Harjoitustyö

Sekamallit, syksy 2021

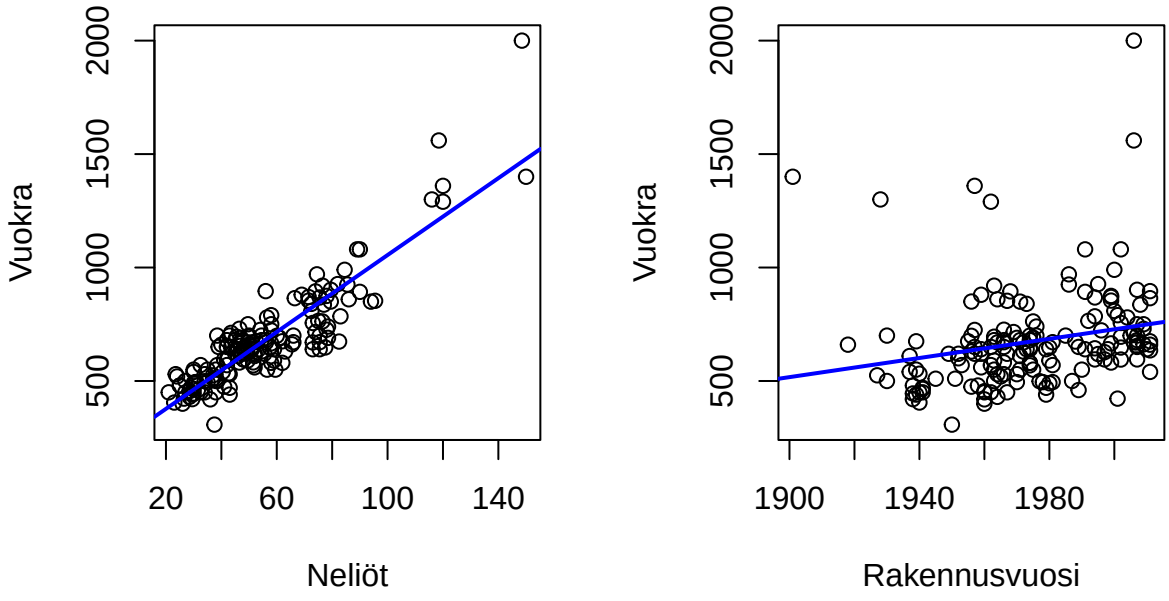
Maria Seppänen

Aineiston alkuperä

Harjoitustyössä käytetty aineisto kuvastaa Tampereella olevia vuokra-asuntoja. Aineisto löytyy Tampereen Yliopiston Moodlesivulta "Tilastomenetelmien työkalupakki". Sieltä löytyy "Tilastotieteen peruskurssien hajoitusaineistoja 2003-2018", josta lopulta työssä käytetty aineisto "Tre_vuokra-asunnot_2011". Analyysiin on käytetty aineiston kaikkia muuttujia. Muuttuja "Kaupunginosa" on kategorinen muuttuja, joka sisältää eri kaupunginosia Tampereella. Muuttuja "Huoneet" on asunnossa olevien huoneiden määrä, "Vuokra" on asunnon vuokra ja "Rakennusvuosi" on milloin asunto on rakennettu. Rakennusvuodesta tehdään myöhemmässä osassa analyysiä kolmen eri ajanjakson kategorinen muuttuja jakamalla havoinnot neljän kymmenen vuoden ajanjaksoihin analyysin helpottamiseksi. Havaintoja on yhteensä 177.

Harjoitustyön edetessä, tarkoituksena on löytää sopiva sekamalli vuokran mallintamiselle muiden aineiston muuttujien avulla. Jotta voi tehdä mitään, otetaan käyttöön kirjasto "nlme", josta löytyy erinäisiä funktioita sekamallieihin liittyen ja haetaan aineisto. Analyysien helpottamiseksi, 10 NULL-arvon sisältävää havaintoa poistettiin, mikä voi johtaa hieman harhaanjohtaviin tuloksiin. Tämän ei pitäisi kuitenkaan vaikuttaa käytettyihin menetelmiin tai johtopäätöksien tekemiseen.

Kun aineiston muuttujia tarkastellaan tarkemmin, voidaan huomata, että muuttujien Neliöt ja Vuokra, sekä Rakennusvuosi ja Vuokra välillä on havaittavissa selvää lineaarista regressiota.



Kiinteiden ja satunnaisten vaikutusten malli.

Kiinteiden vaikutusten malli

Tarkastellaan huoneiden määrän vaikutusta vuokran suuruuteen. Luodaan aineistosta hyvin yksinkertainen kiinteiden vaikutusten malli

$$y_{ij} = \mu_i + \epsilon_{ij},$$

jossa ϵ_{ij} on satunnaisvirheet. Oletetaan, että satunnaisvirheet ovat riippumattomia ja normaalisti jakautuneita parametrein $N(0, \sigma^2)$. Testataan mallilla hypoteesia

