Tampereen yliopisto

Aino Satalahti

REGRESSIOANALYYSIN VIRHEPÄÄTELMÄT

Informaatioteknologian ja viestinnän tiedekunta
Tilastollisen data-analyysin kandidaattitutkielma
Huhtikuu 2021

Tiivistelmä

Aino Satalahti: Regressioanalyysin virhepäätelmät
Tilastollisen data-analyysin kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastollisen data-analyysin tutkinto-ohjelma
Opinnäytetyön ohjaajat: Tapio Nummi, Jyrki Ollikainen
Huhtikuu 2021

Regressioanalyysi on tilastollinen menetelmä, jolla voidaan tarkastella muuttujien välisiä syy- ja seuraussuhteita. Tässä tutkielmassa käsitellään lineaarisessa regressioanalyysissä mahdollisia virhepäätelmiä sekä niihin johtavia tekijöitä. Aiheita mallinnetaan kuvaajien sekä sanallisten esimerkkien avulla, jotka on muodostettu julkisista aineistoista sekä tutkielmaa varten luoduista kuvitteellisista aineistoista.

Tutkielman alussa määritellään regressioanalyysi käsitteenä, esitellään pienimmän neliösumman menetelmä, jolla regressiomalli estimoidaan, sekä esitellään regressioanalyysin vaiheet. Sen jälkeen esitellään yhdeksän oletusta, joiden tulee päteä, jotta virheellisiltä tutkimustuloksilta vältyttäisiin regressioanalyysissä. Virhepäätelmiin johtavien tekijöiden käsittely eli regressiodiagnostiikka on jaettu neljään aihepiiriin.

Ensimmäinen aihepiiri on jäännöksiin liittyvät rikkomukset. Niitä ovat asetetun normaalisuusoletuksen rikkoutuminen ja jäännösten odotusarvon virheellisyys. Myös aineiston heteroskedastisuus eli jäännösvarianssin eriävyys selittävän muuttujan luokkien välillä sekä autokorrelaatio eli jäännösten välinen systemaattisuus ovat jäännöksiin liittyvää regressiodiagnostiikkaa.

Toiseksi käsitellään mallin muuttujien oikeellisuutta. Regressiomallista ei saa löytyä multikollineaarisuutta eli muuttujien välistä korrelaatiota. Mallin muuttujien oletetaan myös olevan virheettömästi muodostettuja.

Kolmanneksi käsitellään regressiomallin oikeellisuuteen liittyviä oletuksia. Lineaarisuusoletuksen mukaan malli tulee voida esittää muodossa, joka mallintaa suoraviivaisia kausaalisuhteita muuttujien välillä. Puuttuvan muuttujan harha taas nimensä mukaisesti tarkoittaa mallista puuttuvan tarpeellisen muuttujan aiheuttamaa virheellistä tulosta. Lisäksi mallin käyttöön liittyy mallin tulkinnan vaiheessa ekologisen virhepäätelmän mahdollisuus, mikä tarkoittaa korrelaation virheellistä yleistämistä.

Viimeinen regressiodiagnostiikan aihepiiri on havaintojen rooli virhepäätelmien aiheuttajana. Havaintojen oletetaan olevan virheettömästi kerättyjä ja tallennettuja ja täten luotettavia. Lisäksi tutkielmassa käsitellään poikkeuksellisen suurten tai pienten havaintojen oikeaoppista tarkastelua, sekä niiden aiheuttaman vipuvaikutuksen huomioimista.

Avainsanat: regressiodiagnostiikka, normaalisuusoletus, lineaarisuusoletus, heteroskedastisuus, autokorrelaatio, multikollineaarisuus

Sisällys

1 Johdanto

Regressioanalyysi on merkittävä ja monipuolinen tilastollinen menetelmä muuttujien välisten syy- ja seuraussuhteiden tarkastelussa, ja täten laajasti sovellettu kaikilla aloilla. Sovellustasolla regressioanalyysissä kuitenkin päädytään herkästi virhepäätelmiin, mikäli tiettyjä asioita ei oteta huomioon.

Tässä kanditutkielmassa esittelen regressioanalyysin toimivuuden takaamiseksi laadittuja oletuksia, ja mallinnan konkreettisten esimerkkien avulla tilanteita, jotka aiheuttavat virhepäätelmiä. Esittelen yhdeksän regressioanalyysin oikeelliseen toimintaan liittyvää oletusta, joiden tulee päteä, jotta vältyttäisiin virhepäätelmiltä. Näiden oletusten tarkastelua kutsutaan regressiodiagnostiikaksi. Regressiodiagnostiikkaa tehdään tilastografiikan, diagnostisten testien sekä diagnostisten tunnuslukujen avulla (Seppälä 2015). Käsittelen virhepäätelmiin johtavat tekijät neljässä aihepiirissä: jäännöksiin liittyvät, mallin muuttujiin liittyvät, regressiomalliin liittyvät sekä havaintoihin liittyvät virhepäätelmät.

Käsittelen tutkielmassa lineaarista regressiota, ja muodostan esimerkit yleisesti esillä olevista aineistokokoelmista sekä luomistani kuvitteellisista aineistoista. Lukijalta edellytetään regressioanalyysin toiminnan sekä siihen liittyvien termien tuntemusta, sillä tilastollisena menetelmänä se esitellään tutkielmassa vain pääpiirteittäin. Tutkielmani konkreettisuus ja yleistajuisuus tekevät tutkielmastani hyödyllisen minkä tahansa alan asiantuntijalle, taaten entistä totuudenmukaisempia johtopäätöksiä regressioanalyysiä soveltavalle.

Ohjelmisto

Esimerkkianalyysejä muodostan käyttäen R-ohjelmaa, joka on ilmaiseksi verkosta ladattavissa oleva avoimen lähdekoodin ohjelmistoympäristö. R perustuu S-ohjelmointikieleen. R:llä voidaan käsitellä erilaisia tilastollisia ja graafisia tekniikoita, ja se soveltuu täten myös hyvin regressioanalyysin mallintamiseen tässä kandityössä. (R Development Core Team 2010.) Liitteessä 1 R-koodit sekä käytetyt aineistot.

2 Regressioanalyysin määritelmä ja vaiheet

2.1 Regressioanalyysin määritelmä

Regressioanalyysi on menetelmä, jonka avulla tutkitaan muuttujien välisiä suhteita. Esimerkiksi jäätelön myynnistä kertovasta aineistosta voidaan tutkia, onko säätilalla vaikutusta jäätelön myyntituloksiin. Tällöin sää on selittävä ja jäätelön myynti selitettävä muuttuja. Selittäviä muuttujia voi olla useampiakin, esimerkiksi jäätelökioskin sijainti voi vaikuttaa jäätelön myyntiin säätilan kanssa. Muuttujien suhde esitetään yhtälön muodossa, jota kutsutaan regressiomalliksi. Selitettävää muuttujaa merkitään Y:llä, ja selittäviä muuttujia $X_1, X_2, ..., X_n$:llä, jossa n on muuttujien määrä. Muuttujien suhdetta kuvaava malli on

$$Y = f(X_1, X_2, ..., X_n) + \epsilon, \tag{2.1}$$

jossa $f(X_i)$ kuvaa X:n ja Y:n suhdetta, ja ϵ on satunnainen jäännös, joka kuvaa mallista löytyvää epäjohdonmukaisuutta. (Chatterjee ja Hadi 2006.)

2.2 Pienimmän neliösumman menetelmä

Tässä tutkielmassa käsittelen lineaarista regressiomallia

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon, \tag{2.2}$$

jossa β_i:t ovat tuntemattomia parametrejä, ja ne voidaan estimoida malliin pienimmän neliösumman menetelmällä.

