna. Kuitenkaan hieman monimutkaisempaan malliin, jossa sekä auton ikä että mittarilukema olisi neljännen asteen polynomina, ei ollut tilastollisesti merkitsevä ($p = 0.2775$).

Edellä mainitun syyn lisäksi lopulliseen regressiomalliin päädyttiin siitä syystä, että kyseisessä mallissa oli varsin hyvä selitysaste, 0.89. Saatu tilastollinen malli selittää hyvin vastemuuttujan vaihtelua, eli auton iällä ja kilometreillä on tilastollisesti merkitsevää vaikutusta auton arvoon. Selitysaste on suunnilleen sama, eli pyöristet-

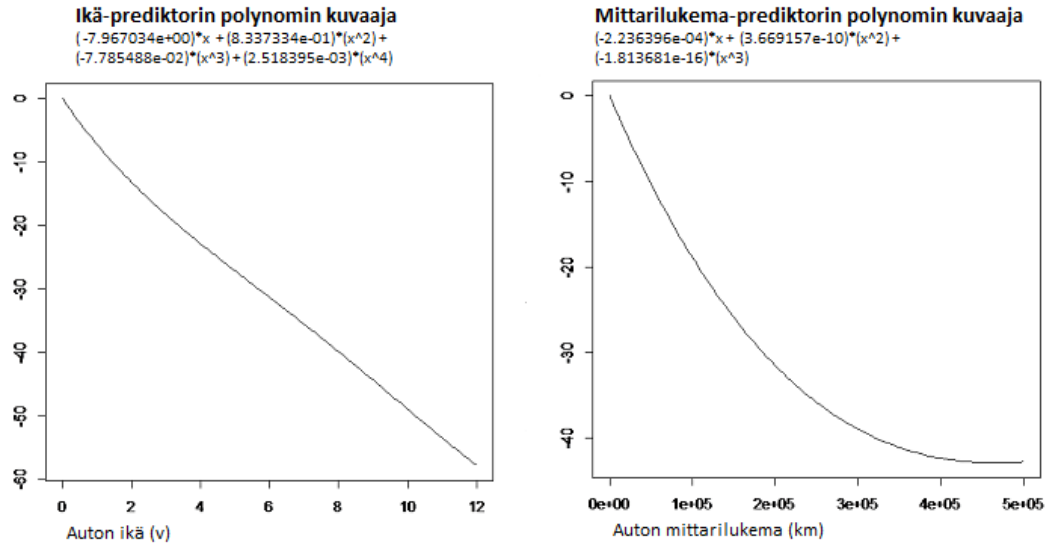
Taulukko 4.1. Sovitettu malli, R-tuloste

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
intercept	1.02e+02	1.82e-01	562.8	<2e-16
age	-7.97e+00	2.30e-01	-34.6	<2e-16
age^2	8.34e-01	8.32e-02	10.0	<2e-16
age^3	-7.79e-02	1.17e-02	-6.65	2.93e-11
age^4	2.52e-03	5.471e-04	4.60	4.18e-06
km	-2.24e-04	3.80e-06	-58.83	<2e-16
km^2	3.67e-10	2.25e-11	16.31	<2e-16
km^3	-1.81e-16	3.81e-17	-4.76	1.98e-06

tynä 0.89, myös monimutkaisemmassa mallissa, jossa molemmat selittävät tekijät ovat viidennen asteen polynomina. Sen pohjalta valittu malli on tarpeeksi hyvä ja siinä jokainen prediktori on tilastollisesti merkitsevä.

4.1.2 Mallin tarkastelu

Edellisessä kohdassa muodostetusta tilastollisesta regressiomallista voidaan erotella iän ja kilometrien polynomit erikseen, ja nämä polynomit voidaan piirtää kuvaajiksi. Näistä kuvaajista näkee regressiomallin laskemat prediktorien vaikutukset auton arvoon.



Kuvio 4.1. Auton iän ja mittarilukeman regressiokertoimien polynomifunktiot

Polynomeista voidaan tarkastella myös sen derivaattaa. Derivaatta on käyrän kulmakerroin kussakin kohdassa. Esimerkiksi iän käyrä ei ole kovin mutkikas, vaan on hyvin lähellä suoraa. Iän polynomin derivaatta onkin melko sama koko tarkasteltavalla välillä [0, 11]. Tämän vanhempiin autoihin tätä tilastollista mallia ei voi käyttää, koska aineistossa ei ollut tätä vanhempia autoja. Derivaatan nollakohta paljastaa, mil-

loin käyrän suunta vaihtuu. Avoimella välillä 0-11 tämän polynomin derivaatalla ei ole nollakohtaa, vaan sillä on vain yksi nollakohta noin kohdassa $ik\ddot{a} = 15.9$. Auton ikä laskee sen arvoa siis melko tasaisesti koko ajan.

Mittarilukeman vaikutus on tässä regressiomallissa kolmannen asteen polynomi, joka kuitenkin kuvassa muistuttaa enemmänkin paraabelin puolikasta. Tämän perusteella siis auton kilometrit vaikuttavat auton arvoon eniten alentavasti silloin, kun autolla on ajettu vain vähän. 300 tuhannen kilometrin jälkeen käyrä on melkein samalla tasolla kuin 400 ja 500 tuhannen kilometrin kohdalla, joten sen rajapyykin jälkeen auton mittariin kertyvät kilometrit eivät tämän mallin perusteella näytä juurikaan vaikuttavan auton arvoon.

Toisin kuin prediktorin $ik\ddot{a}$ kohdalla, mittarilukeman polynomille löytyy derivaatan nollakohta tarksteltavalla välillä. Tällä polynomilla on kaksi nollakohtaa, kohdissa $km = 465250$ sekä $km = 883446$. Näistä siis ensimmäinen osuu tarkasteltavalle välille, minkä mukaan mittarilukemaan kertyvien kilometrien vaikutus kääntyy auton arvoa nostavaksi noin 465 tuhannen kilometrin jälkeen. Kannattaa kuitenkin muistaa, että aineistossa oli hyvin vähän havaintoja enää näin suurissa kilometrimäärissä. Useat näin paljon ajetut autot eivät päädy enää uudestaan myyntiin. Lisäksi näin paljon ajetut autot ovat usein myös vanhoja, joten niiden kohdalla ikä alentaa yhä auton arvoa.

Kannattaa kuitenkin huomioida, että tämä tilastollinen regressiomalli on luotu tarkastelemalla sekä ikää että kilometrejä samaan aikaan selittävissä tekijöissä, joten auton arvon määrittämisessä on syytä ottaa huomioon molemmat tekijät.

