Diagnostiset menetelmät regressioanalyysissä

Aino-Maria Rapala

LuK-tutkielma
Maaliskuu 2025

matematiikan ja tilastotieteen laitos

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Aino-Maria Rapala: Diagnostiset menetelmät regressioanalyysissä
LuK-tutkielma -tutkielma, 26 s.
Tilastotiede
Maaliskuu 2025

Tässä työssä esitellään diagnostisia menetelmiä, joita käytetään regressioanalyysissä muun muassa mallin sopivuuden ja oletusten voimassaolon tarkasteluun. Regressioanalyysin käyttö on hyödyllistä, mutta vaatii huolellista mallin arviointia luotettavien johtopäätösten tekemiseksi. Diagnostinen tarkastelu antaa tärkeää tietoa sovitetun mallin sopivuudesta ja mahdollistaa näin mallin parantamisen tai vaihtoehtoisen mallin käytön.

Työ alkaa yleisen lineaarisen mallin ja oletusten esittelyllä. Tämän jälkeen käydään läpi sovitteet, residuaalit ja standardoidut residuaalit, jotka ovat oleellisessa osassa mallin suurimman uskottavuuden estimoinnissa. Näiden jälkeen seuraa yleisimpien diagnostisten menetelmien läpikäyminen, joita ovat muun muassa mallin harhattomuus, homoskedastisuus ja virhetermin normaalisuus. Menetelmien käytön soveltamista ja tulkintaa havainnollistetaan esimerkkiaineistojen avulla. Yhteenvedossa mainitaan asioita, joita työn rajauksen vuoksi jouduttiin jättämään pois. Osiossa on kirjattu aiheita, joita voisi tulevissa töissä käsitellä laajemmin.

Diagnostisten menetelmien käyttö ja osaaminen on tärkeää kaikille, jotka käyttävät regressioanalyysiä tutkimuksissaan. Mikäli diagnostiikkaa ei suoriteta, riski virheellisten johtopäätösten tekemiselle ja ennusteiden epätarkkuudelle kasvaa merkittävästi. Työn lukija saa alustavaa perustietoa ja ymmärrystä regressioanalyysistä ja diagnostisista menetelmistä sekä niiden tärkeydestä tulosten analysoinnissa ja johtopäätösten tekemisessä.

Asiasanat: diagnostiset menetelmät, regressioanalyysi, oletukset, harhattomuus, homoskedastisuus, normaalisuus, korreloimattomuus, multikollineaarisuus, poikkeavat havainnot

Sisällys

1 Johdanto

Regressioanalyysi on yksi keskeisimmistä menetelmistä tilastotieteessä ja data-analyysissä. Sillä mallinnetaan muuttujien välisiä riippuvuussuhteita ja pyritään selittämään, miten yksi tai useampi numeerinen selittävä muuttuja vaikuttaa jatkuvaan vastemuuttujaan. Regressioanalyysin diagnostisten menetelmien osaaminen on hyödyllistä, sillä se on käytössä laajasti eri tieteenaloilla, esimerkiksi lääke- ja taloustieteessä. Tässä työssä käytetään lineaarista mallia, joka on regressioanalyysin yksinkertaisin ja yleisin muoto. Oikein käytettynä malli kertoo kattavasti aineistosta ja sen avulla voidaan tehdä hyviä päätelmiä muuttujien riippuvuussuhteista.

Lineaariseen malliin liitetään monia oletuksia, joiden voimassaolo on syytä tarkastaa. Oletuksia ovat muun muassa vastemuuttujan ja selittävien muuttujien lineaarinen suhde ja virhetermien normaalisuusoletus. Olennaisena osana regressioanalyysissä on mallin oletusten tarkistaminen ja mallin soveltuvuuden arviointi diagnostisten menetelmien avulla. Diagnostiikan avulla voidaan tarkastella mm. multikollineaarisuutta, poikkeavia havaintoja ja mallin harhattomuutta. Oletusten pätiessä aineistoon sovitettua mallia voidaan pitää luotettavana ja tehtyjä johtopäätöksiä oikeina.

Työn lähteinä käytetty pääasiassa Nyberg, H. *Lineaariset ja yleistetyt lineaariset malli* -luentomonistetta, josta seurattu jaksoja 1.2, 2.1, 3.1, 3.2 ja 3.6 [10] ja Agresti, A. *Foundations of Linear and Generalized Linear models* -teosta, josta viitattu lukuihin 2 ja 2.5. [1] Muita lähteitä on ollut Mellin, I. teokset *Lineaarinen regressioanalyysi* (jaksot 16 ja 18)[7] ja *Tilastollinen päättely* (jakso 8.1). [8] Diagnostiikassa ja hajontakuvioiden tulkinnassa on käytetty apuna myös verkkolähteitä Bobbit, Z. *How to create a Residual Plot in R.* [2] ja University of Virginia: *Understanding Diagnostic Plots for Linear Regression Analysis.* [4]

2 Lineaarinen malli

Regressioanalyysi on ehkä eniten sovellettu ja tunnetuin tilastotieteen menetelmä. Eräs regressioanalyysin yleisimmistä malleista on yleinen lineaarinen malli, joka pyrkii selittämään vastemuuttujan havaittujen arvojen vaihtelun yhden tai useamman selittävän muuttujan arvojen avulla.

2.1 Yleinen lineaarinen malli

Yleisen lineaarisen mallin voi määritellä seuraavalla tavalla. Lineaarisessa mallissa oletetaan selittävien muuttujien ja vastemuuttujan välisen yhteyden olevan lineaarista. Mallissa oletetaan olevan n havaintoyksikköä, jotka ovat toisistaan riippumattomia. Usean selittäjän lineaarisen mallin malliyhtälö on muotoa:

$$\begin{aligned} Y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \\ &= \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i, \qquad\qquad i = 1, \ldots, n \end{aligned}$$

jossa

$$\begin{aligned} Y_i &= \text{vastemuuttujan arvo (havaitut arvot merkitään } y_1, \ldots, y_n) \\ \beta_1, \ldots, \beta_p &= \text{tuntemattomat parametrit} \\ x_{i1}, \ldots, x_{ip} &= \text{selittävien muuttujien havaitut arvot} \\ \varepsilon_i &= \text{satunnainen ja ei-havaittava jäännös- tai virhetermi.} \end{aligned}$$

Malliyhtälö voidaan esittää myös matriisimuodossa:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

jossa

$$\begin{aligned} \mathbf{Y} &= n \times 1 \text{ -matriisi} \\ \mathbf{X} &= n \times p \text{ -matriisi} \\ \boldsymbol{\beta} &= p \times 1 \text{ -matriisi} \\ \boldsymbol{\varepsilon} &= n \times 1 \text{ -matriisi} \end{aligned}$$

Tiukan tilastollisen linjauksen mukaan kyseessä ei vielä ole tilastollinen malli, sillä se vaatii määrittelyyn mukaan havaintojen yhteistodennäköisyysjakauman ja parametriavaruuden spesifioinnin. Havaintojen $Y_1, \ldots, Y_n$ yhteistodennäköisyysjakauma saadaan spesifioitua, kun virhetermeihin liitetään seuraava oletus:

$$\varepsilon_1, \ldots, \varepsilon_n \perp\!\!\!\perp , \ \varepsilon_i \sim \mathrm{N}(0, \sigma^2). \tag{1}$$

Kun merkitään $\boldsymbol{x}_i = \begin{bmatrix} x_{i1} & \cdots & x_{ip} \end{bmatrix}'$ ja $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \cdots & \beta_p \end{bmatrix}'$, saadaan lineaariselle mallille esitys

$$Y_1, \ldots, Y_n \perp\!\!\!\perp , \ Y_i \sim \mathrm{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2), \ \boldsymbol{\beta} \in \mathbb{R}^p, \ \sigma^2 > 0.$$

Virhetermien oletus (1) voidaan lausua myös satunnaisvektorin ε avulla. Satunnaisvektori $\boldsymbol{\varepsilon}$ noudattaa multinormaalijakaumaa odotusarvona nolla ja kovarianssimatriisina $\sigma^2\boldsymbol{I}_n$ toisin sanoen $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$, jossa $\boldsymbol{I}_n$ on $n \times n$ yksikkömatriisi. Näin lineaarinen malli voidaan kirjoittaa muodossa:

$$\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n), \ \boldsymbol{\beta} \in \mathbb{R}^p, \ \sigma^2 > 0.$$

2.2 Lineaarisen mallin oletukset

Lineaarisella mallilla on monia oletuksia, jotka tulee ottaa huomioon mallia käytettäessä ja tulkittaessa. Oletusten pätevyyttä on oleellista arvioida ennen mallin soveltamista. Jos oletukset eivät täyty riittävästi, tuloksiin ja tulkintoihin tulee suhtautua varauksella tai toisen menetelmän käyttöä kannattaa harkita. Oletuksia ovat:

- Satunnaismuuttujat noudattavat normaalijakaumaa:

$$Y_i \sim \mathrm{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2). \tag{2}$$

  Eksakti normaalisuus ei kuitenkaan ole kriittinen oletus lineaarisen mallin sopivuudelle.

- Mallin vastemuuttujan ja selittävien muuttujien välinen yhteys oltava *odotusarvoisesti lineaarista*. Mielenkiinto kohdistuukin juuri satunnaismuuttujien Y_i odotusarvoihin. Selittävien muuttujien ja odotusarvojen $E(Y_i)$ välinen suhde tulisi olla muotoa:

$$\mathrm{E}(Y_i) = \mu_i = \sum_{j=1}^{p} \beta_j x_{ij} = \beta_i x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{x}_i'\boldsymbol{\beta},\ i = 1, \ldots, n \tag{3}$$

  Tällöin malliyhtälö voidaan esittää muodossa:

$$Y_i = \mu_i + \varepsilon_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,\ i = 1, \ldots, n$$

- Varianssin oletetaan olevan sama kaikille havaintoyksiköille i. Toisin sanoen selittävien muuttujien arvot ja odotusarvo μ_i eivät vaikuta satunnaismuuttujien Y_i tuntemattomiin variansseihin. Tällöin oletetaan, että $\mathrm{Var}(Y_i) = \sigma^2$. Näin ollen myös kaikilla jäännöstermeillä tulee olla sama varianssi. Tätä oletusta kutsutaan *jäännösvarianssin homoskedastisuusoletukseksi.*

$$\mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2,\ i = 1, \ldots, n \tag{4}$$

- Täydellistä *multikollineaarisuutta* ei voida sallia eli selittävien muuttujien välillä ei voi olla täydellistä lineaarista riippuvuutta. Tämä tarkoittaa, että matriisin $\boldsymbol{X}$ $(n \times p)$ oletetaan olevan täysiasteinen eli

$$r(\boldsymbol{X}) = p \tag{5}$$

  jolloin pätee $n \geq p$. Tämän oletuksen pätiessä taataan, että odotusarvovektorilla $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Y})$ on yksikäsitteinen esitys $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$.