Pienimmän neliösumman menetelmällä estimointi tapahtuu minimoimalla jäännösten neliösumma

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [(Y - f(X_i, \beta_i)]^2, \tag{2.3}$$

jossa $f(X_i, \beta_i)$ on tuntemattomien parametrien β_i funktio (Freund, Wilson ja Sa 2006).

Esimerkki 2.1. Olkoon

$$Y = \beta_0 + \beta_1 X_i + \epsilon, i = 1, ..., n \tag{2.4}$$

lineaarinen regressiomalli, jonka parametrit estimoidaan pienimmän neliösumman menetelmällä. Sen neliösumma on

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2. \tag{2.5}$$

Neliösumma minimoidaan derivoimalla osittain

$$\frac{\partial(\sum_{i=1}^{n} \epsilon_i^2)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \tag{2.6}$$

$$\frac{\partial(\sum_{i=1}^{n}\epsilon_i^2)}{\partial\beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i), \tag{2.7}$$

ja asettamalla derivaatat nolliksi

$$\sum_{i=1}^{n} Y_i - n\hat{\beta_0} - \hat{\beta_1}\sum_{i=1}^{n} X_i = 0 \tag{2.8}$$

$$\sum_{i=1}^{n} X_i Y_i - \hat{\beta_0}\sum_{i=1}^{n} X_i - \hat{\beta_1}\sum_{i=1}^{n} X_i^2 = 0. \tag{2.9}$$

Pienimmän neliösumman estimaateiksi saadaan

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n} \tag{2.10}$$

ja

$$\hat{\beta_0} = \bar{Y_i} - \hat{\beta_1}\bar{X_i}. \tag{2.11}$$

Yleensä pienimmän neliösumman estimointi esitetään matriisinotaatiota käyttäen, sillä useampien muuttujien mallissa se on yksinkertaisempi esitysmuoto. (Freund, Wilson ja Sa 2006.)

2.3 Regressioanalyysin vaiheet

Regressioanalyysissä vaiheita on useita, joista missä tahansa voi aiheutua virhepäätelmiä, ellei regressioanalyysin oletusten pätevyyttä huomioida (Chatterjee ja Hadi 2006).

1. Ongelman määritteleminen
2. Potentiaalisesti merkittävien muuttujien valinta
3. Aineiston kerääminen
4. Mallin tarkka määrittely
5. Parametrien estimointi
6. Mallin sovittaminen
7. Mallin kriittinen tarkastelu ja hyväksyminen
8. Mallin käyttäminen ongelman ratkaisuun

3 Lineaarisen regression oletukset

Kun ongelma on määritelty, potentiaalisesti merkittävät muuttujat valittu, sekä aineisto kerätty, estimoidaan regressiomalli pienimmän neliösumman menetelmällä. Seuraavat oletukset on asetettu takaamaan regressiomallin oikeellista toimintaa:

Oletus 3.1. Oletetaan, että jäännökset ϵ_i noudattavat normaalijakaumaa.

Oletus 3.2. Oletetaan, että jäännösten ϵ_i odotusarvo on nolla.

Oletus 3.3. Oletetaan, että jäännösvarianssi σ^2 on suunnilleen sama kaikille X:n luokille, eli aineisto on homoskedastinen.

Oletus 3.4. Oletetaan, että jäännös ϵ_i ei ole assosioitavissa aineiston muihin jäännöksiin, vaan jäännökset edustavat täydellistä satunnaisuutta, jota malli ei ennusta. Jäännökset ovat riippumattomasti ja identtisesti jakautuneita.

Oletus 3.5. Oletetaan, että muuttujat X_i ovat keskenään lineaarisesti riippumattomia, eli aineistossa ei ole multikollineaarisuutta.

Oletus 3.6. Oletetaan, että muodostetun mallin muuttujissa X_i ja Y_i ei ole mittavirheitä, eli että arvoja ei ole laskettu väärin.

Oletus 3.7. Oletetaan, että regressiomalli on lineaarinen, eli että malli on määritelmän 2.2. mukaista muotoa. Tällöin i. havainto voidaan kirjoittaa muodossa

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_n X_{in} + \epsilon, i = 1, 2, \ldots, p. \tag{3.1}$$

Oletus 3.8. Oletetaan, että kaikki tarvittavat muuttujat on otettu huomioon mallissa.

Oletus 3.9. Oletetaan, että kaikki aineiston havainnot ovat yhtä luotettavia.

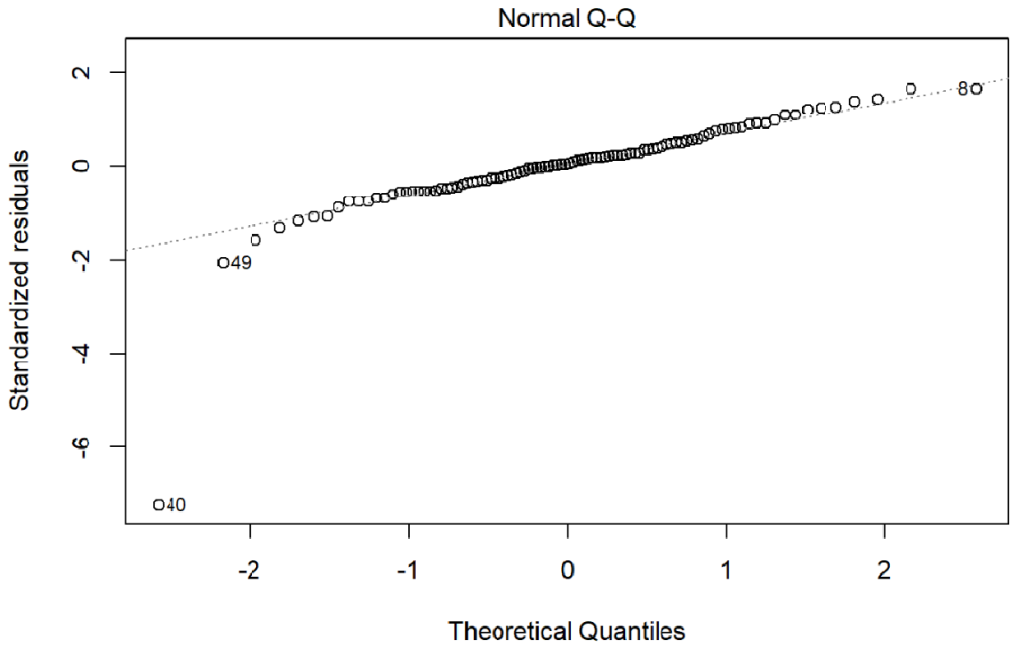
Mikäli nämä oletukset eivät päde, voidaan regressioanalyysissä päätyä virhepäätelmiin (Kahane 2008), joita seuraavat kappaleet esimerkkeineen mallintavat. Pienimmän neliösumman menetelmää käytettäessä mallin estimointiin, pienet rikkeet oletuksissa eivät välttämättä haittaa mallin toimintaa. Kuitenkin on tärkeää tarkastella mallin pätevyyttä kuvaajien avulla, jotta muodostettu malli olisi mahdollisimman totuudenmukainen.

4 Jäännösten tarkastelu

4.1 Normaalisuusoletus

Oletusta 3.1. kutsutaan normaalisuusoletukseksi. Sen mukaan jäännökset ϵ_i noudattavat normaalijakaumaa. Normaalisuusoletuksen paikkaansapitävyyttä on vaikea validoida, mutta sen pätevyyttä voidaan tutkia tarkastelemalla tiettyjä kuviaajia jäännöksistä, kun malli on sovitettu aineistoon. Mallin estimointivaiheessa normaaliusoletusta ei vielä tarvita, vaan sen tärkeys näkyy vasta mallin testausvaiheessa. (Chatterjee ja Hadi 2006.)

