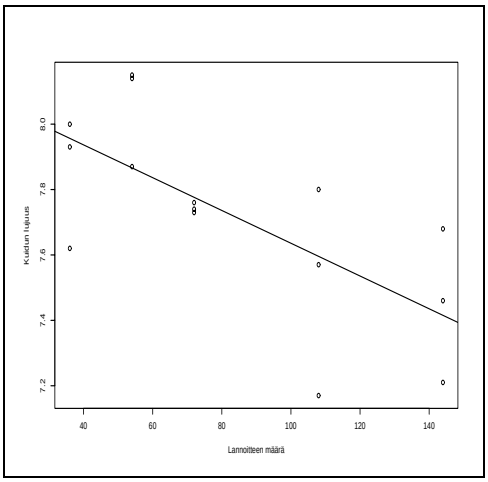
Potaskan määrän vaikutus puuvillakuidun lujuuteen

Tilastollinen mallintaminen
Worksheet 3
Tampereen yliopisto
13.10.2000
Iltanen Olli ja Laakso Jouni

1 Johdanto

Tässä raportissa tarkastellaan *potaska*-aineistoa, jonka alkuperä on hukkunut aikojen saatossa. Aineisto liittyy worksheet 3:een ja siinä vertaillaan viiden eri käsittelyn vaikutusta puuvillakuitujen lujuuteen. Käsittelyt ovat käytettävän lannoitteen, potaskan, määriä (paunaa/aari). Jokainen käsittely toistetaan kolme kertaa. Kuvasta 1 voidaan havaita, että regressiosuora kuvaa pisteparvea vain kohtalaisesti. Yksittäisistä havainnoista suuri osa on kaukana regressiosuorasta.



Kuva 1: Lannoitteen määrän vaikutus puuvillakuidun lujuuteen

Sovitettaessa aineistoon lineaarista regressiomallia

$$H_0 : y_{ij} = \alpha + \beta x_i + \epsilon_{ij}, \tag{1}$$

jossa i=1,...,5 ja j=1,...,3 sekä virhetermit ϵ_{ij} ovat toisistaan riippumattomia ja noudattavat normaalijakaumaa $\epsilon_{ij} \sim \mathrm{N}(0,\sigma^2)$, saadaan taulukon 1 mukainen tulostus. Koska lineaarinen regressiomalli ei ole tämän parempi, yritetään löytää aineistoon paremmin sopiva malli.

2 Varianssianalyysi

Kaikista viidestä käsittelystä on käytettävissä kolme havaintoa, joten on luontevaa tutkia onko käsittelyjen keskiarvojen välillä eroa. Tätä voidaan testata yksisuuntaisella varianssianalyysillä. Nollahypoteesina varianssianalyysissa on se, että tarkastelun kohteena olevan satunnaismuuttujan odotusarvot ovat eri ryhmissä yhtä suuria. Jos nollahypoteesi jää voimaan, perusjoukkoa voidaan tarkastella kyseessä olevan satunnaismuuttujan kannalta yhtenä ryhmänä.

Taulukko 1: Lineaarisen regressiomallin tuloksia

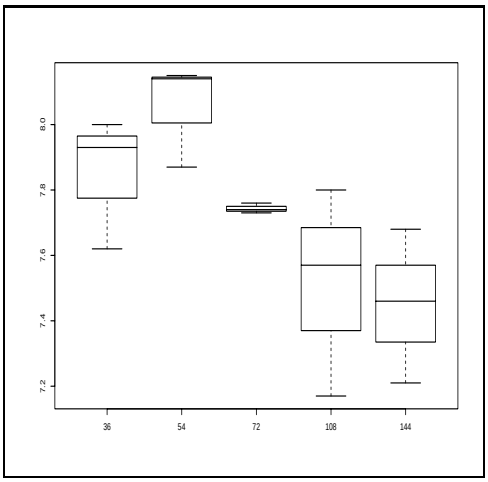
```
------------------------------------------------------------------------------
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.136925   0.132184  61.558  < 2e-16 ***
x           -0.005011   0.001446  -3.466  0.00418 **
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.2171 on 13 degrees of freedom
Multiple R-Squared: 0.4803,     Adjusted R-squared: 0.4403
F-statistic: 12.01 on 1 and 13 degrees of freedom,      p-value: 0.004177
------------------------------------------------------------------------------
```

Komennoilla

```
potaska <- factor(x)
plot(potaska,y)
teeny <- lm(y ~ potaska)
anova(teeny)
summary(teeny)
```

tehdään varianssianalyysi potaska-aineistolle, sekä piirretään box-plot kuvio varanssianalyysin tilanteesta, kuva 2. Komento factor(x) muodostaa viisi luokkaa lannoitemäärien mukaan ja kertoo R:lle, että kyseessä on varianssianalyysi.