- Virhetermien odotusarvon tulee olla nolla. Toisin sanoen

$$\mathrm{E}(\varepsilon_i) = 0,\ i = 1, \ldots, n \tag{6}$$

- Virhetermille on voimassa seuraava oletus:

$$\varepsilon_1, \ldots, \varepsilon_n \perp\!\!\!\perp, \ \varepsilon_i \sim \mathrm{N}(0, \sigma^2) \tag{7}$$

Oletus (3) merkitsee, että virhetermin $\boldsymbol{\varepsilon}$ yhteistiheysfunktio on muotoa

$$f(\varepsilon_1, \ldots, \varepsilon_n; \sigma^2) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \boldsymbol{\varepsilon'\varepsilon} \right\}$$

ja näin ollen vastemuuttujan yhteistiheysfunktio on muotoa

$$f(y_1, \ldots, y_n; \boldsymbol{\beta}, \sigma^2) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \right\}$$

- Virhetermien riippumattomuudesta seuraa korreloimattomuus. Tällä tarkoitetaan *jäännöstermien korreloimattomuusoletusta*, jossa on kyse siitä, että jäännöstermit eivät korreloi keskenään:

$$\mathrm{Cor}(\varepsilon_i, \varepsilon_j) = 0, \ i \neq j \tag{8}$$

2.3 Yleisen lineaarisen mallin suurimman uskottavuuden estimointi

Suurimman uskottavuuden estimoinnissa käytetty lähteenä [11, jakso 1]. Suurimman uskottavuuden estimointi lähtee liikkeelle uskottavuusfunktion määrittelemisestä. Yhteistiheysfunktiosta seuraa, että havaintojen $y_i, i = 1, \ldots, n$ *uskottavuusfunktio* on

$$L(\boldsymbol{\beta}, \sigma^2; y_1, \ldots, y_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \right\}$$

Kun tästä muodostetaan *logaritminen uskottavuusfunktio*, se on muotoa:

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma^2; y_1, \ldots, y_n) &= \log L(\boldsymbol{\beta}, \sigma^2; y_1, \ldots, y_n) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \end{aligned}$$

Tämä funktio maksimoidaan $\boldsymbol{\beta}$ ja σ^2 suhteen. Funktio derivoidaan vektorin $\boldsymbol{\beta}$ suhteen ja merkitään derivaatta nollaksi:

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y})}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2\boldsymbol{X'y} + \boldsymbol{X'X\beta}) = \boldsymbol{0}$$

Aiemmin tehtiin oletus (5) matriisin $\boldsymbol{X}$ täysiasteisuudesta, joten näin ollen $\boldsymbol{\beta}$ voidaan ratkaista yllä olevasta yhtälöstä. Ratkaisuna saadaan suurimman uskottavuuden estimaattori

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

Ratkaisu on uskottavuusfunktion maksimi, sillä matriisi

$$-\frac{\partial l(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \frac{1}{2\sigma^2} \boldsymbol{X}'\boldsymbol{X}$$

on positiivisesti definiitti.

$\hat{\boldsymbol{\beta}}$ yhtyy lineaarisen mallin regressiokertoimien pienimmän neliösumman (PNS) estimaattiin, sillä logaritmisen uskottavuusfunktion maksimointi on ekvivalentti virhetermin ε_i neliösumman

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_2 x_{i3} - \cdots - \beta_p x_{ip})^2$$

minimoinnin kanssa.

Parametrin σ^2 suurimman uskottavuuden estimaattori saadaan kaavalla:

$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Estimaattori on kuitenkin harhainen, sillä $\mathrm{E}(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2$. Tämän vuoksi sille kannattaa tehdä muunnos:

$$S^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{n}{n-p}\hat{\sigma}^2.$$

Muunnoksen jälkeen estimaattori on harhaton eli $\mathrm{E}(S^2) = \sigma^2$.

2.4 Sovitteet ja residuaalit

Mallin sovituksen jäännökset ovat virhetermissä, joten ne sisältävät tiedon aineistosta, jota malli ei kykene selittämään. Tästä syystä ne ovat hyödyllisiä tutkittaessa mallin sopivuutta.

2.4.1 Sovite

Sovite eli ennuste on estimoidun mallin vastemuuttujalle antama arvo, kun selittäjällä on arvo $\boldsymbol{x_i}$. Sovite on mallin systemaattisen osan $\boldsymbol{X\beta}$ empiirinen vastine. Yksittäisen havainnon i sovite on $\hat{\mu}_i = \boldsymbol{x_i}'\hat{\boldsymbol{\beta}}$. Yksittäisten havaintojen sovitteista muodostuva sovitevektori on:

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

jossa $\mathbf{P}$ on $n \times n$ ortogonaalinen projektiomatriisi

$$\boldsymbol{P} = \left[p_{ij}\right] = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

2.4.2 Residuaalit

Vastemuuttujan havaitun arvon y_i ja sovitteen $\hat{\mu}_i$ erotusta kutsutaan residuaaliksi. ***Residuaali*** on virhetermin ε_i empiirinen vastine $\hat{\varepsilon}_i$:

$$\hat{\varepsilon}_i = y_i - \hat{\mu}_i = y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}},\ i = 1, \ldots, n$$

Residuaalivektori voidaan ilmaista vektoreiden ja matriisien avulla muodossa: $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Tämän voi esittää myös muodossa:

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}} &= (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} \\ \hat{\boldsymbol{\varepsilon}} &= (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\varepsilon} = \boldsymbol{M}\boldsymbol{\varepsilon}\end{aligned}$$

jossa $\mathbf{P}$ ja $\mathbf{M} = \boldsymbol{I}_n - \boldsymbol{P}$ ovat $n \times n$ ortogonaalisia projektiomatriiseja. Tästä havaitsee nyt residuaalien ja mallin ei-havaittavien virhetermien välisen yhteyden: $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{\varepsilon}$. Edelleen seuraa, että residuaalineliösumma voidaan kirjoittaa

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}$$

Oletusten ollessa voimassa residuaaleille pätee:

$$\mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0} \tag{9}$$

ja

$$\mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\boldsymbol{I}_n - \boldsymbol{P}) = \sigma^2\boldsymbol{M} \neq \sigma^2\boldsymbol{I}_n.$$

Erityisesti

$$\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - p_{ii}), \quad i = 1, \ldots, n \tag{10}$$

jossa $p_{ii} = [\boldsymbol{P}]_{ii}$ on projektiomatriisin $\mathbf{P}$ i. lävistäjäalkio.

Residuaalit voivat näin ollen olla korreloituneita ja niiden varianssit voivat vaihdella havaintoyksiköstä toiseen. Jos varianssit vaihtelevat, puhutaan heteroskedastisuudesta. Jotta residuaalit kuvaavat virhetermiä hyvin, $\boldsymbol{P}$:n lävistäjäalkioiden tulee olla pieniä ($0 \leq p_{ii} \leq 1$). Lisäksi niiden tulee olla suurin piirtein samansuuruisia, jotta homoskedastisuus toteutuu.

Residuaalit sisältävät tietoa virheiden varianssista eli parametristä $\sigma^2 = \mathrm{Var}(\varepsilon_i)$ ja mallin oletusten paikkansapitävyydestä, joten niiden tarkastelu on oleellisessa osassa regressiodiagnostiikassa.

2.4.3 Standardoidut residuaalit

Kun tutkitaan mallin oletuksia, kannattaa residuaalit standardoida varianssiltaan yhtäsuuriksi. Standardoitu residuaali havaintoyksikölle i voidaan määritellä alla olevalla tavalla:

$$r_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1 - p_{ii}}},\ p_{ii} = \left[\boldsymbol{P}\right]_{ii}, \tag{11}$$

jossa $s^2 = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{n-p}SSE$. SSE on jäännösneliösumma, joka on muotoa:

$$SSE = \sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}})^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Standardoitujen residuaalien varianssi on nyt $\mathrm{Var}(r_i) \approx 1$. Tämä standardoi varianssin likimain johtuen aiemmin mainitusta residuaalien ominaisuudesta (10) oletusten ollessa voimassa.

3 Regressiodiagnostiikka

Diagnostiikassa on käytetty aiempien lähteiden lisäksi lähteitä [2, 4]. Regressiodiagnostiikan tavoitteena on selvittää, kuvaako muodostettu malli selitettävän muuttujan ja selittäjien riippuvuutta oikein. Tämän yleisen ongelman tärkeä osaongelma on se, pätevätkö mallista tehdyt oletukset. Regressiodiagnostiikka on yksi osa mallista tehtyjen oletusten tarkistamista.

3.1 Regressiodiagnostiikan menetelmät

Diagnostiikalla tarkoitetaan mallin sopivuutta koskevia tarkasteluja, jotka tyypillisesti ovat jäännösvarianssin tarkastelua laajempia mallin sopivuuden arviointeja verrattuna mallista tehtäviin oletuksiin. Diagnostiset menetelmät pohjautuvat residuaaleihin ja eräisiin diagnostisiin mittoihin, joita tehdään graafisia apuvälineitä ja menetelmiä hyödyntäen. Tuloksia tulkittaessa ajatellaan, että spesifioinnissa tapahtuneet virheet nousevat esiin sovitetun mallin tuloksissa. Malli tulkitaan hyväksi, mikäli oletukset ja diagnostiikan avulla saadut tulokset ovat saman suuntaiset. Tällöin malli on luotettava ja kuvaa aineistoa hyvin. Selvitettäviä kohteita ovat muun muassa:

1. Mallin harhattomuus
2. Jäännösvarianssin homoskedastisuusoletus
3. Mallin virhetermin normaalisuus
4. Malln virhetermin korreloimattomuus
5. Vaikutusvaltaiset ja mahdolliset poikkeavat havainnot

3.2 Esimerkkiaineistojen esittely

Käytän diagnostisten menetelmien esittelyssä R:n valmiita aineistoja. Ensimmäinen aineisto on `palmerpenguins`-paketista löytyvä `penguins`-aineisto. Se koostuu yhteensä kahdeksasta muuttujasta ja 344 havainnosta. Aineiston tiedot on kerätty vuosina 2007-2009 kolmelta saarelta Palmerin saaristosta Antarktikselta. Esimerkeissäni olen muodostanut lineaarisen mallin, jossa vastemuuttujana on pingviinin paino ja selittävinä muuttujina ovat pingviinin nokan ja räpylän pituus. Muuttujien

nimet ovat `bill_length_mm` = nokan pituus, `flipper_length_mm` = räpylän pituus ja `body_mass_g` = paino.