$$H_0 : \mu_1 = \mu_2 = ... = \mu_5,$$
$$H_1 : \mu_1 \neq \mu_2 \neq ... \neq \mu_5.$$

```
data.kiintv <- aov(Vuokra ~ factor(Huoneet), data = data)
summary(data.kiintv)
```

```
##                   Df  Sum Sq Mean Sq F value Pr(>F)
## factor(Huoneet)    4 5372646 1343162   91.42 <2e-16 ***
## Residuals        162 2380171   14692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(data.kiintv)
```

```
## Analysis of Variance Table
##
## Response: Vuokra
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## factor(Huoneet)    4 5372646 1343162  91.419 < 2.2e-16 ***
## Residuals        162 2380171   14692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nähdään, että p-arvo on hyvin pieni, jolloin nollahypoteesi hylätään. Huoneiden määrällä on siis vaikutusta vuokran suuruuteen.

Satunnaisvaikutusten malli

Otetaan seuraavaksi huomioon mahdolliset satunnaisvaikutukset. Oletetaan, että satunnaisvirheet ϵ_{ij} ovat riippumattomia ja normaalisti jakautuneita parametrein $N(0, \sigma^2)$. Tehdään aineistolle seuraavanlainen sekamalli

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

jossa α_i on satunnaisvaikutukset ja ϵ_{ij} on satunnaisvirheet.

```
data.satv <- lme(Vuokra ~ 1, random=~1|factor(Huoneet), data=data)
```

Tarkastellaan vielä mallia, jossa kiinteidiin vaikutuksiin on lisätty Neliöt ja pidetään satunnaisvaikutukset samana. Verrataan tätä mallia edelliseen malliin anova-funktiolla.

```
data.satv1 <- lme(Vuokra ~ Neliöt, random=~1|factor(Huoneet), data=data)

# Muutetaan menetelmäksi maximum likelihood, jotta vertailu onnistuu.
data.satv <- update(data.satv, method="ML")
data.satv1 <- update(data.satv1, method="ML")
anova(data.satv,data.satv1)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## data.satv      1  3 2107.234 2116.588 -1050.617
## data.satv1     2  4 2009.699 2022.171 -1000.850 1 vs 2 99.53513  <.0001
```

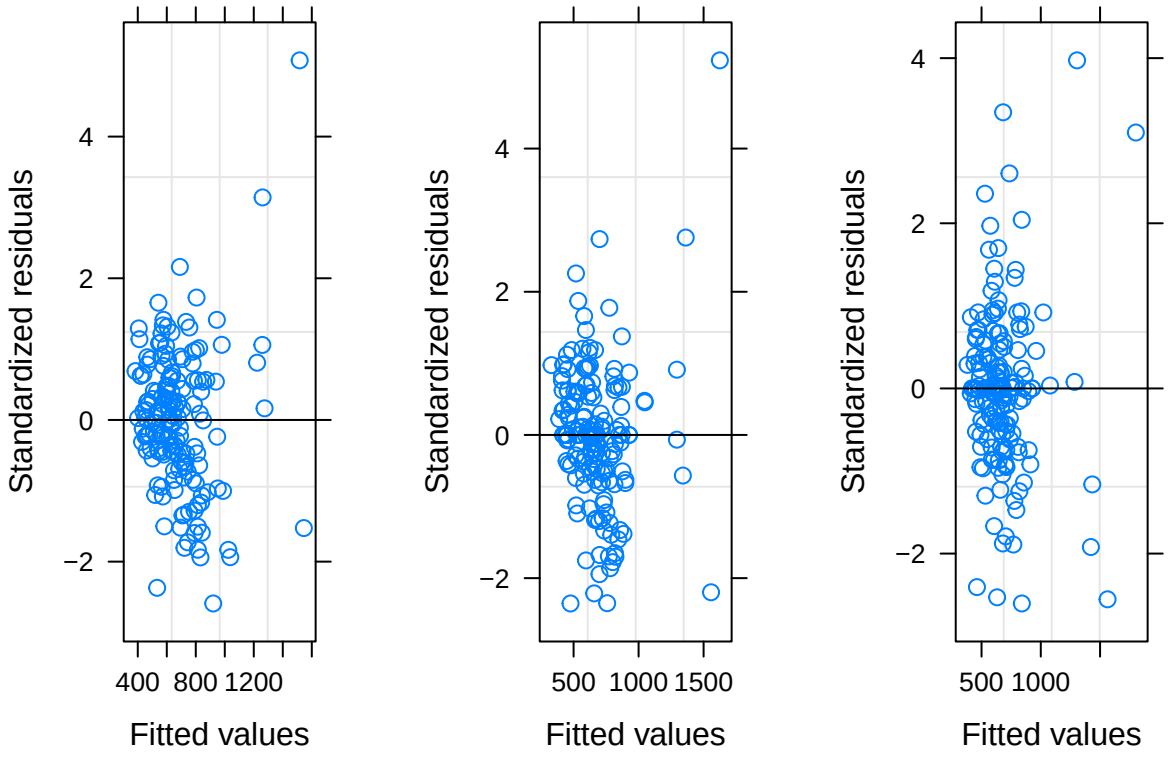
Testaamisesta huomataan, että uusi malli on parempi kuin edellinen, yksinkertaisempi, malli. Tämä näkyy sekä informaatioarvoista, jotka ovat data.satv1-mallilla pienemmät sekä todella pienestä p-arvosta. Yritetään löytää vielä sopivampi malli lisäämällä kiinteään osaan Kaupunginosa ja satunnaisvaikutuksiin Neliöt.