Kuva 2: Puuvillakuidun lujuus käsittelyittäin

Varianssianalyysissä sovitetaan aineistoon mallia

$$H_1 : y_{ij} = \alpha + \theta_i + \epsilon_{ij}, \tag{2}$$

jossa i=1,...,5 ja j=1,...,3 sekä virhetermit ϵ_{ij} ovat toisistaan riippumattomia ja noudattavat normaalijakaumaa $\epsilon_{ij} \sim \mathrm{N}(0,\sigma^2)$. Regressiomalli on tämän mallin erikoistapaus (regressiomallissa θ_i on βx_i). Taulukon 2 tulostuksesta nähdään, että mallin (2) selitysaste on parempi kuin mallilla (1). Varianssianalyysissä testataan tavallisesti hypoteesia

$$H_2 : \theta_i = 0, \tag{3}$$

jossa i=1,...,5. Tämän hypoteesin olessa voimassa käsittelyiden välillä ei ole eroa. Tämä tarkoittaisi sitä, että potaskan määrällä ei ole vaikutusta puuvillakuidun lujuuteen. Malli ei kuitenkaan ole sellaisenaan identifioituva, vaan tarvitaan identifioituvuusrajoite. Yleisten lineaaristen mallien standardirajoite on

$$\theta_1 = 0.$$

Rajoitteen avulla verrataan kaikkien muiden käsittelyjen keskiarvojen poikkeamaa ensimmäisen käsittelyn keskiarvosta. Taulukosta 2 ja kuvasta 2 nähdään kuinka toisen käsittelyn (54 paunaa/aari) keskiarvo on muita suurempi. Tästä voidaan päätellä, että sopivin lannoitetaso puuvillakuidun lujuutta ajatellen on juuri 54 paunaa/aari.

2.1 Testi odotusarvojen yhtäsuuruudelle

Neliösummahajoitelmassa havaintoarvojen havaintojen vaihtelua kokonaiskeskiarvon ympärillä kuvaava kokonaisneliösumma

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}\right)^2$$

on jaettu kahteen osaan (SST = total sum of squares): Ryhmäneliösumma

$$SSG = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\overline{Y_i} - \overline{Y}\right)^2 = \sum_{i=1}^{k} n_i \left(\overline{Y_i} - \overline{Y}\right)^2$$

kuvaa ryhmäkeskiarvojen vaihtelua kokonaiskeskiarvon ympärillä (SSG = sum of squares due to groups). Jäännösneliösumma

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y_i}\right)^2$$

kuvaa havaintojen vaihtelua ryhmäkeskiarvojen ympärillä (SSE = sum of squares due to errors). Yllä esitetty neliösummahajoitelma voidaan esittää edellämainittuja määritelmiä käyttäen muodossa SST = SSG + SSE.

Taulukko 2: Varianssianalyysin tuloksia

```
------------------------------------------------------------------------
Response: y
          Df  Sum Sq Mean Sq F value  Pr(>F)
potaska    4 0.73244 0.18311  4.1001 0.03202 *
Residuals 10 0.44660 0.04466
---
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8500     0.1220  64.338    2e-14 ***
potaska54     0.2033     0.1725   1.178   0.2659
potaska72    -0.1067     0.1725  -0.618   0.5503
potaska108   -0.3367     0.1725  -1.951   0.0796 .
potaska144   -0.4000     0.1725  -2.318   0.0429 *
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.2113 on 10 degrees of freedom
Multiple R-Squared: 0.6212,     Adjusted R-squared: 0.4697
F-statistic:   4.1 on 4 and 10 degrees of freedom,      p-value: 0.03202
------------------------------------------------------------------------
```

Testisuure nollahypoteesille H_2 ryhmäkohtaisten odotusarvojen yhtäsuuruudesta voidaan perustaa edellämainitulle neliösummahajoitelmalle. Varianssianalyysin tulkinnnan kannalta on keskeistä, että ryhmäneliösumma SSG kuvastaa ryhmien välistä vaihtelua, kun taas jäännösneliösumma SSE kuvastaa ryhmien sisäistä vaihtelua. Jos ryhmien välinen vaihtelu on suurta verrattuna ryhmien sisäiseen vaihteluun, nollahypoteesi on syytä asettaa epäilyksen alaiseksi. Muodostetaan testisuure

$$F = \frac{n-k}{k-1} \cdot \frac{SSG}{SSE}.$$

Jos testisuureen F arvo on kyllin paljon odotusarvoaan suurempi, hylätään nollahypoteesi. Taulukosta 2 saadaan potaska-aineiston tapauksessa F-testisuure

$$F = \frac{15-5}{5-1} \cdot \frac{0.73244}{0.4466} = 4.1.$$

Koska 5 prosentin merkitsevyystasoa vastaava F-jakauman kriittinen arvo vapausastein 4 ja 10 on 3.478, nollahypoteesi voidaan hylätä 5 prosentin merkitsevyystasolla: odotettavissa oleva puuvillakuidun lujuus riippuu käytettävästä lannoitemäärästä. Kuitenkin, 1 prosentin merkitsevyystasolla nollahypoteesi jää voimaan, kriittisen arvon ollessa 5.994.