Toinen aineisto on R:n `dplyr`-paketista löytyvä `mtcars` -aineisto, jossa on yhteensä 11 muuttujaa ja 32 havaintoa. Aineisto on kerätty Yhdysvalloissa vuoden 1974 Motor Trend -lehdestä. Aineisto koostuu 32 auton polttoaineen kulutuksesta ja 10:stä siihen mahdollisesti vaikuttavasta tekijästä. Autojen mallit ovat vuosilta 1973-1974. Esimerkkinä olen muodostanut lineaarisen mallin, jossa vastemuuttujana on gallonalla ajettavien mailien määrä ja selittävinä tekijöinä hevosvoimien lukumäärä ja sylinterin koko. Muuttujien nimet ovat `mpg` = mailia per gallona, `hp` = hevosvoimien lukumäärä ja `cyl` = sylinterien lukumäärä. Vakiotermi on sisällytetty malleihin.

Muodostin oman kolmannen esimerkkiaineiston, jossa otin satunnaisotoksen normaalijakaumasta alla olevalla koodilla. Aineisto havainnoi tilannetta, jossa oletukset lineaarisuudesta ja jäännösvarianssin homoskedastisuudesta eivät toteudu. Luodussa aineistossa muuttujien välinen suhde on neliöllinen.

Diagnostiikassa käytetyt kuvaajat ja tiivistelmät on tulostettu R:ssä alla olevalla koodilla.

```
# Muodostetaan lineaariset mallit
Malli1 <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
Malli2 <- lm(mpg ~ hp + cyl, data = mtcars)
# Tulostetaan sovitettujen mallien ominaisuuksia kuvaavat tulokset
summary(Malli1)
summary(Malli2)
# Tulostetaan diagnostiikassa käytetyt kuvaajat
plot(Malli1)
plot(Malli2)

# Muodostetaan satunnaisotos normaalijakaumasta
x <- rnorm(100)
y <- x^2 + rnorm(100)
data <- data.frame(x = x, y = y)
# Muodostetaan lineaarinen malli
Malli3 <- lm(y ~ x, data = data)
# Tulostetaan diagnostiikassa käytetyt kuvaajat
plot(Malli3)
```

Ennen siirtymistä diagnostiikan tarkempaan tarkasteluun, tutkitaan `summary`:n antamia tuloksia sovitetun mallin ominaisuuksista. Ensimmäisen aineiston mallissa tilastollisesti erittäin merkitsevä vaikutus painoon on räpylän pituudella ($p < 0.001$), mutta ei nokan pituudella ($p = 0.244$). Tuloksien mukaan räpylän pituuden kasvaessa yhdellä yksiköllä, paino kasvaa keskimäärin 48.145 grammaa. Selitysaste R^2 on 0.76 ja korjattu R^2 on 0.7585 eli malli selittää noin 76 % vastemuuttujan arvojen vaihtelusta. Jäännöskeskivirhe (Residual standard error) 394.1 vaikuttaa melko korkealta suhteessa arvojen vaihteluväliin, joka on 2700-6300 g. Tämä voi johtua esimerkiksi siitä, että mallissa ei ole kaikkia tarpeellisia selittäjiä. Tuloksien mukaan sovitettu malli ei ehkä ole paras mahdollinen ennustamaan pingviinin painoa, mutta tehdään diagnostiikan tarkastelu siitä huolimatta kyseiselle mallille.