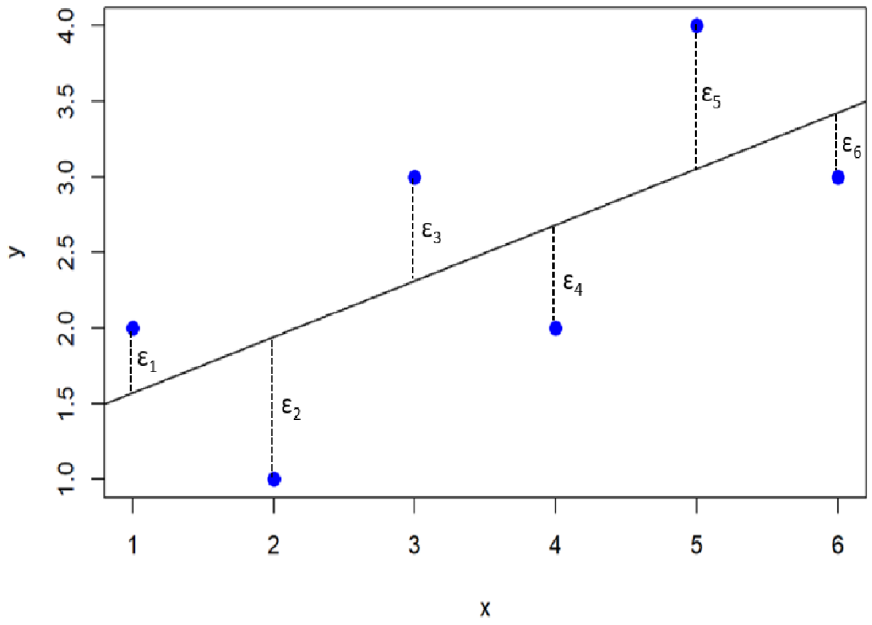
Esimerkki 4.1. Kuvassa 4.1. on esimerkki tilanteesta, jossa normaalisuusoletus pätee. On muodostettu lineaarinen malli lasten pituuden vaikutuksesta painoon, Tampereen Yliopiston Tilastotieteen peruskurssien harjoitusaineistosta Lapset_2007 (Liite 1). Kuvaajassa mallin standardoidut jäännökset sekä niiden teoreettiset kvanttiilit ovat vastakkain, ja havainnot asettuvat suunnilleen suoran mukaisesti. Tämän (lähes) suoran viivan vakiotermi on nolla ja kulmakerroin yksi, toisin sanoen standardoitujen jäännösten odotusarvo on nolla ja keskihajonta yksi. Siispä aineiston jäännökset ovat normaalijakautuneita. Mikäli havainnot eivät asettuisi kuvaajalla suoran mukaisesti, kuvaajasta nähtäisiin, että jäännökset eivät ole normaalijakautuneita, ja että mitkä aineiston havainnot tämän aiheuttaisi.



Kuva 4.1. Standardoidut jäännökset asettuvat suunnilleen regressiosuoran lähettyville, joten normaalisuusoletuksen voidaan tulkita pätevän.

4.2 Jäännösten odotusarvo

Oletuksen 3.2. mukaan jäännösten ϵ_i odotusarvon tulee olla nolla. Graafisesti tarkasteltuna tämä tarkoittaa sitä, että havaintojen jakautuessa regressiosuoran ala- ja yläpuolelle tulee jäännösten kumota toisensa niin, että niiden keskiarvo on nolla. Mikäli jäännösten odotusarvo ei ole nolla, ei regressiosuoraa ole estimoitu oikein, ja se tulee estimoida uusiksi. (Kahane 2008.)

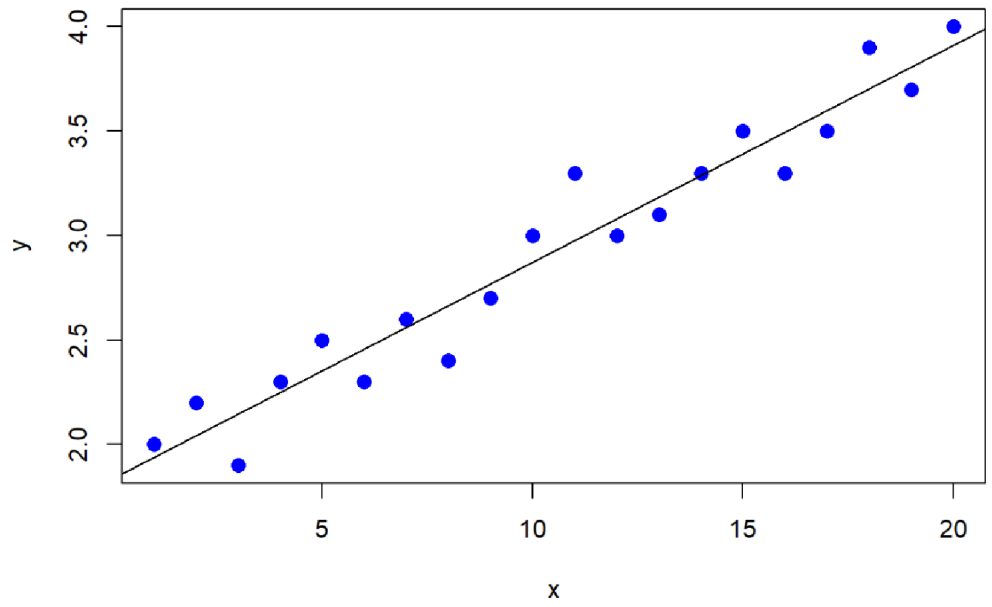


Kuva 4.2. Suoran alapuolelle jäävien havaintojen jäännöksiä kutsutaan positiivisiksi, ja yläpuolelle jäävien havaintojen jäännöksiä negatiivisiksi havainnoiksi. Yhteenlaskettuna niiden tulisi olla nolla.