```
data.satv2 <- lme(Vuokra ~ Neliöt+factor(Kaupunginosa), random=~1|factor(Huoneet),
                  data=data, method="ML")
data.satv3 <- lme(Vuokra ~ Neliöt+factor(Kaupunginosa), random=~1+Neliöt|factor(Huoneet),
                  data=data, method="ML")
anova(data.satv1, data.satv2, data.satv3)
```

```
##            Model df      AIC      BIC     logLik   Test   L.Ratio p-value
## data.satv1     1  4 2009.699 2022.171 -1000.8496
## data.satv2     2 47 1994.607 2141.152  -950.3033 1 vs 2 101.09257  <.0001
## data.satv3     3 49 1976.553 2129.334  -939.2763 2 vs 3  22.05392  <.0001
```

Testauksen perusteella uusin malli, jossa kiinteissä vaikutuksissa Neliöt+Kaupunginosa ja satunnaisvaikutuksissa Neliöt on molempia edellisiä malleja parempi sekä informaatiokriteereiltään, että hypoteesitestauksessa. Tarkastellaan mallien residuaalikaavioita.

```
p1 <- plot(data.satv1)
p2 <- plot(data.satv2)
p3 <- plot(data.satv3)
grid.arrange(p1, p2, p3, ncol = 3)
```



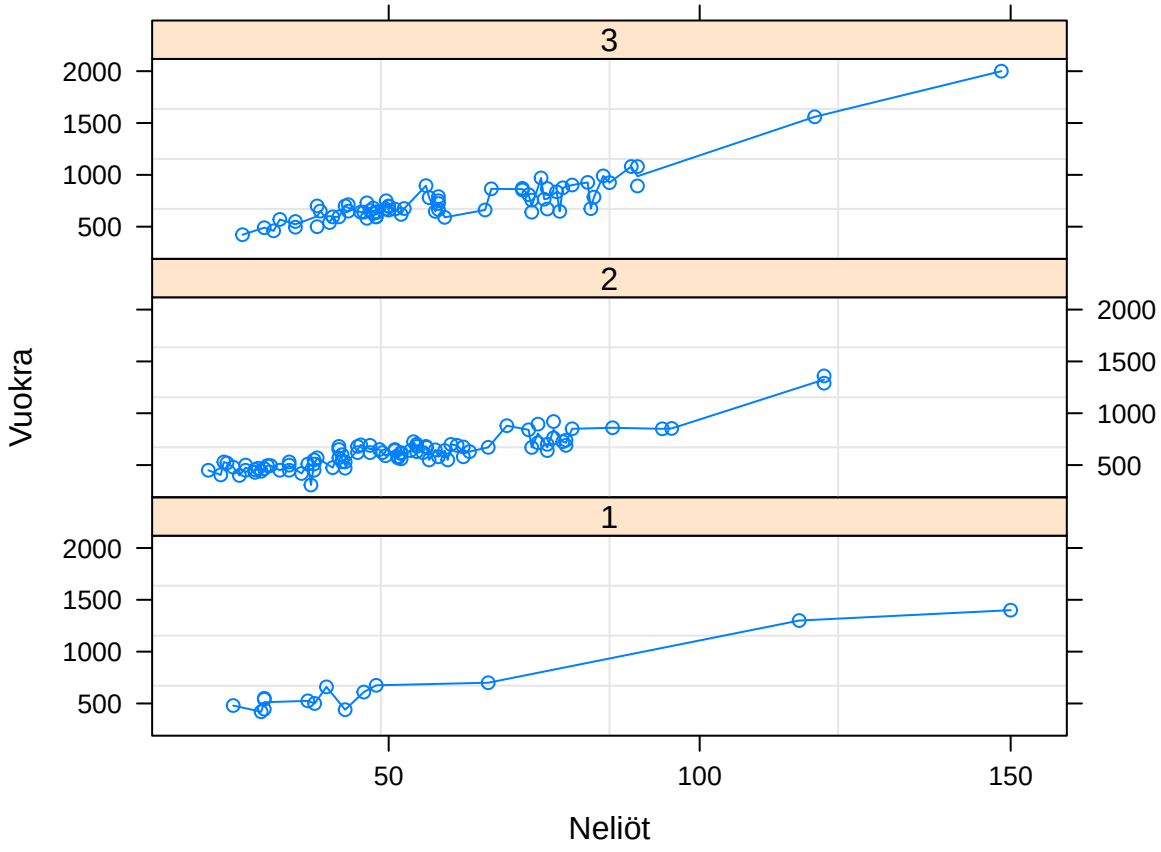
Silmämääräisesti eroa on vaikea huomata, erityisesti ison vuokran havainnot ovat hyvin kaukana toivotusta.

Kolmannen mallin residuaalikuvio on kuitenkin jo huomattavasti “tasaisempi” kuin ensimmäisen mallin residuaalikaavio. Kuitenkin, mallien suhteellisen heikko soopivuus kuvastuu myös residuaalikaavioista.

Korrelaatio ja varianssi

Tarkastellaan seuraavaksi tarkemmin muuttujaa Vuokra, kun Rakennusvuosi on otettu huomioon. Muutetaan Rakennusvuosi kategoriseksi muuttujaksi niin, että vuodet 1901-1940 koodataan numerolla 1, vuodet 1941-1980 numerolla 2 ja vuodet 1981-2020 numerolla 3. Tarkastellaan myös erilaisia varianssi- ja kovarianssirakenteita. Varianssin ja kovarianssin huomioiminen mallissa voi parantaa sen sopivuutta aineistoon. Piirretään ensin Rakennusvuoden kolme kategoriaa ja luodaan kaksi mallia, joissa molemmissa Neliöt on kiinteässä osassa. Ensimmäisessä Rakennusvuosien kategoriat ovat satunnaisessa osassa, toisessa Neliöt Rakennusvuosi-kategorioittain.