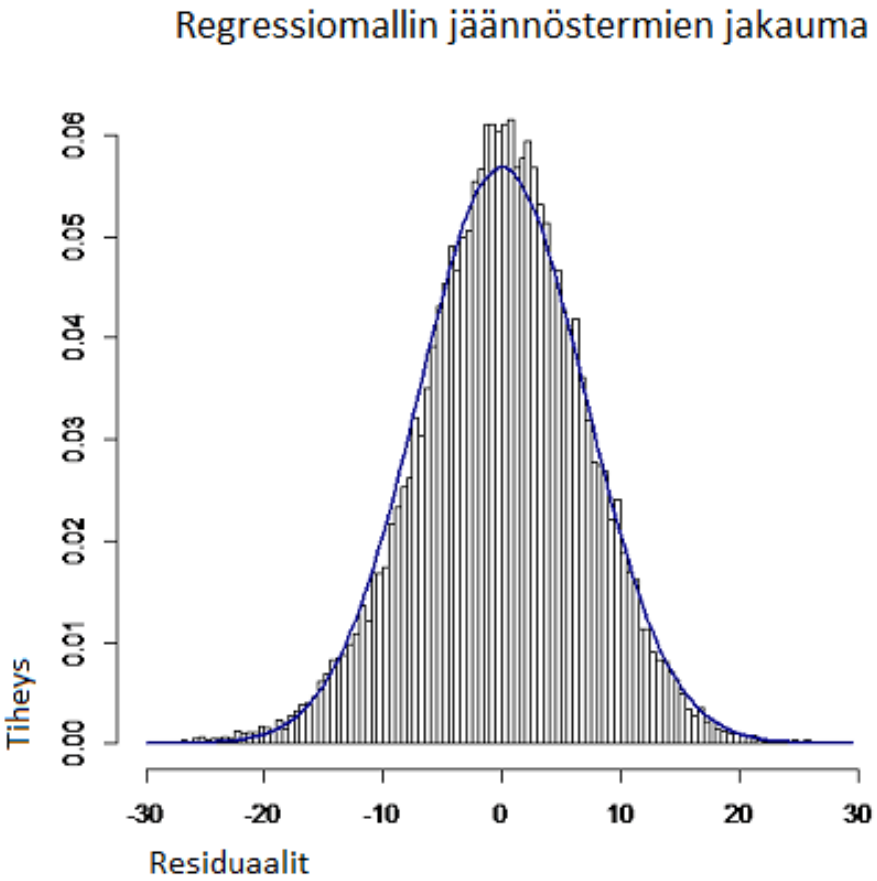
Muodostetun tilastollisen mallin jäännösten histogrammista (4.2) näkee, että ne ovat jakautuneet normaalijakauman mukaan, kuten onnistuneessa regressiomallinuksessa yleensä onkin. Kuvaan on piirretty myös normaalijakauman tiheysfunktio, jossa keskiarvo on nolla ja varianssi on sama kuin residuaalien varianssi. Nämä piirtyvät päällekkäin nätisti niin kuin regressiomallinnuksen oletukset odottavatkin.

Residuaalien ensimmäinen kvartiilikohta on -4,44 ja kolmas 4,61, kun taas mediaani on luonnollisesti lähellä nollaa eli 0,14. Ala- ja yläkvartiilien välille jää keskimmäinen puolikas, eli residuaalien keskimmäinen puolikas sijaitsee välillä [-4,44; 4,61]. Tämä tarkoittaa sitä, että puolet ennusteista erosi enintään 4,6 prosenttiyksikköä havaitusta. Suurin osa ennusteista oli siis alle viiden prosenttiyksikön päästä siitä, kuinka monta prosenttia auton alkuperäisestä arvosta todellisuudessa oli jäljellä. Yli 90 prosenttia residuaaleista jäi välille [-12,12], joten melkein kaikille mallin antama ennuste erosi alle 12 prosenttiyksikköä havaitusta arvosta.

Residuaalien tarkastelu osoittaa sen, mitä myös mallin saama hyvä selityaste kertoo, että valitut prediktorit auton ikä ja mittarilukema selittävät hyvin vastemuuttujan vaihtelua, eli auton nykyistä arvoa alkuperäiseen hintaan verrattuna.

Jäännöskuvioita tarkastellessa on hyvä piirtää myös jäännökset sovitteen suhteen. Mikäli olettamukset ovat kunnossa, mitään systemaattista kuviota ei pitäisi näkyä. Jäännösten tulisi olla satunnaisesti nollan molemmilla puolilla, eikä niiden hajonta saisi olla riippuvainen sovitteen arvoista, eli esimerkiksi hajonta ei saisi kasvaa sovitteen arvon kasvaessa.

Koska aineistossa oli yli 34 tuhatta havaintoa, koko aineiston jäännösten piirtäminen tuottaa hyvin epämääräisen kuvan. Tarkastellaan siis esimerkin vuoksi vain yhtä