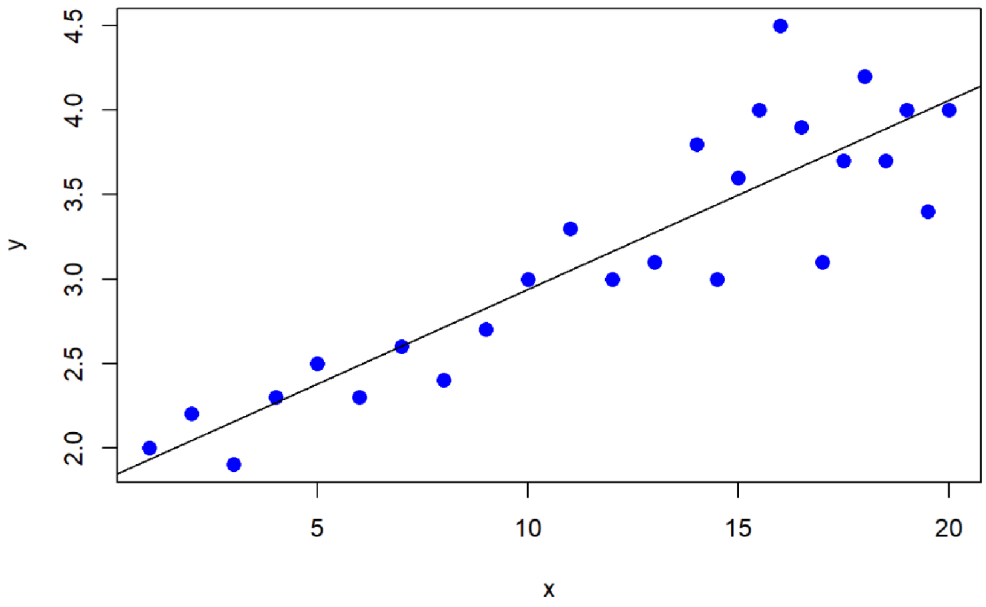
4.3 Heteroskedastisuus

Oletuksen 3.3. mukaan jäännösvarianssin σ^2 tulee olla suunnilleen sama kaikilla mallin selittävän muuttujan luokilla, eli aineiston tulee olla homoskedastinen. Tällöin kuvaajaa tarkasteltaessa havainnot Y:lle annetuista X:n arvoista asettuvat regressiosuoran ylä- ja alapuolelle tasaisesti. Jos näin ei ole, kohdataan heteroskedastisuuden ongelma. Tällaisessa tilanteessa, vaikka keskivirheiden estimaatit ovat harhattomia, ne ovat tehottomia, mikä johtaa esimerkiksi luottamusvälien testauksessa virheellisiin tuloksiin. (Kahane 2008.)

Heteroskedastisuutta voidaan testata Goldfeld-Quandtin testillä (ks. Webster 2013), jossa aineisto jaetaan luokkiin, ja vertaillaan luokkien jäännösvarianssien suuruuksia. Heteroskedastisuutta voidaan poistaa muuttamalla mallia esimerkiksi logaritmi- tai neliöjuurimuunnoksilla.



Kuva 4.3. Homoskedastisen aineiston jäännökset jakautuvat regressiosuoran ympärille tasaisesti.



Kuva 4.4. Heteroskedastisen aineiston jäännökset jakautuvat regressiosuoran ympärille epätasaisesti: Suuremmilla X:n arvoilla, eli kuvaajan oikealla laidalla, jäännösvarianssi on selkeästi suurempi kuin X:n pienemmillä arvoilla vasemmalla laidalla.

4.4 Autokorrelaatio

Oletuksen 3.4. mukaan jäännökset ϵ_i eivät saa olla regressioanalyysissä korreloituneita keskenään. Jos jäännösten väliltä havainnoidaan systemaattisuutta, tarkoittaa se sitä, että aineistosta on löydettävissä informaatiota, jota ei vielä ole hyödynnetty regressiomallissa. Tällaista ilmiötä kutsutaan autokorrelaatioksi. (Kahane 2008.)

Autokorrelaatio voi johtaa virheellisiin tilastollisiin testeihin, sekä harhaisiin luottamusväleihin. Se voi myös liioitella estimoitujen regressioparametrien tilastollisen merkitsevyyden tarkkuutta. Kun autokorrelaatiota ilmenee, jäännökset eivät ole täydellisen satunnaisia, eikä käytössä ole parhain mahdollinen malli.

Autokorrelaatio voi johtua monista syistä. Vierekkäisten havaintojen välinen autokorrelaatio saattaa johtua yhteisistä ulkoisista tekijöistä. Esimerkiksi aikasarja-analyyseissä, eli analyyseissä jotka muodostetaan aineistosta, jossa käsitellään havaintoja jollain aikavälillä, autokorrelaatiota havaitaan usein. Suuret positiiviset jäännökset saattavat johtua aineiston muista positiivisista jäännöksistä, ja sama pätee negatiivisten jäännösten kanssa.

Autokorrelaatiota voidaan testata Durbin-Watson menetelmällä (ks. Webster 2013), ja poistaa mallia muuntamalla. Jonkin selittävän muuttujan puuttuminen mallista voi myös aiheuttaa autokorrelaatiota, mutta kun se lisätään, autokorrelaatio katoaa.

Esimerkki 4.2. Säätä kuvaavassa aineistossa peräkkäisten päivien havaintojen väliltä voidaan löytää autokorrelaatiota. On kyse aikasarja-analyysistä, jossa esimerkiksi monta päivää kestänyt sadesää aiheuttaa peräkkäisten päivien havaintojen välille yhtäläisyyttä. Tällaisen asian huomioimattomuus regressiomallia luodessa aiheuttaa autokorreloituneen mallin, jonka käyttö johtaa virhepäätelmiin.

5 Mallin muuttujien oikeellisuus

5.1 Multikollineaarisuus

Multikollineaarisuus tarkoittaa tapausta, jossa kahden tai useamman selittävän muuttujan välillä on korkeaa korrelaatiota. Oletuksen 3.5. mukaan muuttujat X_i ovat keskenään lineaarisesti riippumattomia, eli multikollineaarisuutta ei tule ilmetä mallista. Muuttujien välinen korkea korrelaatio vaikeuttaa muuttujien yksittäisten vaikutusten tarkastelua, ja on täten mahdollinen virhepäätelmiin johtava tekijä.

Multikollineaarisessa mallissa pienimmän neliösumman estimaatti voi olla tilastollisesti merkitsemätön, vaikka selitysaste R^2 olisikin korkea. Täydellisen multikollineaarisuuden tapauksessa pienimmän neliösumman estimaattia ei voida esittää. Tapaus tarkoittaa sitä, että muuttujan voi esittää lineaarisena kombinaationa toisesta muuttujasta, kuten

$$X_2 = 2 * X_1. \tag{5.1}$$

Multikollineaarisuutta on mahdollista vähentää tai jopa poistaa kokonaan muokkaamalla regressiomallia. Uusien muuttujien lisääminen, aikaisemman informaation hyödyntäminen tai muuttujien funktionaalisten suhteiden muokkaaminen voi olla ratkaisu. Myös vahvan korrelaation omaavista muuttujista toisen pois pudottaminen voi parantaa mallin selittävyyttä. Mallia muokatessa on hyvä luottaa myös maalaisjärkeen muuttujien merkittävyyttä pohtiessa. (Freund, Wilson ja Sa 2006.)

Esimerkki 5.1. Muodostetaan kuvitteellinen regressiomalli kuvaamaan eri asioiden vaikutusta naisen palkkaan rahoitusalalla kuvitteellisesta aineistosta. Olkoon mallissa selittävinä muuttujina ikä, syntymävuosi ja työkokemus alalta vuosina. Huomataan, että naisen ikä ja syntymävuosi ovat laskettavissa toisistaan, eli muuttujat eivät tuo yhdessä lisäarvoa malliin. Muuttujien välillä multikollineaarisuus on siis täydellistä, joten toinen muuttujista voidaan pudottaa tarpeettomuutensa vuoksi pois.

Esimerkki 5.2. Luodaan regressiomalli Tampereen Yliopiston Tilastotieteen peruskurssien harjoitusaineistosta asunnot.xls, jossa selitetään asuntojen hintaa kahdella muuttujalla: neliömäärä sekä huonelukumäärä. Käytetty aineisto sekä tarvittavat tunnusluvut liitteessä 1.

Asetetaan hypoteesit:
$H_0 : \beta_i = 0$, eli neliömäärällä tai huoneiden lukumäärällä ei ole vaikutusta hintaan.
$H_1 : \beta_i \neq 0$, eli neliömäärällä tai huoneiden lukumäärällä on vaikutusta hintaan.

Huomataan, että asunnon neliöiden määrä sekä asunnon huoneiden lukumäärä ovat keskenään vahvasti korreloituneita korrelaatiokertoimella 0.83, eli mallissa on havaittavissa multikollineaarisuutta. Mitä enemmän asunnossa on neliöitä, sitä enemmän sinne mahtuu huoneitakin.