Toisen aineiston mallissa tilastollisesti erittäin merkitsevä vaikutus mailien määrään on sylinterien lukumäärällä ($p < 0.001$), mutta ei hevosvoimilla ($p = 0.2125$). Sylinterien lukumäärän kasvu yhdellä yksiköllä vähentää keskimääräistä mailien määrää 2.264 yksikköä. Selitysaste R^2 on 0.74 ja korjattu R^2 on 0.7228 eli malli selittää noin 74 % vastemuuttujan arvojen vaihtelusta. Jäännöskeskivirhe on 3.173, mikä ei ole kovin matala, mutta ei myöskään kovin korkea suhteessa vastemuuttujan arvojen vaihteluväliin, joka on 10.4-33.9. Tässäkään tapauksessa sovitettu malli ei välttämättä ole paras mahdollinen, mutta katsotaan diagnostiikkaa myös tämän mallin avulla.

```
Malli1:
    Call:
lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
    data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-1090.5  -285.7   -32.1   244.2  1287.5

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
bill_length_mm        6.047      5.180   1.168    0.244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 394.1 on 339 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:   0.76,Adjusted R-squared:  0.7585
F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16

Malli2:
Call:
lm(formula = mpg ~ hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4948 -2.4901 -0.1828  1.9777  7.2934

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.90833    2.19080  16.847  < 2e-16 ***
hp          -0.01912    0.01500  -1.275  0.21253
cyl         -2.26469    0.57589  -3.933  0.00048 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

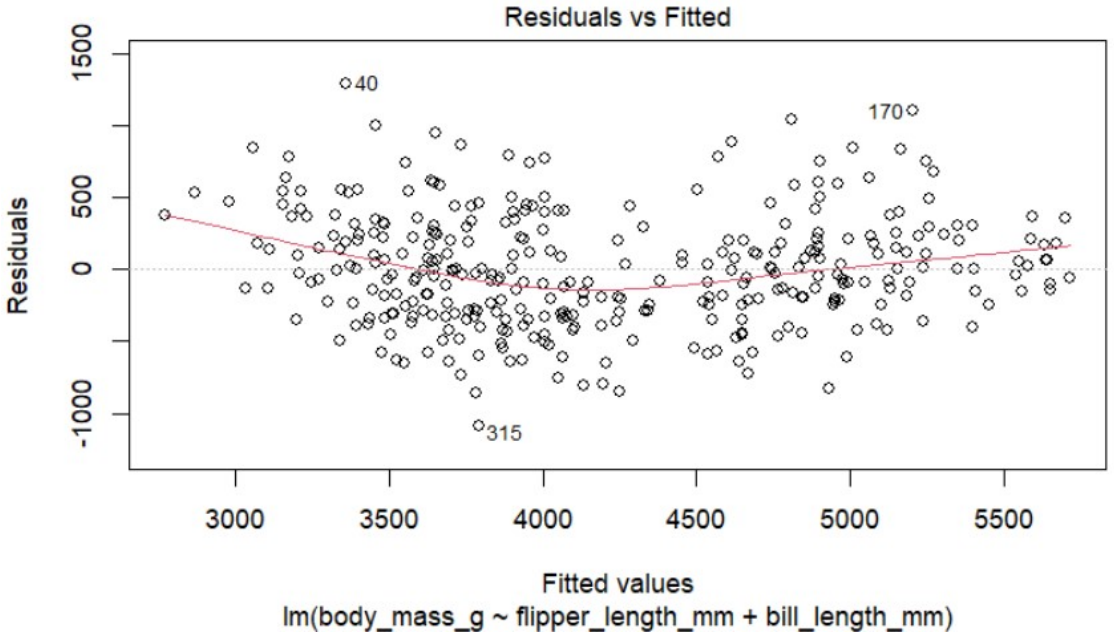
```
Residual standard error: 3.173 on 29 degrees of freedom
Multiple R-squared:  0.7407,Adjusted R-squared:  0.7228
F-statistic: 41.42 on 2 and 29 DF,  p-value: 3.162e-09
```

3.3 Mallin harhattomuus

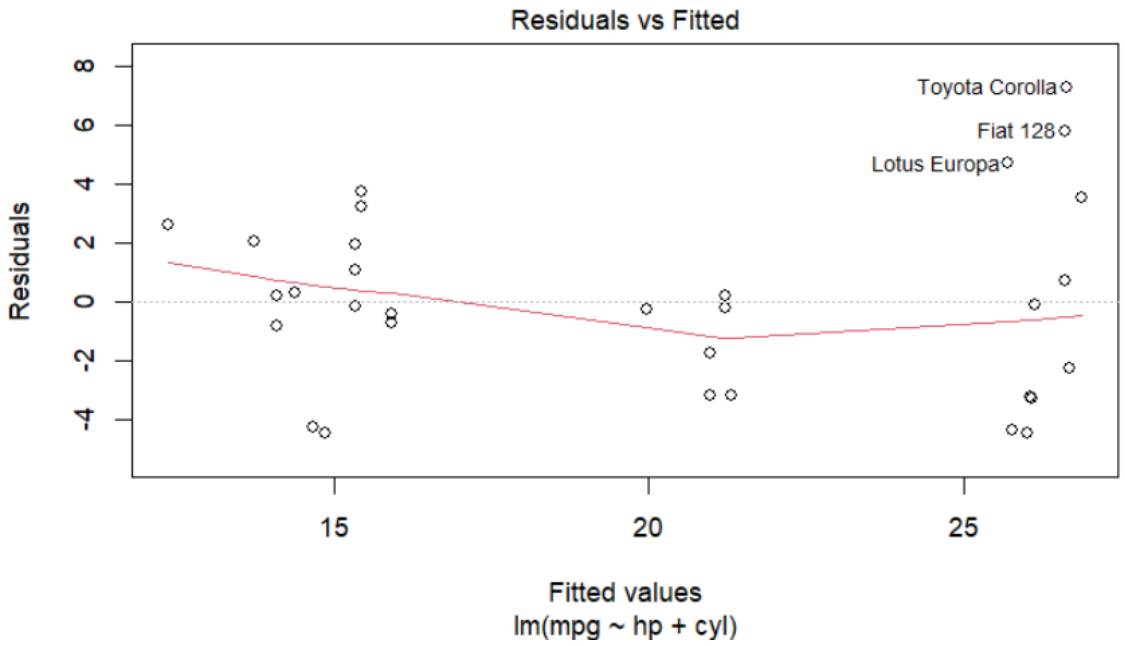
Mallin harhattomuus saadaan selvitettyä useimmiten graafisesti piirtämällä hajontakuvio residuaalien ja sovitteiden kesken. Näin saadaan selville, pitääkö mallin lineaarisuusoletus $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ paikkansa. Mallin ollessa harhaton, lineaarisuusoletus pätee. Vastemuuttujan ja selittävien muuttujien välinen suhde voi olla muutakin kuin lineaarista, ja residuaalien ja sovitteiden välinen hajontakuvio näyttää, mikäli malli ei huomioi suhdetta oikein.

Residuaalien tulee olla jakautunut keskimäärin nollan ympärille sovitteen arvosta riippumatta, koska oletuksena (6) on, että virhetermien ja niiden empiiristen vastineiden, residuaalien, odotusarvo on nolla. Toisin sanoen lineaarisen mallin ennusteiden keskimääräinen virhe on nolla. Mikäli residuaaleissa ei ole havaittavissa systemaattista kaavaa tai selkeää kuviota (esimerkiksi alaspäin aukeava paraabeli), se on merkki siitä, että suhde on lineaarinen niin kuin pitääkin.

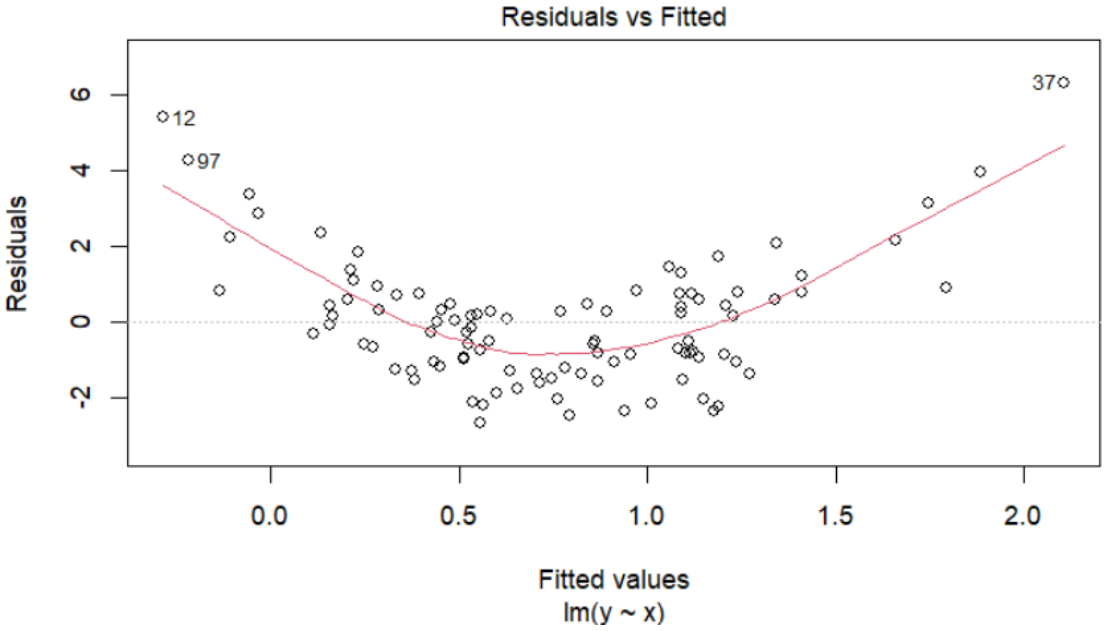
Alla olevien kahden ensimmäisen esimerkkiaineiston tapauksessa (`Malli1` ja `Malli2`) voimme ajatella lineaarisuuden olevan voimassa, sillä residuaalit ovat jakautuneet tasaisesti nollan molemmin puolin. Kolmas aineisto (`Malli3`) havainnollistaa tilannetta, jossa lineaarisuus ei toteudu, sillä pisteparvessa on havaittavissa paraabelin muotoa. Aineistoa tulisi kuvata toisella mallilla.



Kuva 1: Lineaarisuuden tarkastelu residuaalien ja sovitteiden hajontakuvion avulla (`Malli1`).



Kuva 2: Lineaarisuuden tarkastelu residuaalien ja sovitteiden hajontakuvion avulla (`Malli2`).

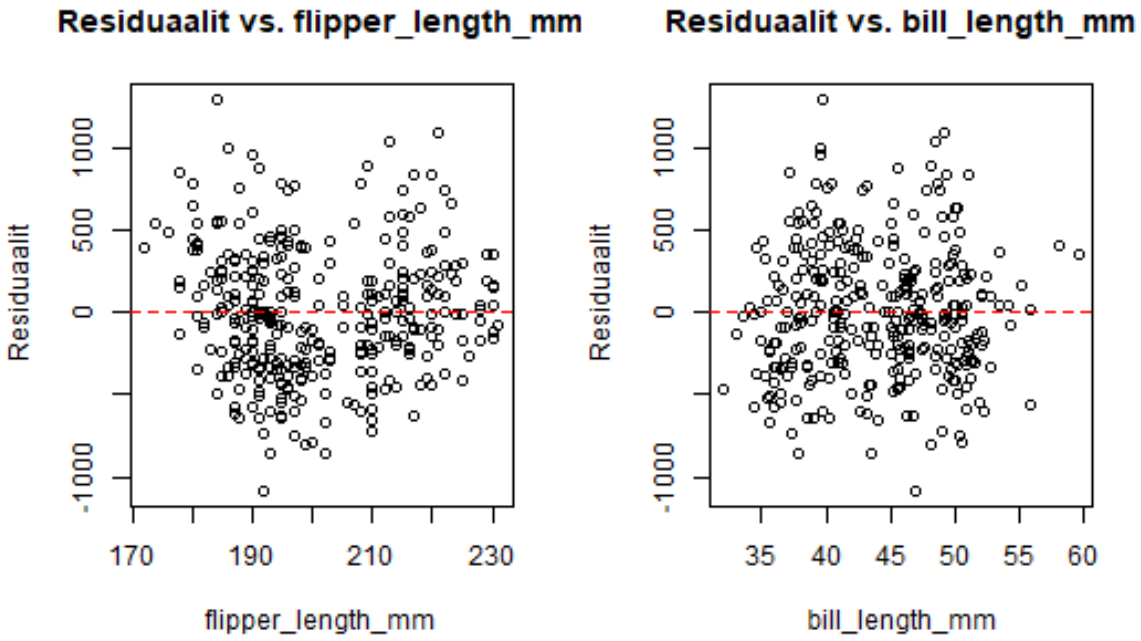


Kuva 3: Lineaarisuuden tarkastelu residuaalien ja sovitteiden hajontakuvion avulla (`Malli3`).

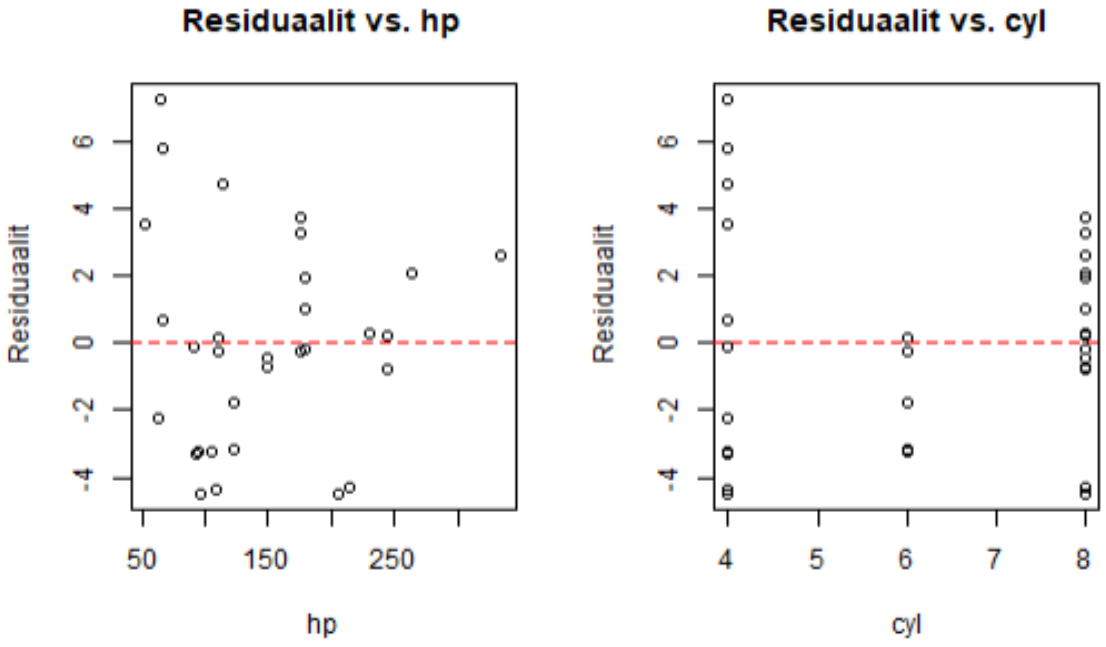
Harhattomuutta kuvataan usein myös yksittäisten selittäjien ja residuaalien välisellä hajontakuviolla. Tämä auttaa myös tarkastamaan lineaarisuusoletuksen (3) voimassaoloa. Kun residuaalien odotusarvo on nolla riippumatta selittäjien arvosta, pisteparvi on jakaantunut satunnaisesti nollan molemmin puolin ja näin ollen harhattomuusoletus on voimassa ja malli kuvaa lineaarista suhdetta hyvin. Alla olevat hajontakuviot tukevat aiemmin tehtyjä johtopäätöksiä lineaarisuusoletuksen voimassaolosta.