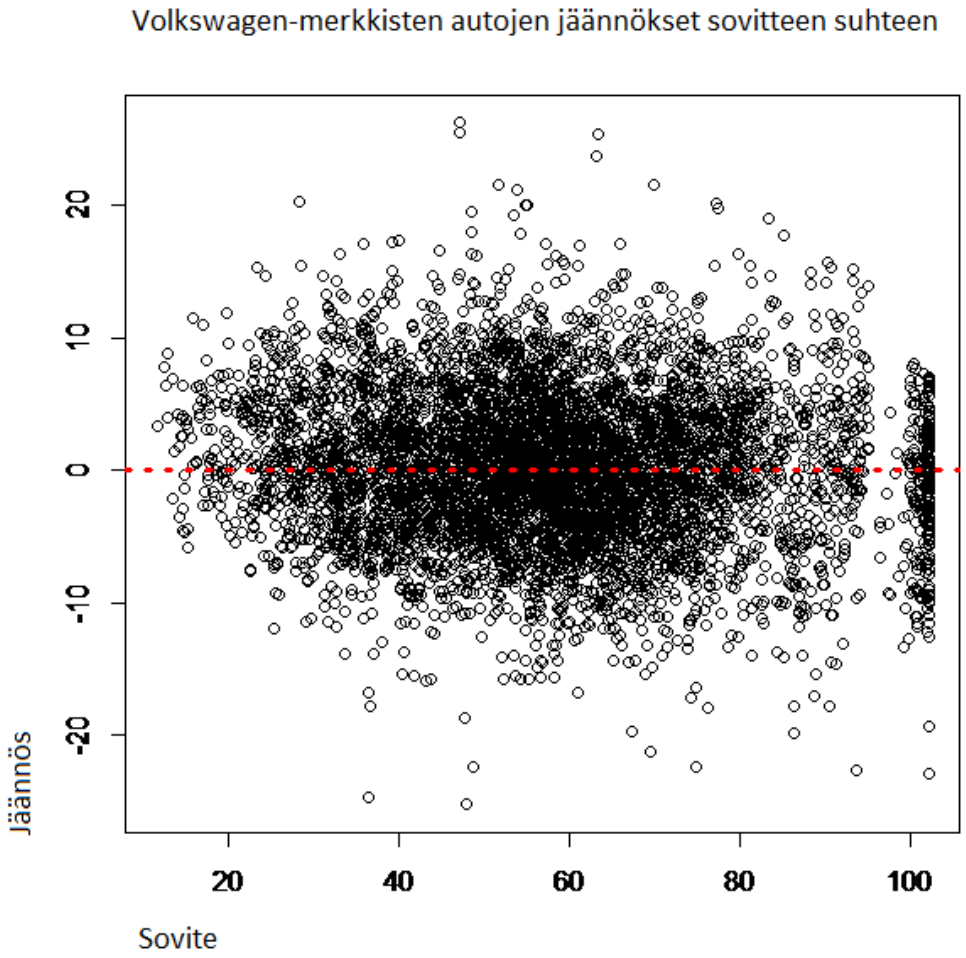


Kuvio 4.2. Regressiomallin jäännökset

automerkkiä, ja valitaan tarkastelun alle aineiston yleisin automerkki eli Volkswagen. Kuvaaja (4.3) näyttää hyvältä, sillä jäännösten varianssi ei näytä kasvavan sovitteen mukana. Kuvaajasta näkee myös sen, että isolle ryhmälle autoja sovite oli lähellä arvoa 100, eli mukana oli iso joukko uusia, lähes käyttämättömiä autoja, mutta toisaalta sovitteen arvon 95 kohdalla ei ollut montaa havaintoa. Tämä ei kuitenkaan ole merkki siitä, että tilastollinen regressimalli olisi huono, vaan se johtuu aineiston hajonnasta. Melko uusia, mutta jo vähän käytettyjä autoja ei siis ollut montaa, vaan havaintoja on ollut sitten taas runsaammin jo hieman enemmän käytetyistä. Mallin hyvyyden tarkastelun kannalta siis tämä diagnostiikkakuvio on kunnossa, eikä ole syytä ryhtyä muokkaamaan mallia, vaan oletuksien voidaan katsoa olevan kunnossa ja analyysin tuloksia voidaan käyttää.

Jäännöskuvioiden lisäksi on hyvä tarkastella myös vasteen arvoja sovitteen suhteen. Pisteparven pitäisi silloin olla keskittynyt sellaisen suoran ympärille, jonka kulmakerroin on yksi. Tarkastellaan taas tässä tapauksessa pelkästään aineiston yleisintä automerkkiä eli Volkswagenia, jotta saadaan rajattua havaintojen lukumäärää taas sellaiseksi, että kuviossa ei ole koko aineiston 34 tuhatta havaintoa.

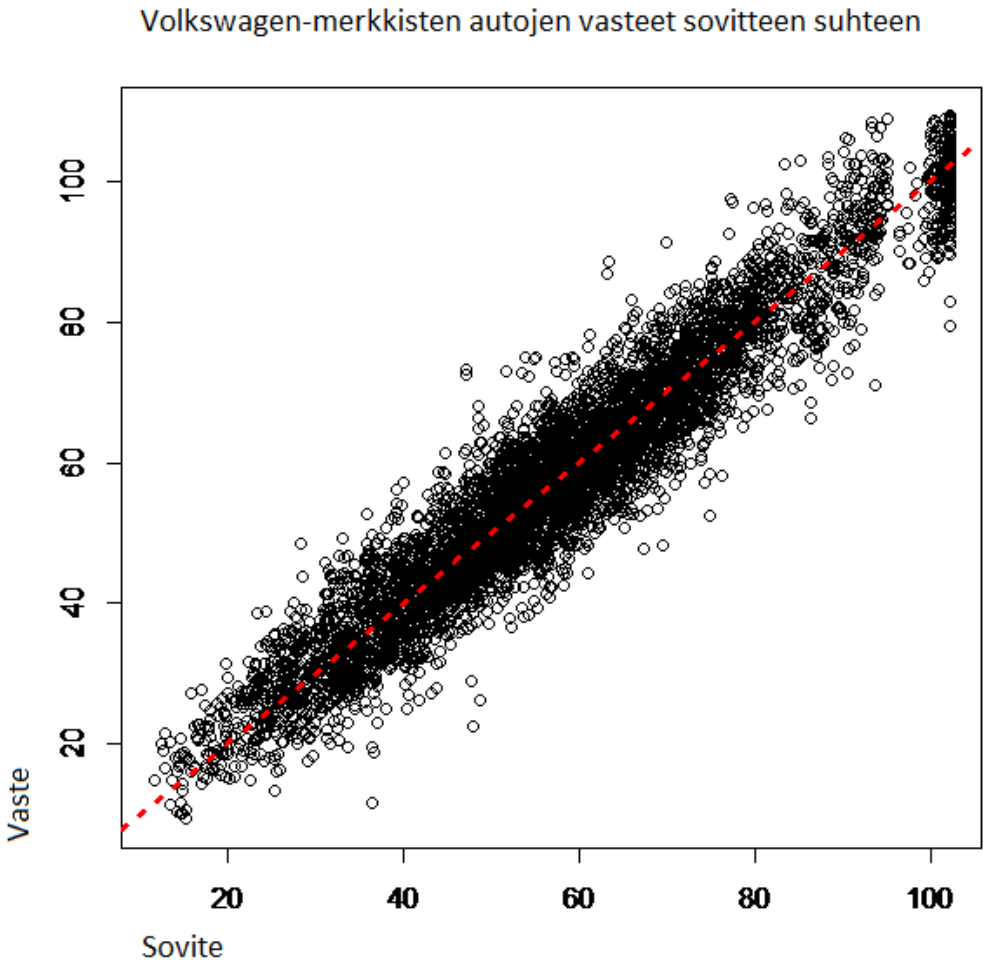
Kuvaaja (4.4) näyttää hyvältä, eli vasteen arvot sovitteen suhteen näyttävät olevan



Kuvio 4.3. Mallin jäännökset sovitteen suhteen, Volkswagen

sellaiset, kuin oletuksien mukaan pitääkin.