Tarkastellaan muodostettua mallia

$$HINTA = -110697 + 12036 * NELIOT - 42943 * HUONELKM. \tag{5.2}$$

Mallin vakiotermin testisuure on -2.387 ja p-arvo 0.0206. Nollahypoteesia ei hylätä, sillä p-arvo ei ole tarpeeksi pieni. NELIOT-muuttujan testisuure on 10.331 ja p-arvo $2.66e-14$. Nollahypoteesi hylätään tarpeeksi pienen p-arvon nojalla. Muuttujan HUONELKM testisuure on -1.321 ja p-arvo 0.1922. Nollahypoteesia ei hylätä.

Voidaan tulkita, että malli ei ole parhain mahdollinen, kun p-arvot eivät ole tarpeeksi pieniä. Muuttujan HUONELKM p-arvo on suurempi kuin muuttujan NELIOT p-arvo, joten pudotetaan muuttuja HUONELKM mallista pois.

Nyt muodostetun mallin

$$HINTA = -127528.6 + 10754.9 * NELIOT. \tag{5.3}$$

vakiotermin testisuureeksi saatiin -2.48 ja p-arvo on 0.00634. Nollahypoteesi hylätään. NELIOT-muuttujan testisuure on 16.55 ja p-arvo alle $2e-16$. Taas nollahypoteesi hylätään.

Vahvan korrelaation omaavista muuttujista toisen pudottua pois, mallin selitysaste kasvoi. Nyt mallin voidaan todeta olevan pätevä.

Huomautus. Multikollineaarisuus ei välttämättä ole ongelma, mikäli mallin tarkoitus on ennustaa, ja ennustetulla jaksolla multikollineaarisuus on sama kuin mallissakin.

5.2 Mittavirheet

Oletuksen 3.6. mukaan mallin muuttujat ovat muodostettu oikein. Virheet muuttujissa viittaavat siis tilanteeseen, jossa regressiomallin muuttujien valinta on suoritettu virheellisesti. Tämän olettamuksen tärkeys on selkeästi ymmärrettävissä: Jos X_i tai Y_i on laskettu väärin, niin pienimmän neliösumman menetelmällä lasketut estimaatit $\hat{\beta}_i$ eivät ole paikkaansa pitäviä, eikä niiden käyttö tuota oikeellisia tuloksia.

Mittavirheiden tarkasteluun ei ole virallista testiä. Ongelma voi löytyä ekonometrian teorian avulla tai tietämyksellä siitä, miten aineisto on kerätty. (Kahane 2008)

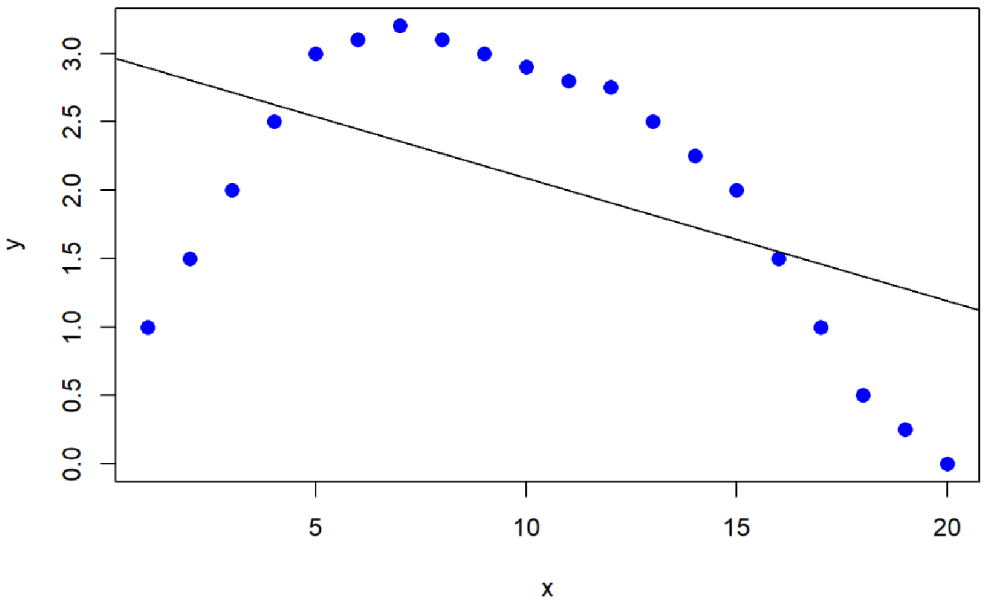
6 Regressiomallin tarkastelu

6.1 Lineaarisuusoletus

Oletusta 3.7. kutsutaan lineaarisuusoletukseksi. Lineaarisuusoletuksen mukaan regressiomallin on oltava määritelmän 2.2 mukaista muotoa. Jos näin ei ole, malli voi olla esimerkiksi polynominen tai logaritminen, ja mallia tulee muuntaa. Lineaariselle mallille tehtävät testit ja siitä lasketut luottamusvälit tuottavat luotettavampia tuloksia kuin ei-lineaariselle mallille tehtävät testit, joten mallin muokkaaminen kannattaa.

Lineaarinen malli mallintaa suoraviivaisia kausaalisuhteita, ja mikäli lineaarisuusoletus ei täyty, ei suoraviivaisia suhteita selittäjien ja selitettävän muuttujan välillä ole. Kuitenkin mallissa voi olla epälineaarisia suhteita, kuten kuvasta 6.1. näkee.

Kun muuttujia on vähän, lineaarisuusoletus voidaan validoida helposti tarkastelemalla tiettyjä kuvaajia, mutta useampien muuttujien tapauksessa validointi on haasteellisempaa. Se on kuitenkin mahdollista jäännöksiä tarkastelemalla, kun malli on sovitettu aineistoon. Jos epälineaarisuus on lievää, sitä voidaan korjata logaritmi- tai neliöjuurimuunnoksilla, ja mallin selitysaste paranee. Vahvan epälineaarisuuden korjaamiseksi voidaan malliin esimerkiksi lisätä muuttujia. (Chatterjee ja Hadi 2006.)



Kuva 6.1. Epälineaariseen dataan sovitettu lineaarinen malli.

6.2 Puuttuvan muuttujan harha

Oletuksen 3.8. mukaan mallista ei puutu merkittäviä selittäviä muuttujia. Jos merkittäviä muuttujia kuitenkin puuttuu, estimoitu malli ei ole parhain mahdollinen. Koska tämäkin oletus on vaikea validoida, mallin muodostusvaiheessa on oltava tarkkana, ja pyrittävä huomioimaan kaikki tarpeelliset muuttujat maalaisjärkeä käyttäen. (Kahane 2008.)

Esimerkki 6.1. Esimerkin 5.1. kuvitteellinen regressiomalli naisten palkasta rahoitusalalla sisältää nyt muuttujat naisen iästä sekä työkokemuksesta vuosina. Huomataan, että myös naisen koulutustaustalla on huomattavaa vaikutusta tienattuun palkkaan. Kun tämä muuttuja lisätään malliin, on mallin selitysaste korkeampi, ja puuttuvan muuttujan aiheuttama harha katoaa.

Huomautus. Jos harha ei ole suuri, voi sen jättää malliin, kunhan sen olemassaolon muistaa tulosten tulkintavaiheessa.

6.3 Ekologinen virhepäätelmä

Viimeisessä regressioanalyysin vaiheessa, *Mallin käyttäminen ongelman ratkaisuun*, on ekologisen virhepäätelmän mahdollisuus. Ekologinen virhepäätelmä tarkoittaa yksilötason päätelmien yleistämistä virheellisesti (Herger 2017).