```
# Yksittäisen selittäjän ja residuaalien välinen hajontakuvio (Malli1)
residuaalit1 <- resid(Malli1)
par(mfrow = c(1, 2))
plot(penguins$flipper_length_mm, residuaalit1,
xlab = "flipper_length_mm", ylab = "Residuaalit",
main = "Residuaalit vs. flipper_length_mm")
abline(h = 0, col = "red", lty = 2)
plot(penguins$bill_length_mm, residuaalit1,
xlab = "bill_length_mm", ylab = "Residuaalit",
main = "Residuaalit vs. bill_length_mm")
abline(h = 0, col = "red", lty = 2)

# Yksittäisen selittäjän ja residuaalien välinen hajontakuvio (Malli2)
residuaalit2 <- resid(Malli2)
par(mfrow = c(1, 2))
plot(mtcars$hp, residuaalit2, xlab = "hp",
ylab = "Residuaalit", main = "Residuaalit vs. hp")
abline(h = 0, col = "red", lty = 2)
plot(mtcars$cyl, residuaalit2, xlab = "cyl",
ylab = "Residuaalit", main = "Residuaalit vs. cyl")
abline(h = 0, col = "red", lty = 2)
```



Kuva 4: Yksittäisen selittäjän ja residuaalien välinen hajontakuvio (`Malli1`)



Kuva 5: Yksittäisen selittäjän ja residuaalien välinen hajontakuvio (`Malli2`)

3.4 Jäännösvarianssin homoskedastisuusoletuksen tarkastelu

Jäännösvarianssin homoskedastisuusoletuksella (4) tarkoitetaan, että standardioletuksen mukaan kaikilla mallin jäännöstermeillä ε_i tulee olla sama varianssi. Mikäli oletus ei ole voimassa, jäännöstermit ovat heteroskedastisia ja silloin:

$$\mathrm{Var}(\varepsilon_i) = \sigma_i^2, \quad i = 1, \ldots, n$$

Jäännöstermien homoskedastisuusoletuksen (4) voimassaoloa voidaan havainnoida hajontakuviosta, johon piirretään standardoidut residuaalit (11) sovitteita vastaan. Kuvio auttaa havaitsemaan, päteekö oletus vakiovarianssista. Mikäli kuvio ei ole tasainen, regressiomallin jäännöstermi saattaa olla heteroskedastinen ja tällöin varianssi vaihtelee.

Standardoitujen residuaalien satunnaisvaihtelun suuruus tulisi olla riippumaton sovitteiden arvosta. Mikäli standardoitujen residuaalien satunnaisvaihtelu on samansuuruista kaikilla sovitteiden arvoilla, homoskedastisuus näyttää pätevän. Tällöin pisteet ovat jakautuneet satunnaisesti nollan ympärille ja malli toimii oikein. Mikäli pisteissä on havaittavissa vaihtelua sovitteiden arvojen muuttuessa, se on merkki siitä, että residuaaleilla on eri varianssi.

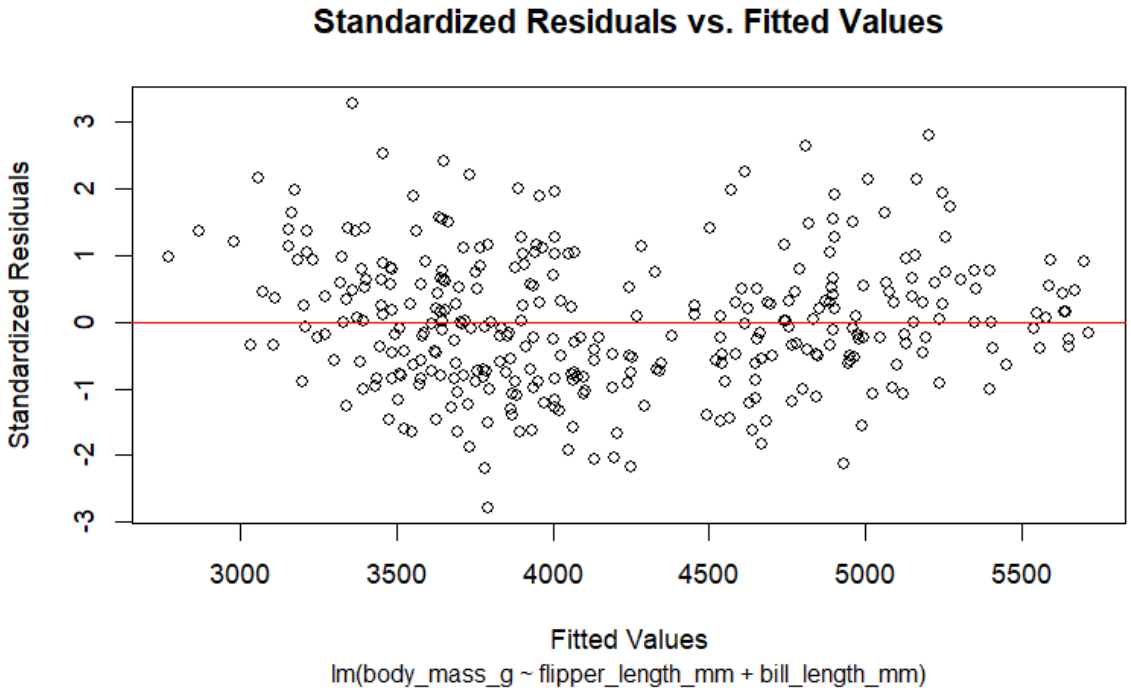
Esimerkkejä epätasaisesta kuviosta ovat muun muassa pisteparven leveneminen, suppilo ja viuhkamainen muoto, jossa pisteet lähtevät heti jyrkkään nousuun. Tämä saattaa olla merkki siitä, että varianssi suurenee ennusteiden kasvaessa. Mallin tulosten luotettavuus heikentyy eikä johtopäätöksiin voi tukeutua, mikäli mallissa on heteroskedastisuutta. Heteroskedastisessa tilanteessa regressiokertoimien estimaatit ovat harhattomia, mutta ne eivät ole tehokkaita. Toisin sanoen niissä ei ole enää mahdollisimman pientä varianssia. Tämän seurauksena kertoimet voivat vaihdella eri otoksissa. Varianssien arvioinnin vääristyessä myös p-arvot ja luottamusvälit ovat epäluotettavia.

Heteroskedastisiin tapauksiin ei ole yhtä selkeää ja varmaa toimintatapaa. Ongelmaa pystyy kuitenkin vähentämään kahdella eri keinolla. Aineistoon voidaan käyttää menetelmiä, jotka eivät ole niin ehdottomia oletusten suhteen. Tilanteessa voidaan käyttää esimerkiksi robustia regressiota standardivirheillä tai tehdä vastemuuttujalle esimerkiksi logaritmimuunnos. Robusti regressio mahdollistaa mallin sovittamisen, vaikka varianssi vaihtelee. Vastemuuttujalle tehtävä muunnos vakauttaa usein jäännösvarianssia ja tekee suhteesta lineaarisemman. Näiden menetelmien käyttöä tai teoriaa ei avata työn rajaamisen vuoksi enempää.

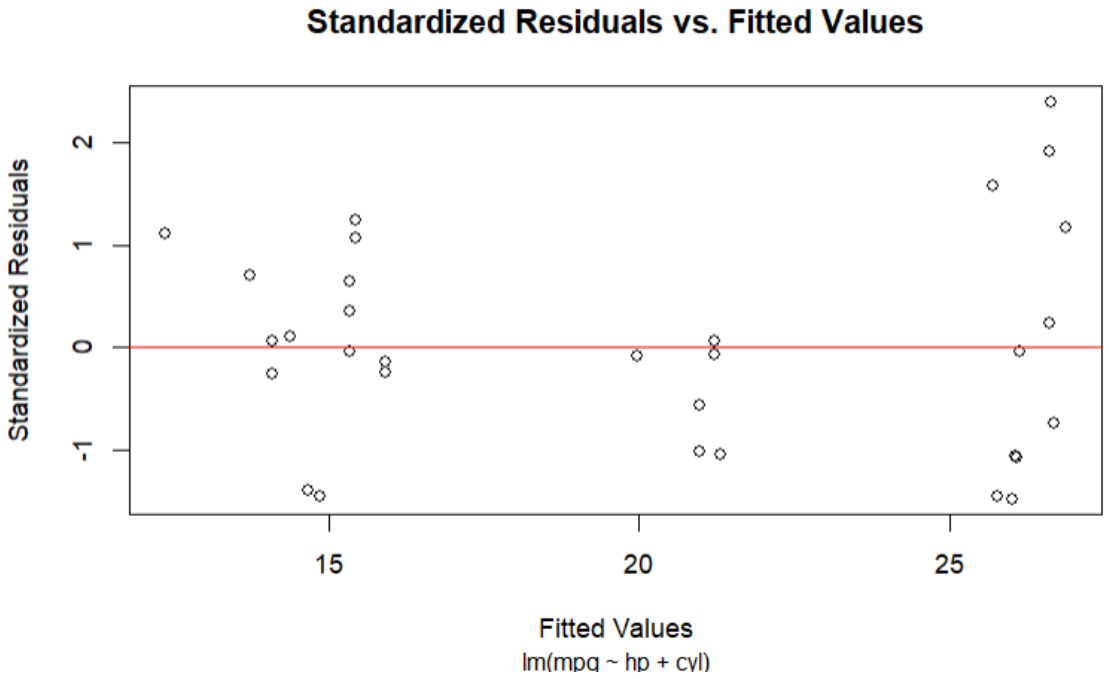
Esimerkkiaineistojen tapauksessa ensimmäisen mallin hajontakuvio on tasainen eikä siinä ole havaittavissa systemaattista kaaviota. Kuvion tasaisuus kuvastaa sitä, että residuaalien varianssi vaikuttaa olevan vakio eikä siinä ole havaittavissa systemaattista riippuvuutta mallin sovitteista, joten homoskedastisuusoletus pätee kyseisessä tilanteessa. Toisen aineiston pisteet ovat keskittyneet kolmeen kohtaan. Tapauksessa on olemassa suurempi mahdollisuus, että varianssi ei ole vakio vaan vaihtelee havaintoyksilöiden välillä. Varmuutta asiasta ei voida sanoa, sillä vaikutelma tulee muutamasta havaintopisteestä eikä otoksen koko ole kovin suuri. Kolmannen aineiston hajontakuviossa huomaa selkeän paraabelin, joten tässä voidaan olla jo lähes varmoja, ettei oletus vakiovarianssista päde. Epäselvissä tapauksissa residuaalit voidaan testata homoskedastisuuden suhteen esimerkiksi Breuschin ja Paganin testillä [15].