Regressiomallinnusta kokeiltiin myös ortogonaalisena polynomimallina, koska prediktorit korreloivat keskenään ja ortogonaalinen polynomimalli sovittaa nimensä mukaisesti regressiotermit ortogonaalisesti. Tilastollisen mallin selitysaste pysyi kuitenkin samana, eikä ortogonaalinen regressiomallinnus myöskään parantanut residuaaleja. Ortogonaalisen polynomin sovittaminen ei myöskään osoittanut tarvetta muokata regressiotermien polynomien astetta, vaan edelleen tilastollinen malli oli parhaimmillaan, kun siinä on iän vaikutus neljännen asteen polynomina ja mittarilukeman vaikutus kolmannen asteen polynomina. Edellä mainituista seikoista johtuen analyysissä pysyttiin tavanomaisessa polynomisessa regressiomallissa, joka oli alunperin valittu analyysimenetelmäksi. Tällöin regressiokertoimet ovat helpommin tulkittavissa siihen tapaan kuten tutkimuskysymys oli muodostettu ja auton arvoa alentavien tekijöiden piirtäminen on mahdollista kuten kuvissa (4.1). Ortogonaalisen polynomisen regressiomallin sovittaminen ei siis tuonut lisäarvoa alkuperäisen tutkimuskysymyksen analyysissä.



Kuvio 4.4. Vasteen arvot sovitteen suhteen, Volkswagen

Saatua regressiomallia tarkasteltiin myös sen suhteen, toimiiko se samalla tavalla alkuperäiseltä hinnaltaan hyvin arvokkaille autoille kuin mitä keskivertohintaisille tai normaalia edullisimmille. Tätä tarkastelua tehtiin silmämääräisesti katsomalla kuvaa, jossa residuaalit piirrettiin samaan kuvaan auton alkuperäisen hinnan kanssa. Näille kahdelle muuttujalle laskettiin myös niiden välinen korrelaatio, joka jäi melko alhaiseksi tasolle 0,1. Residuaaleja tarkastellessa ei löydetty syytä pitää auton alkuperäistä hintaa regressiomallia parantavana muuttujana, joten ainakaan jäännösten perusteella ei ole syytä olettaa, että tämä tilastollinen malli pitäisi sovittaa uudestaan eri hintaluokkien autoille.

Auton hintaluokan vaikutusta auton arvon alenemaan voi tutkia kuitenkin myös niin, että auton alkuperäinen hinta laitetaan mukaan tilastollisen mallin regressiotermiksi. Alkuperäinen tutkimuskysymys oli tutkia auton arvon alenemaa yleisellä tasolla liittyen vain auton ikään ja ajettuihin kilometreihin, joten sinällään auton alkuperäisen hinnan vaikutus ei ole tässä lopputyössä oleellinen. Kuitenkin se on mielenkiintoinen osa-alue tutkimuksessa, joten auton alkuperäinen hinta otettiin kokeilun

vuoksi mukaan regressiomalliin selittäväksi tekijäksi.

Kun auton arvioitu alkuperäinen hinta on mukana regressiomallissa, korjattu selitysaste nousee vain hieman. Alkuperäisessä mallissa se on tarkalleen ottaen 0.8854, kun laajemmassa mallissa se on 0.8903, joten auton alkuperäinen hinta ei juurikaan paranna regressiomallia, mutta tuo siihen kuitenkin hieman lisää arvoa. Koska auton alkuperäisen hinnan regressiokerroin on selkeästi tilastollisesti merkitsevä, voidaan sitä käyttää auton arvon aleneman tulkitsemiseen eri hintaluokkien autoille: Auton alkuperäisen hinnan regressiokerroin on tässä tilastollisessa mallissa positiivinen, mikä tarkoittaa sitä, että alkuperältään kalliimpien autojen arvo alenee käytön ja iän myötä keskimäärin vähemmän kuin halvempien autojen. Kannattaa kuitenkin muistaa, että vasteena oli prosenttiosuus alkuperäisestä hinnasta, mikä tarkoittaa tässä tapauksessa sitä, että auton arvo alenee hintavien autojen joukossa keskimäärin vähemmän juurikin prosenteissa laskettuna. Kun auton arvon aleneman muuttaakin rahayksiköihin, voi uutena enemmän maksaneen auton arvo laskea euroissa laskettuna enemmän kuin vähemmän maksaneen, vaikka regressiomallissa auton alkuperäinen hinta olisikin merkitsevä tekijä positiviisella kulmakertoimella, kun ennustetaan auton arvon säilymistä prosenteissa.

4.2 Auton arvon alenemisen laskukaava

Regressiomallin kertoimista saadaan tehtyä keskimääräinen laskukaava auton arvon alenemiselle. Tuloksena tästä laskukaavasta saadaan luku, joka vastaa sitä, kuinka monta prosenttia auton alkuperäisestä arvosta uutena on jäljellä siihen syötetyillä muuttujien ikä ja ja km arvoilla. Auton ikä syötetään kaavaan vuosissa ja km tarkoittaa auton mittariin kertyneitä ajokilometrejä.

$102+(-7.967)*ik\text{ä}+(0.8337334)*(ik\text{ä}^2)+(-0.07785488)*(ik\text{ä}^3)+(0.002518395)*(ik\text{ä}^4)+(-0.0002236396)*km+(3.669157e-10)*(km^2)+(-1.813681e-16)*(km^3)$

Auton sen hetkisen hinnan arvioimiseen tarvitaan siis tieto tai arvio sen hinnasta uutena, iän ja mittarilukeman tietojen lisäksi.

4.2.1 Automerkeittäinen tarkastelu

Analyysiä tehdessä piirrettiin residuaalien kuvia myös erikseen eri automerkeille. Näiden kuvien avulla huomattiin, että yleinen regressiomalli ei sopinut kaikille automerkeille yhtä hyvin kuin toisille, sillä esimerkiksi jäännöskuvioiden histogrammin kuvasta näkyi, etteivät residuaalit olleet täysin normaalijakautuneet keskiarvolla nolla. Tämä innostaa tutkimaan auton arvon alenemista erikseen automerkeittäin, eli aineiston jakamista useiden regressiomallien muodostamiseen.

Aineistosta erotettiin omiksi osa-aineistoiksi eri automerkkien havainnot. Näille tehtiin vastaava regressioanalyysi kuin yllä, eli niihin sovitettiin polynomista regressiomallia, jossa vasteena on auton arvo myyntihetkellä prosenttiosuutena alkuperäisestä hinnasta. Selittävinä tekijöinä on jälleen kerran auton ikä ja mittarilukeman kilometrit myyntihetkellä.