```
newdata <- data

newdata$Rakennusvuosi[Rakennusvuosi < 1940] <- 1
newdata$Rakennusvuosi[Rakennusvuosi < 1980 & Rakennusvuosi >= 1940] <- 2
newdata$Rakennusvuosi[Rakennusvuosi < 2020 & Rakennusvuosi >= 1980] <- 3

data1 <- groupedData(Vuokra~Neliöt|factor(Rakennusvuosi), data=newdata)

plot(data1)
```



```
data1.satv <- lme(Vuokra~Neliöt, data=newdata, random=~1|factor(Rakennusvuosi))
data1.satv1 <- lme(data1) # satunnaisvaikutukset muotoa ~Neliöt|factor(Rakennusvuosi)

anova(data1.satv, data1.satv1)
```

```
##             Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## data1.satv      1  4 1999.658 2012.082 -995.8289
## data1.satv1     2  6 1995.022 2013.658 -991.5111 1 vs 2 8.635691  0.0133
```

Mallien vertaaminen keskenään paljastaa, että toinen malli on hieman parempi. AIC-arvo on himpun verran matalampi ja p-arvo on hyvin pieni. Jatketaan siis mallilla data.satv1. Tarkastellaan tämän mallin kiinteitä vaikutuksia, myös anova-funktiolla.

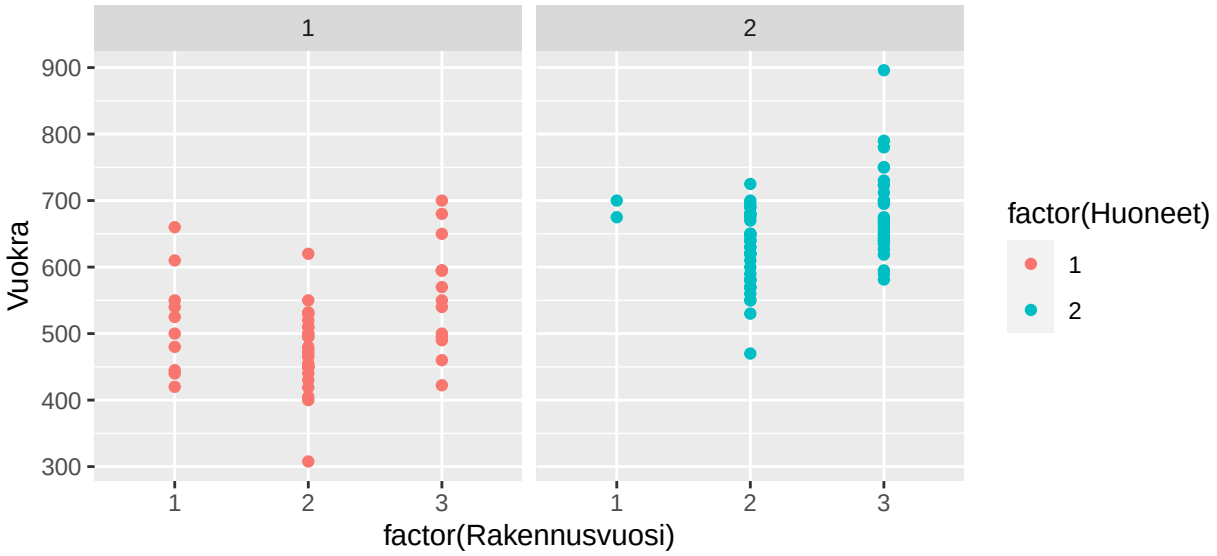
```
anova(data1.satv1)
```

```
##             numDF denDF   F-value p-value
## (Intercept)     1   163 2487.4580  <.0001
## Neliöt          1   163  132.5693  <.0001
```

Kiinteät arvot ovat merkitseviä, sillä p-arvot ovat huomattavasti pienempiä kuin 0,05.

Otetaan seuraavaksi tarkasteluun pelkästään eri rakennusvuosina valmistuneiden yksiöiden ja kaksioiden vuokra, ja katsotaan, onko niillä merkittävää eroa. Luodaan malli, jossa tarkastellaan yksiöitä ja kaksioita, jotka on rakennettu eri rakennusvuosina ilman satunnaisosaa. Tehdään toinen malli, jossa on satunnaisvaikutuksena eri huoneiden määrät. Verrataaan malleja keskenään.