1. A korreloi B:hen.
2. a kuuluu A:han.
3. Siispä a korreloi B:hen.

Esimerkki 6.2. Lämpimällä säällä tapahtuu paljon hukkumisia. Rillu menee uimaan lämpimällä säällä. Siispä Rillu on vaarassa hukkua.

7 Havainnot

7.1 Havaintojen oikeellisuus

Oletuksen 3.9. mukaan kaikki aineiston havainnot ovat yhtä luotettavia. Mikäli havainnot olisivat virheellisiä, päädyttäisiin päätelmissäkin luonnollisesti virheellisiin tuloksiin. Havaintojen oikeellisuuden takaamiseksi aineiston keräämisessä ja käsittelyssä tulee olla tarkkana, ja valmis aineisto kannattaa valita luotettavuuden perusteella. (Kahane 2008.)

7.2 Poikkeavat havainnot ja vipuvaikutus

Havaintoja tarkasteltaessa voidaan törmätä aineistossa poikkeaviin havaintoihin. Ne ovat poikkeuksellisen suuria tai pieniä arvoja verrattuna muihin havaintoihin, ja ne voivat vaikeuttaa mallin valintaa tai estimointia. Poikkeavat havainnot voivat vaikuttaa merkittävästi tutkimustuloksiin. Niiden aiheuttamaa vaikutusta regressioanalyysissä kutsutaan vipuvaikutukseksi, jota voidaan tarkastella poistamalla poikkeavia havaintoja aineistosta, ja vertailemalla kuvaajia sekä saatuja tunnuslukuja. (Freund, Wilson ja Sa 2006.)

Poikkeavia havaintoja voidaan tunnistaa esimerkiksi Cookin etäisyyden avulla. Cookin etäisyyttä merkitään

$$C_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(k+1)\sigma^2}, \tag{7.1}$$

jossa $\hat{Y}_j$ ovat mallin, joka on estimoitu kaikilla havainnoilla, sovitteet, ja $\hat{Y}_{j(i)}$ ovat mallin, joka on estimoitu kaikilla muilla havainnolla paitsi i, sovitteet. Havainto i on poikkeava havainto, jos sitä vastaava Cookin etäisyys C_i on

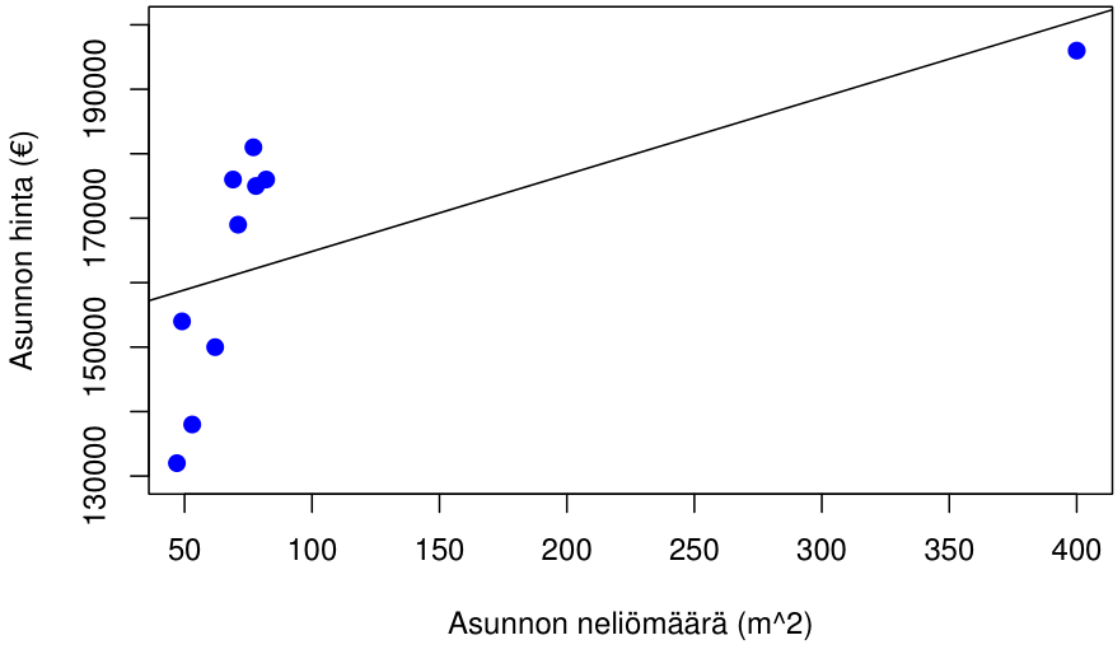
$$C_i > \frac{8}{n - 2(k+1)}, \tag{7.2}$$

jossa n on havaintojen määrä ja k on selitettävien muuttujien lukumäärä. (Seppälä 2015.)

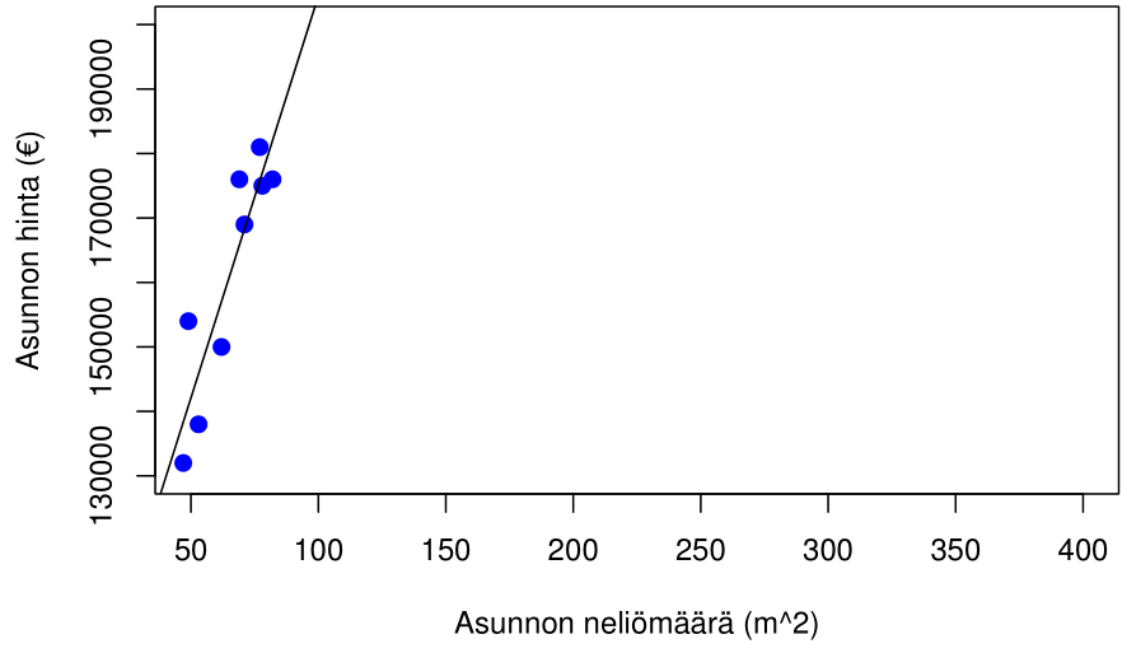
Esimerkki 7.1. Regressiomallissa, joka kuvaa asunnon neliömäärän vaikutusta hintaan, aineiston asuntojen neliömäärät jakautuvat pääasiallisesti viidenkymmenen ja sadan välille. Yksi asunto on kuitenkin suurempi, 400 neliötä. Kuva 7.1. visualisoi aineistosta muodostetun regressiosuoran. Malli ei selvästi ole hyvä, sillä havainnot eivät jakaudu suoran ympärille tasaisesti.