Koska koko aineiston yleisessä mallissa iän vaikutus on lähes lineaarinen, mutta

kilometrien vaikutus selkeästi kolmannen asteen polynomina, mallinnettiin eri automerkkien arvon säilymistä hieman yksinkertaisemmalla mallilla, jossa iän vaikutusta tutkittiin vain lineaarisena, mutta kilometrien vaikutus oli edelleen kolmannen asteen polynomin muodossa. Selitysasteiden erot tällaisten mallien ja monimutkaisempien mallien kanssa eivät olleet suuria, ja automerkeittäin tehdyille malleille saatiin silti keskimäärin yhä noin 90 prosentin selitysasteet. Koska aineiston rajaaminen merkkikohtaiseksi johtaa myös havaintojen pienenemiseen, mallin yksinkertaistamisella vältytään mahdolliselta ylisovittamiselta. Näin myös saatiin mallit mahdollisimman yhdenmukaisiksi, ja mallien regressiokertoimien vertailu kertoo automerkkien keskimääräisistä eroista.

Tällä tavalla saatiin muodostettua viidelle automerkille seuraavat tilastolliset regressiomallit:

Toyota: $100.3562 + (-3.777454) * ikä + (-0.000219996) * km + (7.941122e-11) * (km^2) + (6.222859e-16) * (km^3)$
Adjusted R-Squared: 0.8874

Citroën: $94.89632 + (-4.356972) * ikä + (-0.0003674869) * km + (1.333568e-9) * (km^2) + (1.999831e-15) * (km^3)$
Adjusted R-Squared: 0.9342

Opel: $97.11658 + (-4.36683) * ikä + (-0.0002625216) * km + (3.826696e-10) * (km^2) + (-1.758124e-17) * (km^3)$
Adjusted R-Squared: 0.8983

Mercedez-Benz: $1.017308 + (-4.817878) * ikä + (-0.000246944) * km + (5.662835e-10) * (km^2) + (-4.700490e-16) * (km^3)$
Adjusted R-Squared: 0.9225

Kia: $1.004467 + (-4.506603) * ikä + (-0.0002051714) * km + (1.760252e-10) * (km^2) + (1.516367e-16) * (km^3)$
Adjusted R-Squared: 0.9039

Citroënin ja Opelin kohdalla vakiotermi ei osu kovin lähelle sataa. Koska vakiotermi on se, minkä malli antaisi ennusteeksi silloin, kun muut regressiokertoimet ovat nolla, vakiotermien pitäisi olla mahdollisimman lähellä sataa. Tällöin siis auton alkuperäisestä hinnasta olisi jäljellä täydet sata prosenttia, kun se uusi ja ajamaton. Koska aineistossa ei ollut aivan uusia autoja, regressiokäyrä piirtyy aineiston ulkopuolella niin, että sitä ei saisi tulkita. Tällöin ei siis tarvitse kiinnittää sen suurempaa huomiota vakiotermin kummallisuuteen, vaikka se tulkinnan kannalta tuottaa epäilyksen mallin hyvyydestä. Näiden mallien selitysasteet ovat kuitenkin erittäin hyvät.

Koska jokaiselle eri automerkin regressiomallille sallittiin iän vaikutus vain lineaarisena, voidaan iän kertoimia vertailla. Iän kulmakertoimet ovat välillä (-4.9, -3.7), joten jokaiselle tähän vertailuun valitulle automerkille auton ikääntyminen vuodella vähentää auton arvosta noin 4 tai 5 prosenttia. Tämän lisäksi ajetut kilometrit alentavat myös auton arvoa, mutta niiden vaikutuksen vertailu ei ole yhtä helppoa, koska ne ovat mallissa kolmannen asteen polynomin muodossa. Koska ikä korreloi vah-

vasti ajokilometrien kanssa, lopulliseen ennusteeseen vaikuttaa molemmat muuttujat samaan aikaan, eikä ennuste pätisi esimerkiksi 10 vuotta vanhalle autolle, jolla ei olisi ajettu lainkaan.

5 Johtopäätelmät

Tässä työssä yritettiin löytää vastausta siihen, mikä on auton keskimääräinen arvon alenema. Huomioon otettiin vain auton ikä ja ajetut kilometrit, mutta regressiomalliin voitaisiin ottaa vielä lisää tekijöitä, kuten moottorin tilavuutta tai auton sen hetkistä sijaintipaikkakuntaa. Useamman selittävän tekijän kohdalla pitää kuitenkin olla hyvin tarkka sen suhteen, ettei analyysissä tapahdu ylisovittamista. Esimerkiksi vanhoja, automaattivaihteisia autoja on hyvin vähän verrattuna vaikkapa uudempiin autoihin, joten jos sitä havaintojoukkoa jakaisi osiin vielä moottorin tilavuuden ja polttoaineen suhteen, jäisi jäljelle melko vähän havaintoja. Tämän työn tulokset tyydyttävät kuitenkin alkuperäisen tutkimuskysymyksen.

Lisäksi toisenlaisessa lähestymistavassa voitaisiin yrittää selvittää, mikä on ollut auton lopullinen myyntihinta. Tässä työssä on ollut myynti-ilmoituksessa esitetty pyyntihinta. Pyyntihinnan voidaan olettaa kuitenkin olevan melko lähellä lopullista toteutunutta hintaa, sillä hyvä myyjä osaa arvioida auton hinnan sellaiseksi, että sen saa myydyksi, mutta toisaalta siitä ei haluta suurta tappiota. Pyyntihinta kuvastaa sitä hintaa, mitä myyjä toivoisi saavansa autosta, eikä siitä olla välttämättä halukkaita joustamaan kovin paljoa. Lopullisen myyntihinnan sisältävän datan kerääminen voi olla kuitenkin niin haastavaa, että tarvittavien havaintoyksiköiden lukumäärän saaminen voi olla hankalaa.