```
data.huoneet <- subset(newdata, Huoneet == "1" | Huoneet == "2",
                       select=c(Vuokra, Neliöt, Rakennusvuosi, Huoneet))
data.h <- groupedData(Vuokra~Huoneet|factor(Rakennusvuosi), data=data.huoneet)
ggplot(data.h, aes(x = factor(Rakennusvuosi), y = Vuokra, colour = factor(Huoneet)))+
geom_point() +
facet_wrap( ~ factor(Huoneet))
```



Silmämääräisesti, voidaan sanoa, että yksiöiden ja kaksioiden hinnat ovat eri korkeuksilla näinä kolmena ajankohtana. Yksiöistä ja kaksioista voidaan muodostaa kaksi eri regressiosuoraa. Luodaan edellä mainitut mallit.

```
malli <- gls(Vuokra~factor(Huoneet):factor(Rakennusvuosi)-1, data=data.h, method="ML")
```

```
malli2 <- lme(Vuokra~factor(Huoneet):factor(Rakennusvuosi)-1,
              data=data.h, random=~1|factor(Huoneet), method="ML")
anova(malli, malli2)
```

```
##        Model df      AIC      BIC    logLik   Test      L.Ratio p-value
## malli      1  7 1348.122 1367.634 -667.0608
## malli2     2  8 1350.122 1372.422 -667.0608 1 vs 2 1.568933e-07  0.9997
```

Anova-tarkastelun perusteella yksinkertaisempi malli on parempi.

Tarkastellaan vielä, ovatko yksiöiden ja kaksioiden vuokrat toisistaan erillisiä luottamusvälien avulla.

```
intervals(malli)
```

```
## Approximate 95% confidence intervals
##
##  Coefficients:
##                                                lower     est.    upper
## factor(Huoneet)1:factor(Rakennusvuosi)1 476.6394 517.0000 557.3606
## factor(Huoneet)2:factor(Rakennusvuosi)1 597.2510 687.5000 777.7490
## factor(Huoneet)1:factor(Rakennusvuosi)2 450.4407 474.5607 498.6808
## factor(Huoneet)2:factor(Rakennusvuosi)2 603.8481 624.5526 645.2572
## factor(Huoneet)1:factor(Rakennusvuosi)3 526.0255 560.1364 594.2473
## factor(Huoneet)2:factor(Rakennusvuosi)3 657.6610 681.7811 705.9011
## attr(,"label")
## [1] "Coefficients:"
##
##  Residual standard error:
##    lower     est.    upper
## 55.75752 62.79664 71.88597
```

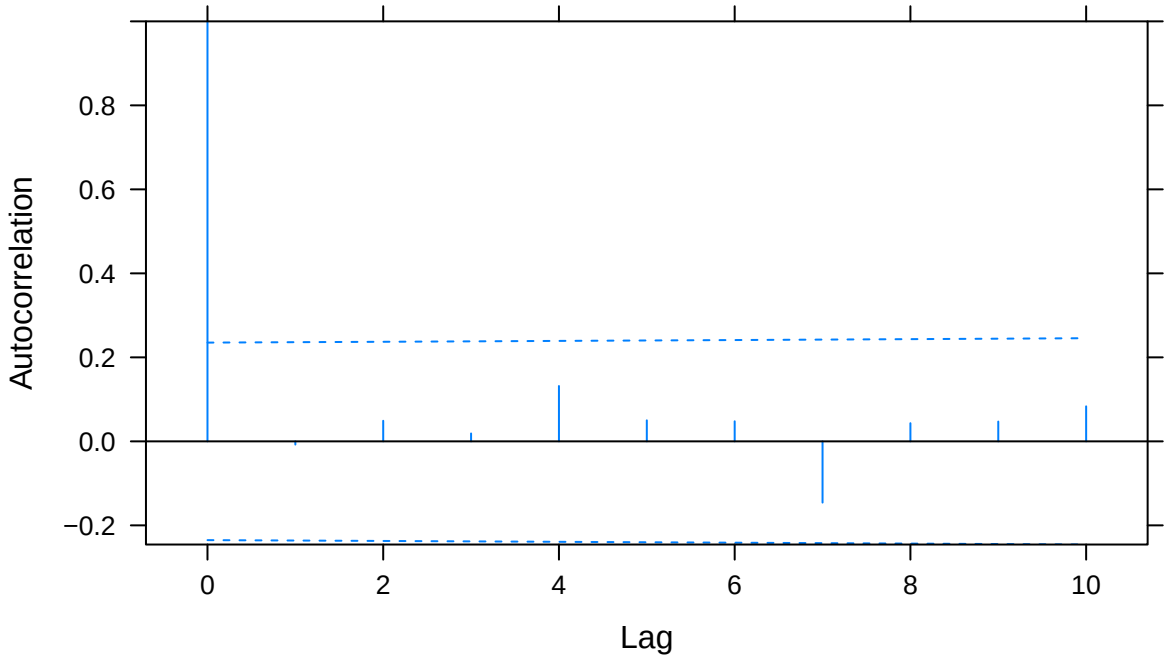
Luottamusvälit eivät mene päällekkäin missään tapauksessa. Niiden on siis oltava toisistaan erillisiä.

Etsitään satunnaisvirheille parhaiten sopiva kovarianssirakenne.