Poistetaan aineistosta poikkeava havainto, ja muodostetaan uusi regressiomalli, kuva 7.2. Nyt kuvaajaa tarkasteltaessa huomataan, että malli on huomattavasti käyttökelpoisempi, sillä havainnot jakautuvat regressiosuoran ympärille tasaisemmin. (Liite 1.)

Huomautus. Poikkeavia havaintoja ei saa poistaa ilman perusteluja, ja tutkimustulosten käsittelyssä havaintojen poistamisesta tulee aina muistaa raportoida.



Kuva 7.1. Regressiosuora aineistossa, jossa on yksi huomattavasti muista havainnoista poikkeava havainto.



Kuva 7.2. Regressiosuora samassa aineistossa kuin kuvassa 7.1., kun poikkeava havainto on poistettu aineistosta.

Lähteet

(1) Chatterjee, S., Hadi, A. (2006). *Regression Analysis by Example*, Wiley, USA.

(2) Freund, R., Wilson, W., Sa, P. (2006). *Regression Analysis - Statistical Modeling of a Response Variable*, Elsevier, USA.

(3) Herger, N. (2017). *When Does the Ecological Fallacy Vanish in Linear Regressions?* - SSRN Electronic Journal.

(4) Kahane, L. (2008). *Regression Basics*, Sage Publications, USA.

(5) Seppälä, H. (2015). *Regressiodiagnostiikka ja regressiomallin valinta - Ennustaminen ja Aikasarja-analyysi*, luentodiat. Matematiikan ja systeemianalyysin laitos, Perustieteiden korkeakoulu, Aalto-yliopisto.

(6) Webster, A. (2013). *Introductory Regression Analysis With Computer Application for Business ans Economics*, Routledge, New York.

(7) *Tilastotieteen peruskurssien harjoitusaineistoja 2003-2018*, Tutkimusmenetelmien työkalupakki, Tampereen Yliopisto.

Liite 1

Liite 1: R-koodit ja aineistot

Esimerkki ja kuva 4.1.

```
lapset <-read.table(file="C:\\Users\\Aino Satalahti\\OneDrive - TUNI.fi\\Työpöytä\\R\\Lapset_2007.txt",
attach(lapset)
lapset
```

```
##     Sukup pituus1 paino1 Esikoine
## 1       1      36    990        0
## 2       0      48   3150        1
## 3       1      48   2755        0
## 4       1      48   2830        0
## 5       0      57   4460        1
## 6       1      46   2820        1
## 7       0      54   4200        1
## 8       0      51   4300        0
## 9       0      51   3685        0
## 10      1      49   2840        1
## 11      0      50   3395        1
## 12      1      48   2710        1
## 13      1      46   2510        1
## 14      1      51   3510        1
## 15      1      50   3330        0
## 16      0      50   3990        0
## 17      0      52   4120        0
## 18      1      52   3900        0
## 19      1      48   2860        1
## 20      0      51   3785        0
## 21      0      53   4210        0
## 22      0      50   3080        0
## 23      1      50   2980        1
## 24      0      53   4270        1
## 25      1      48   2755        0
## 26      1      48   2830        0
## 27      0      53   3770        1
## 28      1      52   3800        0
## 29      0      52   3790        1
## 30      1      50   3250        0
## 31      1      50   3270        0
## 32      0      47   2510        0
## 33      1      50   3600        0
## 34      1      48   2620        1
## 35      1      50   3335        0
## 36      0      54   4125        0
## 37      0      50   3350        1
```

```
## 38      1      47   3025         0
## 39      1      50   3090         0
## 40      0      52    323         0
## 41      1      53   4260         0
## 42      0      51   3625         1
## 43      1      50   3360         0
## 44      1      56   4540         0
## 45      1      50   3325         1
## 46      1      52   4130         0
## 47      0      52   3940         1
## 48      0      51   3315         1
## 49      1      50   2380         1
## 50      1      49   3060         1
## 51      0      49   3455         0
## 52      0      49   3830         0
## 53      0      54   3820         0
## 54      1      49   3325         1
## 55      0      55   4000         0
## 56      0      51   3625         0
## 57      0      51   3500         0
## 58      0      49   2800         0
## 59      0      55   3750         1
## 60      0      52   3385         1
## 61      1      51   3870         0
## 62      0      50   3500         0
## 63      0      49   3930         0
## 64      1      45   2020         1
## 65      0      51   2900         1
## 66      0      52   3590         1
## 67      1      50   3760         1
## 68      1      52   3550         1
## 69      1      50   2600         0
## 70      1      53   3900         0
## 71      0      52   3790         1
## 72      1      51   3305         0
## 73      1      53   3740         0
## 74      0      54   3820         0
## 75      0      52   3930         0
## 76      0      53   4450         1
## 77      0      50   3315         1
## 78      0      52   4160         0
## 79      1      47   2820         0
## 80      1      51   4110         0
## 81      1      52   3520         1
## 82      1      47   2245         0
## 83      1      46   2680         0
## 84      0      47   2835         1
## 85      1      50   3450         1
## 86      1      48   3065         1
## 87      0      55   4760         0
## 88      0      53   3830         1
## 89      0      54   4200         0
## 90      0      53   3390         1
## 91      0      51   4030         1
```

```
## 92      0      51   3890        1
## 93      0      51   3880        0
## 94      0      53   3640        1
## 95      1      51   3600        0
## 96      1      47   2920        0
## 97      1      50   3290        0
## 98      1      52   4290        0
## 99      0      49   3220        0
## 100     1      51   3540        0
```

```r
lapset.lm <- lm(paino1 ~ pituus1)
summary(lapset.lm)
```

```
##
## Call:
## lm(formula = paino1 ~ pituus1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3379.9  -188.5    26.6   216.8   779.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5797.33     868.86  -6.672 1.51e-09 ***
## pituus1       182.70      17.18  10.634  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 470.2 on 98 degrees of freedom
## Multiple R-squared:  0.5357, Adjusted R-squared:  0.531
## F-statistic: 113.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
plot(lapset.lm)
```

Kuva 4.2.

```r
# Aineisto on kuvitteellinen ja luotu esimerkkiä varten.
x <- c(2, 1, 3, 2, 4, 3)
y <- c(1, 2, 3, 4, 5, 6)
plot(y, x, pch = 16, cex = 1.3, col = "blue", main = "Jäännökset", xlab ="x", ylab = "y")
abline(lm(x ~ y))
```

Kuva 4.3.

```r
# Aineisto on kuvitteellinen ja luotu esimerkkiä varten.
c <- c(2, 2.2, 1.9, 2.3, 2.5, 2.3, 2.6, 2.4, 2.7, 3, 3.3, 3, 3.1, 3.3, 3.5, 3.3, 3.5, 3.9, 3.7, 4)
d <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20)
plot(d, c, pch = 16, cex = 1.3, col = "blue", main = "Homoskedastinen aineisto", xlab ="x", ylab = "y")
abline(lm(c ~ d))
```