Havaintojen lukumäärä tietenkin vaikuttaa myös siihen, kuinka monimutkaista tilastollista mallia voidaan sovittaa ilman pelkoa liiallisten parametrien lukumäärästä vapausasteisiin verrattuna. Kuten liitteen hajontakuviosta näkee, jo 300 tuhannen ajokilometrin jälkeen havaintoja on paljon vähemmän kuin sitä ennen. Vanhoja ja paljon ajettuja autoja saatetaan poistaa liikenteestä, eli paljon ajettuja autoja ei välttämättä laiteta enää myyntiin. Ne paljon ajetut autot, jotka päätyvät yhä myyntiin, saattavat olla poikkeuksellisia omassa ryhmässään. Ne voivat olla vaikkapa keskimääräistä autoa paremmin huollettuja. Tällöin luotettavan mallin löytäminen erityisesti vanhoja ja paljon käytettyjen autojen suhteen voi olla vaikeaa, koska kaikki vanhat autot eivät päädy myynti-ilmoituksista kerättyyn aineistoon. Tässäkin työssä näkyi esimerkiksi Citroën-merkkisten autojen kohdalla, että regressiomalli arvioi auton arvon nousevan kilometrien kertyessä jo alle 200 tuhannen kilometrin jälkeen.

Tämän työn tuloksia voi hyödyntää esimerkiksi auton kokonaiskustannuksia laskiessa. Autoiluun liittyy polttoainekulujen lisäksi esimerkiksi huoltokuluja ja vakuutusmaksuja, mutta näiden kulujen lisäksi pitäisi huomioida myös auton hankintahinnan ja mahdollisen jälleenmyyntihinnan välinen arvon erotus, mihin tässä työssä esitelty laskukaava auton arvon alenemasta antaa vastauksen.

Yksityishenkilön lisäksi auton arvon määrittäminen on oleellista myös esimerkiksi vakuutusyhtiöissä, jotka joutuvat pohtimaan korvattavan auton arvoa. Lisäksi auton arvon määrittäminen on osa myös autokauppiaiden työtä, kun he pohtivat auton hyvityshinnan ja tietenkin myös pyyntihinnan suuruutta.

Jatkotutkimuksen aiheena toimisi se, miten pyyntihinnan vaihtelu näkyy siinä, kuinka nopeasti auto tulee myydyksi. Mielenkiinnon kohteena olisi tutkia auton

myyntiaikaa verrattuna siihen, onko se hinnoiteltu ylä- tai alakanttiin. Mikäli auto on ryhmässään hinnoiteltu hyvin alhaiseksi, sen voisi olettaa menevän kaupaksi melko pian. Tutkimuksen lopputuloksena voisi olla tieto siitä, kuinka monta päivää nopeammin auto menee keskimäärin kaupaksi, kun se on hinnoiteltu vaikkapa 20 prosenttia edullisemmaksi kuin viiteryhmänsä keskivertoinen auto. Myös myynnissä olevan auton sijainti olisi mielenkiintoinen lisä tutkimukseen; näkyykö autojen markkinoilla paikkakuntakohtaista eroa.

5.1 Esimerkkitaulukko auton arvon alenemasta

Tässä työssä aiemmin esitetyn kaavan avulla voidaan muodostaa taulukko, jossa lasketaan, paljonko tämän tilastollisen mallin mukaan auton arvo alenee, kun sillä ajetaan tietty määrä ajokilometrejä vuoden aikana.

Tämän aineiston perusteella autoilla ajetaan keskimäärin noin 20 000 kilometriä vuodessa. Tämä luku otettiin aineistosta niin, että laskettiin, kuinka monta kilometriä 5, 6 ja 7 vuotta vanhojen autojen osa-aineistoissa oli keskimäärin ajettuja kilometrejä, ja jakamalla tämä keskimääräinen kilometrimäärä auton iällä.

Näiden avulla voidaan muodostaa taulukko, jossa lasketaan kuvitteellisten 20 000 euron (Auto 1) ja 50 000 euron (Auto 2) arvoisten uusien autojen arvon alenemaa vuosien saatossa.

Taulukko 5.1. Esimerkkitaulukko auton arvon alenemasta, kun autolla ajetaan vuoden aikana 20 tuhatta kilometriä.

Auton ikä (v)	Mittarilukema	Arvosta jäljellä (%)	Arvo autolle 1	Arvo autolle 2
0	0	100	20 000	50 000
1	20 t	90.86	18 200	45 400
2	40 t	80.85	16 200	40 400
3	60 t	71.97	14 400	36 000
4	80 t	63.90	12 800	31 900
5	100 t	56.38	11 300	28 200
6	120 t	49.19	9 800	24 600
7	140 t	42.21	8 400	21 100
8	160 t	35.35	7 100	17 700
9	180 t	28.57	5 700	14 300
10	200 t	21.93	4 400	11 000

Taulukosta näkee esimerkkihintaisten autojen kohdalla, että auton arvo alenee eniten silloin, kun auto on vielä melko uusi. Aluksi auton arvo alenee noin 10 prosentin vuosivauhtia, mutta yli 5 vuotta vanhojen autojen kohdalla auton arvo alenee vuodessa enää noin 7 prosenttia, kun sillä ajetaan keskimääräisesti.

Auton ensimmäiset kilometrit ovat siis ne kalleimmat auton arvon alenemisen kannalta.

5.2 Esimerkkitaulukko eri automerkkien autojen arvon alenemasta

Tässä työssä tutkittiin myös muutaman yleisimmän automerkin välisiä eroja auton arvon alenemisen suhteen. Tehdään näistä nyt vastaava taulukko kuin yllä, mutta nyt vertaillaan alkuperäiseltä hinnaltaan saman arvoisia, ja lasketaan niiden arvon alenema automerkkinsä mukaan.

Otetaan esimerkkiin mukaan japanilainen automerkki Toyota (Auto 1), saksalainen Opel (Auto 2) ja ranskalainen Citroën (Auto 3). Oletetaan näiden kuvitteellisten autojen hinnan uutena olevan 25 000 euroa.

Pidetään yhä kiinni siitä oletuksesta, että autolla ajetaan keskimäärin noin 20 tuhatta kilometriä vuodessa. Näin saadaan lasketuksi näille esimerkkiautoille tämän työn antamien laskukaavojen mukaisesti hinnat eri ikäisinä.

Taulukko 5.2. Esimerkkitaulukko auton arvon alenemasta, kun autolla ajetaan vuoden aikana 20 tuhatta kilometriä.