```
malli2 <- update(malli, correlation=corCompSymm())
malli3 <- update(malli, correlation=corAR1())
malli5 <- update(malli, correlation=corLin())
malli6 <- update(malli, correlation=corCompSymm())

anova(malli,malli2,malli3,malli5,malli6)
```

```
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## malli      1  7 1348.122 1367.634 -667.0608
## malli2     2  8 1319.661 1341.960 -651.8303 1 vs 2 30.46103  <.0001
## malli3     3  8 1350.114 1372.414 -667.0568
## malli5     4  8 1350.122 1372.422 -667.0608
## malli6     5  8 1319.661 1341.960 -651.8303
```

Testauksen perusteella, alkuperäinen malli on edelleen paras. Tämä näkyy myös, kun tarkastellaan empiiristä autokorrelaatiota. Ainoastaan ensimmäinen arvo kohoaa 0.01-merkitsevyystason yli.



Tarkastellaan myös erilasia varianssirakenteita.

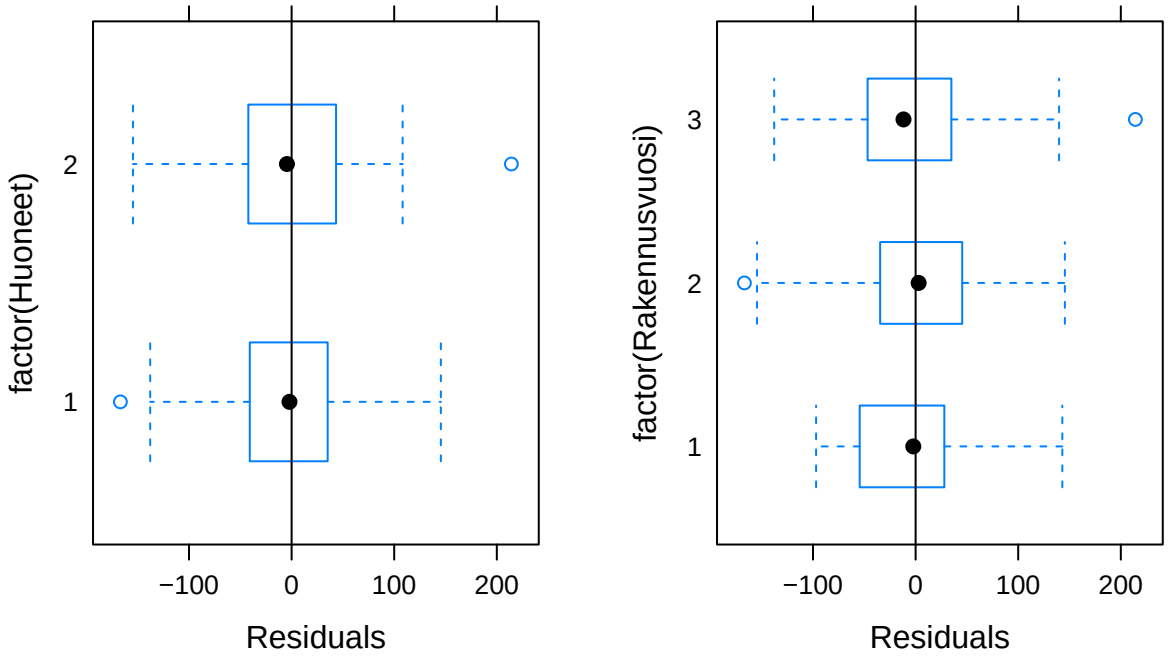
```
malli12 <- update(malli, weights = varPower())
malli13 <- update(malli, weights = varIdent(form= ~1|factor(Huoneet)))
malli14 <- update(malli, weights = varConstPower())