Loin homoskedastisuuden tarkastelussa käytetyt hajontakuviot koodilla, joka näkyy alla. Esimerkkinä on vain `Malli1`, mutta sekä `Malli2` että `Malli3` on luotu samalla koodilla.

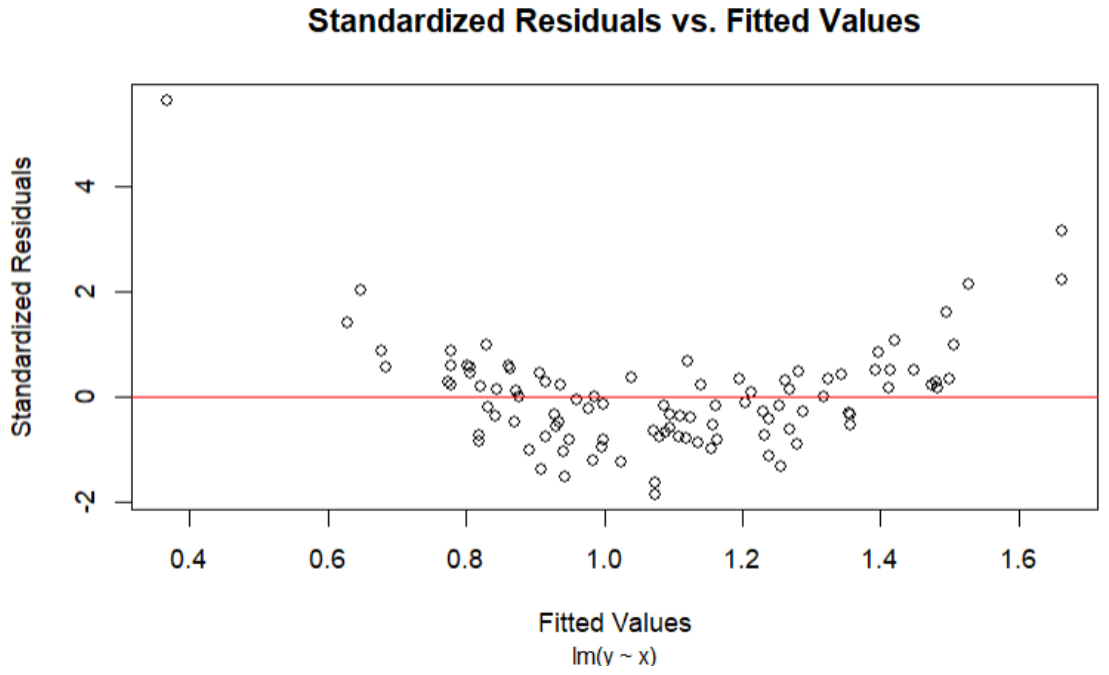
```
#Malli1
sovitteet1 <- fitted(Malli1)
standardoidut1 <- rstandard(Malli1)
plot(sovitteet1, standardoidut1,
     xlab = "Fitted Values",
     ylab = "Standardized Residuals",
     main = "Standardized Residuals vs. Fitted Values")
abline(h = 0, col = "red", lwd = 1)
mtext("lm(mpg ~ hp + cyl)", side = 1, line = 4, cex = 0.9)
```



Kuva 6: Hajontakuvio homoskedastisuusoletuksen tarkasteluun (`Malli1`).



Kuva 7: Hajontakuvio homoskedastisuusoletuksen tarkasteluun (`Malli2`).



Kuva 8: Hajontakuvio homoskedastisuusoletuksen tarkasteluun (`Malli3`).

3.5 Mallin virhetermin normaalisuus

Normaalisen lineaarisen mallin tapauksessa oletetaan y_i:n jakauma normaaliksi. Tällöin residuaalit, jotka ovat lineaarisia y:n suhteen, ovat myös normaalisti jakautuneita.

Normaalisuusoletus on tärkeä, koska t- ja F-testin p-arvot olettavat virhetermien jakaumat normaaleiksi. Mikäli näin ei ole, p-arvot ovat mahdollisesti virheellisiä ja mallin tulkinta heikentyy. Mallin regressiokertoimien luottamusvälit vaativat myös virhetermien normaalisuuden ollakseen luotettavia ja virheettömiä.

Lineaarisen mallin oletuksena mainittiin, että virhetermien ε_i tulee noudattaa normaalijakaumaa. Mallia voidaan kuitenkin käyttää, vaikka normaalisuus ei täysin toteutuisi. Kun vastemuuttujan arvoja vastaavat satunnaismuuttujat $Y_1, \ldots, Y_n$ ovat normaalisti jakautuneet, myös virhetermit noudattavat normaalijakaumaa oletuksen (1) mukaisesti. Tästä syystä vastemuuttujan normaalisuusoletus voidaan varmistaa tarkastelemalla virhetermien jakaumaa.

Normaalisuusoletuksen voimassaolo voidaan tarkistaa tilastollisilla testeillä ja graafisesti residuaalien histogrammia tai normaalijakaumakuviota (ns. QQ-plot) hyväksi käyttäen. Normaalijakaumakuvio on hyödyllinen, kun tarkastellaan residuaalien normaalisuusoletusta. Normaalijakaumakuviossa on kyse siitä, että standardoitujen residuaalien jakaumaa verrataan normaalijakaumaan piirtämällä vastakkain niiden kvantiilit. Kuviossa tehdään pisteparvi, jossa toisella akselilla on arvo $\Phi^{-1}((i-0.5)/n)(i = 1, 2, \ldots, n)$, jossa Φ^{-1} on N(0,1) -jakauman kertymäfunktion käänteisfunktio ja toisella akselilla on suuruusjärjestykseen järjestetyt standardoidut residuaalit $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$. Näin ollen, jos normaalisuusoletus pitää paikkansa, pisteet ovat suunnilleen origon kautta kulkevalla suoralla, jonka kulmakerroin on $45\,^\circ$.

Normaalisuusoletuksen voimassaoloa voidaan tarkastella myös piirtämällä standardoiduista residuaaleista histogrammi ja tarkastelemalla, muistuttaako se kellokäyrää, jolloin voidaan ajatella, että residuaalit noudattavat normaalijakaumaa. Mikäli näin on, myös virhetermit noudattavat normaalijakaumaa ja näin ollen myös satunnaismuuttuja Y noudattaa normaalijakaumaa. Tämä ei kuitenkaan ole kovin luotettava menetelmä eikä normaalisuusoletusta tule tarkastella pelkän histogrammin avulla.

Pieni poikkeama normaalisuusoletuksesta ei ole tuhoisaa. Tuloksia voidaan pitää luotettavina ja perustella asymptoottisina approksimaatioina myös silloin, kun normaalisuusoletus ei toteudu. Oleellista on, että selvästi ei-normaalisiin tilanteisiin lineaarista mallia ei pidä soveltaa suoraan. Jos vastemuuttuja saa diskreettejä arvoja, yleistetyt lineaariset mallit saattavat tulla kyseeseen.

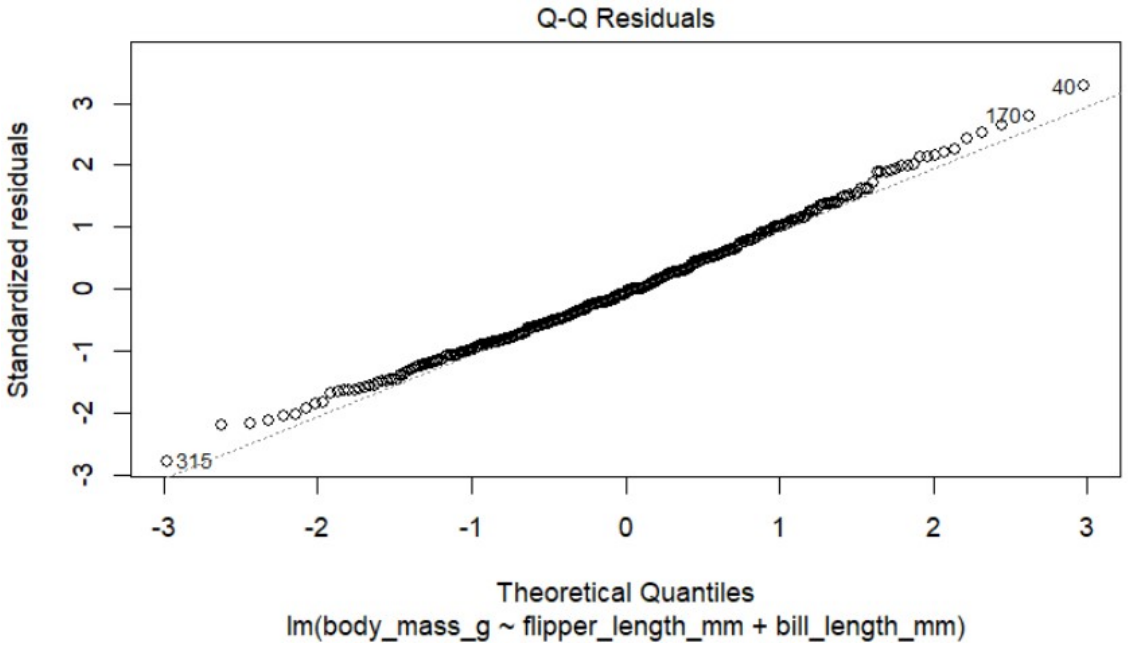
Histogrammin R-koodi:

```
# Malli1
# Tulostetaan histogram
hist(rstandard(Malli1),
        ylab = "Frekvenssi",
     xlab = "Standardoidut residuaalit",
     main = "Standardoitujen residuaalien histogrammi
     (Paino ~ Räpylän pituus + Nokan pituus)
```

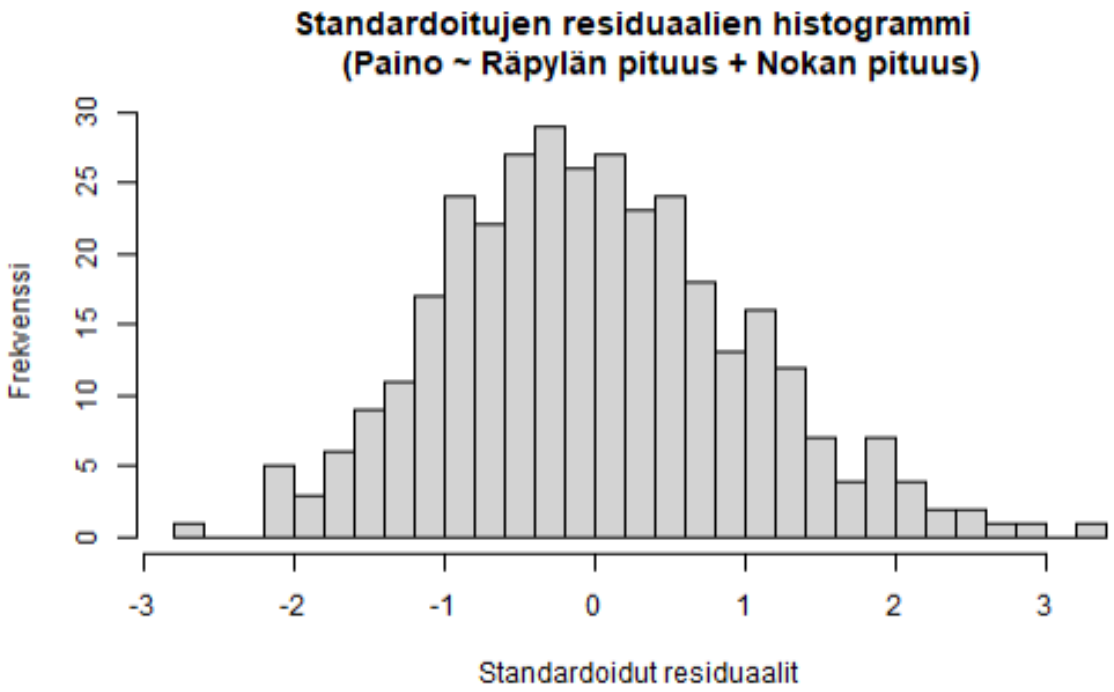
```
breaks = 30")
```

`Malli2` esimerkkiaineistosta en tehnyt histogramia, sillä havaintoja oli niin vähän, joten histogramin antama informaatio on melko olematon. Molemmat esimerkkiaineistot näyttävät noudattavan suhteellisen hyvin normaalijakaumaa. Tarkastellaan normaalisuusoletuksen voimassaoloa kuitenkin myös testeillä. Suoritetaan Shapiron ja Wilkin testi esimerkkiaineistojen standardoiduille residuaaleille. Toinen suosittu testi normaalisuusoletuksen testaukseen on Kolmogorovin ja Smirnovin testi. Ensimmäisen esimerkkiaineiston testitulokseksi saadaan arvo 0.9934 ja p-arvo 0.14, joten tämä tukee normaalisuusoletuksen voimassaoloa ja voimme ajatella, että normaalioletus on voimassa. Toisen esimerkkiaineiston tulokseksi saadaan arvo 0.9559 ja p-arvo 0.2124, joten myös tässä voidaan ajatella, että havainnot noudattavat normaalijakaumaa.

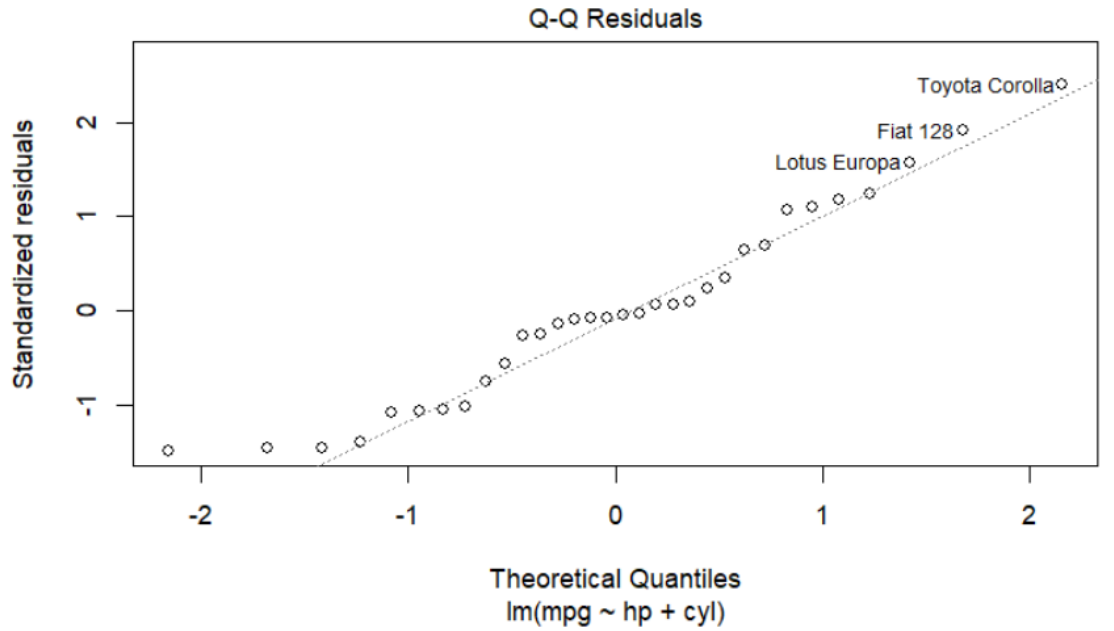
Mikäli virhetermit poikkeavat normaalijakaumasta, asiaa voidaan korjata muutamalla eri tavalla. Vastemuuttujalle voidaan tehdä esimerkiksi logaritmi- tai neliöjuurimuunnos. Toinen vaihtoehto on käyttää ei-parametrisia menetelmiä, joissa normaalisuusvaatimusta ei vaadita.



Kuva 9: Normaalisuusoletuksen tarkastelu normaalijakaumakuvion avulla (`Malli1`)



Kuva 10: Normaalisuusoletuksen tarkastelu residuaalien histogrammin avulla (`Malli1`)



Kuva 11: Normaalisuusoletuksen tarkastelu normaalijakaumakuvion avulla (`Malli2`)

3.6 Mallin virhetermin korreloimattomuus

Oletuksessa (8) on mainittu virhetermien riippumattomuus, josta seuraa niiden korreloimattomuus. Virhetermien ei tule korreloida keskenään, mikä tarkoittaa, että tällöin yhden virhetermin arvo ei vaikuta muihin virhetermeihin. Virhetermien korrelaatiosta seuraa virheellisiä p-arvoja ja luottamusvälejä.

Virhetermien korrelaatiota voidaan tarkastella testien avulla. Durbinin ja Watsonin testi kertoo, onko virhetermeissä autokorrelaatiota. Korreloimattomuushypoteesin ollessa voimassa, testisuureen odotusarvo on kaksi. Mikäli testisuureen havaittu arvo poikkeaa siitä merkittävästi, se viittaa autokorrelaation olemassaoloon. Poikkeamaa voidaan arvioida p-arvojen avulla. Testisuureen arvo voi vaihdella 0 ja 4 välillä. [5] Aikasarja-aineistojen korrelaation tarkasteluun voidaan käyttää esimerkiksi Ljungin ja Boxin testiä [3].