Auton ikä (v)	Mittarilukema	Arvo, Toyota	Arvo, Opel	Arvo, Citroën
0	0	25 000	25 000	25 000
1	20 t	23 054	21 913	20 935
2	40 t	21 042	19 623	18 436
3	60 t	19 061	17 410	16 252
4	80 t	17 118	15 272	14 407
5	100 t	15 221	13 210	12 925
6	120 t	13 378	11 223	11 829
7	140 t	11 595	9 312	11 144
8	160 t	9 880	7 476	10 893
9	180 t	8 240	5 714	11 102
10	200 t	6 684	4 028	11 793

Koska aineiston jakamisessa automerkeittäin havaintoja oli yhdessä osa-aineistossa huomattavasti vähemmän kuin alkuperäisessä, ja alkuperäisessäkin havaintoja oli vähiten havaintojen arvoalueen päissä, toimivat nämä luvut parhaiten välin keskiosassa. Regressiomallin käyrä sovittuu siis parhaiten ja luotettavimmin aineiston havaittujen arvojen keskiarvon lähellä. Tällöin hyvin uusissa ja hyvin vanhoissa autoissa ennuste ei ole kovin luotettava. Taulukko kuvastaa kuitenkin hyvin eri automerkkien keskimääräisiä eroja esimerkiksi viisi vuotta vanhojen autojen arvoissa.

Esimerkiksi merkin Citroën regressiomallissa vakiotermi oli tavallista pienempi, mikä johtaa siihen, että regressiomalli antaa hyvin nuorelle autolle liian pienen ennusteen, joka on selkeästi normaalista poikkeava. Ennustekäyrä on kuitenkin luotettavampi jo vähän vanhemmille autoille, joille on jo ehtinyt kertyä ajokilometrejäkin. Koska näissä merkkikohtaisissa regressiomallissa iälle sallittiin vain lineaarinen vaikutus neljännen asteen polynomin sijaan, ei regressiomallin iän kerroin voi kuitenkaan olla myöskään ylisovitettu iän ääripäiden lähistöllä. Koska havaittuja arvoja vain oli niin vähän hyvin uusissa autoissa, regressiomallin vakiotermi Cit-

roënin kohdalla tuottaa hieman hassusti tuloksen, että uudelle autolle vasteen arvo eli ennustettu arvo olisi vain 95 prosenttia (auton alkuperäisestä hinnasta). Lisäksi kyseisen automerkin kilometrien polynomi sisältää derivaatan nollakohdan taulukon sisältämien prediktorien arvojen sisällä, eli ennustettu arvo alkaa nousta tarpeeksi suurilla kilometriluvuilla jo melko pian keskimääräisen mittarilukeman jälkeen. Tällöin regressiomallin ennuste ei toimi myöskään paljon ajetuilla autoilla.

Kuten yllä on sanottu, ääriarvojen ennusteiden tulkintaa ei suositella. Vertaillaan siis näiden kolmen eri automerkin regressiomallin ennusteita taulukon keskivaiheessa. 5 vuotta vanhan ja 100 tuhatta kilometriä ajetun auton ennustettu arvo on Toyotalla 15 221 euroa, Opelilla 13 210 euroa ja Citroënilla 12 925 euroa. Näistä kolmesta autosta siis Toyota säilytti arvonsa parhaiten, ja Citroënin arvo aleni näistä kolmesta eniten. Kuitenkin vuotta vanhemmalle ja 120 tuhatta kilometriä ajetuille autoille Opelin ennustettu arvo on jo hieman penempi kuin Citroënin. Toyotan ennustettu arvo on yhä suurin.

Siitä taas vuotta vanhempaan ja 20 tuhatta kilometriä enemmän ajettuun autoon siirtyessä ennusteet muuttuvat niin, että Citroënin ennustettu arvo on jo melkein yhtä suuri kuin Toyotan. Opelin arvo laski yhä tasaista noin parin tuhannen euron vauhtia. Tässä kohtaa kilometrimäärää törmätään jo aiemmin mainittuun ongelmaan, jossa Citroënin kilometrien prediktorin polynomi ei ennusta hyvin paljon ajetuille autoillle. Sen sijaan jokaisessa näissä kolmessa automerkissä iän vaikutus regressiomallissa oli noin 4 prosentin arvon alenema vuositasolla. Iän vaikutuksen tulkinnassa pitää kuitenkin ottaa huomioon se, että ikä ja kilometrit ovat yhteydessä toisiinsa, joten niiden vaikutuksia ei voi tarkastella täysin itsenäisesti. Tähän tarkasteluun valituista automerkeistä Toyota näyttää säilyttävän arvonsa parhaiten.

Tutkielman lopuksi haluan antaa kiitokseni Autotalli.comille aineistosta ja työn mahdollistamisesta. Kiitän myös Tampereen yliopiston puolelta apuna olleita Jaakko Peltosta ja Tapio Nummea.

Lähteet

Aalto-Setälä, V. & Halonen, M. (2003), ”Käytettyjen ajoneuvojen markkinahinnat”, Työselosteita ja esitelmiä 76, Kuluttajatutkimuskeskus 2003. Saatavilla Internetistä: https://helda.helsinki.fi/bitstream/handle/10138/152311/Kaytettyjen_ajoneuvojen_markkinahinnat.pdf?sequence=1.

Autoalan tiedotuskeskus (2017), Autoalan liikevaihto: Käytettyjen autojen kauppa, Internetsivusto tammikuussa 2017: http://www.autoalantiedotuskeskus.fi/autoala_suomessa/autoalan_liikevaihto.

Hyttinen, M. (2016), ”Hyötyvaraosat”, Opinnäytetyö, Metropolia Ammattikorkeakoulu, Auto- ja kuljetustekniikan koulutusohjelma. Saatavilla Internetistä: http://theseus32-kk.lib.helsinki.fi/bitstream/handle/10024/12575/hyotyvar.pdf?sequence=1.

Kangosjärvi, M. & Sassi, T. (2016), ”Käytetyn auton hankintaan vaikuttavat tekijät”, Opinnäytetyö, Lapin ammattikorkeakoulu, Yhteiskuntatieteiden, liiketalouden ja hallinnon ala. Saatavilla Internetistä: https://www.theseus.fi/bitstream/handle/10024/115613/Kaytetyn%20auton%20hankintaan%20vaikuttavat%20tekijat.pdf?sequence=1.

Kanniainen, T. (2014), ”Laskuri paljastaa autosi arvon”, Uutinen 3.5.2014, Ilta-Sanomat. Saatavilla Internetistä: http://www.is.fi/taloussanomat/art-2000001835459.html.

Keurulainen, R. (2010), ”Yksityinen henkilöautoilu Suomessa, Auton kustannukset ja tarpeellisuus pääkaupunkiseudulla”, Opinnäytetyö, Metropolia Ammattikorkeakoulu, Liiketalouden koulutusohjelma. Saatavilla Internetistä: http://theseus32-kk.lib.helsinki.fi/bitstream/handle/10024/25175/Opinnaytetyo.pdf?sequence=1.

