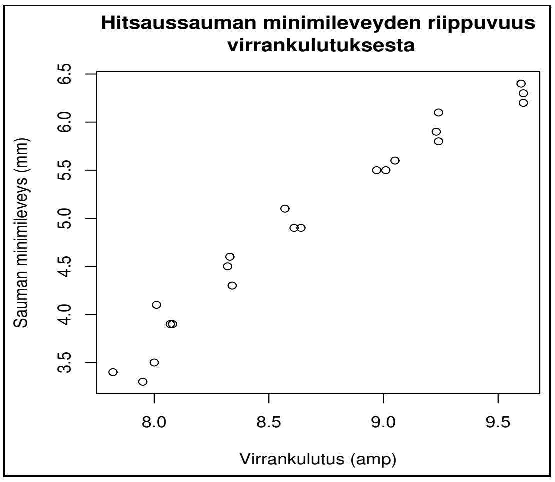
Hitsaussauman minimileveyden riippuvuus virrankulutuksesta

Tilastollinen mallintaminen
Work Sheet 1
Tampereen yliopisto
27.9.2000

Jarkko Isotalo, Aki Koskinen ja Jari Niemelä

1 Lineaarinen malli

Tarkastellaan siis *Weld.txt* tiedoston mukaista aineistoa, jossa hitsaussauman minimileveyttä selitetään hitsauksessa kulutetun virran määrällä. Kuvassa 1 onkin kuvattu pisteparvena hitsaussauman minimileveyden y_i riippuvuus hitsaukseen kulutetun virran määrästä x_i.



Kuva 1: Sauman leveyden riippuvuus kulutetun virran määrästä

Kuvan 1 perusteella voidaan ajatella, että hitsaussauman minimileveyden y_i ja virran kulutuksen x_i välillä vallitsee lineaarinen riippuvuus. Sovitetaankin siten aineistoon ensiksi seuraava malli

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, \ldots, n, \tag{1}$$

missä virhetermit ε_i ovat toisistaan riippumattomia ja noudattavat normaalijakaumaa $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ kaikille $i = 1, 2, \ldots, n$.

Kaavan (1) mukaisen mallin tulokset saadaan laskettua seuraavien käskyjen avulla:

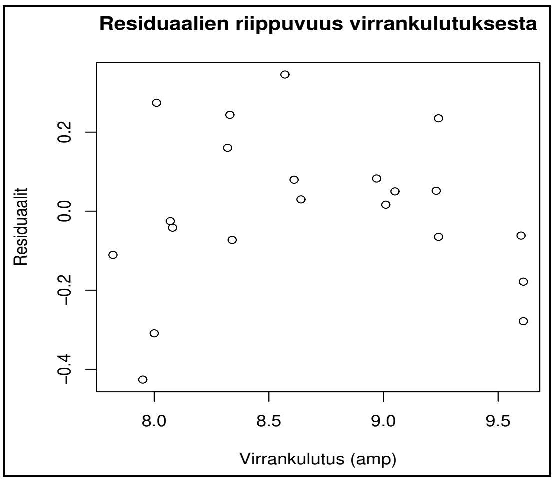
```
teeny <- lm(y~x)
summary(teeny),
```

jolloin saadaan taulukon 1 mukaiset tulokset.

```
-----------------------------------------------
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.45427    0.65526  -14.43 1.09e-11
x            1.65793    0.07531   22.01 5.55e-15
---
Residual standard error: 0.2012 on 19 df
Multiple R-Squared: 0.9623,
Adjusted R-squared: 0.9603
F-statistic: 484.6 on 1 and 19 df,
p-value: 5.551e-015
-----------------------------------------------
```

Taulukko 1: Lineaarisen mallin tuloksia

Taulukosta 1 nähdään, että kaavan (1) mukaisen mallin selitysaste R^2 on suuri, joten lineaarinen yhden selittäjän malli näyttäisi selitysasteen puolesta sopivan aineistoon *Weld.txt* hyvin. Tarkastellaan kuitenkin kaavan (1) mukaisesta mallista syntyviä residuaaleja $e_i, i = 1, 2, \ldots, n$. Kuvassa 2 on esitetty pisteparvena residuaalien e_i riippuvuus hitsaukseen kulutetun virran määrästä x_i.



Kuva 2: Residuaalien riippuvuus kulutetun virran määrästä

Kuvasta 2 huomataan, että residuaalit eivät ole täysin satunnaisesti jakautuneet vaan, että kuvan 2 pisteparveen olisi mahdollista piirtää toisen asteen yhtälön kuvaaja. Residuaalien ja virrankulutuksen välisestä riippuvuudesta johtuen tarkastellaan seuraavaksi toisen asteen polynomimallia.