```
    Durbin-Watson test

data:  Malli1
DW = 2.1129, p-value = 0.8362
alternative hypothesis: true autocorrelation is greater than 0


Durbin-Watson test

data:  Malli2
DW = 1.6668, p-value = 0.1349
alternative hypothesis: true autocorrelation is greater than 0
```

Virhetermien korreloimattomuutta tarkasteltiin nyt vain Durbinin Watsonin testillä työn rajaamisen vuoksi. Ensimmäisen aineiston mallissa testi antaa tulokseksi 2.1129, joka on hieman yli kaksi. Tämä viittaa siihen, että virhetermeissä voi mahdollisesti olla negatiivista autokorrelaatiota. P-arvo on kuitenkin korkea (0.8362), joten korreloimattomuushypoteesi jää voimaan. Tämä voidaan tulkita siten, että malli ei todennäköisesti kärsi autokorrelaatiosta. Toisen aineiston tapauksessa testitulos on 1.66668 ja p-arvo 0.1349. Testin antama tulos on pienempi kuin kaksi, mikä tarkoittaa positiivista autokorrelaatiota. P-arvo on kuitenkin sen verran korkea, joten korreloimattomuushypoteesia ei hylätä. Tulkinta on, että mallissa ei ole autokorrelaatiota.

3.7 Vaikutusvaltaiset ja mahdolliset poikkeavat havainnot

Estimoidun mallin residuaalien $\hat{\varepsilon}_i$ avulla voidaan tunnistaa poikkeavia havaintoja. Jos jotkut residuaalit poikkeavat selvästi muista residuaaleista, ne saattavat viitata poikkeaviin havaintoihin. Mikäli havaittu vastemuuttujan arvo poikkeaa selvästi vastaavasta sovitteesta, mutta sillä on pieni vipuvaikutus, se ei merkittävästi vaikuta $\hat{\boldsymbol{\mu}}$ ja $\hat{\boldsymbol{\beta}}$ -arvoihin. Vipuvaikutuksella tarkoitetaan tilannetta, jossa havainnolla on poikkeava x-arvo, mutta pystyakselin arvo noudattaa annettua regressiosuoraa [13]. Jos pisteellä on suuri vipuvaikutus, se saattaa olla vaikutusvaltainen. Suuri vipuvaikutus ei kuitenkaan automaattisesti tarkoita vaikutusvaltaista havaintoa. Pistees-

tä tulee vaikutusvaltainen, kun havainto poikkeaa kauas pienimmän neliösumman suorasta. Tällöin havainto on regressiopoikkeama tai ”regressioerhe”. Jos suuren vipuvaikutuksen omaava havainto on sopusoinnussa muun aineiston kanssa, se ei ole vaikutusvaltainen. Jotta piste on vaikutusvaltainen, sillä on oltava suuri vipuvoima ja suuri standardoitu residuaali.

Mittarit, jotka kuvaavat havainnon vaikutusvaltaisuutta, yhdistävät tietoa vipuvaikutuksesta ja residuaaleista. Mitä suurempi havainnon vaikutus on, sen suurempi on mittarin antama arvo. Cookin etäisyys perustuu $\hat{\boldsymbol{\beta}}$:n muutokseen, kun havainto poistetaan aineistosta. Olkoon $\hat{\boldsymbol{\beta}}_{\boldsymbol{(i)}}$ PNS-estimaatti $\boldsymbol{\beta}$:lle, kun havainto i on poistettu. Sijoittamalla $\mathrm{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$:n estimaatti $\hat{\mathrm{Cov}}(\hat{\beta}) = s^2(X'X)^{-1}$ Cookin etäisyyden määritelmään saadaan Cookin etäisyydeksi havainnolle i:

$$\mathrm{D}_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'[\hat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p} = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'(\boldsymbol{X'X})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{ps^2}$$

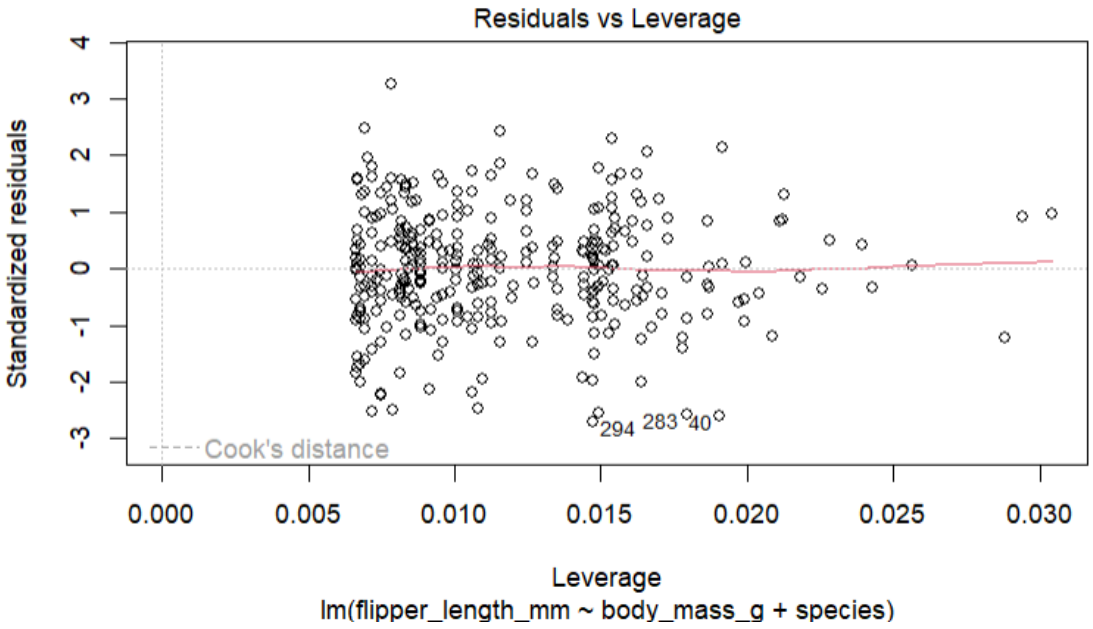
joka voidaan ilmaista standardoitujen residuaalien (11) avulla seuraavasti:

$$\mathrm{D}_i = r_i^2 \left[\frac{p_{ii}}{p(1-p_{ii})}\right] = \frac{(y_i - \hat{\mu}_i)^2 p_{ii}}{ps^2(1-p_{ii})^2}$$

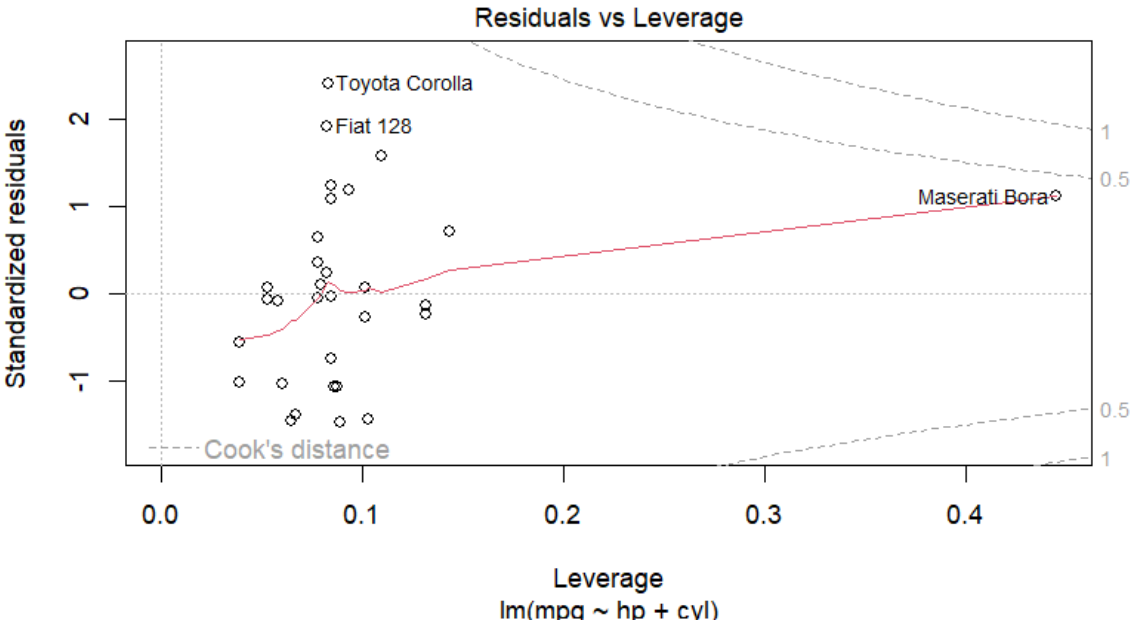
Kun D_i on suhteellisen suuri (yleensä suuruusluokkaa 1), sekä standardoidut residuaalit että vipuvaikutus ovat suhteellisen suuria.

Poikkeavia havaintoja pyritään havaitsemaan hajontakuviolla, jossa vaaka-akselilla on Cookin etäisyydet ja pystyakselilla standardoidut residuaalit. Kuten aiemmin mainittiin, kaikki poikkeavat havainnot eivät välttämättä ole vaikutusvaltaisia. Tällä tarkoitetaan sitä, että vaikka aineisto sisältäisi poikkeavia havaintoja, tulokset eivät muutu riippuen siitä, ovatko poikkeavat havainnot mukana vai eivät. Jotkut tapaukset voivat olla hyvinkin vaikuttavia, vaikka ne olisivat tiettyjen arvojen sisällä. Kun havainnot poistetaan ja tulokset muuttuvat, kyseessä on poikkeava havainto, joka on vaikutusvaltainen.

Esimerkkiaineistoissa ensimmäinen kaavio osoittaa, ettei aineisto sisällä poikkeavia havaintoja, sillä kaikkien havaintojen Cookin etäisyys D_i on pienempi kuin tietty kynnysarvo, minkä seurauksena kaikki pisteet ovat Cookin etäisyyden rajojen sisäpuolella. Toisen aineiston arvot ovat myös viitteiden sisäpuolella, mutta ne voivat mahdollisesti olla vaikutusvaltaisia. Aineisto tulee analysoida ja tuloksia tulee verrata siten, että arvot ovat mukana sekä niin, että arvot on poistettu. Mikäli tulokset eroavat, niin havainnot ovat vaikutusvaltaisia. Tämä analyysi jätetään työn rajauksen vuoksi pois.



Kuva 12: Poikkeavat havainnot (`Malli1`)



Kuva 13: Poikkeavat havainnot (`Malli2`)

4 Yhteenveto

Työssä käytiin läpi lineaarinen malli ja siihen liittyvät oletukset. Oleellisimmat diagnostiset menetelmät esiteltiin esimerkkiaineistojen avulla.

Työn rajauksen vuoksi osa tärkeistä menetelmistä jäi käymättä läpi. Yksi mielenkiintoinen diagnosointiin liittyvä menetelmä on esimerkiksi osittainen regressiokuvio. Sen avulla voi arvioida yksittäisen selittäjän vaikutusta vastemuuttujaan, kun muiden muuttujien vaikutus on poistettu. Tämä menetelmä on hyödyllinen erityisesti monimuuttujaregressiossa. Regressioanalyysin diagnostisten menetelmien aihetta voisi jatkaa perehtymällä esimerkiksi tähän kuvioon.

Hyvä jatke työlle olisi myös selostaminen, miten tulee toimia, kun oletukset eivät täyty. Työssä sivuttiin heteroskedastisuutta ja mitä menetelmiä niissä tilanteissa voisi käyttää. Aiemmin mainittiin robusti regressio standardivirheillä tai vastemuuttujalle tehtävä muutos, esimerkiksi logaritmimuunnos. Muunnoksia on kuitenkin monia muitakin, ja näitä menetelmiä voisi käsitellä laajemmin seuraavassa työssä.

Työtä voisi jatkaa myös muilla mielenkiintoisilla aiheilla, kuten esimerkiksi vaikutusvaltaiset arvot ja vipuvaikutus. Työssä käsittelin niitä pintapuoleisesti, joten näiden laajempi käsittely olisi hyödyllistä. Työn toinen esimerkkiaineisto saattaa sisältää vaikutusvaltaisia arvoja. Aineisto tulisi analysoida siten, että arvot ovat mukana sekä niin, että arvot on poistettu. Tämä analyysi jätettiin työstä pois, mutta jatkoa voisi tehdä siten, että aineistolle suoritettaisiin kyseinen analyysi.

Eräs mahdollinen aihe on multikollineaarisuus ja vif-arvo. Työstä rajattiin aihe pois, sillä se soveltuu paremmin niihin tilanteisiin, joissa on enemmän selittäjiä. Työssä ei puhuttu tarkemmin regressiomallilla ennustamisesta ja selitysasteesta, joten näiden tarkempi avaaminen voisi myös olla paikallaan.

Lähteitä, joissa on käsitelty yllä olevia aiheita, olisi muun muassa [9] ja [14].

Viitteet

[1] Agresti, A.: *Foundations of Linear and Generalized Linear models*, Wiley, 2015

[2] Bobbit, Z.: How to create a Residual Plot in R.
https://www.statology.org/residual-plot-r/
Viitattu: 23.2.2024

[3] Bobbit, Z.: Ljung-Box test: Definition + Example
https://www.statology.org/ljung-box-test/
Viitattu: 3.2.2025

[4] Bommae, K. 2015. "Understanding Diagnostic Plots for Linear Regression Analysis."UVA Library StatLab.
https://library.virginia.edu/data/articles/diagnostic-plots
Viitattu: 28.2.2024

[5] CFI: Durbin Watson Statistics
https://corporatefinanceinstitute.com/resources/data-science/durbin-watson-statistic/
Viitattu: 3.2.2025

[6] Krasser, R.: Explore mtcars.
https://cran.r-project.org/web/packages/explore/vignettes/explore_mtcars.html
viitattu: 14.3.2024

[7] Mellin, I.: *Tilastolliset menetelmät: Lineaarinen regressioanalyysi* (2006). Aalto yliopisto.
https://math.aalto.fi/opetus/sovtoda/oppikirja/Regranal.pdf
Viitattu: 7.2.2024

[8] Mellin, I.: *Tilastollinen päättely: Yleinen lineaarinen malli* (2010). Aalto yliopisto.
https://math.aalto.fi/opetus/mellin/tilpaat/luennot/Viikko11.pdf
Viitattu: 23.2.2024

[9] Montgomery, D. C., Peck, E. A., & Vining, G. G.: *Introduction to Linear Regression Analysis* (2012).

[10] Nyberg, H.: *Lineaariset ja yleistetyt lineaariset malli* -luentomoniste. Turun yliopisto.
helmikuu 2021

[11] Nyberg, H.: *Matriisilaskenta tilastotieteessä* -luentomoniste. Turun yliopisto.
tammikuu 2024

[12] Twomey, M.: Exploring Palmer Penguins.
https://rpubs.com/michelle10128/923430
viitattu: 14.3.2024

[13] Shaheen, A.: Unusual Observations. Outlier, Leverage and Influential Point. The Open Educator.
https://www.theopeneducator.com/doe/Regression/outlier-leverage-influential-points viitattu: 7.2.2025

[14] Wiley ja Kutner, M. H., Nachtsheim, C. J., & Neter, J.: *Applied Linear Regression Models* (2004).

[15] Struck, J. 2024. Homoscedasticity. University of Wisconsin-Madison.
https://sscc.wisc.edu/sscc/pubs/RegDiag-R/homoscedasticity.htmlstatistical-tests-2
Viitattu: 6.2.2025