anova(malli12, malli13, malli14) # malli13 on paras.
```

```
##         Model df      AIC      BIC    logLik   Test   L.Ratio p-value
## malli12     1  8 1350.121 1372.421 -667.0607
## malli13     2  8 1349.341 1371.641 -666.6708
## malli14     3  9 1352.121 1377.209 -667.0607 2 vs 3 0.7798846  0.3772
```

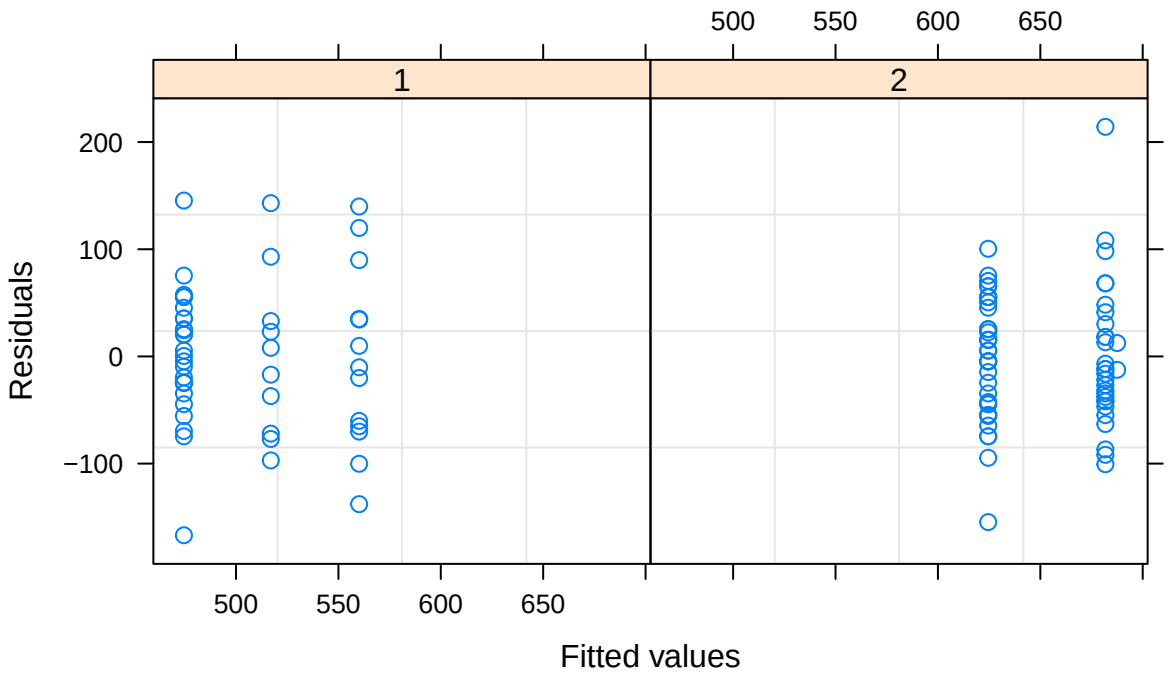
Nyt huomataan, että malli, jossa on varIdent()-varianssirakenne on paras tapauksessamme. VarIdent()-rakenteessa havainnoilla voi olla eri variannssit.

Tarkastellaan mallin sopivuutta aineistoon

Tarkastellaan graafisesti saamamme mallin sopivuutta aineistoon eri tavoin. Ensiksi residuaalikaavio sekä Huoneet- että kategorisen Rakennusvuosi -muuttujien osalta. Huoneet on hyvin tasaisen näköinen, vaikkakin kaksioit on vasemmalle kallistunut. Rakennusvuodet ovat vielä tasaisempia, paria hyvin suurta ja hyvin pientä havaintoa lukuunottamatta.

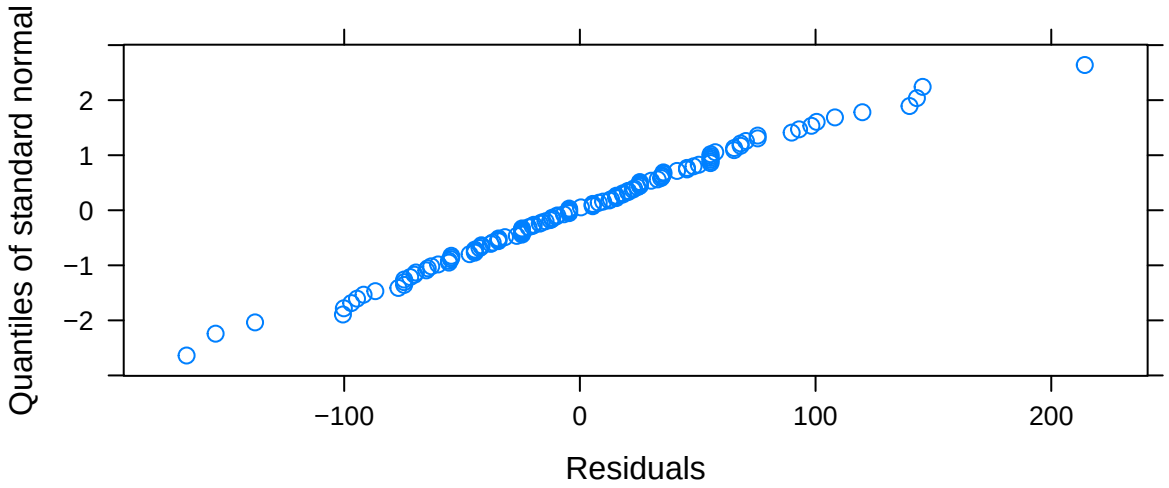


Residuaalit ovat kutakuinkin tasaisesti jakautuneita sekä yksiöiden, että kaksioiden osalta - kuitenkin hyvin eri arvoissa. Kaksioiden residuaalit ovat hieman lähempänä toisiaan, kun taas yksiöiden residuaalit ovat laajemmalle levittäytyneitä.

