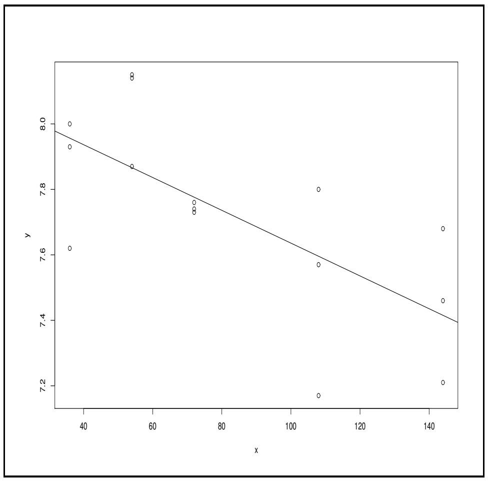
Yksisuuntainen varianssianalyysi

Tilastollinen mallintaminen
Worksheet 3
Tampereen yliopisto
1.12.2000
Iltanen Olli ja Laakso Jouni

1 Johdanto

Pellolle levitettävän potaskan määrän vaikutusta puuvillakuidun lujuuteen on tutkittu kokeellisesti. Käytetyt lannoitemäärät olivat 36, 54, 72, 108 ja 144 paunaa/aari. Kuhunkin lannoitemäärään liittyy kolme arvoa, jotka kuvaavat kuidun lujuutta. Lujuuden mittayksikkö ei valitettavasti ole tiedossa. Mitatut 15 arvoa ja lyhyt aineiston kuvaus löytyvät *worksheet 3*:sta.



Kuva 1: Lannoitteen määrän vaikutus puuvillakuidun lujuuteen

Pisteparven perusteella ihanteellinen lannoiteen määrä on noin 60 paunaa/aari (Kuva 1). Kuvaajaan on piirretty myös regressiosuora, joka on saatu mallilla

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

jossa i=1,...,5 ja j=1,...,3 sekä virhetermit ϵ_{ij} ovat toisistaan riippumattomia ja noudattavat normaalijakaumaa $\epsilon_{ij} \sim \mathrm{N}(0,\sigma^2)$. Regressiomalli selittää 48 % kuidun lujuuden varianssista, mikä toimii vertailukohtana varianssianalyysille.

2 Varianssianalyysi

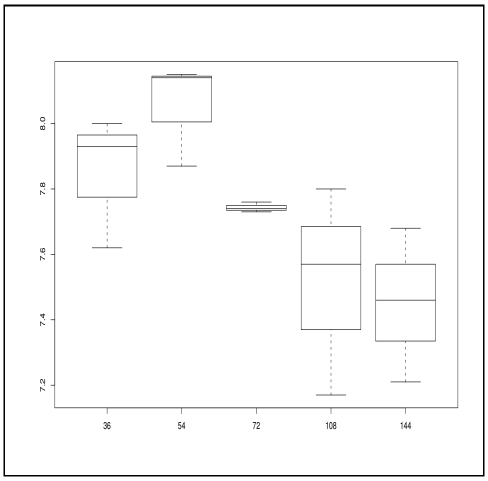
2.1 Lähtökohta

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.

R:n komennoilla

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

tehdään varianssianalyysi potaska-aineistolle, sekä piirretään box-plot kuvio varianssianalyysin tilanteesta (Kuva 2). Kuvio havainnollistaa oivallisesti mittaustulosten eroavaisuudet paitsi luokkien välillä, myös luokkien sisällä. Komento factor(x) jakaa lujuusarvot viiteen luokkaan lannoitemäärien mukaan.



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). Mikäli hypoteesi

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

on voimassa, ei käsittelyiden välillä 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.

2.2 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.

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.

2.3 Varianssianalyysi ja yleinen lineaarinen malli

Yksisuuntaista varianssianalyysiasetelmaa kuvaava yleinen lineaarinen malli voidaan esittää muodossa

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

jossa i=1,...,5 ja j=1,...,3 sekä virhetermit ϵ_{ij} ovat toisistaan riippumattomia ja noudattavat normaalijakaumaa $\epsilon_{ij} \sim$ N(0,σ^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 regessioanalyysimuodossaan seuraava

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

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.

3 Johtopäätökset

Varianssianalyysi (2) antaa paremman selitysasteen kuin regressioanalyysi (1), joten ainakin tältä osin varianssianalyysi täytti odotukset (Taulukko 1). Toisen käsittelyn (54 paunaa/aari) keskiarvo on muita suurempi. Tästä voidaan päätellä, että sopivin lannoitetaso puuvillakuidun lujuutta ajatellen on noin 54 paunaa/aari. Lannoitemäärää lisättäessä puuvillakuidun lujuus pienenee. Heikompi laatu saattaa johtua nopeutuneesta kasvusta.

Taulukosta 1 saadaan potaska-aineiston tapauksessa F-testisuureen arvoksi

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

Koska 5 prosentin merkitsevyystasoa vastaava F-jakauman kriittinen arvo vapausastein 4 ja 10 on 3.478, nollahypoteesi (3) voidaan hylätä 5 prosentin merkitsevyystasolla: odotettavissa oleva puuvillakuidun lujuus riippuu käytettävästä lannoitemäärästä. Kuitenkin, 1 prosentin merkitsevyystasolla nollahypoteesi jää voimaan, kriittisen arvon ollessa 5.994. Tulokseen vaikuttaa osaltaan aineiston pieni koko. Pienetkin erot ovat usein tilastollisesti merkitseviä, kun havaintojen lukumäärä on suuri. Tässä tapauksessa ainoastaan 144:n paunan lannoitemäärän saaneiden puuvillakasvien lujuus poikkeaa merkitsevästi vertailuryhmästä.

Taulukko 1: 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
-----------------------------------------------------------------------
```

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 ,
```

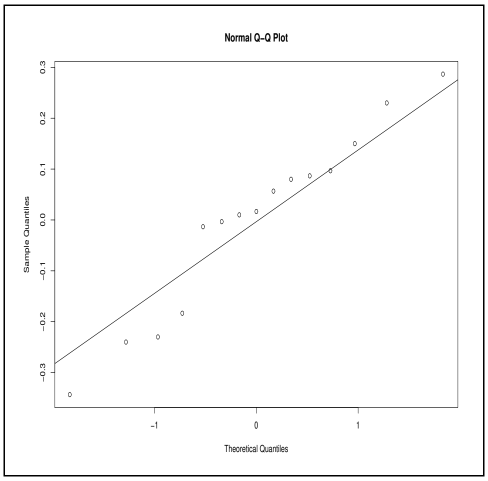
josta nähdään kutakin lannoitemäärää vastaava otoskeskiarvo. Toisin sanoen tapply suorittaa annetun funktion halutun lukujoukon osajoukoille. Muita esimerkkejä tapplyn käytöstä:

```
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 lähes 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. Esimerkkitapauksessamme residuaalit voidaan ajatella normaalisti jakautuneiksi (Kuva 3).



Kuva 3: Q-Q plot anova-mallin residuaaleille