Lee, A. L. & Seber A. F. (2003), *Linear Regression Analysis* (Wiley Series in Probability and Statistics), Wiley

Leppikangas, S. (2016), ”Yksityisautoilun kustannusrakenteen ja kokonaiskustannusten selvittäminen”, Opinnäytetyö, Hämeen Ammattikorkeakoulu, Liikennealan koulutusohjelma. Saatavilla Internetistä: https://theseus32-kk.lib.helsinki.fi/bitstream/handle/10024/96984/Leppikangas_Sauli.pdf?sequence=1.

Liimatainen, H., Mäkelä, T., Mäntynen, J., Nykänen, L. & Pöllänen, M. (2015), ”Liikenteen markkinat Suomessa”, Trafin tutkimuksia 16-2015, Liikenteen turvallisuusvirasto Trafi. Saatavilla Internetistä: https://www.trafi.fi/filebank/a/1452675021/34e771ac250db32ab331b2d71ae92ffc/19497-Liikennemarkkinat_raportti_2015-12-10.pdf.

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2006), *Introduction to Linear Regression Analysis* (Wiley Series in Probability and Statistics), Wiley

Nurmela, T. (2009), ’Näin auton arvo laskee käytettynä - lue IS:n jättiselvitys!”, Uutinen 23.7.2009, Ilta-Sanomat. Saatavilla Internetistä: http://www.is.fi/autot/art-2000000063585.html.

Nyblom, J. (2015), *Yleistetyt lineaariset mallit*, Jyväskylän yliopisto, Matematiikan ja tilastotieteen laitos. Saatavilla Internetistä: http://users.jyu.fi/~junyblom/JTMprujub.pdf.

Pulkkinen, T. (2016), ”Autot sijoituksen kohteena”, Opinnäytetyö, Haaga-Helia Ammattikorkeakoulu Oy, Liiketalouden koulutusohjelma. Saatavilla Internetistä: http://www.theseus.fi/bitstream/handle/10024/105952/Pulkkinen_Timi.pdf?sequence=1.

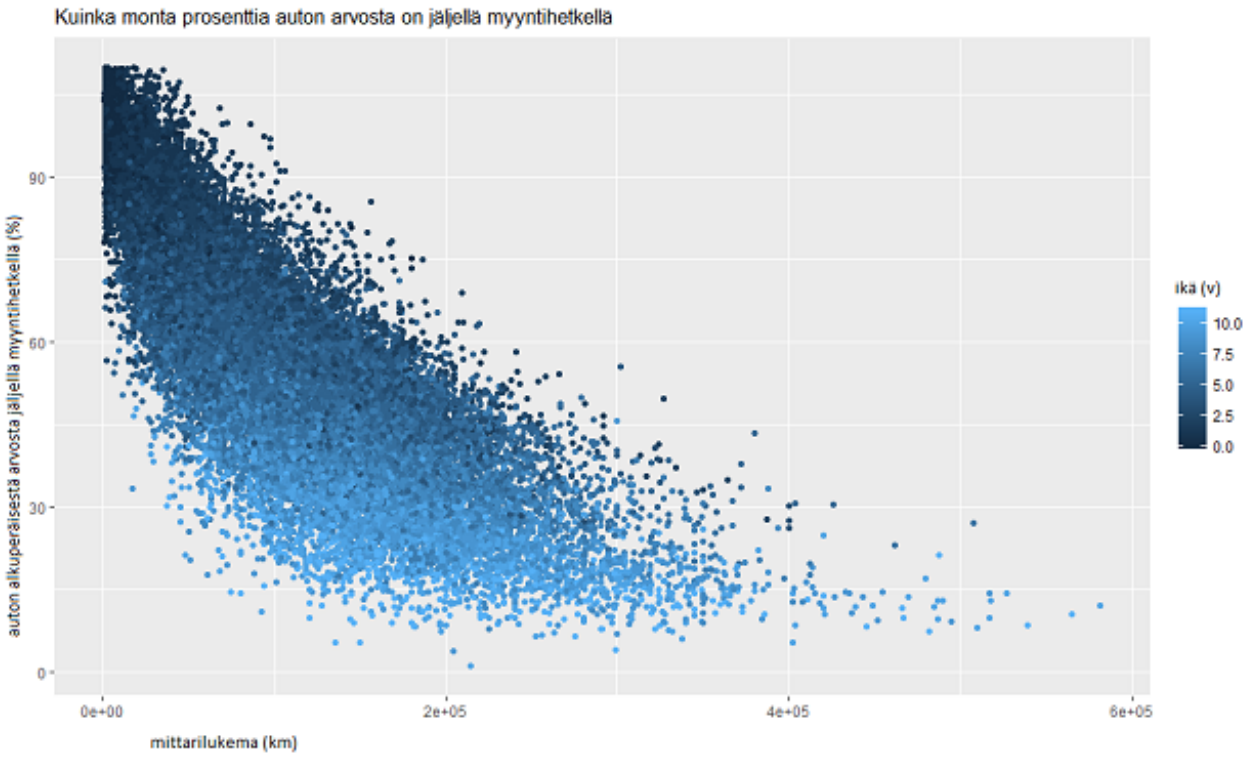
Suomen virallinen tilasto SVT. (2015a), ”Moottoriajoneuvokanta 2015”, Ajoneuvokanta-tilastot 2015, Suomen virallinen tilasto, Tilastokeskus. Saatavilla Internetistä: http://www.stat.fi/til/mkan/2015/mkan_2015_2016-03-23_fi.pdf.

Suomen virallinen tilasto SVT. (2015b), ”Väestön tieto- ja viestintätekniikan käyttö 2015”, Tiede, teknologia ja tietoyhteiskunta 2015, Suomen virallinen tilasto, Tilastokeskus. Saatavilla Internetistä: http://www.tilastokeskus.fi/til/sutivi/2015/sutivi_2015_2015-11-26_fi.pdf.

Toivanen, O. (2005), ”Autoliikkeiden tietojärjestelmät: Käytetyn auton arvon määrittely”, pro gradu -tutkielma, Tampereen yliopisto, Tietojenkäsittelytieteiden laitos. Saatavilla Internetistä: https://tampub.uta.fi/bitstream/handle/10024/93209/gradu00941.pdf?sequence=1.

Weisberg, S. (2005), *Applied Linear Regression* (Wiley Series in Probability and Statistics), Wiley

Liite: Aineiston hajontakuvio



Kuva 1. Aineiston havaintojen hajontakuvio.