3 Varianssianalyysi ja yleinen lineaarinen malli

Yksisuuntaista varianssianalyysiasetelmaa kuvaava yleinen lineaarinen malli voidaan esittää muodossa

$$Y_{ij} = \mu_i + \epsilon_{ij}, \tag{4}$$

jossa i=1,...,5 ja j=1,...,3 sekä virhetermit ϵ_{ij} ovat toisistaan riippumattomia ja noudattavat normaalijakaumaa $\epsilon_{ij} \sim \mathrm{N}(0,\sigma^2)$. Malli vastaa sellaista lineaarista regressiomallia, jossa selitettävänä muuttujana on Y ja selittäjinä ovat indikaattorimuuttujat I_i, jotka kuvaavat sitä, mihin ryhmään havainto Y_{ij} kuuluu. Indikaattorimuuttujat I_i, i=1,...,5 voidaan määritellä kaavalla

$$I_{ij} = \begin{cases} 1 & \text{, jos } Y_{ij} \text{ kuuluu ryhmään } i \\ 0 & \text{, jos } Y_{ij} \text{ ei kuulu ryhmään } i \end{cases}$$

Mallin regressiokertoimina ovat ryhmäkohtaiset odotusarvot μ_i. 1-suuntaista varianssianalyysia vastaava yleinen lineaarinen malli on regressioanalyysimuodossaan seuraava

$$Y_{ij} = \mu_1 I_{1j} + \mu_2 I_{2j} + \cdots + \mu_k I_{kj} + \epsilon_{ij}.$$

Edellä sanottu merkitsee sitä, että yksisuuntainen varianssianalyysi on tilastollisena mallina erikoistapaus yleisestä lineaarisesta mallista. Tämä pätee yleisemminkin: myös m-suuntainen varianssianalyysi on lineaarisen mallin erikoistapaus.

4 Joitakin R-komentoja

Worksheet 3:ssa käytetään myös R-komentoja tapply, qqnorm ja rt. Komennolla

```
tapply(y,x,mean)
```

R laskee y:n keskiarvon x:n eri tasoilla. Esimerkkitapauksessamme komennon suoritus antaa tulokseksi

```
      36       54       72      108      144
7.850000 8.053333 7.743333 7.513333 7.450000 ,
```

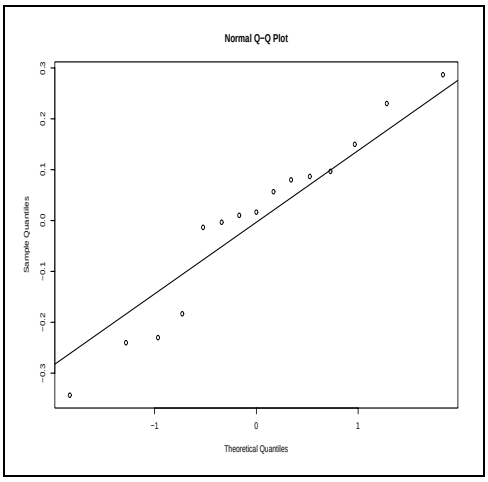
josta nähdään kutakin lannoitemäärää vastaava otoskeskiarvo. Toisin sanoen tapply suorittaa annetun funktion halutun lukujoukon osajoukoille. Muita esimerkkejä tapplyn käytöstä:

```
tapply(y,x,max)
tapply(y,x,median).
```

Komennolla qqnorm saadaan tehtyä Quantile-Quantile Plot, jossa havaitut arvot plotataan teoreettisia kvantiileja vastaan. Mikäli teoreettinen jakauma sopii hyvin havaintoihin, asettuvat havaitut arvot suoraan linjaan. Esimerkkitapauksessamme komennot

```
qqnorm(resid(teeny))
qqline(resid(teeny))
```

tekevät Q-Q Plotin anova-mallin residuaaleille. Kuvaajan avulla voidaan tarkastella ovatko residuaalit normaalisti jakautuneita. Kuvasta 3 havaitaan, että residuaalit noudattavat kohtalaisesti normaalijakaumaa.



Kuva 3: Q-Q plot anova-mallin residuaaleille

Komennoilla

```
y <- rt(150,5)
```

saadaan generoitua 150 arvoa studentin t-jakaumasta 5:llä vapausasteella.