2 Toisen asteen polynomimalli

Tarkastellaan siis seuraavaksi, sopiiko toisen asteen polynomimalli merkitsevästi paremmin aineistoon *Weld.txt* kuin kaavan (1) mukainen yhden selittäjän lineaarinen malli. Sovitetaan siis aineistoon malli

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,\ i = 1, 2, \ldots, n. \tag{2}$$

Kaavan (2) mukaisen kvadraattisen mallin tulokset ja parametrien estimaattien väliset korrelaatiot saadaan laskettua seuraavien käskyjen avulla

```
teeny2 <- lm(y ~ x + xx)
summary(teeny2,correlation=TRUE)
```

jolloin saadaan taulukon 2 mukaiset tulokset.

```
------------------------------------------------
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.5662     9.9392  -4.182 0.000560
x             9.0430     2.2833   3.960 0.000917
xx           -0.4227     0.1306  -3.236 0.004589
---
Residual standard error: 0.1644 on 18 df
Multiple R-Squared: 0.9761,
Adjusted R-squared: 0.9735
F-statistic: 368.3 on 2 and 18 df,
p-value: 2.554e-015

Correlation of Coefficients:
   (Intercept)       x
x     -0.9996
xx     0.9985 -0.9996
------------------------------------------------
```

Taulukko 2: Kvadraattisen mallin tuloksia

Taulukosta 2 nähdään, että kvadraattisen mallin selitysaste R^2 on vielä suurempi kuin lineaarisella mallilla. Aikaisempien parametrien estimaattien arvot ovat muuttuneet radikaalisti ja niiden hajonnat ovat huomattavasti suurempia. Korrelaatiomatriisista havaitaan, että mallin selittäjien välillä on voimakasta kollineaarisuutta. Kun $\mathbf{X}$:n sarakkeet ovat lineaarisesti riippuvia on kyseessä kollineaarisuus. $\mathbf{X}'\mathbf{X}$:n pienin ominaisarvo on lähellä nollaa. Tästä voimme päätellä, että $\mathbf{X}$:n sarakkeet ovat lineaarisesti riippuvia. Kollineaarisuusongelmaa voidaan myös havainnollistaa laskemalla matriisin $\mathbf{X}'\mathbf{X}$ suurimman ja pienimmän ominaisarvon $\mathrm{ch}_i(\mathbf{X}'\mathbf{X})$ avulla ns. condition number $\mathcal{K} = \frac{\sqrt{\mathrm{ch}_{max}(\mathbf{X}'\mathbf{X})}}{\sqrt{\mathrm{ch}_{min}(\mathbf{X}'\mathbf{X})}}$. Seuraavassa on esitetty käsky ja tulos condition numberin laskemiseen:

```
kappa(teeny2)
[1] 28257.94.
```

Koska siis condition number on hyvin suuri, matriisin $\mathbf{X}'\mathbf{X}$ sarakkeiden välillä on suurta lineaarista riippuvuutta ja siten matriisi $\mathbf{X}'\mathbf{X}$ on melkein singulaarinen. Singulaariongelman välttämiseksi siirrytäänkin seuraavaksi tarkastelemaan toisen asteen ortogonaalista polynomimallia.

3 Ortogonaalinen polynomimalli

Kvadraattisessa regressiomallissa esiintyneen singulaarisuusongelman välttämiseksi sovitetaan *Weld.txt* aineistoon seuraavaksi toisen asteen ortogonaalinen polynomimalli

$$y_i = \alpha_0 P_0(x_i) + \alpha_1 P_1(x_i) + \alpha_2 P_2(x_i) + \varepsilon_i,\ i = 1, 2, \ldots, n, \tag{3}$$

missä $P_\nu(x_i)$ on sellainen astetta ν oleva polynomi, että

$$\sum_{i=1}^{n} P_j(x_i) P_k(x_i) = 0,$$

kun $j \neq k$. Oletetaan lisäksi, että $\sum_{i=1}^{n} P_j(x_i)^2 = 1, j = 1, 2$.