Kuva 4.4.

```r
# Aineisto on kuvitteellinen ja luotu esimerkkiä varten.
e <- c(2, 2.2, 1.9, 2.3, 2.5, 2.3, 2.6, 2.4, 2.7, 3, 3.3, 3, 3.1, 3.8, 3, 3.6, 4, 4.5,
```

```
      3.9, 3.1, 3.7, 4.2, 3.7, 4, 3.4, 4)
f <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 14.5, 15, 15.5, 16, 16.5, 17,
       17.5, 18, 18.5, 19, 19.5, 20)
plot(f, e, pch = 16, cex = 1.3, col = "blue", main = "Heteroskedastinen aineisto", xlab ="x", ylab = "y
abline(lm(e ~ f))
```

Esimerkki 5.2.

```
asunnot <-read.table(file="C:\\Users\\Aino Satalahti\\OneDrive - TUNI.fi\\Työpöytä\\R\\asunnot.txt", hea
attach(asunnot)
asunnot
```

```
##    ID   HINTA NELIOT HUONELKM SAUNA ALUE    NELIOHIN ASUNTO KESKUSTA
## 1   6  435000     83        4     0    3 5240,963855      4        0
## 2   7  390000     78        3     0    2        5000      3        0
## 3   9  635000     77        3     1    2 8246,753247      3        0
## 4  11  450000     74        3     1    2 6081,081081      3        0
## 5  16  375000     64        2     0    3    5859,375      2        0
## 6  17  325000     62        2     0    3 5241,935484      2        0
## 7  19  510000     56        2     1    3 9107,142857      2        0
## 8  21  285000     51        2     0    3 5588,235294      2        0
## 9  22  348000     48        2     0    3        7250      2        0
## 10 23  295000     45        2     0    3 6555,555556      2        0
## 11 26  285000     36        1     0    3 7916,666667      1        0
## 12 28  258000     35        1     0    3 7371,428571      1        0
## 13 29  255000     29        1     0    3 8793,103448      1        0
## 14 30  250000     30        1     0    3 8333,333333      1        0
## 15 35  910200     76        3     1    3 11976,31579      3        0
## 16 39  478000     74        3     1    3 6459,459459      3        0
## 17 40  415000     73        3     0    2 5684,931507      3        0
## 18 41  349000     70        3     0    3 4985,714286      3        0
## 19 42  545000     63        2     1    3 8650,793651      2        0
## 20 43  352000     63        2     1    3 5587,301587      2        0
## 21 44  539000     60        2     1    3 8983,333333      2        0
## 22 49  297000     39        1     0    2 7615,384615      1        0
## 23 50  219000     27        1     0    2 8111,111111      1        0
## 24 53  198000     23        1     0    3 8608,695652      1        0
## 25 54  198000     30        1     0    3        6600      1        0
## 26 55  234000     27        2     0    3 8666,666667      2        0
## 27  1 1400000    148        4     1    1 9459,459459      4        1
## 28  2 1390000    133        3     1    1 10451,12782      3        1
## 29  3 1365000    118        5     1    1 11567,79661      4        1
## 30  4  950000     87        3     1    1 10919,54023      3        1
## 31  5  920000     86        4     1    1 10697,67442      4        1
## 32  8  695000     78        3     0    1  8910,25641      3        1
## 33 10  835000     76        3     1    1 10986,84211      3        1
## 34 12  550000     73        3     0    1 7534,246575      3        1
## 35 13  625000     67        3     0    1 9328,358209      3        1
## 36 14  550000     67        3     1    1 8208,955224      3        1
## 37 15  698000     65        3     1    1 10738,46154      3        1
## 38 18  545000     61        2     0    1  8934,42623      2        1
## 39 20  486000     54        2     0    1        9000      2        1
## 40 24  360000     43        1     0    1 8372,093023      1        1
```

```
## 41 25  269000     40        1     0    1        6725      1        1
## 42 27  295000     35        1     0    1 8428,571429      1        1
## 43 31  440000     43        1     0    1 10232,55814      1        1
## 44 32  535000     45        2     1    1 11888,88889      2        1
## 45 33  570000     61        2     0    1 9344,262295      2        1
## 46 34  575000     63        2     0    1 9126,984127      2        1
## 47 36  900000     78        2     1    1 11538,46154      2        1
## 48 37 1400000    133        3     1    1 10526,31579      3        1
## 49 38  839000     90        4     0    1 9322,222222      4        1
## 50 45  549000     55        2     0    1 9981,818182      2        1
## 51 46  439000     54        2     0    1  8129,62963      2        1
## 52 47  480000     51        2     0    1 9411,764706      2        1
## 53 48  452000     46        2     0    1 9826,086957      2        1
## 54 51  450000     46        1     0    1 9782,608696      1        1
## 55 52 1350000    122        4     1    1 11065,57377      4        1
## 56 56  350000     44        2     0    1 7954,545455      2        1
```

```
cor(NELIOT,HUONELKM)
```

```
## [1] 0.8325038
```

```
malli <- lm(HINTA ~ NELIOT + HUONELKM)
malli
```

```
##
## Call:
## lm(formula = HINTA ~ NELIOT + HUONELKM)
##
## Coefficients:
## (Intercept)       NELIOT     HUONELKM
##     -110697        12036       -42943
```

```
summary(malli)
```

```
##
## Call:
## lm(formula = HINTA ~ NELIOT + HUONELKM)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -309281  -50589   15212   65764  270166
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -110697      46378  -2.387   0.0206 *
## NELIOT         12036       1165  10.331 2.66e-14 ***
## HUONELKM      -42943      32513  -1.321   0.1922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 131700 on 53 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8345
## F-statistic: 139.7 on 2 and 53 DF,  p-value: < 2.2e-16
```

```
malli2 <- lm(HINTA ~ NELIOT)
malli2
```

```
##
## Call:
## lm(formula = HINTA ~ NELIOT)
##
## Coefficients:
## (Intercept)       NELIOT
##     -127529        10755
```

```
summary(malli2)
```

```
##
## Call:
## lm(formula = HINTA ~ NELIOT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -330131  -47650   25017   83301  223446
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -127528.6    44899.2   -2.84  0.00634 **
## NELIOT        10754.9      649.9   16.55  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 132600 on 54 degrees of freedom
## Multiple R-squared:  0.8353, Adjusted R-squared:  0.8322
## F-statistic: 273.9 on 1 and 54 DF,  p-value: < 2.2e-16
```

Kuva 6.1.

```
# Aineisto on kuvitteellinen ja luotu esimerkkiä varten.
a <- c(1, 1.5, 2, 2.5, 3, 3.1, 3.2, 3.1, 3, 2.9, 2.8, 2.75, 2.5, 2.25, 2, 1.5, 1, 0.5, 0.25, 0)
b <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20)
plot(b, a, pch = 16, cex = 1.3, col = "blue", main = "Epälineaarisuus", xlab ="x", ylab = "y")
abline(lm(a ~ b))
```

Kuva 7.1.

```
# Aineisto on kuvitteellinen ja luotu esimerkkiä varten.
hinta2 <- c(176000, 154000, 138000, 196000, 132000, 176000, 181000, 169000, 150000, 175000)
neliot2 <- c(82, 49, 53, 400, 47, 69, 77, 71, 62, 78)
plot(neliot2, hinta2, pch = 16, cex = 1.3, col = "blue", main = "Regressiosuora neliömäärän vaikutukses
abline(lm(hinta2 ~ neliot2))
```

Kuva 7.2.

```
# Aineisto on kuvitteellinen ja luotu esimerkkiä varten.
hinta <- c(176000, 154000, 138000, 132000, 176000, 181000, 169000, 150000, 175000)
neliot <- c(82, 49, 53, 47, 69, 77, 71, 62, 78)
plot(neliot, hinta, pch = 16, cex = 1.3, col = "blue", main = "Regressiosuora neliömäärän vaikutuksesta
abline(lm(hinta ~ neliot))
```