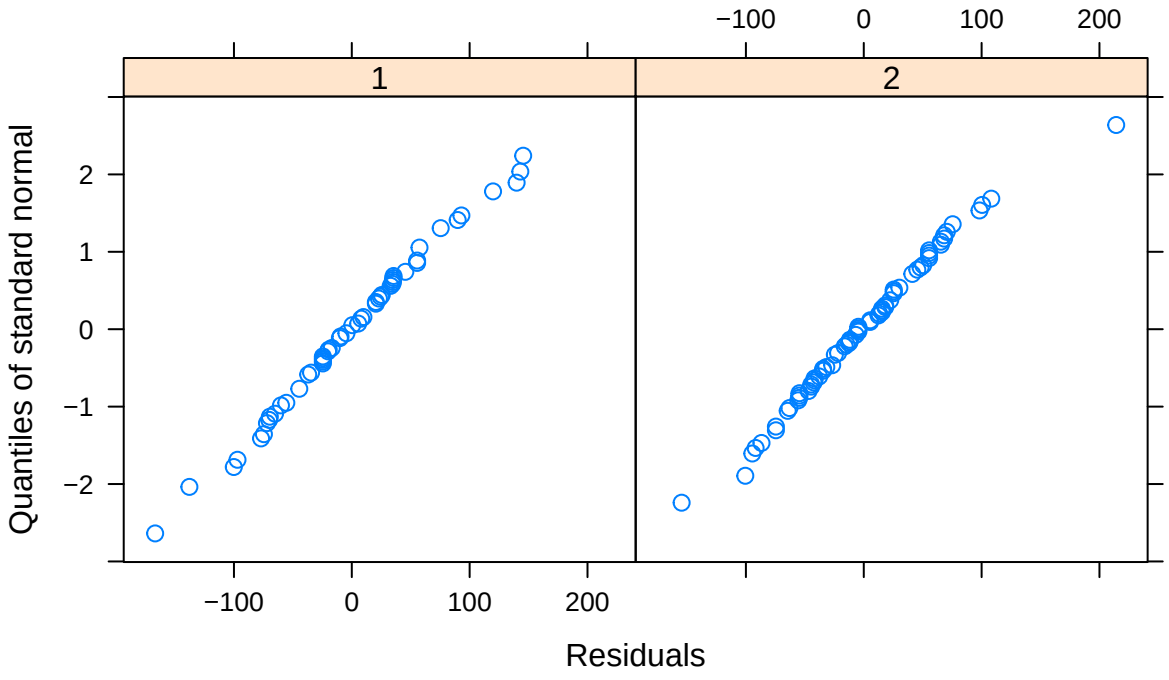


Jäännösten normaalisuus täyttyy todella hyvin, sillä pisteet ovat hyvin lähellä suoraa. Ainoastaan pienimmissä

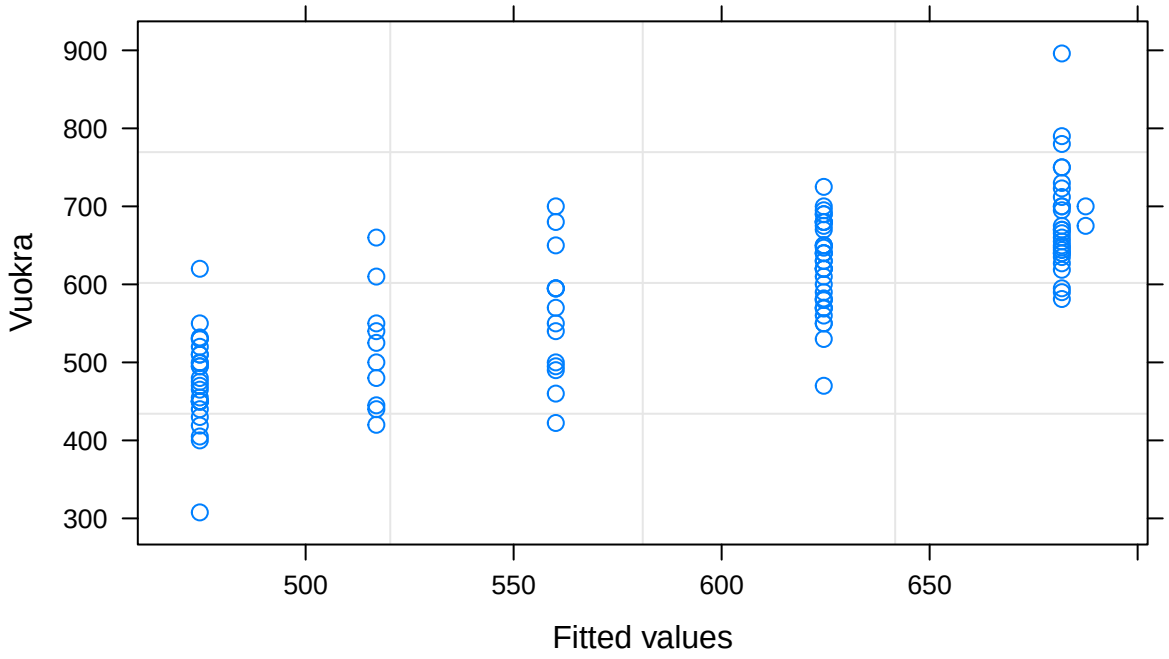
ja suurimmissa arvoissa on näkyvissä eroavaisuutta.



Yksiöittäin ja kaksioittan huomataan suurinpiirtein samanlaista kuin edellisessäkin kaaviossa - keskellä olevat arvot ovat hyvin lähellä suoraa, mutta pienimmissä ja suurimmissa arvoissa on nähtävissä eroavaisuutta suorasta.



Kun katsotaan havaittuja ja sovitettua arvoja, huomataan, että ne ovat aika kaukana toisistaan. Malli ei siis ole vielä kovin hyvä käytetylle aineistolle.



Tarkastelujen perusteella, valittu malli ei ole kovin hyvin sopiva aineistoon. Vaikka muuttujien ja mallin jäännösten normaalisuus täyttyy suhteellisen hyvin, ei malli vielä kuvaa aineiston muutoksia hyvin, eli parannettavan varaa olisi. Vaihtelua aiheuttaa jokin, mitä ei ole osattu ottaa tämän tutkimuksen parhaassa mallissa huomioon. Mallia saattaisi voida parantaa olemassaolevilla muuttujilla, mutta on myös mahdollista, että vaihtelun aiheuttaa jokin taustamuuttuja, jota ei aineistosta edes löydy.