Kaavan (3) mukainen ortogonaalinen polynomimalli voidaan esittää matriisitermein siten, että havaintovektori $\mathbf{y}$ muodostuu seuraavasti

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{4}$$

missä siis mallimatriisin $\mathbf{Z} = (\mathbf{1} : \mathbf{z}_1 : \mathbf{z}_2)$ varsinaiset muuttujat $\mathbf{z}_1$ ja $\mathbf{z}_2$ ovat ykkösen pituisia ja kaikki vektorit $\mathbf{1}$, $\mathbf{z}_1$ ja $\mathbf{z}_2$ keskenään ortogonaalisia, eli

$$\mathbf{Z}'\mathbf{Z} = \begin{pmatrix} n & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Lisäksi oletetaan edelleen, että virhetermit ε_i ovat toisistaan riippumattomia ja noudattavat normaalijakaumaa eli $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

Nyt kaavan (3) mukaisen ortogonaalisen polynomimallin estimointitulokset saadaan laskettua seuraavien käskyjen avulla

```
teenyp2 <- lm(y ~ poly(x,2))
summary(teenyp2,correlation=TRUE),
```

ja siten saadaan taulukon 3 mukaiset estimointitulokset.

```
-----------------------------------------------------
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.93810    0.03587 137.664  < 2e-16
poly(x, 2)1  4.42948    0.16438  26.947 4.44e-16
poly(x, 2)2 -0.53186    0.16438  -3.236  0.00459
---
Residual standard error: 0.1644 on 18 df
Multiple R-Squared: 0.9761, Adjusted R-squared: 0.9735
F-statistic: 368.3 on 2 and 18 df,
p-value: 2.554e-015

Correlation of Coefficients:
            (Intercept) poly(x, 2)1
poly(x, 2)1   1.148e-16
poly(x, 2)2  -1.697e-17    4.25e-17
-----------------------------------------------------
```

Taulukko 3: Ortogonaalisen polynomimallin tuloksia

Taulukosta 3 nähdään estimoitujen parametrien korrelaatiokertoimia katsottaessa, että kaavan (2) mukaisessa kvadraattisessa regressiomallissa esiintynyt kollineaarisuusongelma on nyt tavallaan hävinnyt ortogonaalisessa polynomimallissa. Samaan johtopäätökseen päädytään, jos lasketaan condition number $\mathcal{K}$ matriisille $\mathbf{Z}'\mathbf{Z}$

```
kappa(teenyp2)
[1] 4.230139
```

(Todellisuudessa tässä tapauksessa $\mathcal{K} = \frac{\sqrt{\mathrm{ch}_{max}(\mathbf{Z}'\mathbf{Z})}}{\sqrt{\mathrm{ch}_{min}(\mathbf{Z}'\mathbf{Z})}} = \sqrt{n} = \sqrt{21} \approx 4.5825$)

Taulukosta 3 nähdään myös, että parametrien estimoidut arvot ovat muuttuneet selvästi verrattaessa taulukossa 2 esitettyihin kvadraattisen mallin estimaatteihin. Toisaalta heti huomataan, että vakiotermin α_0 estimaatti on nyt yhtä kuin havaintojen y_i keskiarvo eli $\hat{\alpha}_0 = \bar{y}$. Myös muiden parametrien estimaattiarvojen muutoksia voidaan yrittää ymmärtää geometrisilla kuvilla. Sen sijaan selitysaste R^2 on sama kummassakin mallissa (2) ja (3). Tämä johtuu siittä, että matriisien $\mathbf{X}$ ja $\mathbf{Z}$ muodostamat sarakeavaruudet ovat samat eli $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{Z})$, jolloin havaintovektori $\mathbf{y}$ projisoituu samaan paikkaan kummassakin mallissa.

4 Yhteenvetoa ja Johtopäätökset

Kun verrataan toisen asteen polynomimallia ja ortogonaalista polynomimallia huomataan, että ortogonalisointi ei siis vaikuttanut selitysasteeseen. Parametrit eivät kuitenkaan korreloi ortogonaalisessa polynomimallissa. Myöskään toisen asteen t-arvo ei muutu ortogonalisoidessa. Sen sijaan ensimmäisen asteen t-arvo muuttuu huomattavasti. Ortogonalisoimisen vaikutus parametrien estimaattien varіansseihin oli pienentävä. Tämän ansiosta estimoinnin luottamusvälit kapenivat.

Hitsaussauman minimileveys riippuu hitsaukseen kulutetun virran määrästä siten, että mitä enemmän virtaa sitä leveämpi hitsaussauma. Hitsaussauman minimileveyden estimoinnin kannalta toisen asteen polynomimallin käyttäminen on näistä tarkastelluista malleista käyttökelpoisin.